

## BIROn - Birkbeck Institutional Research Online

Osborne, Andrew and Thalassinou, Konstantinos and Davies, Heledd (2016) Expansion of Lysine-rich repeats in Plasmodium proteins generates novel localisation sequences that target the periphery of the host Erythrocyte. *Journal of Biological Chemistry* 291 , pp. 26188-26207. ISSN 0021-9258.

Downloaded from: <https://eprints.bbk.ac.uk/id/eprint/17577/>

*Usage Guidelines:*

Please refer to usage guidelines at <https://eprints.bbk.ac.uk/policies.html>  
contact [lib-eprints@bbk.ac.uk](mailto:lib-eprints@bbk.ac.uk).

or alternatively

Expansion of Lysine-rich Repeats in *Plasmodium* Proteins Generates Novel Localisation Sequences that Target the Periphery of the Host Erythrocyte

Heledd M. Davies, Konstantinos Thalassinos, and Andrew R. Osborne

From the Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck and University College London, London, UK

Running title: Targeting Role of Repetitive *Plasmodium* Sequences

\*For correspondence, contact Andrew Osborne: E-mail [a.osborne@ucl.ac.uk](mailto:a.osborne@ucl.ac.uk); Telephone (+44) 0207 679 3155; Fax (+44) 0207 679 7046.

**Keywords:** Malaria, host-pathogen interaction, protein targeting, intrinsically disordered protein, protein evolution, cytoskeleton, *Plasmodium*, intracellular trafficking, tandem repeats, low complexity sequences.

---

## ABSTRACT

Repetitive low-complexity sequences, mostly assumed to have no function, are common in proteins that are exported by the malaria parasite into its host erythrocyte. We identify a group of exported proteins containing short lysine-rich tandemly repeated sequences that are sufficient to localise to the erythrocyte periphery where key virulence-related modifications to the plasma membrane and the underlying cytoskeleton are known to occur. Efficiency of targeting is dependent on repeat number, indicating that novel targeting modules could evolve by expansion of short lysine-rich sequences. Indeed, expression GARP fragments from different species shows that two novel targeting sequences have arisen via the process of repeat expansion in this protein. In the protein Hyp12, the targeting function of a lysine-rich sequence is masked by a neighbouring repetitive acidic sequence, further highlighting the importance of repetitive low complexity sequences. We show that sequences capable of targeting the erythrocyte periphery are present in at least nine proteins from *Plasmodium falciparum*, and one from *Plasmodium knowlesi*. We find these sequences in proteins known to be involved in erythrocyte rigidification and cytoadhesion, as well as in previously uncharacterised exported proteins. Together, these data suggest that expansion and contraction of lysine-rich repeats could generate targeting sequences *de novo* as well as modulate protein targeting efficiency and function in response to selective pressure.

## INTRODUCTION

Tandemly repeating protein sequences are common in most eukaryotes, but are particularly abundant in protozoan parasites such as *Plasmodium falciparum* (1,2), the species responsible for the most severe form of malaria in humans. Repetitive sequences can form through slipped strand mis-pairing during DNA replication or unequal crossover of chromosomes in meiosis (3). This is a dynamic process with repetitive sequences often expanding and contracting at a greater rate than that of single nucleotide mutation (4). Over half the open reading frames in the parasite genome encode repetitive sequences (1), from modular arrays of folded domains to poly-asparagine sequences which are prone to aggregation during malarial fevers (5,6). Hydrophobic residues are under-represented in many *P. falciparum* repetitive sequences (7) and these are therefore predicted to be intrinsically disordered (8). To date, very few repetitive sequences of this variety have been characterised.

The host erythrocyte undergoes drastic changes during the blood stage of the parasite life cycle (9-11). Based on the presence of a conserved *Plasmodium* export element (PEXEL<sup>1</sup>) or host-targeting (HT) motif, over 400 proteins are predicted to be exported by *P. falciparum* into the infected cell (12,13). These proteins, as well as a group of PEXEL-negative exported proteins (PNEPs) (14), mediate erythrocyte modifications necessary for the parasite to survive: the nutrient-permeability of the membrane increases (15) and protrusions referred to as knobs are assembled at the erythrocyte plasma membrane.

These spiral-shaped scaffolds present proteins from the PfEMP1 family on the erythrocyte surface which mediate the adhesion of infected erythrocytes to blood vessel endothelial cells (16-18). The erythrocyte cytoskeleton, which is composed of flexible alpha and beta-spectrin filaments (19), is also rigidified upon infection (20). Cytoadhesion and the increased rigidity of infected cells contribute to parasite sequestration in specific tissues; sequestered parasites evade clearance in the spleen and are linked to severe disease outcomes such as cerebral malaria (21). Many proteins associated with erythrocyte rigidification and cytoadhesion contain tandem repeats (22-29), yet their role in protein function remains unclear. Some repeating sequences appear to be under immune selection (30) and many are highly antigenic (31); it has been proposed that this may allow the parasite to evade the host immune system by diverting B-cell responses towards non-protective epitopes (32) or promoting an inferior T-cell independent maturation of B-cells (33,34). Such general roles for tandemly repeating sequences may explain their broad distribution in parasite proteins.

Repetitive sequences in some proteins may be removed with no consequence for protein function (35), suggesting they are encoded by functionally neutral 'junk DNA' that has expanded due to errors in DNA replication. However, removal of the repetitive regions of other proteins can affect activity: deletion of repeat regions of the parasite circumsporozoite protein (CSP), and ring exported protein 1 (REX1) lead to loss of protein function (36,37). The knob-associated histidine-rich protein, KAHRP, is involved in both rigidifying the host cell (23) and the formation of cytoadherent knob structures (16), and deletion of a C-terminal sequence encompassing two lysine-rich repetitive sequences results in smaller knob structures and reduced cytoadhesion (38). The Lysine-rich membrane-associated PHISTb protein (LYMP) also modulates cytoadhesion (39). PHISTb proteins are a subgroup of the PHIST family of exported proteins which contain a *Plasmodium* RESA N-terminal domain (PRESAN) (40,41). Several PRESAN domain-containing proteins have been shown to localise to the erythrocyte periphery (42,43), and in the case of LYMP this domain has been shown to bind to PfEMP1 (44,45). Its C-terminus, which includes tandem repeats rich in lysine, has been shown to interact with the cytoskeletal component band 3 (44). The role of tandemly

repeating sequences in functionally important regions of both LYMP and KAHRP suggests that these are not erroneous expansions but may be directly involved in modulating the cytoadhesive properties of the infected host cell. Other known cytoskeleton-binding proteins also contain repetitive sequences, many of which are rich in lysine and glutamate residues (46,47). A role for these highly-charged sequences in protein function has yet to be demonstrated, and cytoskeleton-binding sites for the proteins RESA, Pf332, PfEMP3, MESA and the PHISTa protein PF3D7\_0402000 have previously been identified in non-repetitive regions (27,48-53). Here we show that lysine-rich repeating sequences constitute targeting modules that direct a number of exported parasite proteins to the periphery of the infected erythrocyte. Based on the observation that targeting efficiency is dependent upon repeat length, we present a model in which repeat expansion and contraction can generate novel targeting modules or modulate the targeting efficiency of exported parasite proteins.

## RESULTS

### *Multiple lysine-rich repeating sequences within glutamic acid-rich protein (GARP) localise to the infected erythrocyte periphery*

Glutamic Acid-Rich Protein (GARP) is an 80 kDa protein encoded by the *P. falciparum* gene PF3D7\_0113000 (54). It contains an N-terminal signal sequence for targeting to the parasite ER and a PEXEL/HT motif sequence 'RLLNE', enabling the protein to be exported into the host erythrocyte. GARP is a highly charged protein: it contains 24% glutamic acid, 21% lysine and 9% aspartic acid residues. These charged residues are concentrated within six tandemly repeated sequences which each contain a unique repeated motif. The first four repeat sequences are lysine-rich and the C-terminus of the protein contains an acidic stretch composed of two different repeating units (Fig. 1A, 1B and Table 1). Beyond the N-terminal signal sequence, GARP contains very few hydrophobic residues suggesting that it does not contain stable folded domains. Indeed, protein disorder analysis using the program DISOPRED (55) suggests that the entire sequence of GARP is intrinsically disordered (Fig. 1C).

To determine the localisation of GARP, the protein was GFP-tagged and expressed in the blood stage of *P. falciparum* parasites using the calmodulin promoter. GFP-fluorescence was

localised at the periphery of the red blood cell, indicating that the protein is recruited either to the plasma membrane or the adjacent spectrin cytoskeleton of the infected erythrocyte (Fig. 1D). Quantification of relative fluorescence intensity indicated a 3.27 (S.D.  $\pm$  0.86) fold increase in fluorescence intensity at the erythrocyte periphery relative to the cytoplasm (see Supplemental Material 1 and Table 2 for additional images and quantification of all parasite lines).

As GARP is comprised mainly of repetitive, low-complexity, and intrinsically disordered sequences, it is likely that at least some of these sequences constitute novel modules that can target a protein to the periphery of the infected cell. To test this, fragments encoding the three lysine-rich repeating sequences were GFP-tagged and fused to the N-terminal signal sequence and PEXEL-HT motif of the protein REX3 (residues 1-61), which has been used previously to mediate protein export (Fig. 1E-G) (41,43). This Rex3 fragment alone does not target proteins to the erythrocyte periphery (Supplemental Material 1B). The first lysine-rich repeat sequence contains a three-residue motif that is repeated 15 times. The consensus sequence of the repeated motif, defined by the program XSTREAM, is EKK. The first residue in this motif varies (represented by E, D, H or K residues) but the two lysine residues are highly conserved (Fig. 1B). A GFP fusion protein containing the first lysine-rich repeat region (GARP<sub>119-163</sub>) is efficiently exported and localised to the periphery of the infected erythrocyte (Fig. 1E). GARP<sub>253-340</sub>, which contains the second lysine-rich repeat comprising 7 repeats of the degenerate amino acid sequence E-KE--K-KKQ- (- indicates that a gap is most commonly found at a particular position), is also efficiently localised to the erythrocyte periphery (Fig. 1B and 1F). Similarly, GARP<sub>372-446</sub>, encompassing the third and fourth repeats which are immediately adjacent and comprise 9 repeats of the sequence EEHKE followed by 5 repeats of the sequence KGKKD, also exhibits a clear localisation at the periphery of the infected erythrocyte (Fig. 1B and 1G). Conversely, the acidic C-terminus of GARP, GARP<sub>535-673</sub>, remains in the erythrocyte cytosol (Fig. 1H). GFP accumulation in the food vacuole is also seen in some parasites, likely due to endocytosis of the erythrocyte cytoplasm by the parasite. This is also seen in other parasite lines but is generally less apparent when proteins

are localised to the erythrocyte periphery as this likely reduces the efficiency with which these proteins are endocytosed. Likewise, the uncharged N-terminus of the protein, GARP<sub>50-118</sub>, is not peripherally-targeted (Fig. 1I). Expression of all proteins was confirmed by western blotting (Fig. 1K). The full-length GARP protein appears as a blurred band and most constructs migrate at a mass higher than that expected; this is likely due to the highly charged and repeating nature of the proteins. Taken together, these data indicate that at least three lysine-rich repeating and intrinsically disordered regions within GARP are sufficient to form targeting modules that localise to the erythrocyte periphery.

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

As each peripheral-targeting sequence of GARP is repetitive in character, we tested whether the length of the lysine-rich sequence affects its targeting efficiency. The first lysine-rich repeat sequence of GARP was truncated from 45 residues to 30 and 15 residues; containing fifteen, ten, and five repeats, respectively (Fig. 2). An additional linker sequence of 12 residues was inserted between GFP and the GARP fragments, to ensure that proximity to GFP did not compromise potential interactions of the lysine-rich fragments. As expected, GARP<sub>119-163</sub>, which encodes all fifteen repeats, is localised at the erythrocyte periphery, indicating that the addition of the linker sequence does not alter the targeting function of the first lysine-rich repeat sequence (Fig. 2A and D). GARP<sub>134-163</sub>, which contains only 10 repeats, is also localised to the erythrocyte periphery but targeting is less efficient (Fig. 2B and D); fluorescence intensity at the erythrocyte periphery relative to the erythrocyte cytoplasm is reduced. The shortest construct, GARP<sub>149-163</sub>, only encodes 5 repeats and is not efficiently recruited to the periphery; the protein is predominantly localised diffusely in the erythrocyte cytoplasm (Fig. 2C and D). Expression of each of the proteins was confirmed by Western blotting (Fig. 2E). These data show that multiple repeats are required for the targeting of lysine-rich sequences to the erythrocyte periphery, and that the efficiency of targeting increases as the number of repeats increases. In the context of the first GARP repeat, a sequence of approximately 30 amino acids in length is necessary for robust peripheral targeting.

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

Repetitive DNA sequences are highly mutable and are prone to expansion and contraction (4). Given the preceding data, this suggests that sequences with a peripheral targeting function may arise *de novo* simply by expansion of short non-functional lysine-rich motifs.

To test whether this phenomenon can be observed over evolutionary time we compared the GARP sequences of *P. falciparum* to those of closely related *Plasmodium* species (56,57). The *P. falciparum* and *P. reichenowi* genes encoding GARP are syntenic; the latter also encodes an exported protein that contains four lysine-rich repeats and a C-terminal acidic sequence. While the first lysine-rich repeat of the *P. falciparum* protein corresponds to 15 copies of the (E/D/K/H)KK motif, the first lysine-rich repeat of PrGARP contains only five repeats conforming to this consensus (Fig. 3A). Instead, in the *P. reichenowi* protein a more acidic DE(T/K) repeat has expanded in this region (Fig. 3A). Analysis of the equivalent *P. gaboni* GARP sequence indicates that yet another repeat motif, (HDN)KN, has expanded in addition to four repeats of the (E/D/K/H)KK motif (57).

Although the second GARP repeat sequence is similar in all three parasite species, the third and fourth sequences in the *P. gaboni* protein have not expanded and comprise only 1 or 2 highly degenerate repeats (Fig 3A).

To test whether the expansion of the first and the third and fourth repeat sequences has led to the formation of a functional targeting sequence in *P. falciparum* GARP, the localisation of GFP fusion proteins derived from these sequences from different species was compared. Consistent with this model, the GFP-tagged first repeat from *P. falciparum* GARP (PfGARP<sub>119-163</sub>), is localised to the erythrocyte periphery (Fig. 3B and 3F) but the equivalent GFP-tagged *P. reichenowi* GARP fragment (PfGARP<sub>71-130</sub>), which contains fewer lysine-rich repeats, is diffusely localised in the erythrocyte cytoplasm (Fig. 3C and 3F). Likewise, the region of PfGARP (PfGARP<sub>372-446</sub>), comprising the third and fourth repeats is localised to the red cell periphery (Fig. 3D and 3F) but the equivalent region from the *P. gaboni* protein (PgGARP<sub>381-412</sub>), is not (Fig 3E and 3F). Anti-GFP western

blotting confirmed the expression of proteins at the expected size (Fig. 3G).

Although the sequence of the common ancestor of these proteins is not known, these experiments suggest that expansion of non-functional, short lysine-rich repeats can lead to the formation of novel protein modules that can direct the localisation of exported parasite proteins within the infected erythrocyte.

*Lysine-rich Repeat regions from multiple exported P. falciparum proteins confer peripheral localization in the infected erythrocyte.*

Many *Plasmodium* proteins contain repetitive sequences enriched in charged residues. To investigate whether sequences similar to those in GARP are capable of targeting to the erythrocyte periphery, we identified putative exported proteins, characterized by an N-terminal signal sequence or transmembrane domain and an RxL motif, that also contain repeating sequences of at least 30 residues in length and a lysine content of 20% or greater.

Lysine-rich and repetitive sequences were identified in exported protein sequences using a sliding window algorithm and the program XSTREAM, respectively. Thirty-five sequences, including those within GARP, were found to conform to the above criteria, with some proteins containing multiple repeating lysine-rich sequences (Table 1).

Sequences encoding lysine-rich repeat sequences from ten proteins (Table 1 - highlighted) were expressed as GFP fusion proteins and their localisation assessed by fluorescence microscopy. PF3D7\_1102300 protein, like GARP, is predicted to be entirely intrinsically disordered; the majority of the sequence is lysine-rich and repeating (Fig. 4A). A fusion protein which included the N-terminus of Rex3, GFP, and the lysine-rich sequence of PF3D7\_1102300 comprising residues 121-415 (PF3D7\_1102300<sub>121-415</sub>), was expressed in parasites. Within the infected erythrocyte, GFP fluorescence was localised at the periphery of the infected cell (Fig. 4A). GEXP12 contains an N-terminal PRESAN domain belonging to the PHISTc family, and a C-terminal lysine-rich sequence. A similar pattern of peripheral GFP fluorescence is seen in erythrocytes infected with parasites expressing an exported GFP protein that includes this fragment (GEXP12<sub>231-370</sub>) (Fig. 4B); some brighter foci of fluorescence are also seen in some cells.

A number of proteins known to target the erythrocyte cytoskeleton via defined motifs in non-repeating sequences also contain lysine-rich repeating sequences that have not previously been shown to function as independent targeting domains *in vivo*. The lysine-rich repeat regions of the PHISTb proteins LYMP (LYMP<sub>419-528</sub>), and PF3D7\_1476200 (PF3D7\_1476200<sub>443-512</sub>), and the PHISTa protein PF3D7\_0402000 (PF3D7\_0402000<sub>305-428</sub>) were expressed as GFP fusion proteins; peripheral GFP fluorescence was seen in erythrocytes infected with all three parasite lines (Fig. 4C, 4D, and 4E, respectively). A GFP fusion protein encompassing the lysine-rich region of the PHISTb/c protein PF3D7\_1201000 (PF3D7\_1201000<sub>292-397</sub>) exhibited a weak localisation at the periphery that was visible in only a fraction (50-80%) of infected cells (Fig. 4F).

The N-terminus of MESA contains a 20-residue cytoskeleton-binding MEC motif (51). The remainder of the MESA sequence consists of various charged repetitive sequences, three of which have a lysine content of over 20%. 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<sub>850-1147</sub>) also localises to the erythrocyte periphery (Fig. 4G). Similarly, KAHRP contains an N-terminal histidine-rich sequence that is sufficient to target to the erythrocyte periphery (58) but also contains two lysine-rich repeat regions that are important for protein function (38). A GFP fusion protein encompassing the first of the lysine-rich repeats (5' repeats) is also targeted to the periphery of the infected erythrocyte (Fig. 4H).

Hyp12 contains a lysine-rich C-terminal sequence; the repeats in the sequence are highly degenerate. When fused to GFP in the absence of other sequences the repeat sequence localises to the erythrocyte periphery (Fig. 4I).

Protein PF3D7\_0114200 is predicted to contain a C-terminal transmembrane domain as well as a lysine-rich sequence. The lysine-rich sequence was fused to Rex3:GFP and expressed in parasites (PF3D7\_0114200<sub>97-420</sub>). In this case the fluorescence remained localised within the cytosol of the erythrocyte, with no peripheral targeting (Fig. 4J). PF3D7\_1149100.1 contains six repetitions of a 40-residue motif but the lysine content of this sequence is only 17% lysine and this fragment also remained in the erythrocyte cytosol (Fig. 4K). Additionally, the

C-terminal lysine-rich repeat sequence (3' repeats) from KAHRP (KAHRP<sub>540-600</sub>) does not localise to the cell periphery despite having a lysine content of 20% (Supplemental Material 1W). Although it is difficult to interpret a negative result this suggests that a certain threshold of lysine residues is required for peripheral localisation within the erythrocyte, and that the distribution of residues within repeats may also be important. In the case of the KAHRP 3' repeats the lack of peripheral targeting could also be due to partial degradation of the protein as several bands are seen on western blots of parasites expressing this protein (Fig. 4L).

These data indicate that many diverse repetitive lysine-rich sequences, in which the size of the repeating unit can vary from 3-30 residues in length, have a propensity to localise to the periphery of the infected erythrocyte. Of the ten repetitive sequences with a lysine content greater than 20% which were tested, nine were localised to the erythrocyte periphery. 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 1). The two exceptions, MESA (pI of fragment: 4.90) and the PHISTb/c protein PF3D7\_1476200 (pI of fragment: 4.71), both display the least prominent peripheral targeting, and the aspartate-rich repeats of PF3D7\_0114200 (pI: 5.12) remained entirely cytosolic. Acidic residues may therefore interfere with the peripheral localisation of some lysine-rich repeating sequences.

Having determined that lysine-rich repetitive sequences, when fused to GFP, can localise to the periphery of the infected erythrocyte we next tested whether these sequences function similarly in the context of the corresponding full-length proteins. GFP-tagged PF3D7\_1102300, GEXP12, PF3D7\_0402000, and PF3D7\_1201000 were expressed and all showed peripheral localisation (Fig. 5A, 5B, 5C and 5D, respectively). PF3D7\_1201000 showed a very weak localization to the cell periphery which is similar to the localization of the isolated lysine-rich fragment; GFP fluorescence was also accumulated in the parasitophorous vacuole in this case. LYMP, MESA, and KAHRP, have previously been localised to the periphery of infected cells by immunofluorescence (39,45,59,60). GFP-tagged PF3D7\_1476200, when expressed from the

calmodulin promoter has previously been localised to the periphery of the infected erythrocyte (43). When expressed from its own promoter the protein localisation is similar (Fig 5E). Similarly, GARP when expressed from its own promoter is also peripherally localized (Fig 1J). Detection by western blotting of this protein is variable; smeared bands and prominent fragments of the protein are often detected (Fig 1K). Transcripts of GEXP12 and PF3D7\_1102300 are enriched in gametocyte stage parasites relative to the asexual stage (61). To test whether lysine-rich sequences can also target proteins to the erythrocyte periphery in this lifecycle stage we expressed GFP-tagged PF3D7\_1102300 from its own promoter. Within a mixed culture, most brightly GFP-expressing parasites were gametocytes (Fig. 5F and Supplemental Material 1AJ). The GFP localisation is consistent with the protein targeting to the periphery of the gametocyte-infected cell. Expression of proteins was confirmed by western blotting (Fig. 5J). Gametocyte-enriched parasites were purified for western blots of PF3D7\_1102300, which was detected as a smeared band at a molecular weight higher than expected. Other proteins were detected at approximately the expected sizes.

*The targeting function of the lysine-rich sequence in Hyp12 is masked by an acidic sequence.*

We also localised GFP tagged full-length Hyp12 protein. The lysine-rich fragment of Hyp12 is efficiently recruited to the periphery of the red cell (Fig 4I). By comparison, the full-length protein with either a C- or N-terminal GFP tag is not efficiently recruited to the cell periphery (Fig. 5G and 5H, respectively). This localisation for the full-length protein has also been described previously (62). Hyp12 contains a C-terminal lysine-rich sequence but also a highly acidic N-terminal sequence. The acidic sequence is also repetitive and predicted to be intrinsically disordered. To test the possibility that this sequence is able to inhibit the targeting function of the lysine-rich sequence, the C-terminus of Hyp12 protein lacking the acidic sequence was expressed. This protein is robustly recruited to the cell periphery suggesting that the acidic sequence masks the targeting function of the lysine-rich sequence within this protein (Fig. 5I, K, and L).

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

The length of repeat sequences often varies between different parasite strains (63) and the preceding experiments suggest that variation in length of lysine-rich repeats may influence the efficiency with which these sequences can target proteins to the erythrocyte periphery. To determine the extent of repeat length variation seen in lysine-rich repeat sequences we analysed sequences from the genomes of several laboratory strains of parasites (3D7, DD2, HB3, IT, and 7G8) as well as eleven parasites isolated from infected people from diverse geographic locations ('long-read' sequence data generated by the PF3K consortium was used for these analyses to ensure the correct assembly of repetitive regions).

Significant variation in repeat number is seen in many lysine-rich targeting sequences. The C-terminal repeating sequence of LYMP, the first repeat of GARP, and the PHISTa protein PF3D7\_0402000 contain 5 to 7, 12 to 17, and 9 to 14 copies of repeating motifs, respectively (Fig. 6A-C). Although unlikely to lead to a complete loss or gain of peripheral localisation, 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 PHISTB protein PF3D7\_1476200, PHISTc protein GEXP12, and Hyp12 are invariant (Fig. 6D-G). The 5' repeats of KAHRP do not vary but variations in repeat number are observed for the 3' repeats (Fig. 6H), as has been reported previously (64,65). In 3D7 parasites, the protein PF3D7\_1201000 contains two PRESAN domains which are separated by 18 units of the sequence DEKEK. In all other parasites the repeat unit number has increased; in some cases by as much as twofold (Fig. 6I).

In 3D7 parasites, MESA contains five repeat sequences; all except the second repeat sequence vary significantly in length. In the 3D7 genome, the sequence encoding the third repeat region, which is itself variable in length (Fig. 6J), 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. GFP tagged lysine-rich sequences from both MESA and PF3D7\_1201000 display a weak fluorescence signal at the erythrocyte periphery and duplication and extension of the repeat regions

may increase the targeting efficiency of these sequences.

*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, we also searched other parasite genomes for putative exported proteins that contain lysine-rich repeating sequences. The proteins predicted to contain sequences with a targeting function are shown in Supplemental Material 2. The largest numbers of potential periphery-targeting sequences were found in the *P. reichenowi* genome, with 20 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 19, 15 and 6 proteins containing lysine-rich repetitive regions, respectively, while fewer sequences were predicted for *Plasmodium* species infecting rodents; *P. yoelii*, *P. chabaudi* and *P. berghei* (Supplemental Material 2).

To test whether lysine-rich sequences from parasites other than *P. falciparum* have targeting functions, we tested the localisation of the *P. knowlesi* protein PKNH\_1325700 in *P. falciparum* infected erythrocytes. This protein contains a HT-PEXEL sequence, 'RSLSV', and two repetitive lysine-rich stretches at its C-terminus (Fig. 7A). Full-length PKNH\_1325700 was efficiently exported to the erythrocyte, where the GFP signal appears as a partially punctuate distribution around the periphery of the red blood cell (Fig. 7B). In younger parasites, fewer of these puncta were present and a continuous line of fluorescence was apparent around the periphery of the cell (Fig. 7C). 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 regions which have lysine contents of 12.5% and 40%, respectively. This GFP tagged protein formed a continuous ring at the erythrocyte periphery (Fig. 7D) indicating that lysine-rich sequences from multiple parasite species can form modules with a targeting function. Anti-GFP western blotting confirmed the expression of proteins at the expected size (Fig. 7E).

*A conserved protein family containing an EMP3-KAHRP like domain and expanded repeated sequences.*

Notably, PKNH\_1325700 also contains an N-terminal 70-residue sequence which is predicted to form a folded domain (Fig. 8A) and is homologous to the N-terminus of *Plasmodium falciparum* KAHRP (41). Although the repeating motifs found in the C-terminal sequences of PKNH\_1325700 and KAHRP are not related, they are similar in that they are lysine-rich and both sequences target to the erythrocyte periphery. The presence of an N-terminal conserved domain and C-terminal lysine-rich repeating sequences in both KAHRP and PKNH\_1325700 suggests that these proteins may to some extent be functionally related. Given that KAHRP is a key cytoskeleton-associated protein involved in sequestration of *P. falciparum*-infected erythrocytes we searched for proteins that have similar domain architecture in other species. In *P. falciparum*, the conserved N-terminal domain is also found at the N-terminus of the erythrocyte cytoskeleton-associated PfEMP3 protein; the remainder of this protein is also formed of repeating sequences including a central lysine-rich-region (Fig. 8C). As the domain is present in both EMP3 and KAHRP we refer to it as the EMP3-KAHRP-like domain or EKAL domain. KAHRP-like proteins have previously been identified in some species (41); we identify additional EKAL domain-containing proteins in the genomes of the primate-infecting parasites *P. reichenowi*, *P. knowlesi*, *P. vivax*, *P. cynomolgi*, *P. fragile*, *P. ovale*, and *P. inui* (Fig. 8B, 8C and Fig. 9). These proteins can be grouped into seven branches; five branches are closely related to PfKAHRP while two represent homologs of the EMP3 protein (Fig. 8C). Remarkably, each parasite genome encodes at least one protein with a KAHRP-like EKAL domain that is followed by a C-terminal lysine-rich repeating sequence that may target the protein to the periphery of the infected host cell (Fig. 8C). Although sequence homology in PfEMP3 and KAHRP-like proteins is largely restricted to the EKAL domain, it is likely that in many cases the expansion of divergent repetitive lysine-rich sequences has generated protein modules that contribute to the peripheral localization of this protein family in the infected erythrocyte.

## DISCUSSION

Repetitive sequences in many organisms are crucial for protein function (66-73) (reviewed in (4)), yet there are currently few functions assigned to repeats in *Plasmodium*. 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 rapid expansion and contraction of repeating sequences suggests that they can contribute significantly to protein evolution and the generation of novel functional modules (4,63,74,75). Within PfGARP, decreasing the number of repeating units within the N-terminal lysine-rich sequence proportionally decreases the efficiency of targeting. Given this, it is likely that exported parasite proteins can rapidly evolve novel localisation domains by expanding short low-affinity lysine-rich motifs to create high-avidity targeting sequences. Comparison of the repeating sequences of *P. falciparum* GARP with those found in GARP from *P. reichenowi* and *P. gaboni* provides two examples of such repeat expansion occurring. In the first repeating sequence of *P. falciparum* GARP, the repeat EKK has expanded to generate a peripheral-targeting sequence, whereas in *P. reichenowi* a more acidic repeat has expanded which does not efficiently localise to the periphery.

Smaller changes in repeat number may also subtly modulate the targeting efficiency of lysine-rich repeating sequences (Fig. 10). Within proteins which modulate key properties of the host cell such as rigidity, cytoadhesion and nutrient import, such changes could confer a selective advantage. Indeed, correlation between repeat sequence length and phenotype has been observed in other organisms (68-70), and the number of repeat motifs within the functionally important C-terminal domain of *P. falciparum* RNA polymerase II also varies between isolates (76). Analysis of lysine-rich targeting sequences from laboratory and field strains of *P. falciparum* parasites confirms that repeat units can be both lost and gained from these sequences. There is a high level of conservation between repeat motifs within targeting sequences; this is a common feature of disordered repetitive sequences and suggests the repeats were recently expanded and may be particularly dynamic (77,78). This may allow

rapid adaption of parasites under selective pressure.

Although Hyp12 contains a lysine-rich sequence which targets the cell periphery, the targeting function is masked by an acidic repetitive sequence. Expansion or contraction of either sequence in Hyp12 could lead to a change in protein localization. Contraction of the acidic sequence might reduce the inhibitory propensity of this sequence while expansion of the lysine-rich sequence might allow it to overcome the inhibition by the acidic sequence. Over evolutionary time, the localisation of this protein may be determined by two 'competing' repetitive, low complexity, disordered sequences. It remains unclear whether there is a physiological stimulus that might unmask the lysine-rich sequence in Hyp12; proteolytic cleavage, changes in ionic composition or temperature could potentially regulate this process. Notably, deletion of the gene encoding Hyp12 leads to a change in infected cell rigidity (79).

Targeting of proteins by lysine-rich repeating sequences is not restricted to *P. falciparum* proteins. The protein encoded by the *P. knowlesi* gene PKNH\_1325700 contains an N-terminal EKAL domain, with homology to the N-terminus of *P. falciparum* KAHRP (41), and two adjacent lysine-rich repeat sequences at its C-terminus. Although the repeated motifs differ from those in *P. falciparum* KAHRP, the lysine-rich repeats of both proteins localise uniformly to the erythrocyte periphery. The full-length PKNH\_1325700 protein, however, appears as a number of peripherally-located disperse dots, suggesting the N-terminus is prone to self-association. KAHRP is a key component of the electron-dense cytoadherence-related knob structures that are seen in *P. falciparum* infected cells, and which are also observed in *P. fragile*-infected rhesus monkey erythrocytes (80). However, although *P. vivax* and *P. knowlesi* infected cells adhere to specific ligands (81-83), knob-like structures are not seen on erythrocytes infected with these parasites. In addition to PKNH\_1325700, we find at least one KAHRP-like gene characterized by an EKAL domain and a repetitive lysine-rich sequence in the genomes of *P. reichenowi*, *P. vivax*, *P. ovale*, *P. cynomolgi*, *P. fragile*, and *P. inui*. Knob structures in *P. falciparum*-infected cells cluster PfEMP1 proteins. Whilst parasites other than *P. falciparum* and *P. reichenowi* do not express PfEMP1 proteins, other variant surface antigens

have been identified in other species (84); it is possible that the KAHRP homologues in these species play a role in clustering of these proteins on the surface of infected cells in structures which are not morphologically distinctive or electron dense. Notably, EKAL domains and repeating sequences are also found in PfEMP3 and its homologues. Like KAHRP, PfEMP3 is involved in PfEMP1 trafficking, localises to the Maurer's clefts and cytoskeleton of infected cells and affects infected cell rigidity (23,25,48,85). Expansion of different repeat sequences may represent a means of diversifying the function of EKAL domain-containing proteins.

Although several of the identified lysine-rich targeting sequences are found in proteins with known interacting partners, the identity of the binding partner of the lysine-rich sequences remains unclear. We show that a fragment of KAHRP encompassing the 5' lysine-rich repeats is sufficient to target to the erythrocyte periphery *in vivo*. This region is important for the cytoadhesion-modulating function of the protein (38), however the binding partners of the KAHRP repeating sequences remain controversial. It has been suggested that the 5' lysine-rich repeat region interacts with PfEMP1 (86,87), but this interaction was not observed in other studies (88). Fragments of KAHRP that include the 5' lysine-rich repeat sequence also bind to spectrin *in vitro* (89). Although the repeat sequence alone was not sufficient for this interaction under previous experimental conditions (89,90), recent work indicates the 5' repeats are sufficient for spectrin binding<sup>2</sup>. *In vitro*, the C-terminus of LYMP interacts with inside out erythrocyte vesicles (39) and with purified band 3 (44). It is unclear whether the lysine-rich repeats, which are located in the final 100 residues of this fragment, contribute to this interaction but a fragment comprising only the lysine-rich repeats of LYMP does not bind to inside-out erythrocyte vesicles *in vitro* (39). In MESA, the lysine-rich sequence shown here to localise to the erythrocyte periphery was also shown to be insufficient for binding inside-out erythrocyte membranes (51). This may indicate that these lysine-rich repeats interact with *Plasmodium* proteins or cytoskeletal components that are post-translationally modified during infection (91,92). Given the diversity of lysine-rich repeat sequences that can target to the erythrocyte periphery, it is possible that they interact with different host or parasite proteins.

Several proteins that contain lysine-rich targeting sequences also contain other well-characterized cytoskeleton-targeting domains, suggesting that they crosslink multiple components of the erythrocyte cytoskeleton or membrane. Indeed, LYMP functions by linking PfEMP1 and band 3 via its PRESAN domain and C-terminus, respectively (44,45). A lysine-rich repeating C-terminus is also seen in other proteins with PRESAN domains capable of targeting the periphery, including PF3D7\_0936800 (45) and PF3D7\_1476200 (43). Two other uncharacterised proteins with peripherally-localised lysine-rich repeating sequences also contain PRESAN domains: the PHISTC protein GEXP12, and PF3D7\_1201000, which contains N- and C-terminal PRESAN domains from the PHISTb and c families, respectively. It is possible these proteins play similar roles to LYMP at the erythrocyte periphery. However not all PRESAN domains interact with PfEMP1; the PHISTa protein PF3D7\_0402000 binds to band 4.1 (52). Both PF3D7\_0402000 and MESA contain lysine-rich repeat sequences capable of associating with the erythrocyte periphery in addition to band 4.1-binding domains. Although previous immunofluorescence experiments suggest that PF3D7\_0402000 co-localises with band 4.1, a significant fraction of the protein was localised in the parasitophorous vacuole. This is not consistent with the localization that we observe for the GFP-tagged protein; it is possible that the antibody epitope is hidden when the protein is bound to the erythrocyte cytoskeleton (52).

The proteins GARP and PF3D7\_1102300 are predicted to be entirely intrinsically disordered and repeating sequences make up 44% and 66% of the mature proteins, respectively. It is therefore possible that the interaction of the lysine-rich sequences with their target fulfils the function of the protein. Interestingly, expression of GARP is up-regulated in parasites isolated from children with severe malaria (93), and is differentially expressed in parasites selected for adherence to different ligands (94). PF3D7\_1102300 is up-regulated during heat shock (40) and also in parasites selected for cytoadhesion (95). Deletion of the genes encoding GARP and PF3D7\_1102300, as well as the 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

(79). 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 (79).

Some proteins may also function in the gametocyte stage during which the rigidity of the infected cell changes (96). GEXP12 transcripts and peptides are detected in both asexual stage parasites and in gametocytes (97,98). As we have used the calmodulin promoter to express GFP-tagged GEXP12 we are only able to assess its localisation in asexual stages; this shows that the protein has a propensity to localise to the erythrocyte periphery. Notably, when GFP-tagged PF3D7\_1102300 is expressed from its own promoter the protein is localised to the periphery of gametocyte-infected cells indicating that proteins containing lysine-rich sequences can also be similarly targeted during this life cycle stage. Given this, it might be expected that GEXP12 would also localise to the cell periphery in the gametocyte stage.

Electrostatic interactions between the basic lysine residues and a negatively-charged surface, either protein or lipid, are likely responsible for the peripheral localisation of the repeating sequences. Other basic residues may confer a similar localisation; a poly-histidine sequence in KAHRP also targets the erythrocyte periphery (58), however arginine residues are under-represented in the AT-rich parasite genome (3). Interestingly, despite the high predicted isoelectric points of most of the sequences, many peripherally-localised repeats also contain acidic residues, and targeting does not appear to require a strict sequence consensus or repeat length. This makes accurate prediction of sequences with a targeting function difficult. Two lysine-rich proteins tested did not associate with the erythrocyte periphery and, while some untested proteins such as the FIKK kinases FIKK4.1 and FIKK7.1, and the megadalton repeat protein Pf11-1 are implicated in modulating cytoskeletal properties (91,99,100), FIKK4.1 and the PfEMP1 trafficking protein (PTP3) have been shown to localise to Maurer's clefts or the erythrocyte cytoplasm, respectively (79,101).

The observation that repetitive lysine-rich sequences in *Plasmodium* can target proteins to the periphery of the infected erythrocyte suggests that such proteins will perform key functions at the host parasite interface.

Moreover, the potential for expansion and contraction of these sequences to modulate targeting efficiency or to generate novel targeting sequences suggests that they play important roles in evolution of proteins targeted into the host erythrocyte.

## EXPERIMENTAL PROCEDURES

*Plasmids and Parasite Transfection* - Gene sequences were amplified from *P. falciparum* (3D7), *P. knowlesi* (A1H.1), or *P. reichenowi*, genomic DNA and inserted into *P. falciparum* expression plasmids containing an attP site. Gene expression was controlled by the *P. falciparum* calmodulin promoter and *P. berghei* dihydrofolate reductase-thymidylate synthase 3' untranslated region. Gene sequences encoding full-length proteins were cloned in frame with 3' GFP and STREPII tags. Constructs with 5' truncations were fused to a sequence encoding the N-terminal 61 residues of PFI1755c (REX3); this sequence contains the N-terminal signal sequence and HT/PEXEL motif of REX3. These plasmids contained REX3<sub>1-61</sub>, GFP, a linker sequence (LESGSGTGASDV), the lysine-rich sequence encoding fragment, followed by a STREPII tag. The linker was not included in the following constructs shown in Fig. 1: GARP<sub>50-118</sub>, GARP<sub>119-163</sub>, GARP<sub>253-340</sub>, GARP<sub>372-446</sub>, and GARP<sub>535-673</sub>. All cloned *P. falciparum* sequences matched the 3D7 genome sequence; however, two silent base-pair mutations were made in the sequences of GARP<sub>134-163</sub> and GARP<sub>149-163</sub> to facilitate cloning through overlap-PCR. The *P. knowlesi* gene PKNH\_1325700 contained an insertion corresponding to one repeat of the 'KKEQA' motif in both the full-length and truncated constructs. The *P. gaboni* GARP fragment was constructed *de novo* using multiple primers based on a DNA sequence assembled from multiple short sequencing reads (see later for details). Full-length GARP, PF3D7\_1102300, and PF3D7\_1476200 were also expressed under their own promoters; the PfcAM promoter was replaced with sequences starting 932 bp, 967 bp and 1084 bp upstream of the start codon for each gene, respectively. Expression plasmids, together with a plasmid encoding the Bxb1 integrase (102) were transfected (103) into a 3D7attB parasite strain (obtained through BEI Resources, NIAID, NIH: *Plasmodium falciparum* 3D7-attB, MRA-845, deposited by DA Fidock) (104). Transfected parasites were selected using 2.5 nM WR99210 and 2 µg/ml blasticidin.

*Microscopy* - Parasites were fed one day prior to imaging. A drop of culture material in RPMI was placed between a microscope slide and coverslip. Phase contrast and GFP fluorescence images were acquired at room temperature with a ZEISS Axiovert 200M microscope equipped with a HBO100 lamp a 100X oil lens with a numerical aperture of 1,30. Images were taken with an AxioCam MR camera using AxioVision software release 4.8.2. 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. Images were cropped, and automatic brightness and contrast settings applied using Image J.

*Statistical Analysis* - The average fluorescence intensity at the periphery relative to the cytoplasm of infected cells was quantified in ImageJ as described in Supplemental Material 1. Statistical analysis was performed in GraphPad Prism 7 using ordinary one-way ANOVA with each parasite line compared to the GFP-tagged REX3<sub>1-61</sub> fragment to establish whether proteins were significantly enriched at the erythrocyte periphery. Fisher's uncorrected Least Significant Difference (LSD) test was used for multiple comparisons. P-values are reported in Table 2. Labels represent significance; ns, \*, \*\*, \*\*\*, and \*\*\*\* indicate not significant ( $P > 0.05$ ),  $P \leq 0.05$ ,  $P \leq 0.01$ ,  $P \leq 0.001$ , and  $P \leq 0.0001$ , respectively.

*Western Blotting* - Schizonts were purified using 70% percoll, 3% sorbitol in PBS (105). Approximately  $5 \times 10^6$  schizonts were loaded per lane. Blots were probed with rabbit  $\alpha$ -GFP antibody (Torrey Pines, Catalog Number: TP401 Lot Number: 071519) diluted 1:4000 or mouse  $\alpha$ -histoaspartic protease (HAP) (obtained through BEI Resources, NIAID, NIH: Monoclonal Antibody 3F10-6 Anti-*Plasmodium falciparum* Histo-Aspartic Protease (HAP), (MRA-811A, contributed by Daniel E. Goldberg) antibody diluted 1:1000, as indicated (106). Goat anti-rabbit (ThermoFisher, Catalog number: 35568, Lot number: OK195926) and goat anti-mouse (ThermoFisher, Catalog number: 35521, Lot number: LB143097) secondary antibodies were diluted 1:10000. Membranes were imaged with a LI-COR Odyssey imager. The specificity of the  $\alpha$ -GFP antibody was confirmed by western blotting of

parasites not expressing GFP. For western blotting of parasites expressing PF3D7\_1102300 under its own promoter gametocytes were enriched and purified as described previously (107).

#### *Comparison of GARP genes from Closely-Related Parasite Species*

A GARP homologue from *P. gaboni* was assembled from two incomplete protein coding sequences deposited in the NCBI (GenBank accession numbers: KYN95113.1 and KYN95116.1); the connecting region was assembled from sequence reads (GenBank Biosamples SAMN04053641 and SAMN04053639) (57). The *P. reichenowi* GARP gene sequence (PRCDC\_0111200) was from PlasmoDB (Version 26).

*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 were identified by the presence of either a signal sequence (defined by SignalP (108)), or a transmembrane domain within the first 100 residues (defined using MPEX translocon TM analysis (109)), 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.

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 X-STREAM (110). Parameters for XSTREAM were as follows; min word match = 0.6, min consensus match = 0.6, max period = 30, miss penalty = -3, and gap penalty = -3) (111). Another custom perl script was used to interpret the output of XSTREAM and select proteins in which the sequence region comprised of repeats was over 30 residues in length. Multiple lysine-rich repeat sequences were found in some proteins. XSTREAM 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. More stringent parameters were used to reduce the number of gaps in the consensus sequence, with min word match = 0.6, min consensus match = 0.65, miss penalty = -3, and gap penalty = -5. Max period value was set to 30 residues unless shorter repeats were apparent within the predicted consensus sequence; other parameters were set to default values. The consensus sequence of the degenerate repeats of Hyp12 and PF3D7\_0106600 were defined using less stringent criteria. Theoretical isoelectric point values were predicted by PROTPARAM (112).

*Sequence analysis of proteins from different parasite isolates* - 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 (113), and represented with 'Multiple Align Show' (114). 10 out of 141 gene sequences, indicated in Supplemental Material 3A, 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 (see Supplemental Material 3A).

*Sequence analysis of the KAHRP conserved domain* - Proteins with homology to the conserved domain of KAHRP and PfEMP3 were identified by HMMer (115). 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* containing EKAL domains were assembled using the SEQman Ngen software (116). 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 Supplemental Material 3B for details). Sequences were aligned with T-COFFEE (113) in Jalview (117). Maximum likelihood estimation with TREE-PUZZLE (118) was used to create phylogenetic trees based on an extended conserved domain (See Fig. 9 for details), which were assembled with FigTree v1.2.4 (119). Secondary structure predictions and disorder predictions were made by PSIPRED (120) and DISOPRED (55), respectively.

**Acknowledgements:** This work was supported by Wellcome Trust Grants 091095/Z/10/Z and 099764 Z/12/Z. The plasmid pINT No Neo was from Christiaan Van Ooij (NIMR, UK). *P. reichenowi* gDNA was provided by David Conway (London School of Hygiene and Tropical Medicine) and Alan Thomas (Biomedical Primate Research Centre, The Netherlands). Long-read PACBIO sequence data was obtained from Thomas Otto (Wellcome Trust Sanger Institute as part of the Pf3k project ([www.malariagen.net/pf3k](http://www.malariagen.net/pf3k))). Assembled *P. inui* and *P. fragile* sequencing reads were obtained from the Plasmodium 100 Genomes initiative, Broad Institute ([broadinstitute.org](http://broadinstitute.org)). Unassembled *P. ovale* and *P. gaboni* sequencing reads were obtained from the Wellcome Trust Sanger Institute.

**Conflict of Interest:** The authors declare that they have no conflicts of interest with the contents of this article.

**Author Contributions:** Experiments were conducted by HD and AO and designed by HD, KT and AO. HD and AO wrote the paper and all authors approved the final version of the manuscript.

## REFERENCES

1. Mendes, T. A. O., Lobo, F. P., Rodrigues, T. S., Rodrigues-Luiz, G. F., daRocha, W. D., Fujiwara, R. T., Teixeira, S. M. R., and 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
2. Fankhauser, N., Nguyen-Ha, T. M., Adler, J., and Maser, P. (2007) Surface antigens and potential virulence factors from parasites detected by comparative genomics of perfect amino acid repeats. *Proteome Science* **5**, 9
3. DePristo, M. A., Zilvermit, M. M., and Hartl, D. L. (2006) On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. *Gene* **378**, 19-30
4. Gemayel, R., Cho, J., Boeynaems, S., and Verstrepen, K. J. (2012) Beyond Junk-Variable Tandem Repeats as Facilitators of Rapid Evolution of Regulatory and Coding Sequences. *Genes* **3**, 461-480
5. Singh, G. P., Chandra, B. R., Bhattacharya, A., Akhouri, R. R., Singh, S. K., and Sharma, A. (2004) Hyper-expansion of asparagines correlates with an abundance of proteins with prion-like domains in Plasmodium falciparum. *Molecular and Biochemical Parasitology* **137**, 307-319
6. Muralidharan, V., Oksman, A., Pal, P., Lindquist, S., and Goldberg, D. E. (2012) Plasmodium falciparum heat shock protein 110 stabilizes the asparagine repeat-rich parasite proteome during malarial fevers. *Nature Communications* **3**, 10
7. Verra, F., and Hughes, A. L. (1999) Biased amino acid composition in repeat regions of Plasmodium antigens. *Molecular Biology and Evolution* **16**, 627-633
8. Simon, M., and Hancock, J. M. (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biology* **10**, 16
9. Marti, M., and Spielmann, T. (2013) Protein export in malaria parasites: many membranes to cross. *Current Opinion in Microbiology* **16**, 445-451
10. Spillman, N. J., Beck, J. R., and Goldberg, D. E. (2015) Protein Export into Malaria Parasite-Infected Erythrocytes: Mechanisms and Functional Consequences. in *Annual Review of Biochemistry, Vol 84* (Kornberg, R. D. ed.), Annual Reviews, Palo Alto. pp 813-841
11. Boddey, J. A., and Cowman, A. F. (2013) Plasmodium Nesting: Remaking the Erythrocyte from the Inside Out. in *Annual Review of Microbiology, Vol 67* (Gottesman, S. ed.), Annual Reviews, Palo Alto. pp 243-269
12. Hiller, N. L., Bhattacharjee, S., van Ooij, C., Liolios, K., Harrison, T., Lopez-Estrano, C., and Haldar, K. (2004) A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science* **306**, 1934-1937
13. Marti, M., Good, R. T., Rug, M., Knuepfer, E., and Cowman, A. F. (2004) Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* **306**, 1930-1933
14. 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., and 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
15. Nguiragool, W., Bokhari, A. A. B., Pillai, A. D., Rayavara, K., Sharma, P., Turpin, B., Aravind, L., and Desai, S. A. (2011) Malaria Parasite clag3 Genes Determine Channel-Mediated Nutrient Uptake by Infected Red Blood Cells. *Cell* **145**, 665-677
16. 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., and 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
17. Watermeyer, J. M., Hale, V. L., Hackett, F., Clare, D. K., Cutts, E. E., Vakonakis, I., Fleck, R. A., Blackman, M. J., and Saibil, H. R. (2016) A spiral scaffold underlies cytoadherent knobs in Plasmodium falciparum-infected erythrocytes. *Blood* **127**, 343-351

18. Baruch, D. I., Gormley, J. A., Ma, C., Howard, R. J., and 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
19. Mankelov, T. J., Satchwell, T. J., and Burton, N. M. (2012) Refined views of multi-protein complexes in the erythrocyte membrane. *Blood Cells Molecules and Diseases* **49**, 1-10
20. Nash, G. B., O'Brien, E., Gordonsmith, E. C., and Dormandy, J. A. (1989) Abnormalities In The Mechanical-Properties Of Red Blood-Cells Caused by Plasmodium-Falciparum. *Blood* **74**, 855-861
21. Craig, A. G., Khairul, M. F. M., and Patil, P. R. (2012) Cytoadherence and severe malaria. *The Malaysian journal of medical sciences : MJMS* **19**, 5-18
22. Glenister, F. K., Fernandez, K. M., Kats, L. M., Hanssen, E., Mohandas, N., Coppel, R. L., and Cooke, B. M. (2009) Functional alteration of red blood cells by a megadalton protein of Plasmodium falciparum. *Blood* **113**, 919-928
23. Glenister, F. K., Coppel, R. L., Cowman, A. F., Mohandas, N., and Cooke, B. M. (2002) Contribution of parasite proteins to altered mechanical properties of malaria-infected red blood cells. *Blood* **99**, 1060-1063
24. 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., and Cowman, A. F. (2009) Analysis of structure and function of the giant protein Pf332 in Plasmodium falciparum. *Molecular Microbiology* **71**, 48-65
25. Waterkeyn, J. G., Wickham, M. E., Davern, K. M., Cooke, B. M., Coppel, R. L., Reeder, J. C., Culvenor, J. G., Waller, R. F., and 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
26. Pei, X., Guo, X., Coppel, R., Mohandas, N., and An, X. (2007) Plasmodium falciparum erythrocyte membrane protein 3 (PfEMP3) destabilizes erythrocyte membrane skeleton. *J Biol Chem* **282**, 26754-26758
27. Pei, X. H., Guo, X. H., Coppel, R., Bhattacharjee, S., Haldar, K., Gratzer, W., Mohandas, N., and An, X. L. (2007) The ring-infected erythrocyte surface antigen (RESA) of Plasmodium falciparum stabilizes spectrin tetramers and suppresses further invasion. *Blood* **110**, 1036-1042
28. 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., and Suresh, S. (2012) Pf155/RESA protein influences the dynamic microcirculatory behavior of ring-stage Plasmodium falciparum infected red blood cells. *Scientific Reports* **2**, 7
29. 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., and 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
30. 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., and 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
31. Kaur, P., Sharma, P., Kumar, A., and 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
32. Kemp, D. J., Coppel, R. L., and Anders, R. F. (1987) Repetitive Proteins and Genes of Malaria. *Annual Review of Microbiology* **41**, 181-208
33. Schofield, L. (1991) On The Function of Repetitive Domains in Protein Antigens of Plasmodium and other Eukaryotic Parasites. *Parasitology Today* **7**, 99-105
34. Guy, A. J., Irani, V., MacRaild, C. A., Anders, R. F., Norton, R. S., Beeson, J. G., Richards, J. S., and Ramsland, P. A. (2015) Insights into the Immunological Properties of Intrinsically Disordered Malaria Proteins Using Proteome Scale Predictions. *Plos One* **10**, 22

35. Muralidharan, V., Oksman, A., Iwamoto, M., Wandless, T. J., and 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
36. Ferguson, D. J. P., Balaban, A. E., Patzewitz, E. M., Wall, R. J., Hopp, C. S., Poulin, B., Mohammed, A., Malhotra, P., Coppi, A., Sinnis, P., and Tewari, R. (2014) The Repeat Region of the Circumsporozoite Protein is Critical for Sporozoite Formation and Maturation in Plasmodium. *Plos One* **9**, 25
37. 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., and 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
38. Rug, M., Prescott, S. W., Fernandez, K. M., Cooke, B. M., and Cowman, A. F. (2006) The role of KAHRP domains in knob formation and cytoadherence of P falciparum-infected human erythrocytes. *Blood* **108**, 370-378
39. Proellocks, N. I., Herrmann, S., Buckingham, D. W., Hanssen, E., Hodges, E. K., Elsworth, B., Morahan, B. J., Coppel, R. L., and 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
40. 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
41. Sargeant, T. J., Marti, M., Caler, E., Carlton, J. M., Simpson, K., Speed, T. P., and Cowman, A. F. (2006) Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. *Genome Biology* **7**, 22
42. Goel, S., Muthusamy, A., Miao, J., Cui, L. W., Salanti, A., Winzeler, E. A., and Gowda, D. C. (2014) Targeted Disruption of a Ring-infected Erythrocyte Surface Antigen (RESA)-like Export Protein Gene in Plasmodium falciparum Confers Stable Chondroitin 4-Sulfate Cytoadherence Capacity. *Journal of Biological Chemistry* **289**, 34408-34421
43. Tarr, S. J., Moon, R. W., Hardege, I., and 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
44. Oberli, A., Zurbrugg, L., Rusch, S., Brand, F., Butler, M. E., Day, J. L., Cutts, E. E., Lavstsen, T., Vakonakis, I., and Beck, H. P. (2016) Plasmodium falciparum PHIST Proteins Contribute to Cytoadherence and Anchor PfEMP1 to the Host Cell Cytoskeleton. *Cell Microbiol*
45. Oberli, A., Slater, L. M., Cutts, E., Brand, F., Mundwiler-Pachlatko, E., Rusch, S., Masik, M. F. G., Erat, M. C., Beck, H. P., and 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
46. Paila, U., Kondam, R., and Ranjan, A. (2008) Genome bias influences amino acid choices: analysis of amino acid substitution and re-compilation of substitution matrices exclusive to an AT-biased genome. *Nucleic Acids Research* **36**, 6664-6675
47. Dalby, A. R. (2009) A Comparative Proteomic Analysis of the Simple Amino Acid Repeat Distributions in Plasmodia Reveals Lineage Specific Amino Acid Selection. *Plos One* **4**, 15
48. Knuepfer, E., Rug, M., Klonis, N., Tilley, L., and 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
49. Waller, K. L., Stubberfield, L. M., Dubljevic, V., Nunomura, W., An, X. L., Mason, A. J., Mohandas, N., Cooke, B. M., and Coppel, R. L. (2007) Interactions of Plasmodium falciparum erythrocyte membrane protein 3 with the red blood cell membrane skeleton. *Biochimica Et Biophysica Acta-Biomembranes* **1768**, 2145-2156

50. Waller, K. L., Stubberfield, L. M., Dubljevic, V., Buckingham, D. W., Mohandas, N., Coppel, R. L., and 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
51. Bennett, B. J., Mohandas, N., and 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
52. Parish, L. A., Mai, D. W., Jones, M. L., Kitson, E. L., and 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
53. Foley, M., Tilley, L., Sawyer, W. H., and 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
54. Triglia, T., Stahl, H. D., Crewther, P. E., Silva, A., Anders, R. F., and 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
55. Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., and Jones, D. T. (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138-2139
56. Otto, T. D., Rayner, J. C., Bohme, U., Pain, A., Spottiswoode, N., Sanders, M., Quail, M., Ollomo, B., Renaud, F., Thomas, A. W., Prugnonle, F., Conway, D. J., Newbold, C., and Berriman, M. (2014) Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nature Communications* **5**, 9
57. Sundararaman, S. A., Plenderleith, L. J., Liu, W., Loy, D. E., Learn, G. H., Li, Y., Shaw, K. S., Ayoub, A., Peeters, M., Speede, S., Shaw, G. M., Bushman, F. D., Brisson, D., Rayner, J. C., Sharp, P. M., and Hahn, B. H. (2016) Genomes of cryptic chimpanzee Plasmodium species reveal key evolutionary events leading to human malaria. *Nat Commun* **7**, 11078
58. Wickham, M. E., Rug, M., Ralph, S. A., Klonis, N., McFadden, G. I., Tilley, L., and Cowman, A. F. (2001) Trafficking and assembly of the cytoadherence complex in Plasmodium falciparum-infected human erythrocytes. *Embo Journal* **20**, 5636-5649
59. Howard, R. J., Lyon, J. A., Uni, S., Saul, A. J., Aley, S. B., Klotz, F., Panton, L. J., Sherwood, J. A., Marsh, K., Aikawa, M., and Rock, E. P. (1987) Transport of an MR - 300,000 Plasmodium-Falciparum Protein (Pf-EMP-2) From the Intraerythrocytic Asexual Parasite To The Cytoplasmic Face Of The Host-Cell Membrane. *Journal of Cell Biology* **104**, 1269-1280
60. Culvenor, J. G., Langford, C. J., Crewther, P. E., Saint, R. B., Coppel, R. L., Kemp, D. J., Anders, R. F., and Brown, G. V. (1987) Plasmodium falciparum - Identification and Localization Of A Knob Protein Antigen Expressed by a CDNA Clone. *Experimental Parasitology* **63**, 58-67
61. 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., and Marti, M. (2015) Transcriptional profiling defines dynamics of parasite tissue sequestration during malaria infection. *Genome Medicine* **7**, 20
62. Petersen, W., Matuschewski, K., and 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
63. Tan, J. C., Tan, A., Checkley, L., Honsa, C. M., and Ferdig, M. T. (2010) Variable Numbers of Tandem Repeats in Plasmodium falciparum Genes. *Journal of Molecular Evolution* **71**, 268-278
64. Hirawake, H., Kita, K., and 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
65. Triglia, T., Stahl, H. D., Crewther, P. E., Scanlon, D., Brown, G. V., Anders, R. F., and 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

66. Salichs, E., Ledda, A., Mularoni, L., Alba, M. M., and 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
67. Philipps, D., Celotto, A. M., Wang, Q. Q., Tarng, R. S., and Graveley, B. R. (2003) Arginine/serine repeats are sufficient to constitute a splicing activation domain. *Nucleic Acids Research* **31**, 6502-6508
68. Verstrepen, K. J., Jansen, A., Lewitter, F., and Fink, G. R. (2005) Intragenic tandem repeats generate functional variability. *Nature Genetics* **37**, 986-990
69. 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., and Verstrepen, K. J. (2015) Variable Glutamine-Rich Repeats Modulate Transcription Factor Activity. *Molecular Cell* **59**, 615-627
70. Fondon, J. W., and 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
71. Luo, H., and Nijveen, H. (2014) Understanding and identifying amino acid repeats. *Briefings in Bioinformatics* **15**, 582-591
72. Michael, T. P., Park, S., Kim, T. S., Booth, J., Byer, A., Sun, Q., Chory, J., and Lee, K. (2007) Simple Sequence Repeats Provide a Substrate for Phenotypic Variation in the *Neurospora crassa* Circadian Clock. *Plos One* **2**, 10
73. Lin, W. H., and Kussell, E. (2012) Evolutionary pressures on simple sequence repeats in prokaryotic coding regions. *Nucleic Acids Research* **40**, 2399-2413
74. Muralidharan, V., and Goldberg, D. E. (2013) Asparagine Repeats in *Plasmodium falciparum* Proteins: Good for Nothing? *Plos Pathogens* **9**, 4
75. Toll-Riera, M., Rado-Trilla, N., Martys, F., and Alba, M. M. (2012) Role of Low-Complexity Sequences in the Formation of Novel Protein Coding Sequences. *Molecular Biology and Evolution* **29**, 883-886
76. Kishore, S. P., Perkins, S. L., Templeton, T. J., and 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
77. Jorda, J., Xue, B., Uversky, V. N., and Kajava, A. V. (2010) Protein tandem repeats - the more perfect, the less structured. *Febs Journal* **277**, 2673-2682
78. Hughes, A. L. (2004) The evolution of amino acid repeat arrays in *Plasmodium* and other organisms. *Journal of Molecular Evolution* **59**, 528-535
79. Maier, A. G., Rug, M., O'Neill, M. T., Brown, M., Chakravorty, S., Szeszak, T., Chesson, J., Wu, Y., Hughes, K., Coppel, R. L., Newbold, C., Beeson, J. G., Craig, A., Crabb, B. S., and Cowman, A. F. (2008) Exported proteins required for virulence and rigidity of *Plasmodium falciparum*-infected human erythrocytes. *Cell* **134**, 48-61
80. Fujioka, H., Millet, P., Maeno, Y., Nakazawa, S., Ito, Y., Howard, R. J., Collins, W. E., and Aikawa, M. (1994) A Nonhuman Primate Model for Human Cerebral Malaria - Rhesus-Monkeys Experimentally Infected with *Plasmodium fragile*. *Experimental Parasitology* **78**, 371-376
81. Fatih, F. A., Siner, A., Ahmed, A., Woon, L. C., Craig, A. G., Singh, B., Krishna, S., and Cox-Singh, J. (2012) Cytoadherence and virulence - the case of *Plasmodium knowlesi* malaria. *Malaria Journal* **11**, 6
82. 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., and Costa, F. T. M. (2010) On the Cytoadhesion of *Plasmodium vivax*-Infected Erythrocytes. *Journal of Infectious Diseases* **202**, 638-647
83. 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., and 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
84. Frech, C., and Chen, N. (2013) Variant surface antigens of malaria parasites: functional and evolutionary insights from comparative gene family classification and analysis. *Bmc Genomics* **14**, 23
  85. Pei, X., Guo, X., Coppel, R., Mohandas, N., and An, X. (2007) Plasmodium falciparum erythrocyte membrane protein 3 (PfEMP3) destabilizes erythrocyte membrane skeleton. *Journal of Biological Chemistry* **282**, 26754-26758
  86. Waller, K. L., Cooke, B. M., Nunomura, W., Mohandas, N., and 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
  87. Ganguly, A. K., Ranjan, P., Kumar, A., and Bhavesh, N. S. (2015) Dynamic association of PfEMP1 and KAHRP in knobs mediates cytoadherence during Plasmodium invasion. *Scientific Reports* **5**, 9
  88. Mayer, C., Slater, L., Erat, M. C., Konrat, R., and 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
  89. Pei, X. H., An, X. L., Guo, X. H., Tarnawski, M., Coppel, R., and 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
  90. Kilejian, A., Rashid, M. A., Aikawa, M., Aji, T., and Yang, Y. F. (1991) Selective Association of a Fragment of The Knob Protein With Spectrin, Actin and the Red-Cell Membrane. *Molecular and Biochemical Parasitology* **44**, 175-182
  91. Nunes, M. C., Okada, M., Scheidig-Benatar, C., Cooke, B. M., and Scherf, A. (2010) Plasmodium falciparum FIKK Kinase Members Target Distinct Components of the Erythrocyte Membrane. *Plos One* **5**, 8
  92. Doerig, C., Rayner, J. C., Scherf, A., and Tobin, A. B. (2015) Post-translational protein modifications in malaria parasites. *Nature Reviews Microbiology* **13**, 160-172
  93. 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., and 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
  94. Ralph, S. A., Bischoff, E., Mattei, D., Sismeiro, O., Dillies, M. A., Guigon, G., Coppee, J. Y., David, P. H., and Scherf, A. (2005) Transcriptome analysis of antigenic variation in Plasmodium falciparum-var silencing is not dependent on antisense RNA. *Genome Biology* **6**, 12
  95. 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., and 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
  96. 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., and Lavazec, C. (2012) A switch in infected erythrocyte deformability at the maturation and blood circulation of Plasmodium falciparum transmission stages. *Blood* **119**, E172-E180
  97. Llinas, M., Bozdech, Z., Wong, E. D., Adai, A. T., and DeRisi, J. L. (2006) Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. *Nucleic Acids Research* **34**, 1166-1173
  98. Silvestrini, F., Lasonder, E., Olivieri, A., Camarda, G., van Schaijk, B., Sanchez, M., Younis, S. Y., Sauerwein, R., and Alano, P. (2010) Protein Export Marks the Early Phase of

- Gametocytogenesis of the Human Malaria Parasite *Plasmodium falciparum*. *Molecular & Cellular Proteomics* **9**, 1437-1448
99. Scherf, A., Carter, R., Petersen, C., Alano, P., Nelson, R., Aikawa, M., Mattei, D., Dasilva, L. P., and 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
  100. Brandt, G. S., and 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
  101. Nunes, M. C., Goldring, J. P. D., Doerig, C., and 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
  102. van Ooij, C., Withers-Martinez, C., Ringel, A., Cockcroft, S., Haldar, K., and Blackman, M. J. (2013) Identification of a *Plasmodium falciparum* Phospholipid Transfer Protein. *Journal of Biological Chemistry* **288**, 31971-31983
  103. Deitsch, K. W., Driskill, C. L., and Wellems, T. E. (2001) Transformation of malaria parasites by the spontaneous uptake and expression of DNA from human erythrocytes. *Nucleic Acids Research* **29**, 850-853
  104. Nkrumah, L. J., Muhle, R. A., Moura, P. A., Ghosh, P., Hatfull, G. F., Jacobs, W. R., Jr., and Fidock, D. A. (2006) Efficient site-specific integration in *Plasmodium falciparum* chromosomes mediated by mycobacteriophage Bxb1 integrase. *Nat Methods* **3**, 615-621
  105. Radfar, A., Mendez, D., Moneriz, C., Linares, M., Marin-Garcia, P., Puyet, A., Diez, A., and Bautista, J. M. (2009) Synchronous culture of *Plasmodium falciparum* at high parasitemia levels. *Nature Protocols* **4**, 1899-1915
  106. Liu, J., Gluzman, I. Y., Drew, M. E., and Goldberg, D. E. (2005) The role of *Plasmodium falciparum* food vacuole plasmepsins. *Journal of Biological Chemistry* **280**, 1432-1437
  107. Tanaka, T. Q., Dehdashti, S. J., Nguyen, D. T., McKew, J. C., Zheng, W., and Williamson, K. C. (2013) A quantitative high throughput assay for identifying gametocytocidal compounds. *Molecular and Biochemical Parasitology* **188**, 20-25
  108. Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8**, 785-786
  109. Snider, C., Jayasinghe, S., Hristova, K., and White, S. H. (2009) MPEx: A tool for exploring membrane proteins. *Protein Science* **18**, 2624-2628
  110. Newman, A. M., and Cooper, J. B. (2007) XSTREAM: A practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *Bmc Bioinformatics* **8**, 19
  111. Jorda, J., and Kajava, A. V. (2009) T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics* **25**, 2632-2638
  112. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., and Bairoch, A. (2005) Protein Identification and Analysis Tools on the ExpASY Server. in *The Proteomics Protocols Handbook* (Walker, J. M. ed.), Humana Press. pp 571-607
  113. Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**, 205-217
  114. Stothard, P. (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* **28**, 1102-+
  115. Finn, R. D., Clements, J., and Eddy, S. R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**, W29-W37
  116. DNASTAR. SeqMan NGen®. Version 12.0 Ed., Madison, WI
  117. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191
  118. Schmidt, H. A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502-504
  119. Rambaut, A. (2014) FigTree v1.4.2.

120. Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K., and Jones, D. T. (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Research* **41**, W349-W357

## FOOTNOTES

1. The abbreviations used are: PRESAN, *Plasmodium* RESA N-terminal; PHIST, poly-helical interspersed sub-telomeric; GFP, green fluorescent protein; HT, host-targeting; PEXEL, *Plasmodium* export element; PNEP, PEXEL-negative exported protein; KAHRP, knob-associated histidine-rich protein; GARP, glutamic-acid-rich protein; MESA, mature parasite-infected erythrocyte surface protein; LYMP, lysine-rich membrane-associated PHISTb; GEXP12, gametocyte exported protein 12; RESA, ring-infected erythrocyte surface protein; PfEMP1, *Plasmodium falciparum* erythrocyte membrane protein 1; PfEMP3, *Plasmodium falciparum* erythrocyte membrane protein 3; Pf332, *Plasmodium falciparum* protein 332; CSP, circumsporozoite protein; REX1, ring exported protein 1; REX3, ring exported protein 3; PACBIO, Pacific Biosciences sequencing; EKAL, EMP3-KAHRP-like domain.
2. Personal communication with I. Vakonakis *et al.*

## FIGURE LEGENDS

**Figure 1 – Glutamic acid-rich protein (GARP) is targeted to the erythrocyte periphery by three lysine-rich repeat regions.** (A) Representation of the GARP protein. The lysine-rich repeating regions are highlighted in blue and labelled R1-R4, and the number of repeating units in each is indicated. The acidic C-terminus is in red. The export sequence is coloured purple and represents both the signal sequence and PEXEL-HT motif. (B) Sequence logos for the four lysine-rich repeats: residue position is shown on the x-axis and conservation on the y-axis (bits). (C) Disorder prediction for GARP (using DISOPRED (55)). Amino acids are considered disordered if they have a confidence score over 0.5, represented by a dotted line. (D-I) GFP-tagged full-length GARP and truncations expressed using the calmodulin promoter in *P. falciparum* parasites. (J) GFP-tagged full-length GARP expressed using the GARP promoter. GFP fluorescence and phase contrast images are shown in the left and right panels, respectively. A representation of each construct is shown below. Scale bar – 2µm. For quantification of fluorescence see Supplemental Material 1 and Table 2. (K) Anti-GFP western blot (top panel), with anti-HAP used to confirm equal loading (lower panel).

**Figure 2 – 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. GFP fluorescence and phase contrast images are shown in the left and right panels, respectively. Scale bar – 2µm. (D) The ratio of the fluorescence intensity adjacent to the erythrocyte membrane relative to the erythrocyte cytoplasm for the indicated proteins is shown. Error bars show standard error of the mean; ns, \*, \*\*, \*\*\*, and \*\*\*\* indicate not significant ( $P > 0.05$ ),  $P \leq 0.05$ ,  $P \leq 0.01$ ,  $P \leq 0.001$ , and  $P \leq 0.0001$ , respectively. (E) Anti-GFP western blot (top panel), with anti-HAP used to confirm equal loading (lower panel).

**Figure 3 – Repeat expansion and generation of functional targeting sequences in parasite proteins.** (A) Representation of *P. falciparum* GARP (upper), *P. reichenowi* GARP (middle), and *P. gaboni* GARP (lower) with the export sequence in purple and the acidic C-terminus in red. Lysine-rich repeats are shown in blue. Repeated sequences corresponding to the *P. falciparum* first lysine-rich sequence motif (EKDH)KK are also shown in blue and the alternative expanded repeat motifs DE(TK) from *P. reichenowi* and (HDN)KN from *P. gaboni* are shown in pink and turquoise, respectively. Alignment of lysine-rich repeat 1 and lysine-rich repeats 3 and 4 of PfGARP and the homologous regions in PrGARP and PgGARP are shown below. *P. falciparum* lysine-rich repeats are boxed in blue, *P. reichenowi* motifs in pink, and *P. gaboni* motifs in turquoise. Flanking conserved sequences are shaded grey. (B) GFP-tagged PfGARP and (C) PrGARP regions encompassing the first repeat expressed in *P. falciparum*. (D) GFP-tagged PfGARP and (E) PrGARP regions encompassing both the third and fourth repeats expressed in *P. falciparum*. A GFP fluorescence image and a phase contrast image are shown in left and right panels, respectively. A schematic of the cloned region is shown below. Scale bar – 2µm. (F) The ratio of the fluorescence intensity at the erythrocyte periphery relative to the erythrocyte cytoplasm for the indicated proteins is shown; as in Fig. 2. (G) Anti-GFP western blot of parasites (top panel). Anti-HAP was used to confirm equal loading (lower panel).

**Figure 4 - Multiple lysine-rich repeating sequences from *P. falciparum* proteins target the erythrocyte periphery.** (A–K) 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, acidic sequences in red, PRESAN/PHIST domains in orange, predicted transmembrane domains in pink, and the MESA erythrocyte cytoskeleton-binding (MEC) motif 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, with MESA downscaled by ½. Scale bar - 2 µm. (L) Anti-GFP western blots of all parasite lines (top panel). Anti-HAP was used to confirm equal loading (lower panel).

**Figure 5 – Localisation of lysine-rich repeat containing proteins.** (A–I) 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, and acidic regions in red. Expressed fragments are also illustrated below panels G, H and I. Scale bar - 2  $\mu$ m. (J) Anti-GFP western blots of all parasite lines. (K) The ratio of the fluorescence intensity at the erythrocyte periphery relative to the erythrocyte cytoplasm; as in previous figures. \*\*\*\* indicates an extremely significant difference ( $P \leq 0.0001$ ). (L) Model showing the potential masking of the C-terminal lysine-rich sequence by the acidic N-terminus of Hyp12.

**Figure 6 – The number of repeating units in peripheral-targeting proteins varies between different field isolates and lab strains.** (A–J) 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 in a repeat region. 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), acidic repeats in red, other repeats in green and the duplicated MESA domain in yellow. A histogram indicating the number of duplications of the MESA domain is shown below the protein with yellow bars. Schematics are approximately to scale, with MESA downscaled by  $\frac{1}{2}$ .

**Figure 7 – 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 yellow, the histidine-rich region in orange and the export sequence in purple. (B–C) *P. falciparum* parasites 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 – 2  $\mu$ m. (E) Western blots with anti-GFP (top panel). Anti-HAP was used to confirm equal loading (lower panel).

**Figure 8 – 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 (120)). (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: phylogenetic 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 which may contain frameshift mutations are indicated with an asterisk (see Supplemental Material 3B). *P. ovale* proteins were assembled *de novo* and have been named EKAL1-4. *P. fragile* and *P. inui* proteins are named according to their assigned gene names preceded by PFR or PI, respectively. Numbers at each node represent Quartet Puzzling (QP) support values predicted by TREEPUZZLE, where values represent the reliability of groupings (118). Right: diagrams representing each protein sequence, with EKAL domains in yellow. Export sequences are in purple (signal sequence and PEXEL-HT motif). Many proteins contain lysine-rich tandemly repeated sequences (blue), as well as repeating sequences which do not contain over 20% lysine (green). The first repeating sequence of PKNH\_1325700 is shown in light blue as only 12.5% of residues are lysine. The histidine-rich regions of *P. falciparum* and *P. reichenowi* KAHRP are shown in orange. Schematics are approximately to scale, with PVX\_003525, Pf3D7\_0201900, and PO\_EKAL2 scaled down by half. PCYB\_001100 and PFR\_A0A0D9QJA3 sequences are truncated due to gaps in the assembled sequences.

**Figure 9 – Alignment of the EKAL domain in proteins identified as homologs of KAHRP and EMP3 in *P. falciparum*, *P. reichenowi*, *P. knowlesi*, *P. vivax*, *P. cynomolgi*, *P. fragile*, *P. inui*,**

**and *P. ovale*.** Proteins were aligned using T-COFFEE (113). Residues with over 70% identity or similarity are shaded in dark grey and light grey, respectively, using Multiple Align Show (114). 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 phylogenic trees.

**Figure 10 – Evolution of novel targeting domains and modulation of targeting efficiency by repeat expansion.**

## TABLES

Gene ID (Alias/Family)	Consensus Sequence	Position Within Protein	Repeat Unit Length	Number of Repeat Units	Error from Consensus	Theoretical 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 (FIKK4.1)	KEKSKKKHRDDKFNK	85-145	15	4.07	0.2	9.97
PF3D7_0500800 (MESA)	EKND-EKKDKVLGEGDKEDVK	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 (GEXP12)	KECVPNECMK	262-337	10	7.6	0.17	8.90
PF3D7_1201000 (PHISTb/c)	EKDEK	296-387	5	18.4	0.08	4.71
PF3D7_1249600 (LRR12)	NKKEDGD	560-638	7	11.29	0.18	4.68
PF3D7_1401200	KPSKYDDIRCFGEPAQKKK	76-138	19	3.32	0.21	9.88
PF3D7_1476200 (PHISTb)	KEQEKEKERKRKE	451-497	13	3.62	0.18	9.80
PF3D7_1476300 (PHISTb)	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 (PTP3)	KLDSQNGKNEKNEKSIPN	659-814	18	8.67	0.1	9.43
PF3D7_0106600	KNERKKKKKNE-K-KEKRRR-	89-144	18	3.06	0.29	11.72
PF3D7_1301400 (Hyp12)	K-KKEK-QE	300-357	7	8.00	0.36	9.80

**Table 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 grey. Asterisks indicate tested sequences 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 (110). Non-integer numbers of repeat units indicate degeneration at the ends of repetitive sequences. Theoretical pI calculated by PROTPARAM is shown for each fragment (112).

Construct	Fold-difference in Fluorescence at Membrane relative to Cytosol	Standard Deviation	P-value (Significance)
REX3 <sub>1-61</sub>	0.92	0.06	-
Full-Length GARP	3.27	0.86	<0.0001 (****)
GARP <sub>119-163</sub>	2.39	0.43	<0.0001 (****)
GARP <sub>253-340</sub>	3.23	0.61	<0.0001 (****)
GARP <sub>372-446</sub>	1.71	0.29	<0.0001 (****)
GARP <sub>535-673</sub>	0.92	0.07	0.9988 (ns)
GARP <sub>50-118</sub>	0.92	0.06	0.9548 (ns)
GARP (+ promoter)	2.90	0.72	<0.0001 (****)
GARP <sub>119-163</sub> (+linker)	2.29	0.60	<0.0001 (****)
GARP <sub>134-163</sub>	1.76	0.29	<0.0001 (****)
GARP <sub>149-163</sub>	1.12	0.20	0.2323 (ns)
GARP <sub>372-446</sub> (+linker)	1.76	0.37	<0.0001 (****)
<i>P. reichenowi</i> GARP <sub>71-130</sub>	0.90	0.06	0.9181 (ns)
<i>P. gaboni</i> GARP <sub>381-412</sub>	0.94	0.07	0.9248 (ns)
PF3D7_1102300 <sub>121-415</sub>	3.90	0.65	<0.0001 (****)
GEXP12 <sub>231-370</sub>	1.75	0.35	<0.0001 (****)
LYMP <sub>419-528</sub>	2.13	0.34	<0.0001 (****)
PF3D7_1476200 <sub>443-512</sub>	2.50	0.57	<0.0001 (****)
PF3D7_0402000 <sub>305-428</sub>	1.73	0.40	<0.0001 (****)
PF3D7_1201000 <sub>292-397</sub>	1.42	0.45	0.0016 (**)
MESA <sub>850-1147</sub>	1.38	0.20	0.0062 (**)
KAHRP <sub>363-428</sub>	1.49	0.21	0.0007 (***)
KAHRP <sub>540-600</sub>	0.91	0.08	0.9528 (ns)
PF3D7_0114200 <sub>97-240</sub>	0.92	0.04	0.9845 (ns)
PF3D7_1149100.1 <sub>120-416</sub>	0.92	0.05	0.9917 (ns)
Hyp12 <sub>297-381</sub>	2.31	0.33	<0.0001 (****)
PKNH_1325700 <sub>303-445</sub>	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)
Hyp12 (FL)	0.94	0.08	0.9064 (ns)
Hyp12 <sub>51-381</sub>	1.00	0.10	1.82403 (ns)
Hyp12 <sub>158-381</sub>	0.65	2.35	<0.0001 (****)
PF3D7_1102300 (+ promoter)	4.01	1.82	<0.0001 (****)
PF3D7_1476200 (+promoter)	2.35	0.52	<0.0001 (****)

**Table 2 - Quantification and statistical analysis of GFP fluorescence at the periphery of infected erythrocytes.** The fold difference in fluorescence intensity at the erythrocyte membrane relative to the cytosol was calculated as described in Supplemental Material 1. Statistical analysis using one-way ANOVA was performed and multiple comparisons were made between each parasite line and a line expressing GFP-tagged REX3<sub>1-60</sub> only. Images of twenty parasites were quantified per parasite line (n=20). P-values and levels of significance are indicated, from not significant (ns) to extremely significant (\*\*\*) and (\*\*\*\*).

FIGURES

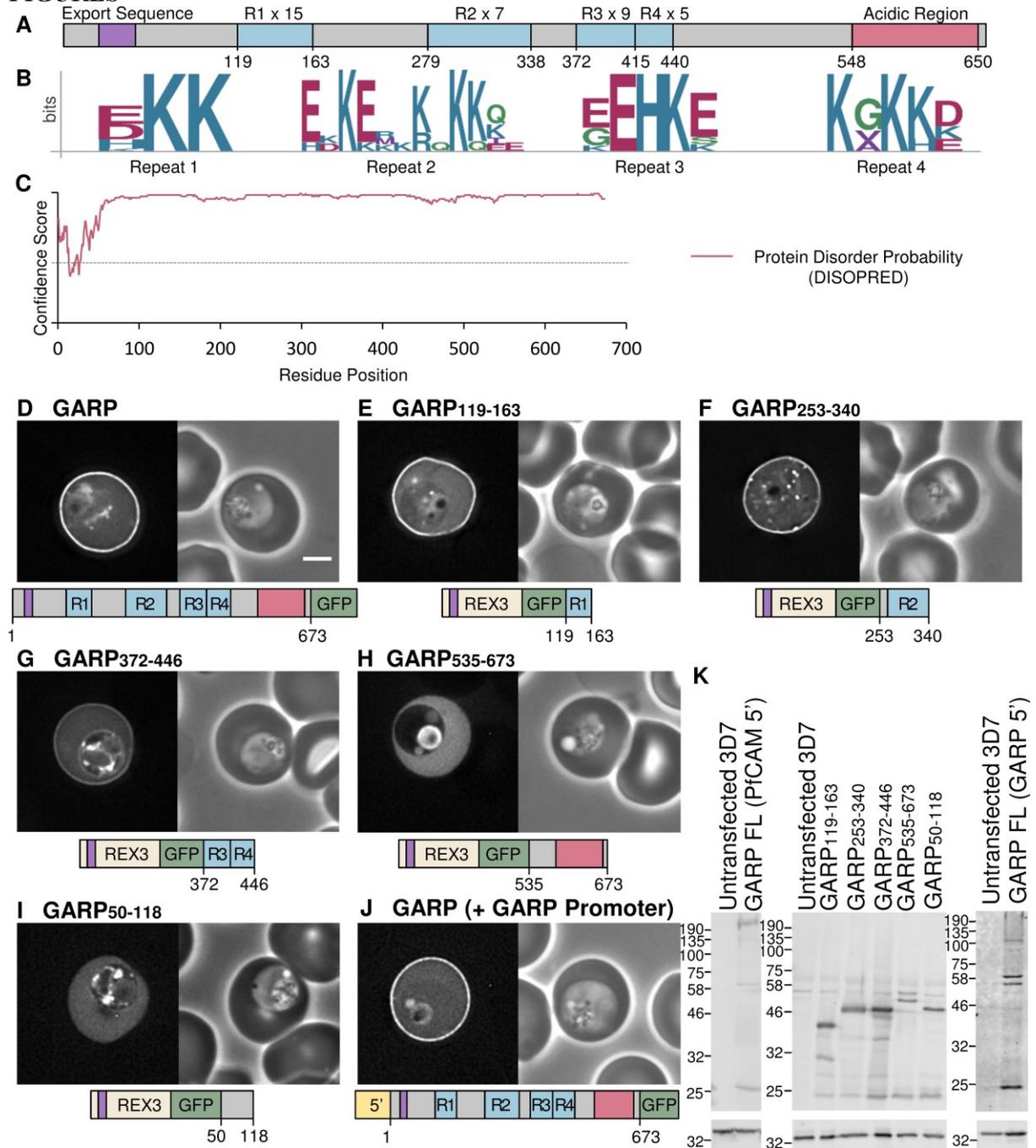


Figure 1

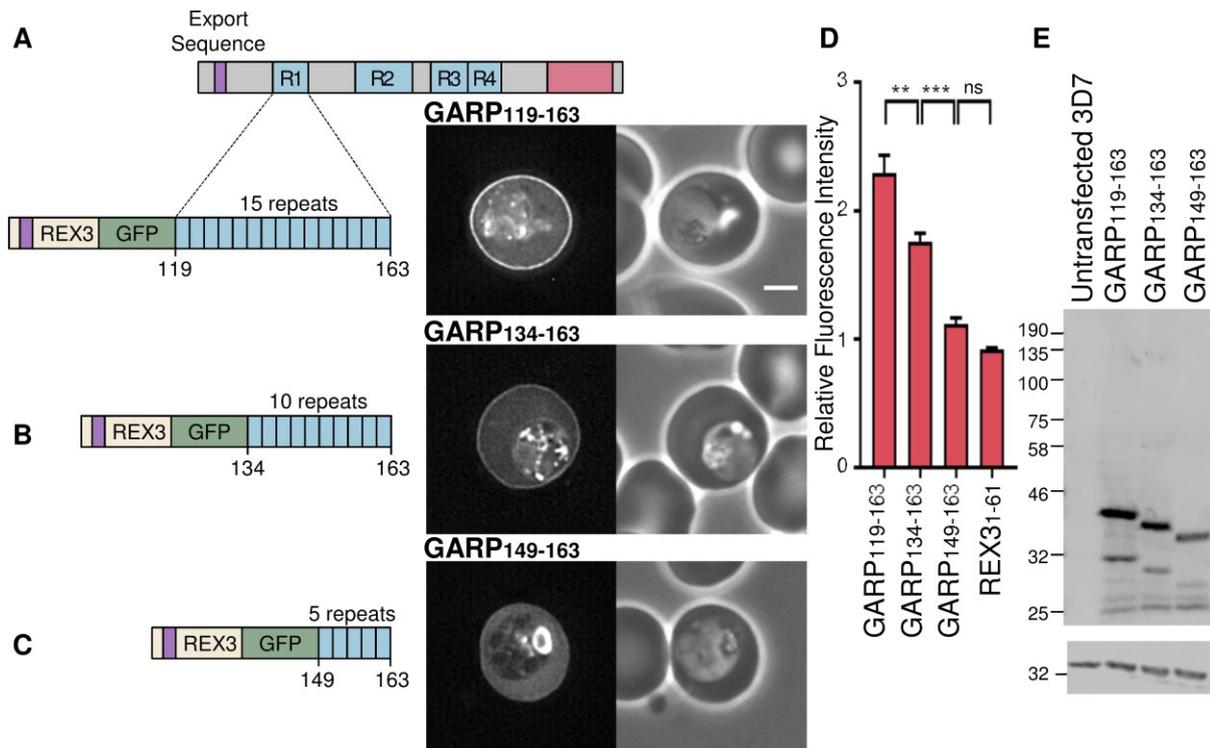


Figure 2

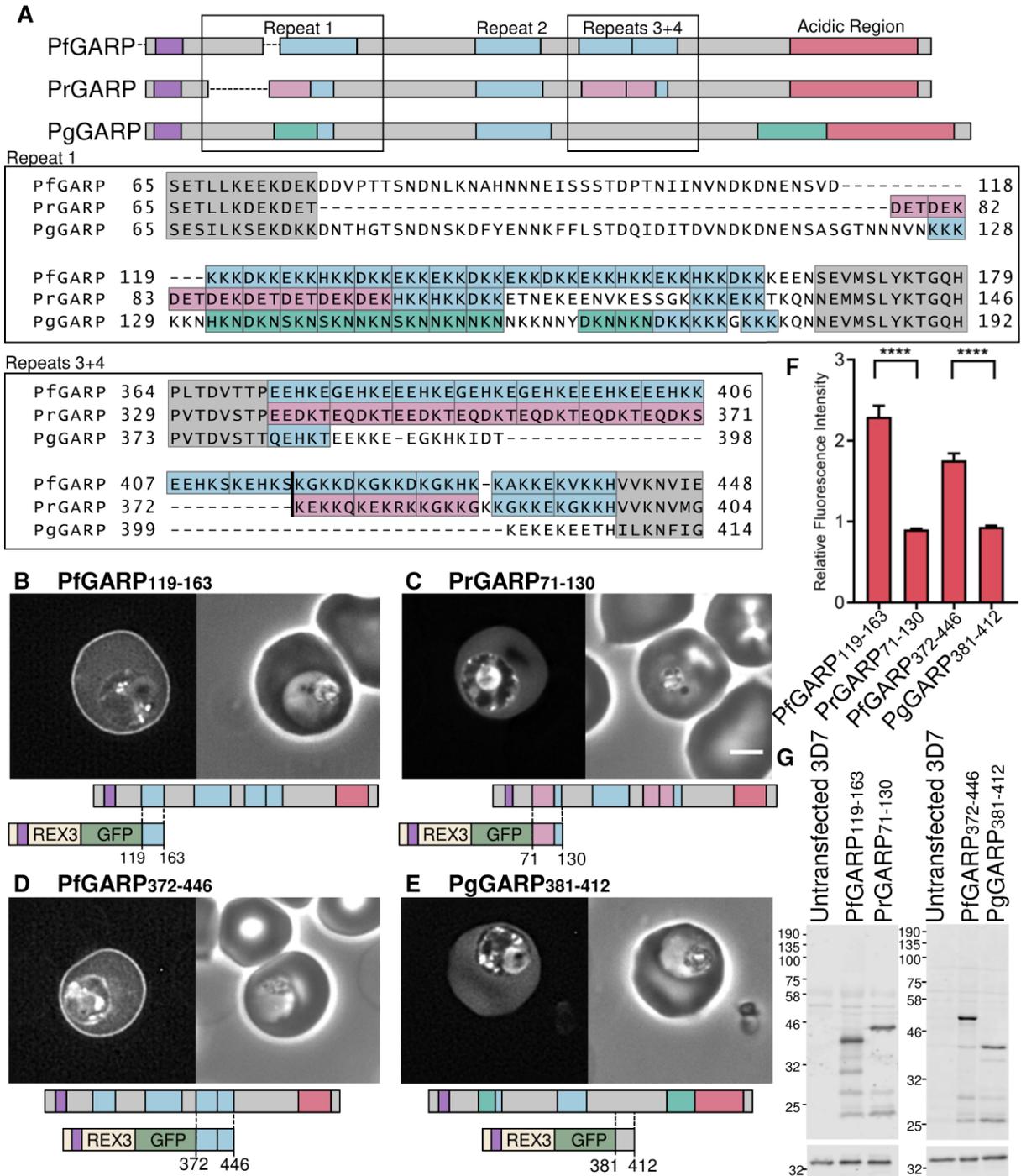


Figure 3

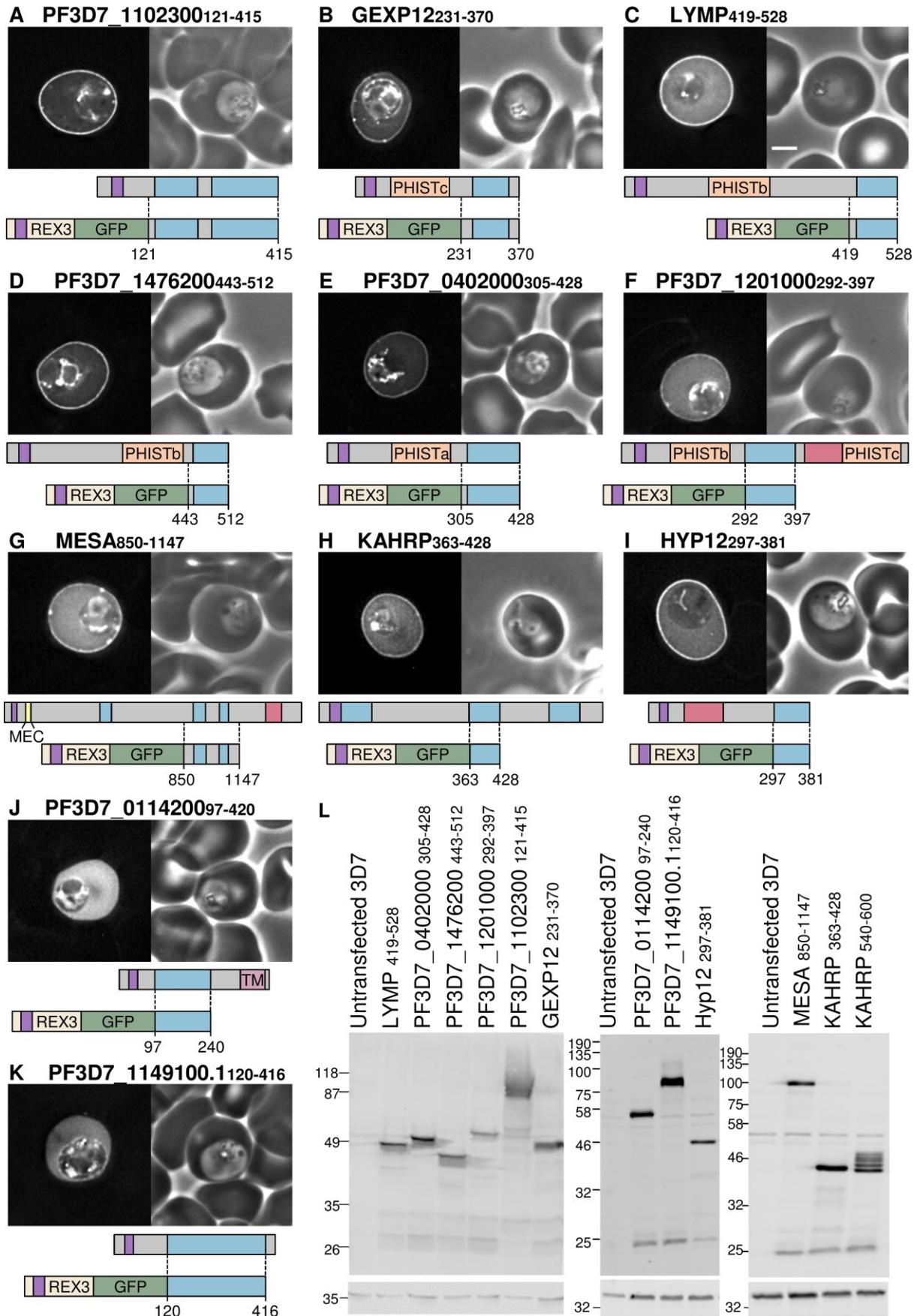


Figure 4

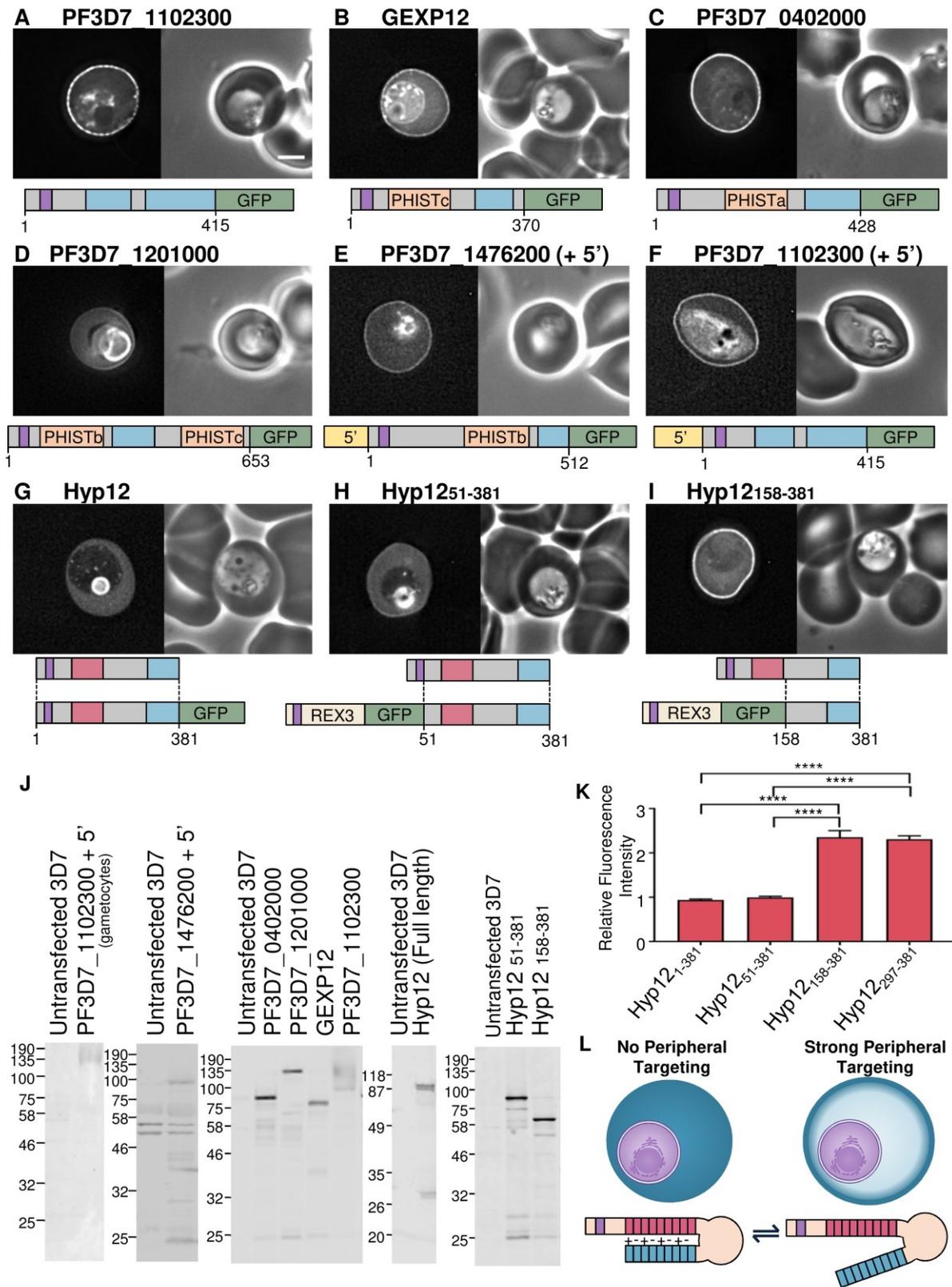


Figure 5

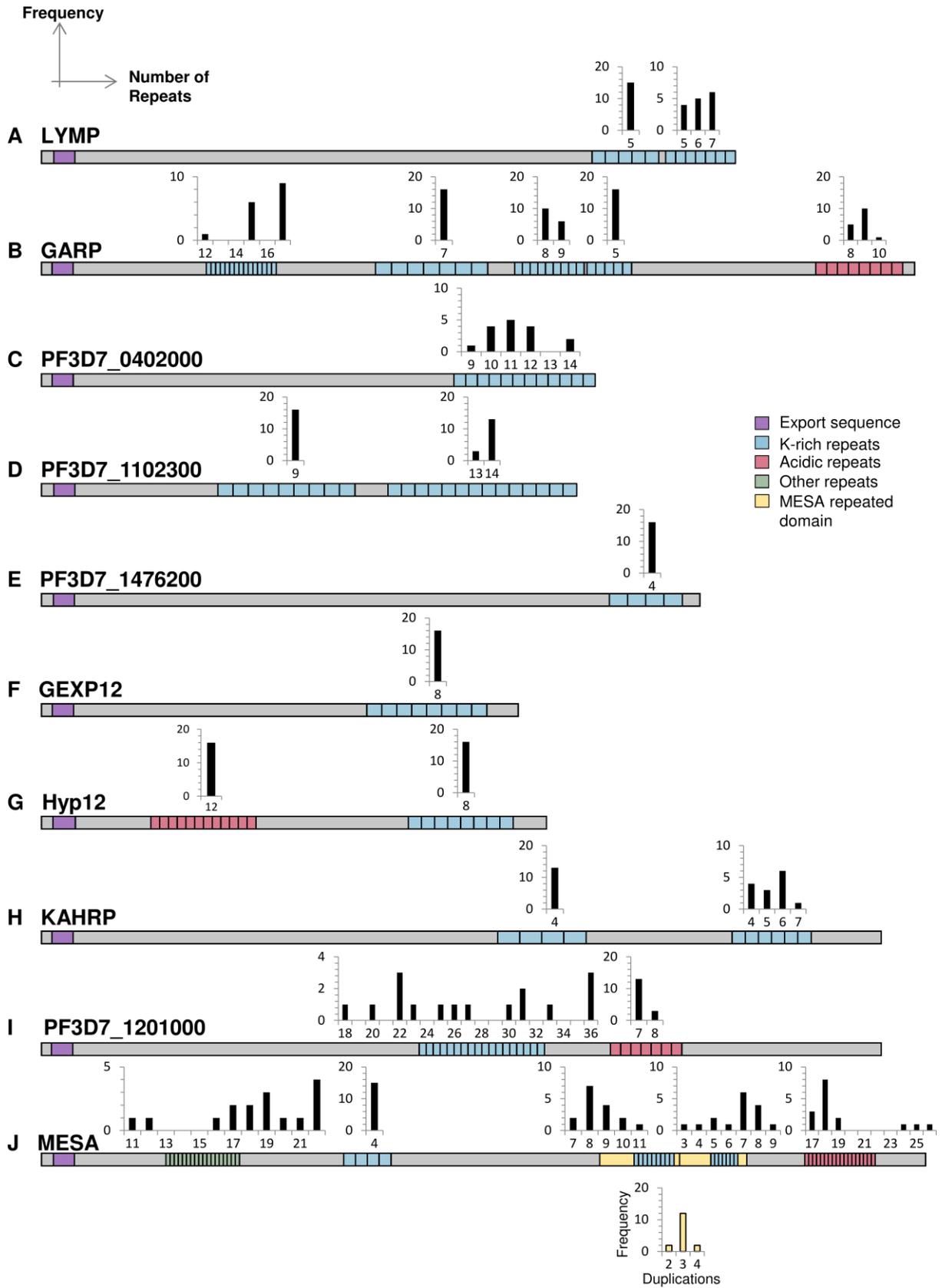


Figure 6

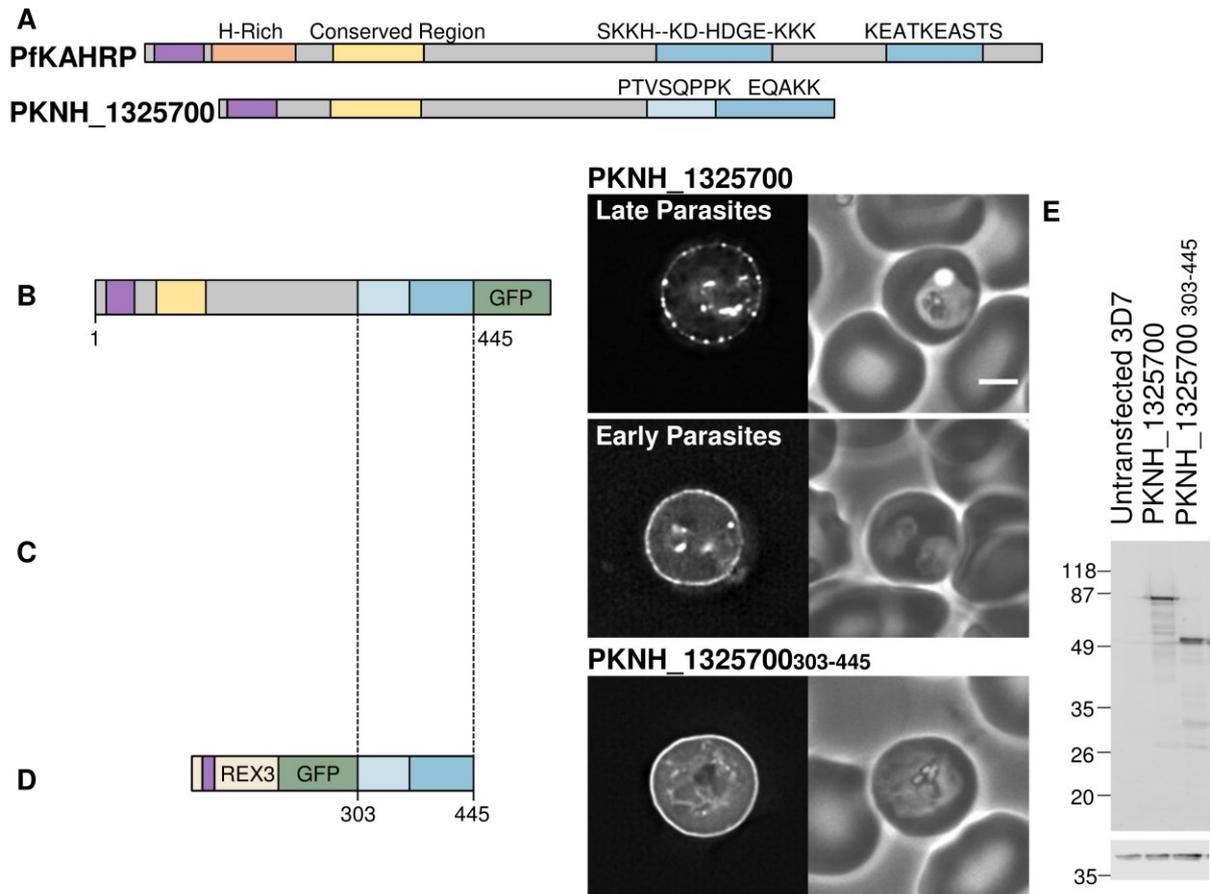


Figure 7

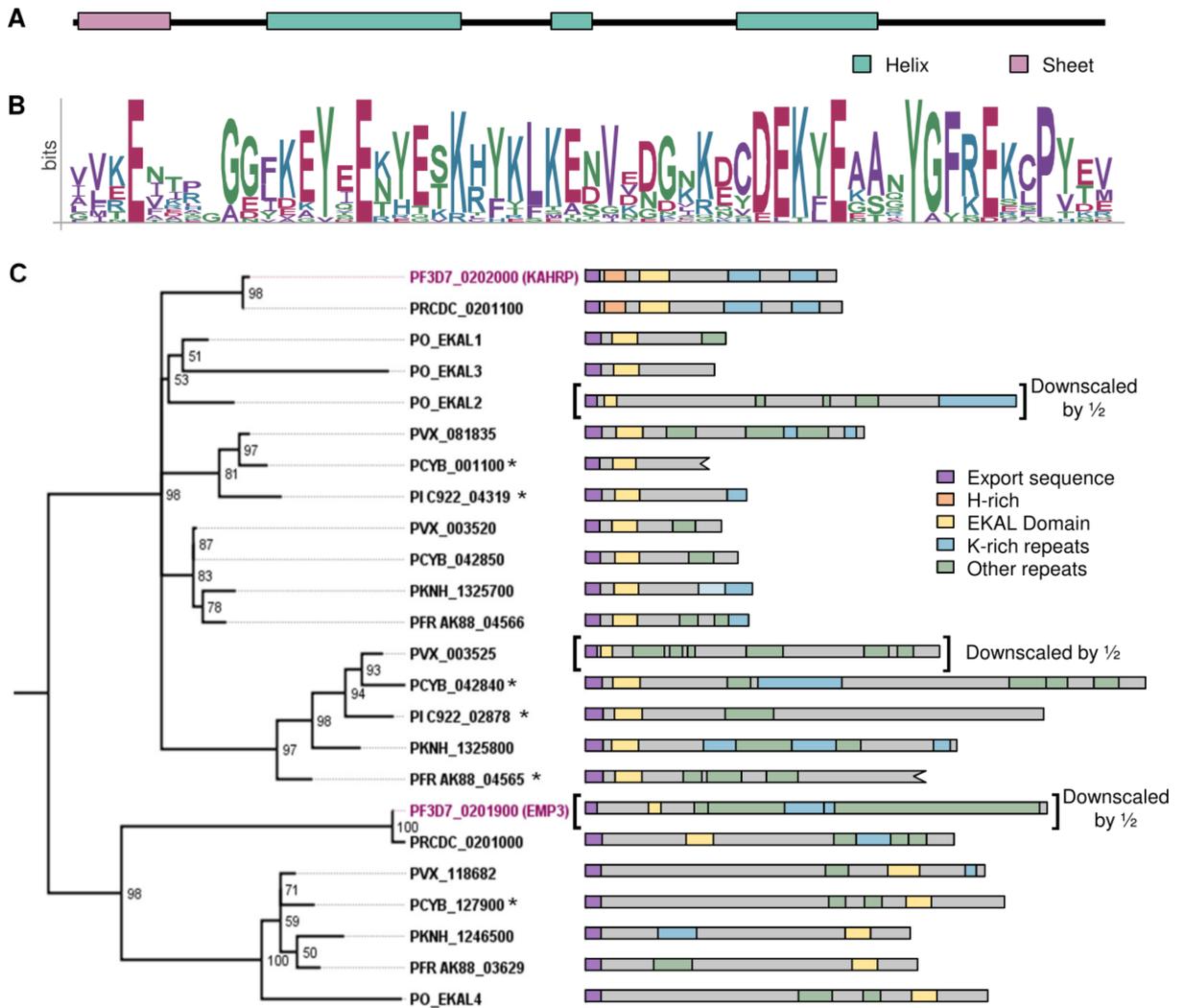


Figure 8

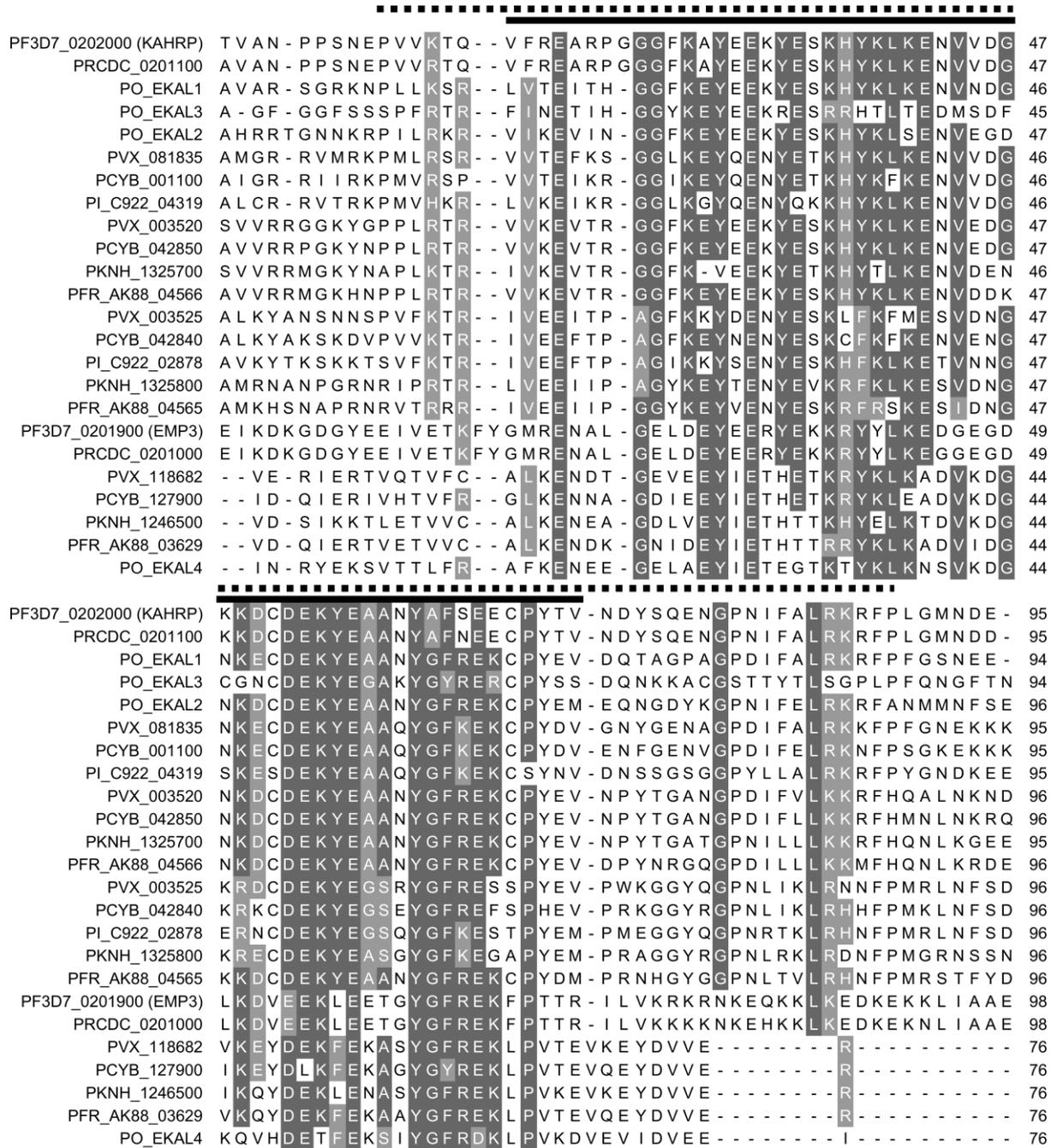


Figure 9

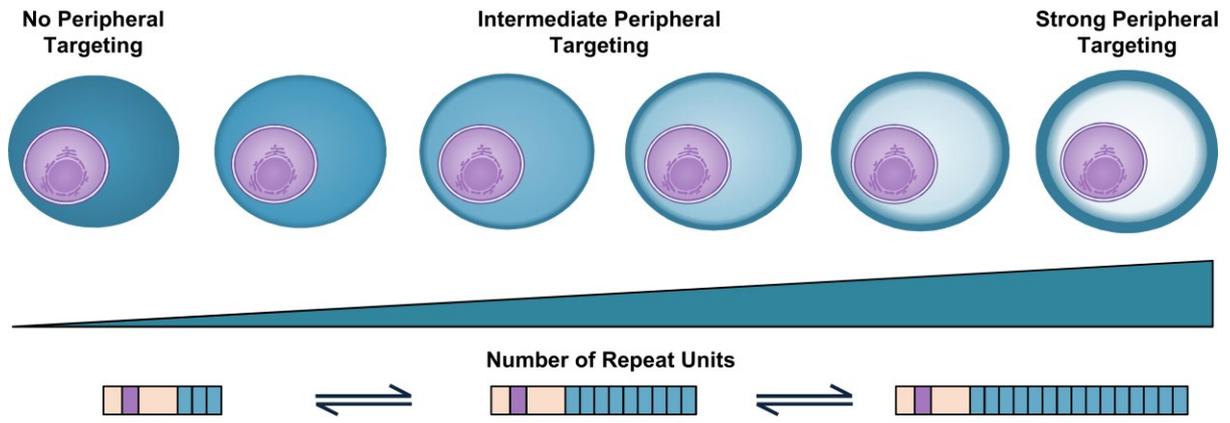
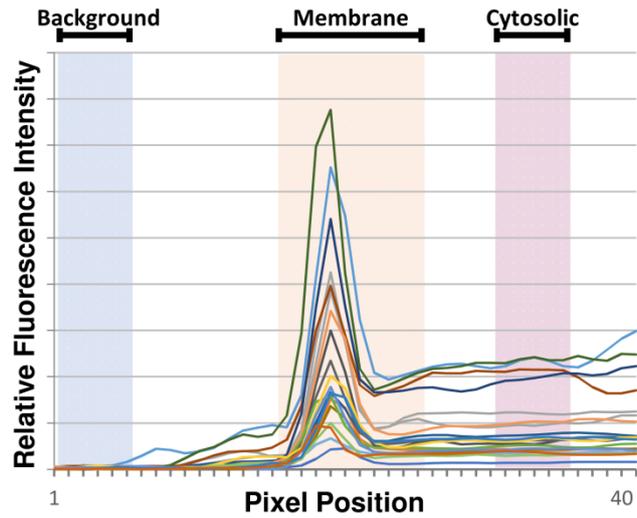
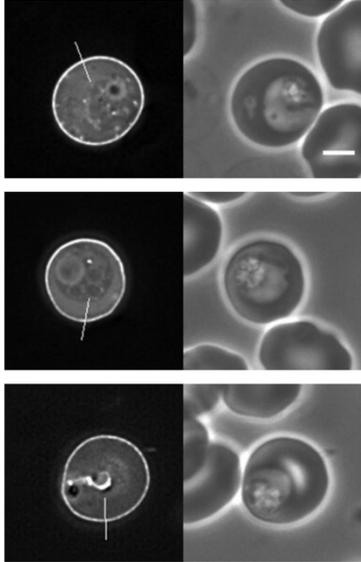


Figure 10

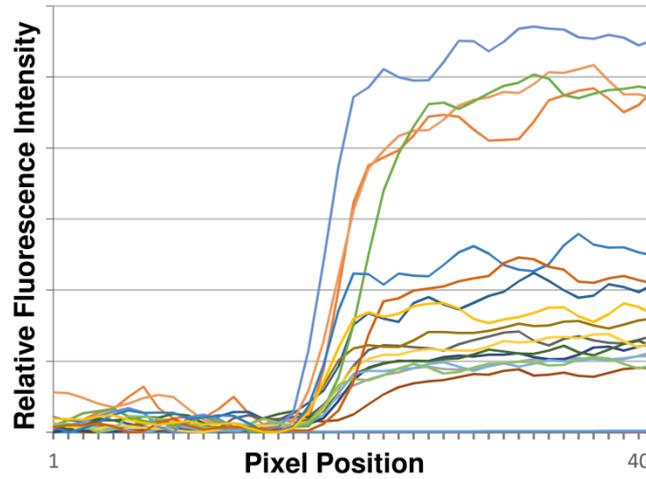
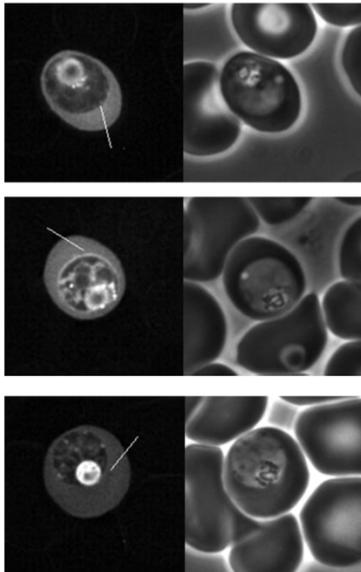


**A GARP:GFP**



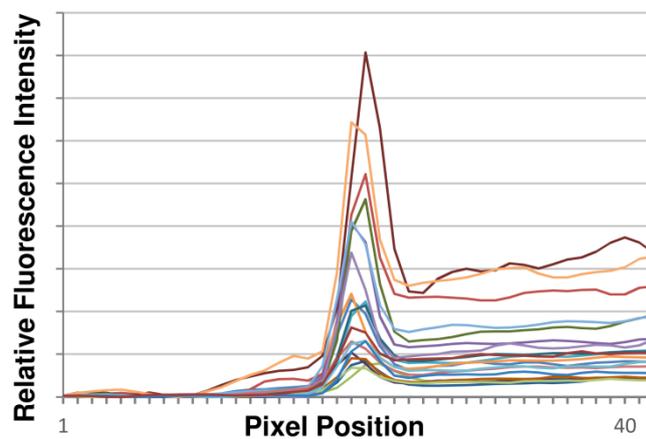
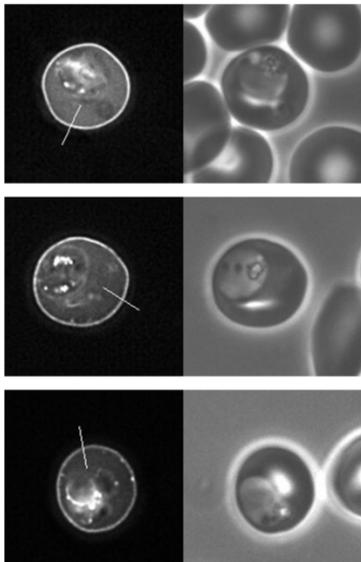
**Fold Difference at Membrane: 3.27**  
**Standard Deviation: 0.86**

**B REX31-61:GFP**



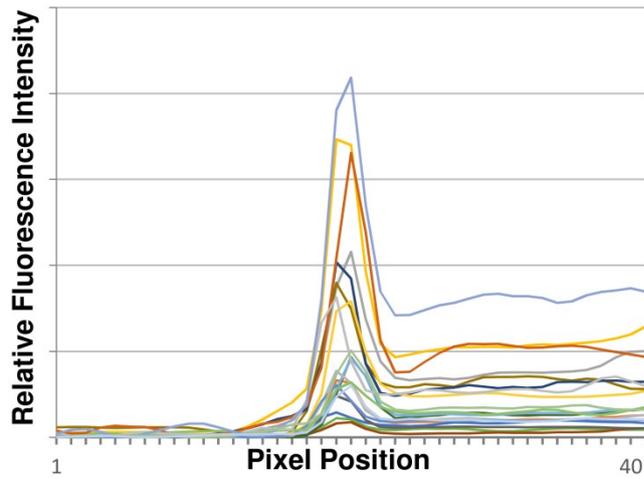
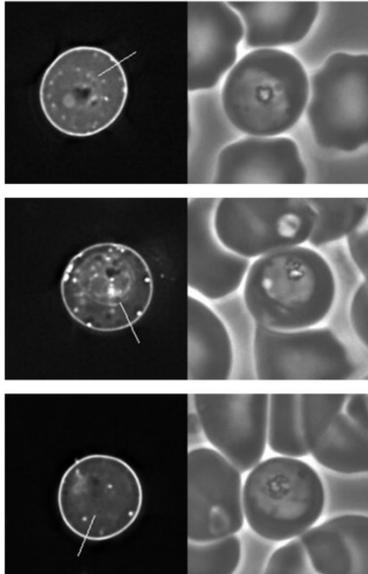
**Fold Difference at Membrane: 0.92**  
**Standard Deviation : 0.06**

**C GFP:GARP119-163**



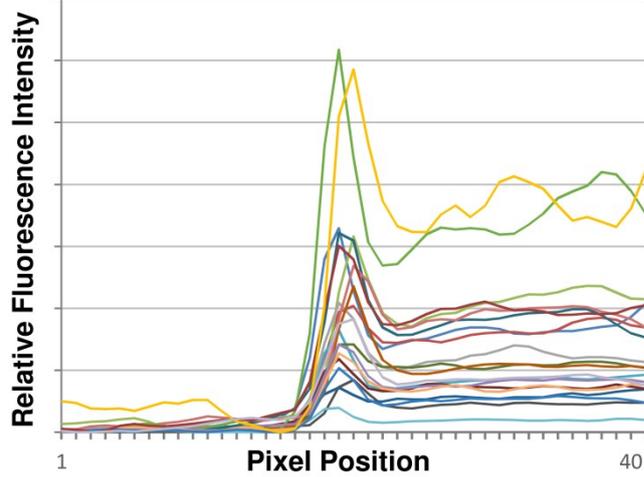
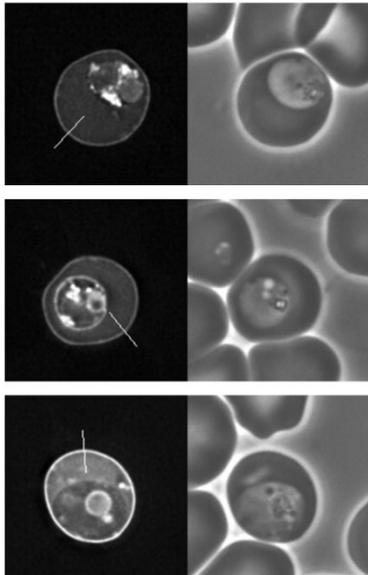
**Fold Difference at Membrane : 2.39**  
**Standard Deviation: 0.43**

**D** GFP:GARP253-340



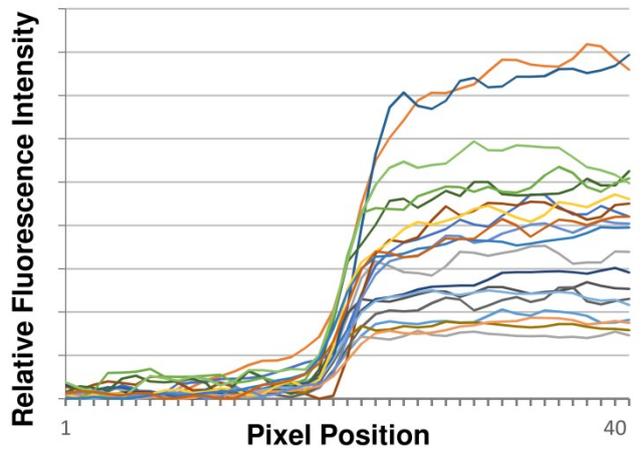
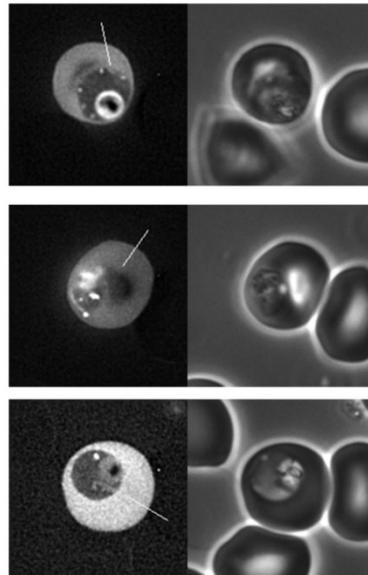
**Fold Difference at Membrane: 3.23**  
**Standard Deviation: 0.61**

**E** GFP:GARP372-446



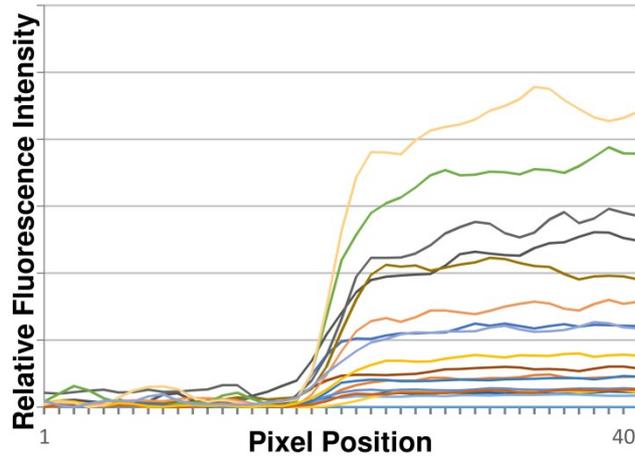
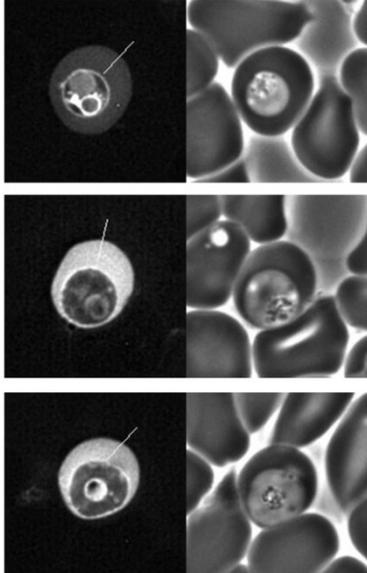
**Fold Difference at Membrane: 1.71**  
**Standard Deviation: 0.29**

**F** GFP:GARP535-673



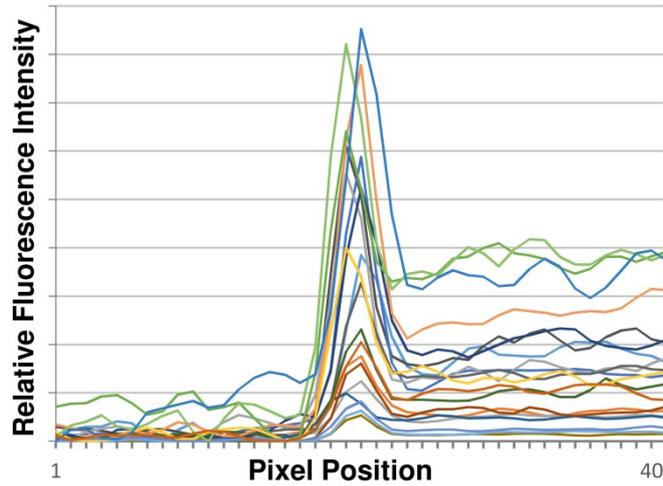
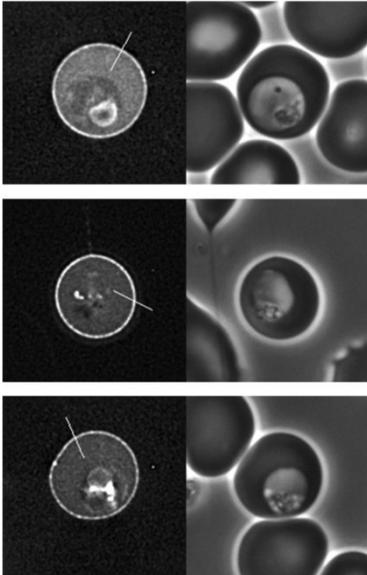
**Fold Difference at Membrane: 0.92**  
**Standard Deviation: 0.07**

**G** GFP:GARP50-118



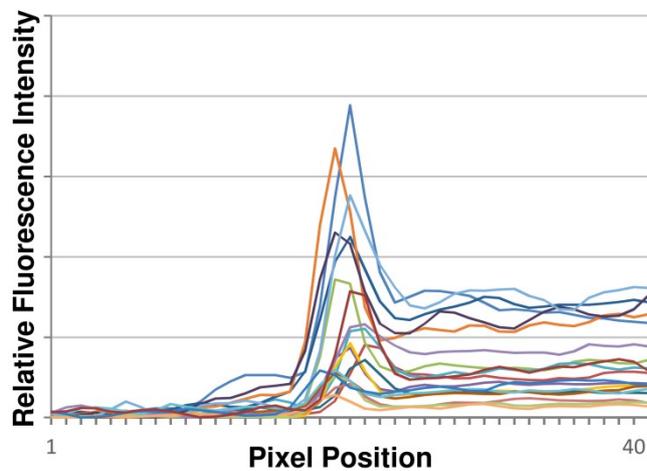
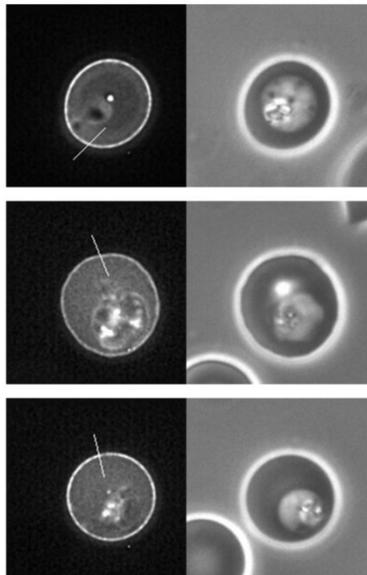
**Fold Difference at Membrane: 0.92**  
**Standard Deviation: 0.06**

**H** GARP:GFP + Promoter



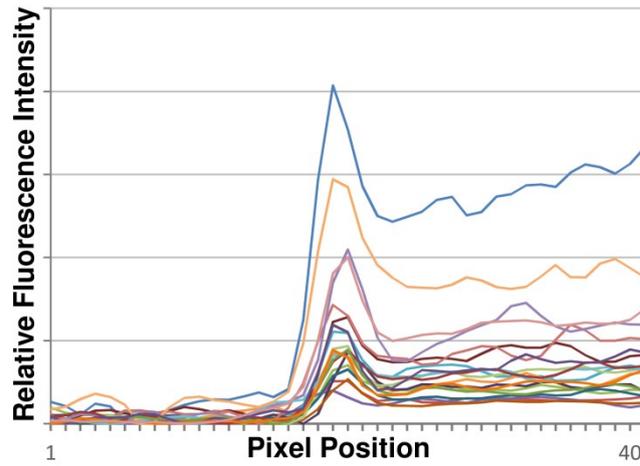
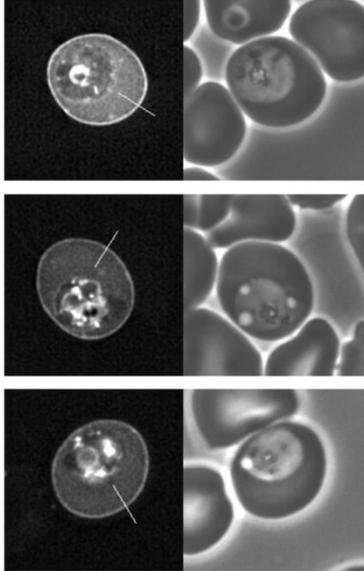
**Fold Difference at Membrane: 2.90**  
**Standard Deviation: 0.72**

**I** GFP:GARP119-163 + linker



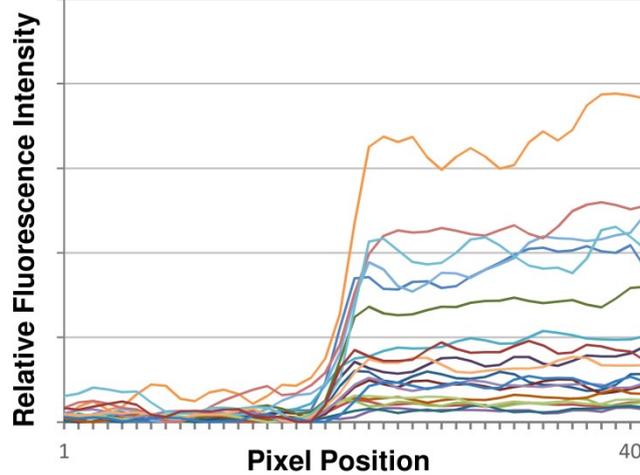
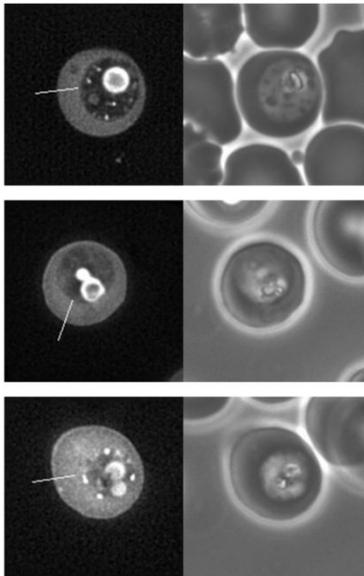
**Fold Difference at Membrane: 2.29**  
**Standard Deviation: 0.60**

**J** GFP:GARP134-163



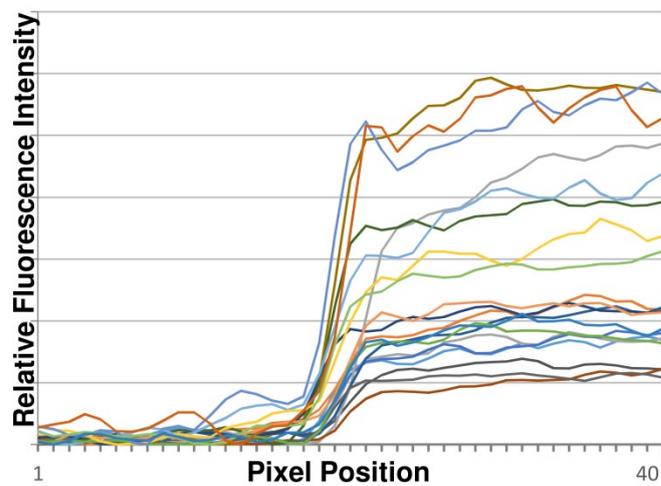
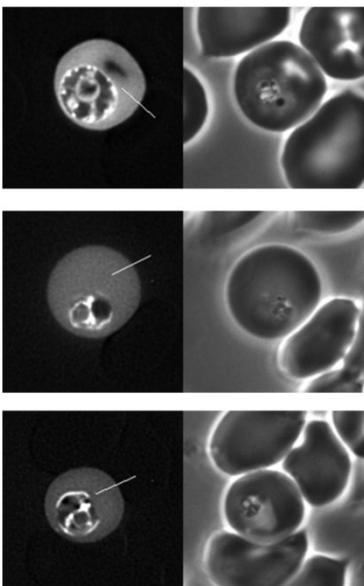
**Fold Difference at Membrane: 1.76**  
**Standard Deviation: 0.29**

**K** GFP:GARP149-163



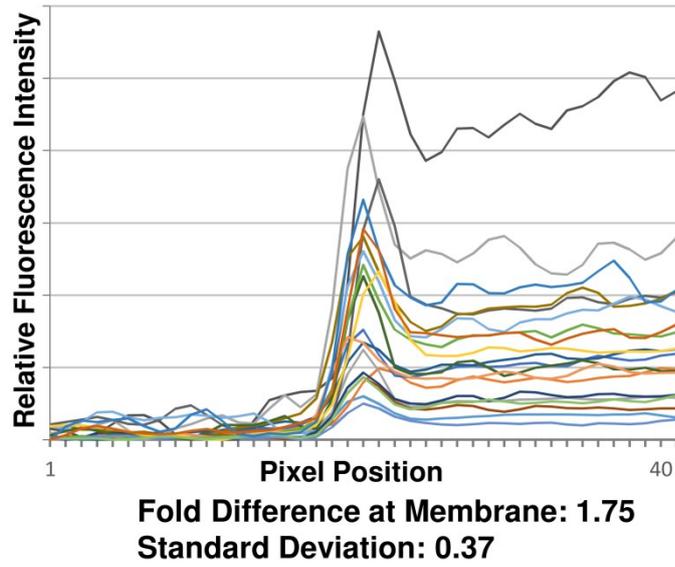
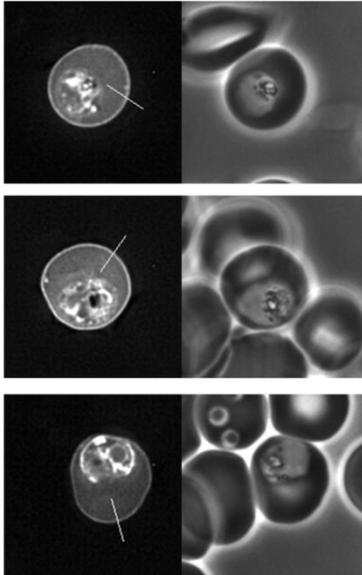
**Fold Difference at Membrane: 1.12**  
**Standard Deviation: 0.20**

**L** GFP:PrGARP71-130

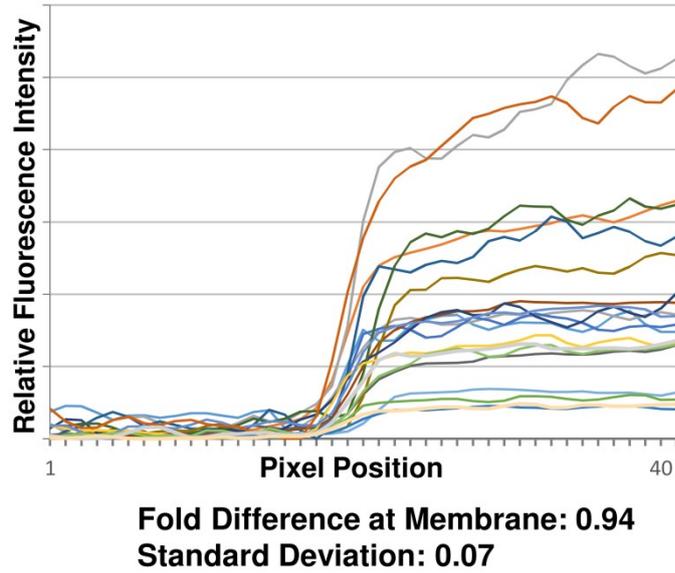
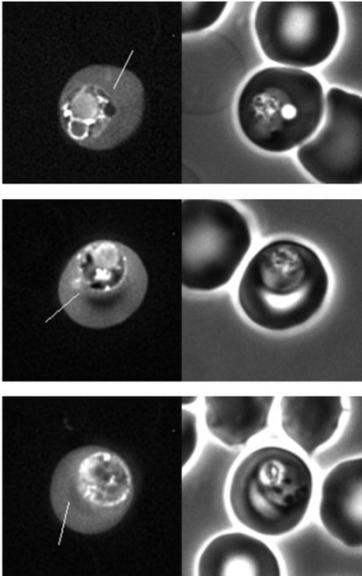


**Fold Difference at Membrane: 0.90**  
**Standard Deviation: 0.06**

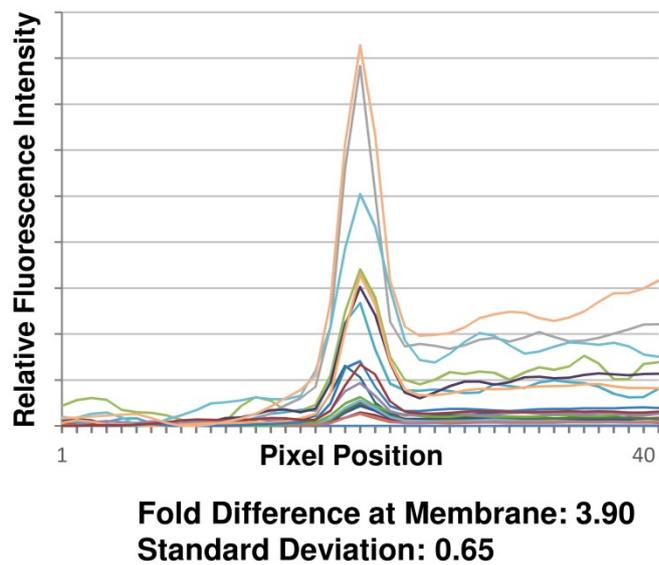
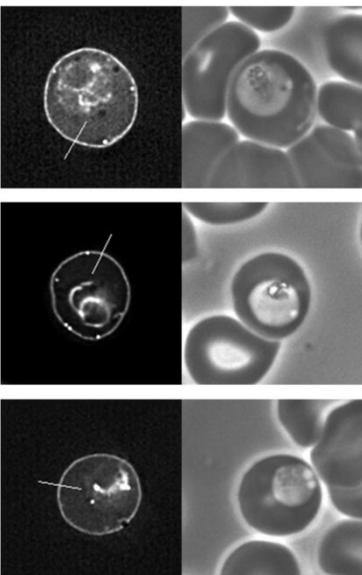
**M** GFP: PfGARP372-446 + linker



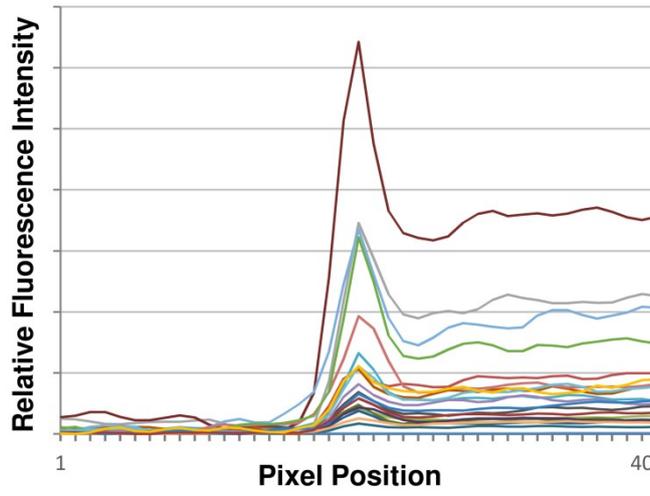
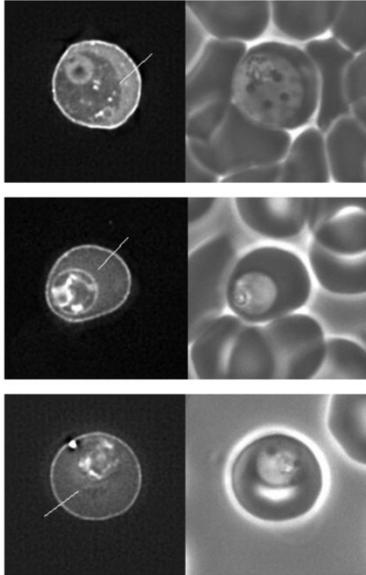
**N** GFP: PgGARP381-412



**O** GFP: PF3D7\_1102300121-415

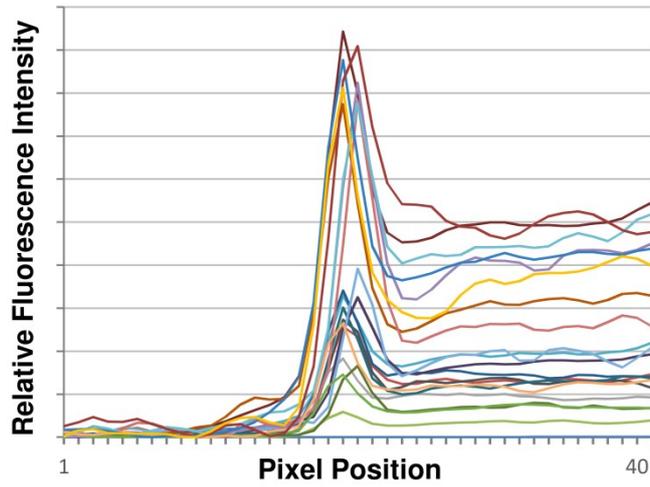
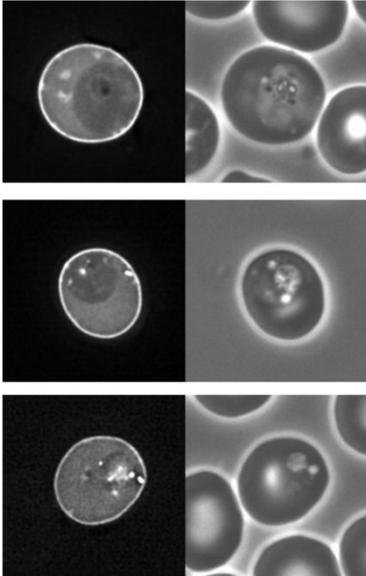


**P** GFP:GEXP12231-370



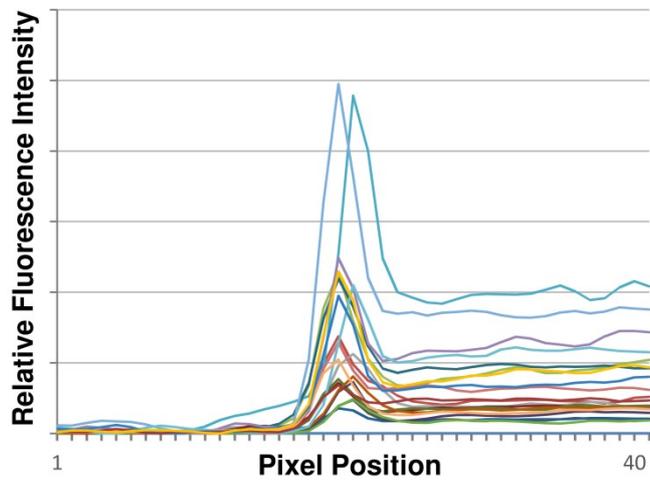
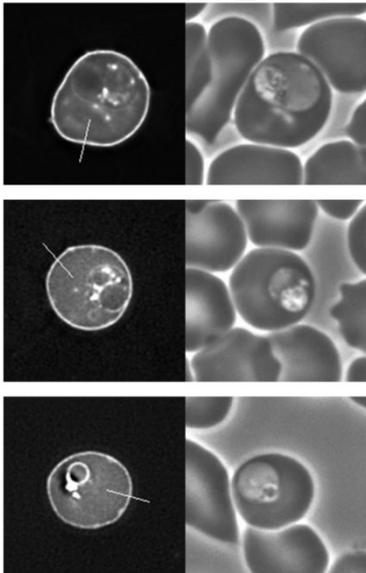
Fold Difference at Membrane: 1.75  
Standard Deviation: 0.35

**Q** GFP:LYMP419-528



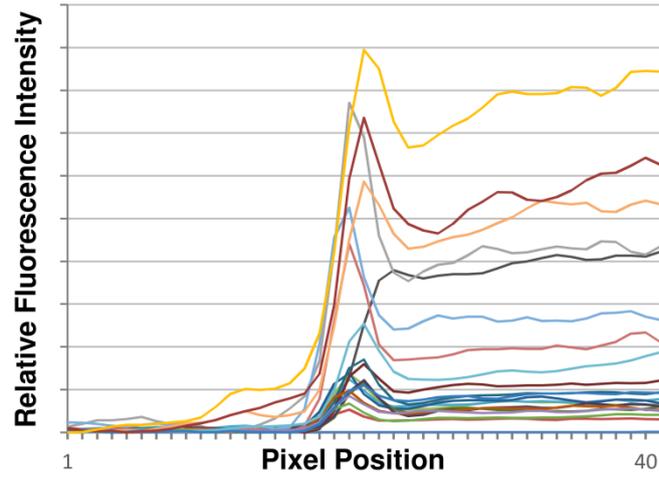
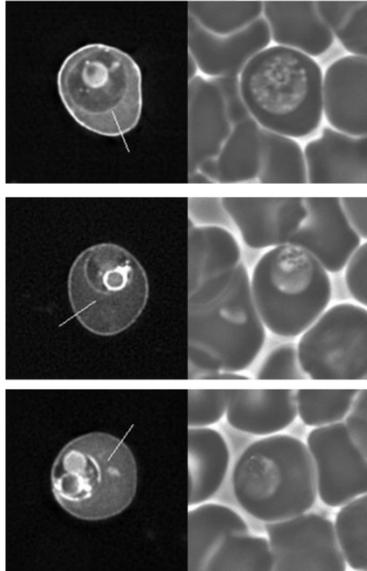
Fold Difference at Membrane: 2.13  
Standard Deviation: 0.34

**R** GFP:PF3D7\_1476200443-512



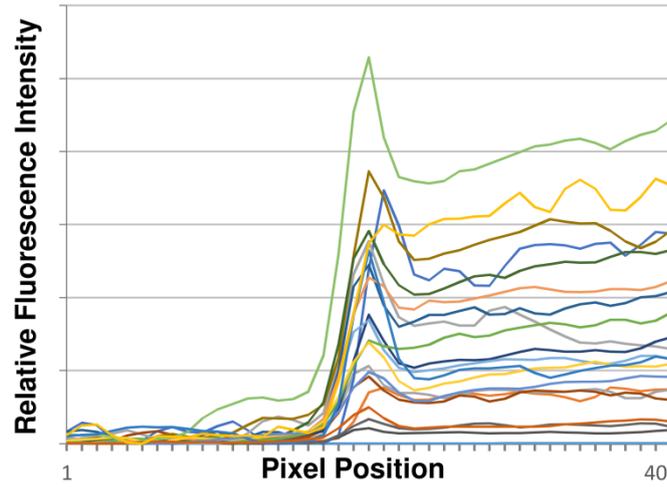
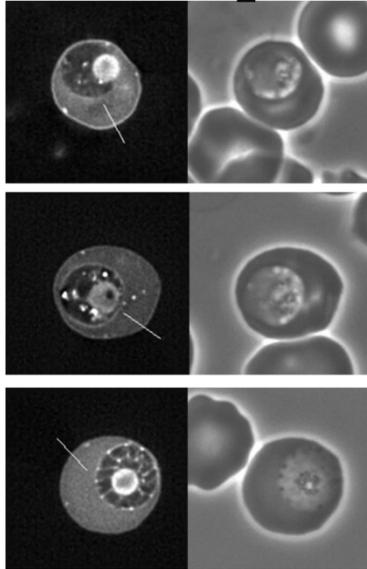
Fold Difference at Membrane: 2.50  
Standard Deviation: 0.57

**S** GFP:PF3D7\_0402000305-428



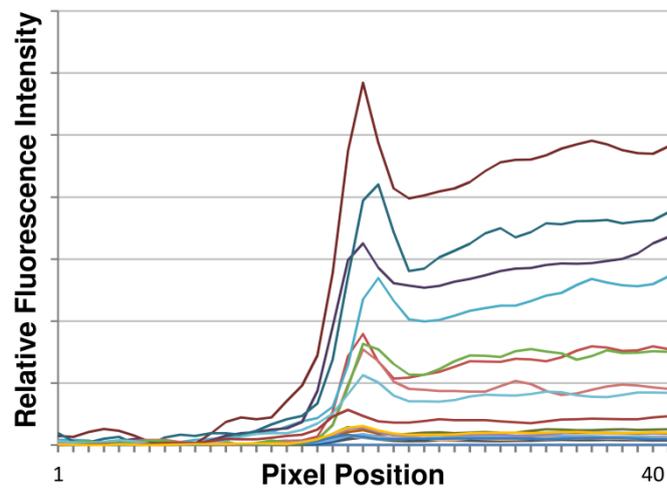
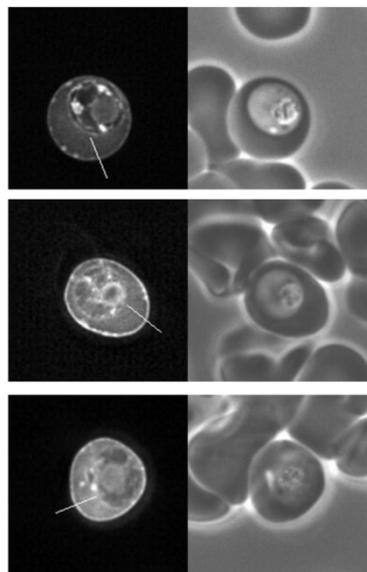
**Fold Difference at Membrane: 1.73**  
**Standard Deviation: 0.40**

**T** GFP:PF3D7\_1201000292-397



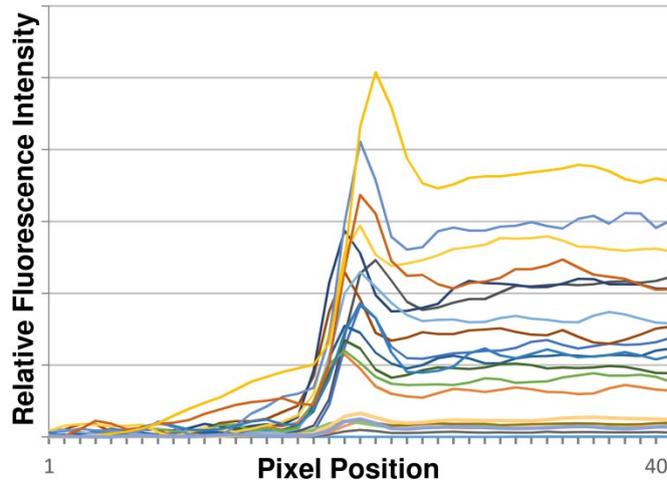
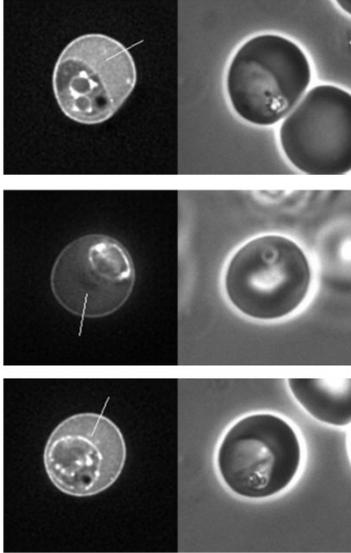
**Fold Difference at Membrane: 1.42**  
**Standard Deviation: 0.45**

**U** GFP:PfMESA850-1147



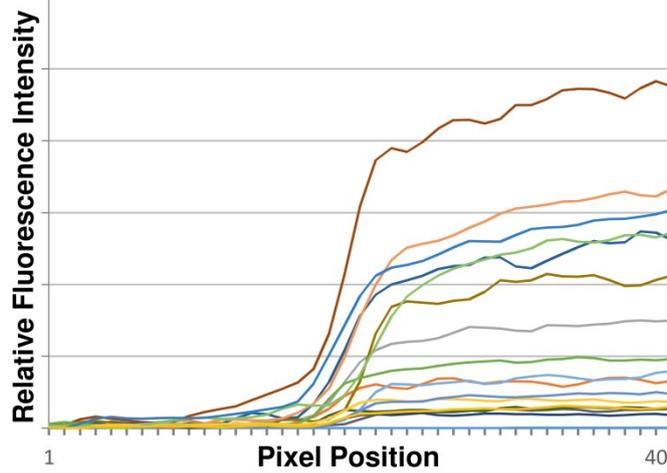
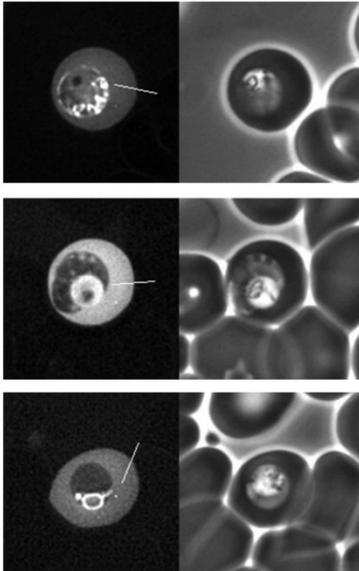
**Fold Difference at Membrane: 1.38**  
**Standard Deviation: 0.20**

V GFP:PfKAHRP363-424



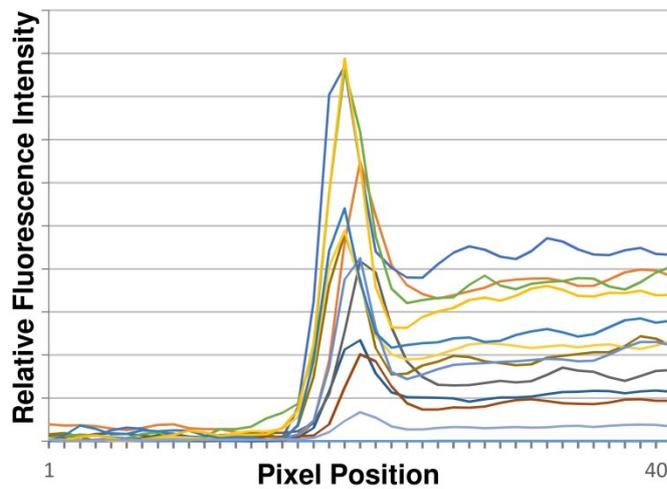
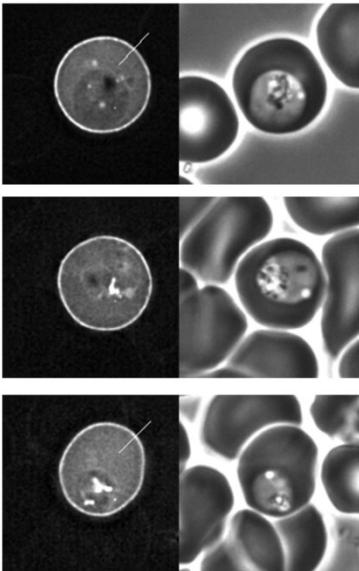
Fold Difference at Membrane: 1.49  
Standard Deviation: 0.21

W GFP:PfKAHRP540-600



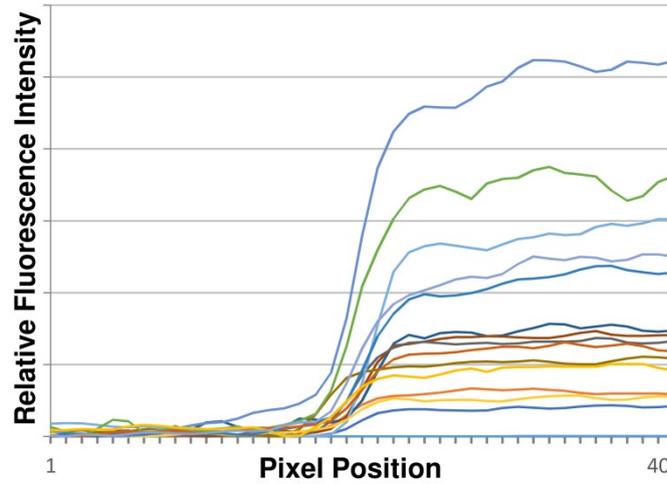
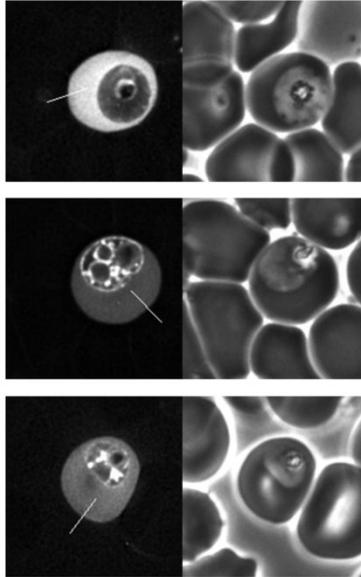
Fold Difference at Membrane: 0.91  
Standard Deviation: 0.08

X GFP:PfHYP12297-381



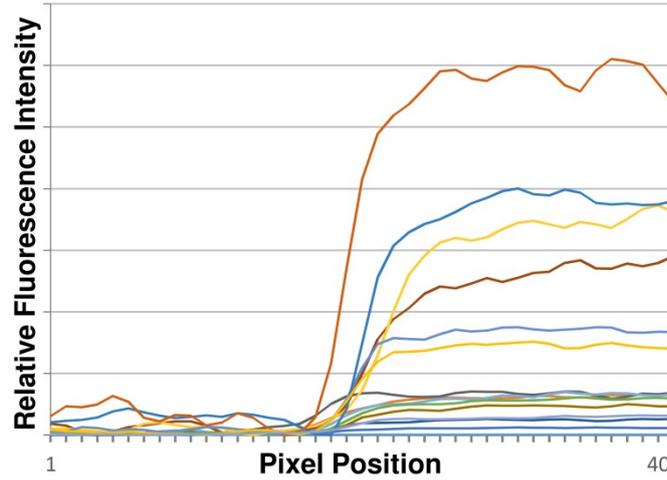
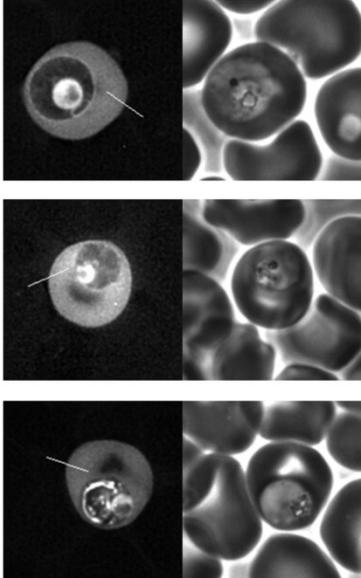
Fold Difference at Membrane: 2.310  
Standard Deviation: 0.33

**Y** GFP:PF3D7\_011420097-420



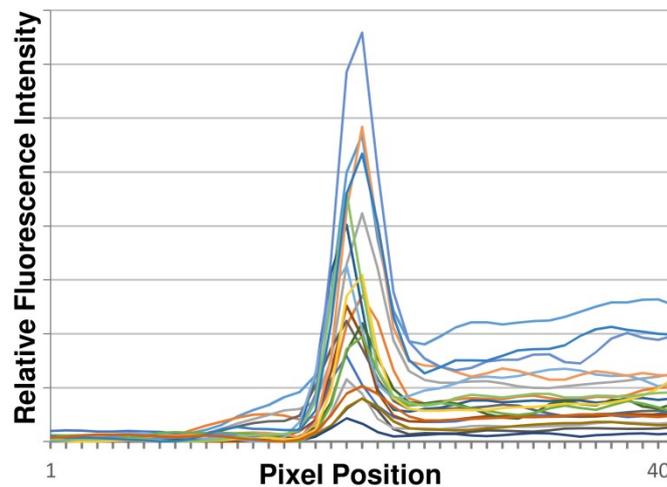
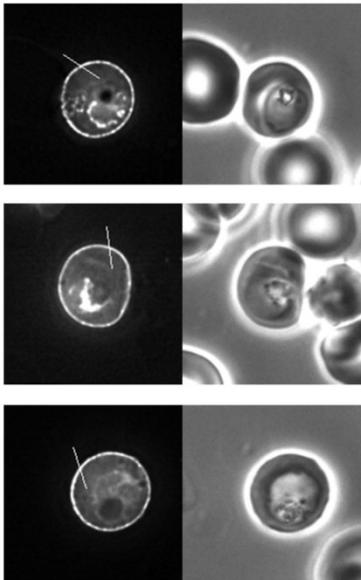
**Fold Difference at Membrane: 0.92**  
**Standard Deviation: 0.04**

**Z** GFP:PF3D7\_1149100.1120-416



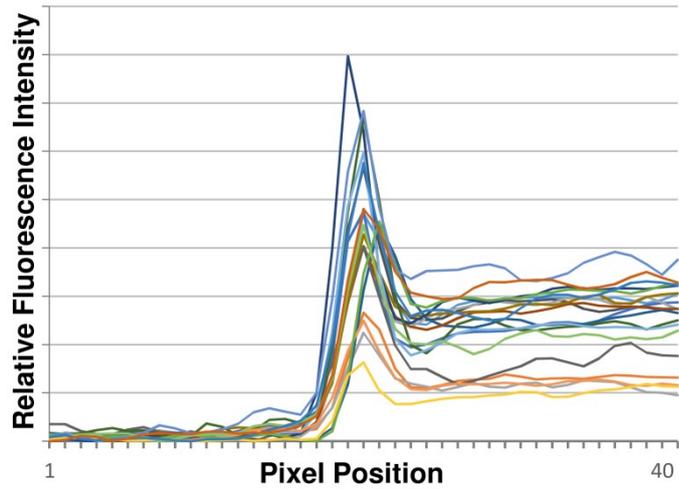
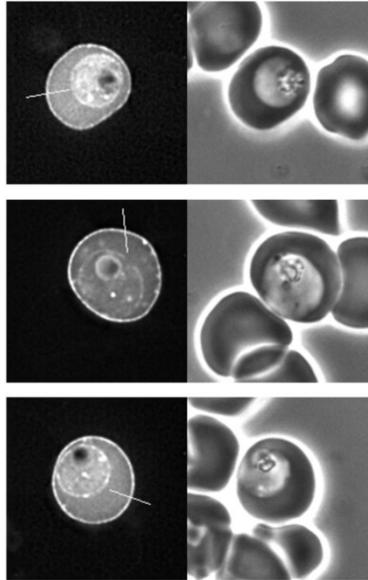
**Fold Difference at Membrane: 0.92**  
**Standard Deviation: 0.05**

**AA** PF3D7\_1102300:GFP



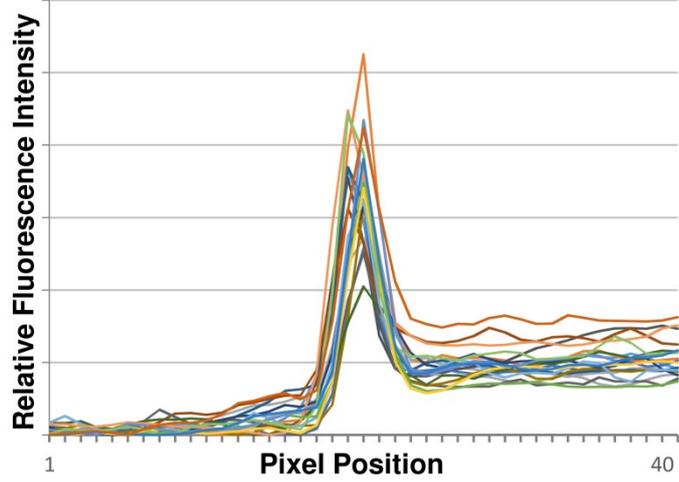
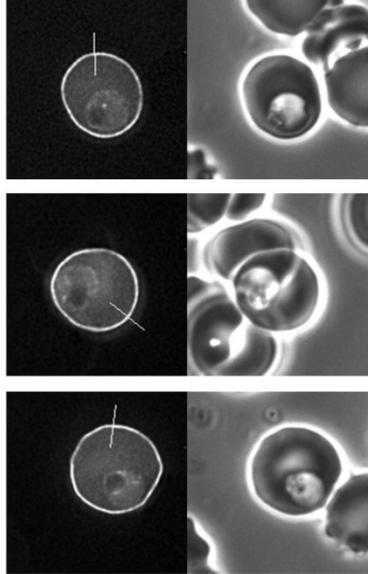
**Fold Difference at Membrane: 4.08**  
**Standard Deviation: 1.06**

AB GEXP12:GFP



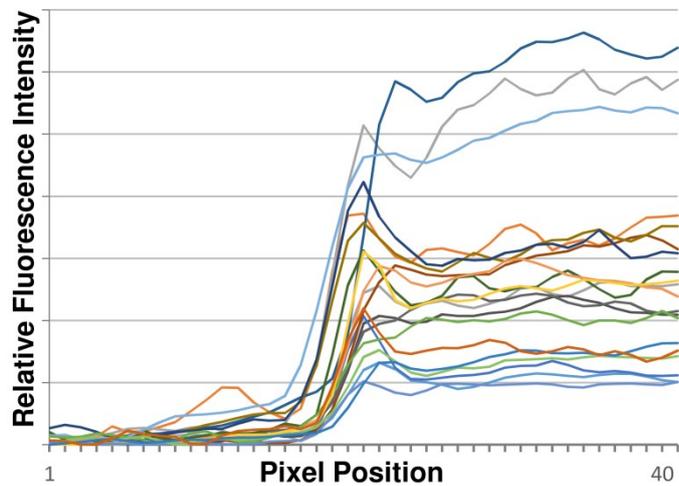
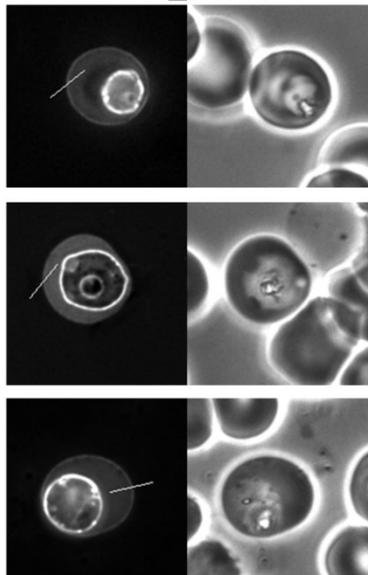
Fold Difference at Membrane: 1.97  
Standard Deviation: 0.44

AC PF3D7\_0402000:GFP



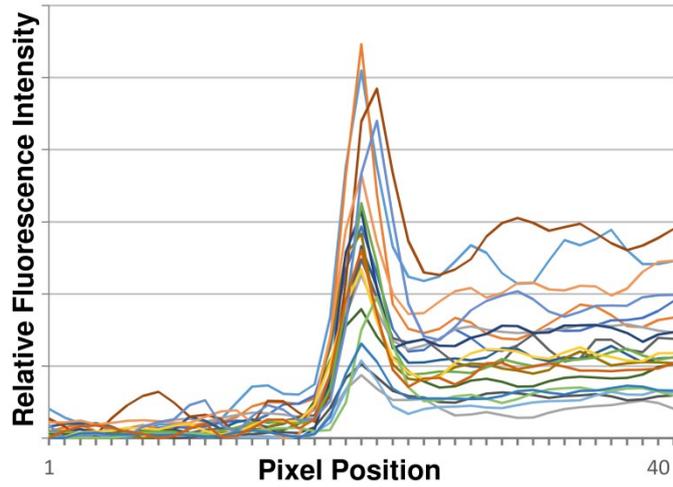
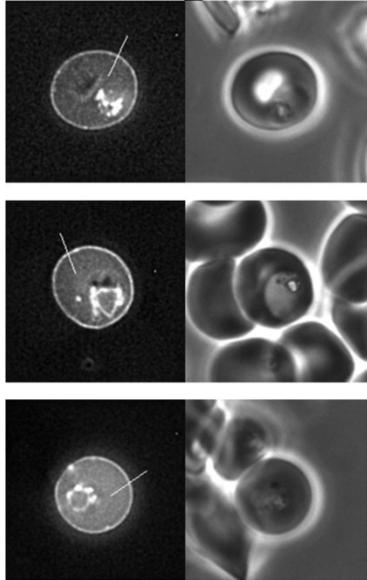
Fold Difference at Membrane: 3.73  
Standard Deviation: 0.80

AD PF3D7\_1201000:GFP



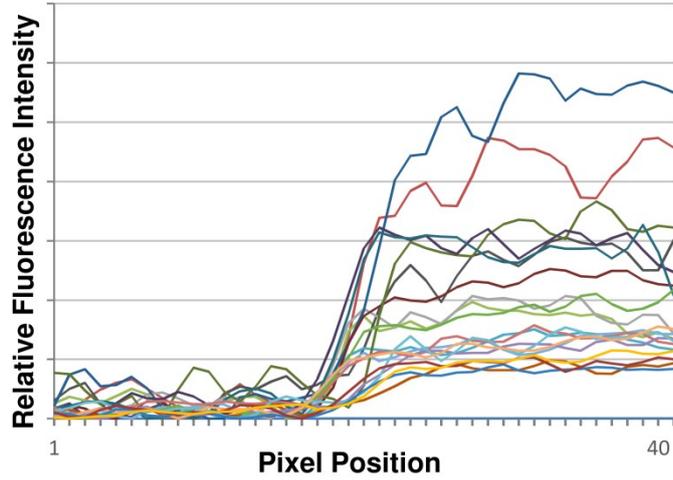
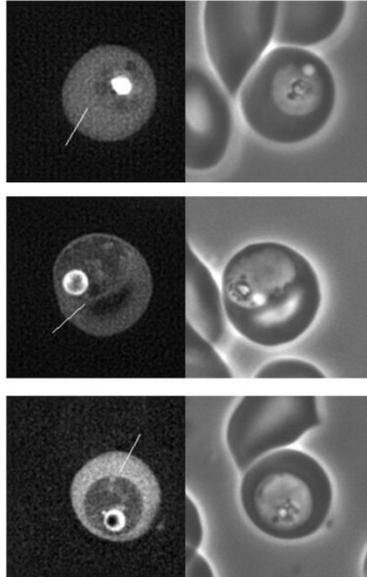
Fold Difference at Membrane: 1.09  
Standard Deviation: 0.20

**AE PF3D7\_1476200:GFP + promoter**



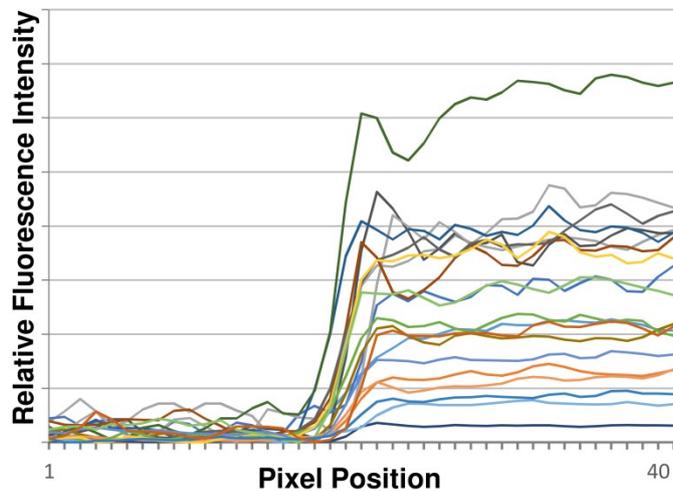
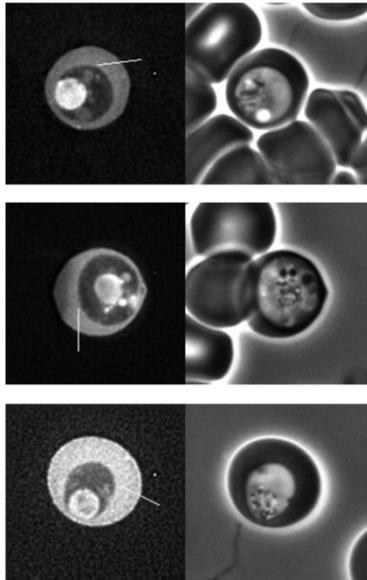
**Fold Difference at Membrane: 2.35**  
**Standard Deviation: 0.52**

**AF HYP12:GFP**



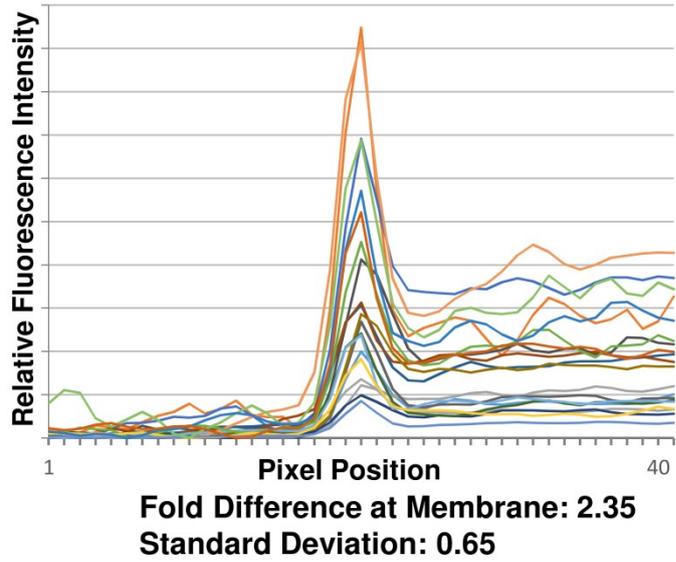
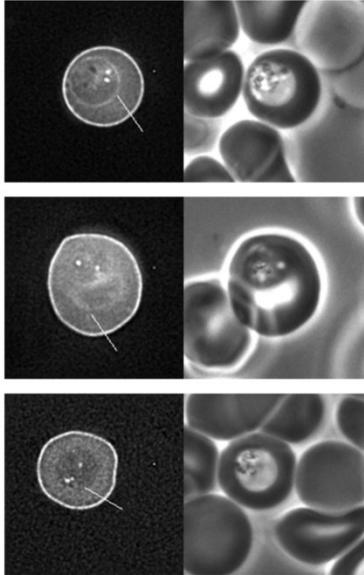
**Fold Difference at Membrane: 0.94**  
**Standard Deviation: 0.08**

**AG GFP:HYP12\_51-381**

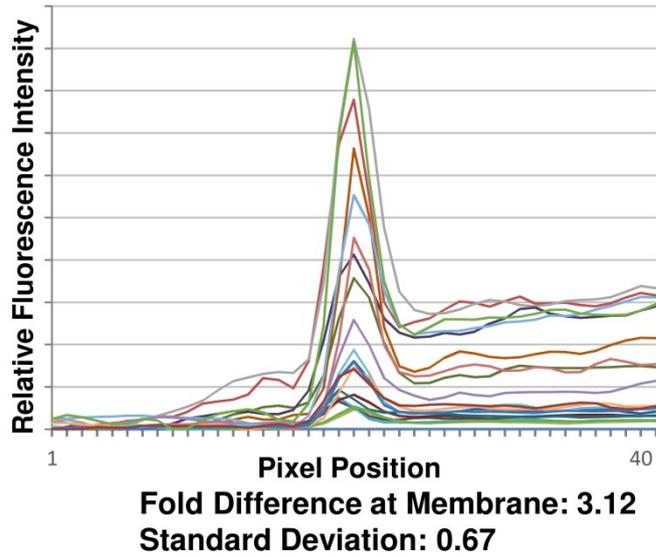
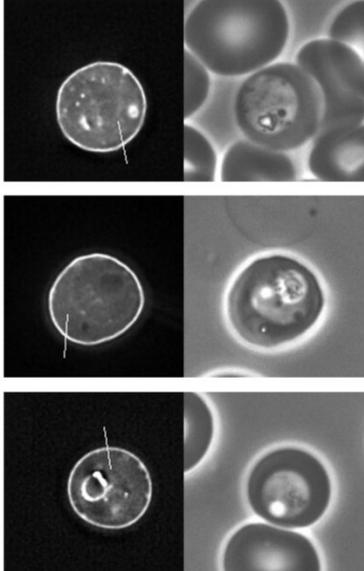


**Fold Difference at Membrane: 1.03**  
**Standard Deviation: 1.00**

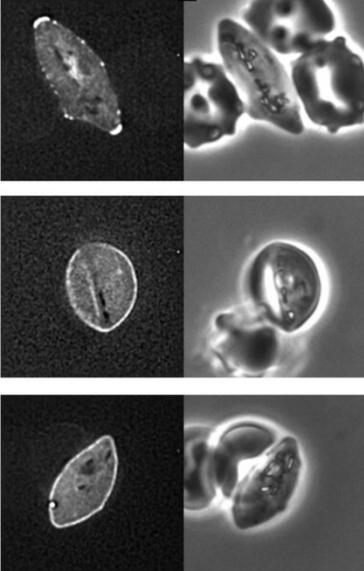
**AH** GFP:HYP12 158-381



**AI** GFP:PkKAHRP303-445



**AJ** PF3D7 1476200:GFP + promoter (gametocytes)



**Supplemental Figure 1 - Quantification of GFP fluorescence at the periphery of infected erythrocytes.** (A-AI) Representative GFP fluorescence and phase contrast images are shown on the left and right panels, respectively. ImageJ was used to plot a profile of fluorescence signal intensity across the periphery of the red blood cell as described previously (1). Images were cropped to 170 x 170 pixels and a 40 pixel line was drawn starting outside the cell and ending within the erythrocyte cytoplasm, avoiding the parasite (shown by white lines in fluorescence images). The centre of the line was placed approximately at the erythrocyte membrane. The background signal was taken as the average intensity of the first 5 pixels, while the cytoplasmic signal was calculated as the average intensity of pixels 30-35. The peripheral fluorescence was determined as the maximal signal intensity between pixels 15-25, as indicated. The background signal was subtracted from both cytoplasmic and membrane signals and the difference in fluorescence at the membrane was calculated as the ratio of the normalised membrane to cytosolic signal. This ratio was calculated for 20 infected cells from two separate experiments; the fluorescence intensity profiles of these cells are shown. The average and standard deviation is indicated. (AJ) Additional images of gametocytes at various stages expressing GFP-tagged PF3D7\_1102300.

Gene ID	Alias	Consensus Sequence	Position Within Protein	Repeat Unit Length	Number of Repeat Units	Error from Consensus
PBANKA_0524700		EIYMENKKEEQQSKKKEKVI	466-577	21	5.33	0
PBANKA_0711200		EIKKTPTT-	233-355	8	15.25	0.09
PBANKA_0112600	fam-b	NDKKTCV	149-193	7	6.43	0.02
PBANKA_1145400		EKSEKSKKKSEKSEKSEKSEKS	217-295	24	3.29	0.14
PBANKA_0214600		LKNDVANKTQK	799-839	11	3.73	0.02
PCHAS_0404100	CSP	PGDK	79-175	4	24.25	0.11
PCHAS_0521700		KENGEKVT	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_1246800		EYKS	179-215	4	9.25	0.16
PCHAS_0318300		EEK-VE	1832-1900	5	16.60	0.01
PCHAS_1370100		DGKKIFEEKKES	169-201	12	2.75	0.15
		EKKSSNEKKTGP	199-231	12	2.75	0.06
PCYB_051130		EKKAET	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_0300600		D-KA-KK-EA	131-284	7	20	0.23
PKNH_0400300		KEEV	191-288	4	24.5	0.07
PKNH_0623000		GKKECPFKAQNSSEKCA	462-703	18	13.44	0.04
PKNH_0807900		KKEE	69-110	4	10.5	0.1
PKNH_0841300		EKGAQKPGQKQVQEKKDSN	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		PPKGTKKKTPTEETEQA	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		ETEKQDKPKYTYGSYKPTVK	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	KRQKHKEQRDKNIQKIIKDKR	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	GLKENAELKNKELRNKGS	694-793	19	5.26	0.03
		KNKDI	796-818	5	4.6	0.26
PRCDC_0201100	KAHRP	GE-KKKSCKKND-	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	ENKKGAS	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	1210-1444	22	10.68	0.2
PRCDC_1001500	PTP5	NETEKKTDQ	224-262	9	4.33	0.05
PRCDC_1037500	GSP	EKEEKIKKKKVVIEKKK	1513-1547	16	2.19	0.28
		E-PK-KEK--AP	1623-1776	8	18.75	0.21
		KDVKAKHK	1550-1566	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
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_1475300	PHISTB	KKEEDV	372-404	6	5.5	0.09
PRCDC_1475600		NKEENKDN	471-505	8	4.38	0.06
PVX_002507	Pv-fam-b	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	1010-1061	13	4	0.13
PVX_110835		EGDQ--D-GK-EDKGEEDGK	258-297	18	2	0.22
		CPYKDQSVDKKE	772-823	12	4.33	0.1
		KKTANVKKGAEP	1200-1226	12	2.25	0.04
		DK-D-KDDK	293-337	7	6	0.32
		EEEAKKL	1146-1195	7	7.14	0.16
PVX_118682	EMP3	EAKKPEVKKT	1001-1030	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	KDDLPEEEK	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		EIDKSIKKEEHIKK-	120-173	15	3.6	0.26
PY17X_1203700		QVTDK	1242-1313	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

**Supplemental Table 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 (2).

<b>Table 2A</b>		
<b>Gene Name</b>	<b>Location of Potential Error</b>	<b>Nature of Potential Error</b>
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)
<b>Table 2B</b>		
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)

**Supplemental Table 2 – Annotation of introns and location of potential frameshift mutations.**

(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). (B) EKAL-domain proteins modified from their deposited protein sequences. 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\_AK88\_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.

## **Supplemental Material References**

1. Tarr, S. J., Moon, R. W., Hardege, I., and 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
2. Newman, A. M., and Cooper, J. B. (2007) XSTREAM: A practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *Bmc Bioinformatics* **8**, 19