Published in final edited form as:

Infect Genet Evol. 2013 December ; 20: 239–248. doi:10.1016/j.meegid.2013.09.011.

The evolution and diversity of a low complexity vaccine candidate, merozoite surface protein 9 (MSP-9), in *Plasmodium vivax* and closely related species

Stella M. Chenet^{a,b,*}, M. Andreína Pacheco^{b,*}, David J. Bacon^c, William E. Collins^d, John W. Barnwell^d, and Ananias A. Escalante^{a,b,§}

Stella M. Chenet: schenet@asu.edu; M. Andreína Pacheco: Maria.Pacheco@asu.edu; David J. Bacon: david.bacon@nrl.navy.mil; William E. Collins: wec1@cdc.gov; John W. Barnwell: wzb3@cdc.gov

^aArizona State University, School of Life Sciences, Tempe, AZ, USA

^bCenter for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, Tempe, Arizona, USA

^cNaval Research Laboratory, Washington, DC, USA

^dMalaria Branch, Division of Parasitic Diseases and Malaria, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

Abstract

The merozoite surface protein-9 (MSP-9) has been considered a target for an anti-malarial vaccine since it is one of many proteins involved in the erythrocyte invasion, a critical step in the parasite life cycle. Orthologs encoding this antigen have been found in all known species of *Plasmodium* parasitic to primates. In order to characterize and investigate the extent and maintenance of msp-9 genetic diversity, we analyzed DNA sequences of the following malaria parasite species: *Plasmodium falciparum*, *P. reichenowi*, *P. chabaudi*, *P. yoelii*, *P. berghei*, *P. coatneyi*, *P. gonderi*, *P. knowlesi*, *P. inui*, *P. simiovale*, *P. fieldi*, *P. cynomolgi* and *P. vivax* and evaluated the signature of natural selection in all msp-9 orthologs. Our findings suggest that the gene encoding MSP-9 is under purifying selection in *Plasmodium vivax* and closely related species. We further explored how selection affected different regions of MSP-9 by comparing the polymorphisms in *P. vivax* and *P. falciparum*, and found contrasting patterns between these two species that suggest differences in functional constraints. This observation implies that the MSP-9 orthologs in human parasites may interact differently with the host immune response. Thus, studies carried out in one species cannot be directly translated into the other.

© 2013 Elsevier B.V. All rights reserved.

[§]Corresponding author: Ananias A. Escalante, Phone: 1- 480-965-3739, Ananias.Escalante@asu.edu.

^{*}Contributed equally

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

ABRA; Binding sites; Genetic diversity; Merozoite Surface Proteins; MSP-9; *Plasmodium*

1. Introduction

Regardless the extraordinary progress achieved in malaria control, this vector-borne disease is still ranking among the most important causes of disease and death worldwide. There are four human malarial parasites; however, two of them, *Plasmodium falciparum* and *P. vivax*, are responsible for most malaria morbidity and mortality (WHO 2011). Although the species exhibit great biological differences (Coatney et al., 1971), all human malarial parasites invade erythrocytes in a complex, multi-step process that exhibits important differences among species (Rayner et al., 2005). For instance, *P. vivax* merozoites invade only reticulocytes, whereas *P. falciparum* merozoites invade mature erythrocytes as well as reticulocytes. Given that this step is critical in the parasite life cycle, proteins involved in this process are considered to be potential vaccine targets. One of these proteins is the merozoite surface protein-9 (MSP-9) (Vargas-Serrato et al., 2002).

MSP-9 was first identified in *P. falciparum* as the 101 kDa acidic-basic repetitive antigen (ABRA); then, orthologous genes were identified in other *Plasmodium* species (Barnwell et al., 1999; Vargas-Serrato et al., 2002; Vargas-Serrato et al., 2003; Lopera-Mesa et al., 2008). It was first described as a hydrophilic protein with a putative 20 amino acid signal peptide, a conserved N-terminal domain with a cluster of four cysteines and a C-terminal region containing species specific blocks of repeated amino acids (Vargas-Serrato et al., 2002; Oliveira-Ferreira et al., 2004). For instance, two repetitive tandems has been reported in the C-terminal domain *P. vivax* MSP-9 (*PvMSP-9*) (Vargas-Serrato et al., 2002); whereas only one has been reported in *P. falciparum* N-terminal (Weber et al., 1988).

The role of MSP-9 in the parasite life cycle is still not clear. It is found on the surface of merozoites as well as in the parasitophorous vacuole within the infected erythrocytes (Weber et al., 1988; Kushwaha et al., 2002; Cowman et al., 2006). Experimental evidence suggests that in *P. falciparum*, MSP-9 (*PfMSP-9*) acts as a serine protease and participates in the release of merozoites from infected erythrocytes (Nwagwu et al., 1992). Furthermore, *in vitro* experiments indicate that *PfMSP-9* binds the human erythrocyte surface through its N-terminal cysteine-rich region or by high binding activity peptides (HBAPs) within positions 121-240 (Curtidor et al., 2001). These peptides overlap with the region labelled 1a (77-235 aa), one of two segments of *PfMSP-9* (the second is 2 between 364-528 aa) that interact with the merozoite surface protein-1₁₉ (MSP-1₁₉) in a complex that binds the erythrocyte protein band 3 (Kushwaha et al., 2002; Li et al., 2004; Kariuki et al., 2005; Lopera-Mesa et al., 2008). Whereas the *P. vivax* MSP-9 (*PvMSP-9*) interaction with *PvMSP-1₁₉* has not been documented; there is evidence that the N-terminal of *PvMSP-9* is immunogenic. Indeed, five T-cell epitopes (pE, pJ, pK, pH and pL) have been identified in the N-terminus that interact with a broad range of HLA class II molecules; these peptides elicited robust cellular responses in individuals naturally exposed to malaria (Lima-Junior et al., 2008).

The role of *PfMSP-9* in the erythrocyte invasion, the fact that there are orthologs in several *Plasmodium* species, and that it is immunogenic as shown in animal models as well as in natural exposed individuals, made MSP-9 a protein of interest for vaccine development (Sharma et al., 1998; Pimtanothai et al., 2000; Kushwaha et al., 2001; Oliveira-Ferreira et al., 2004; Lima-Junior et al., 2010; Lima-Junior et al., 2012). Hence, it is important to further explore its genetic diversity and how such polymorphism is being altered or maintained in *Plasmodium*. In this study, we explored patterns of genetic polymorphism in *P. falciparum* and *P. vivax* and compared those to other *Plasmodium* species in non-human primates. For instance, we included the chimpanzee parasite *P. reichenowi* that is closely related to *P. falciparum* and the species: *P. cynomolgi*, *P. inui*, *P. simiovale*, *P. fieldi*, *P. knowlesi*, *P. gonderi* and *P. coatneyi* that are a part of a clade that include all these non-human primate malarias and the human parasite *P. vivax* (Escalante et al., 1998, 2005). In addition, we also included the rodent malarial species, *P. chabaudi*, *P. yoelii* and *P. berghei*, to study the long term evolution of the gene encoding MSP-9. Overall, we found that the genetic polymorphism varies across *Plasmodium* species. Whereas we detected a pattern consistent with balancing selection in the red blood cell binding region at the N-terminal of *PfMSP-9*, we found a more complex pattern in *P. vivax*. *PvMSP-9* exhibits evidence consistent with purifying selection in the C terminal- and extensive polymorphism in terms of the number and type of motifs in the low complexity regions. However, the motifs themselves within each low complexity region show an excess of synonymous substitutions, which could be indicative of functional constraints. These contrasting polymorphism patterns between *P. falciparum* and *P. vivax* indicate that MSP-9 may be recognized differently across *Plasmodium* spp.-host interactions.

2. Materials and methods

2.1 Samples

In this study, we processed samples from different geographic locations that included laboratory isolates of *P. vivax* (Mauritania, Brazil I, Vietnam Palo Alto, Vietnam II, Sumatra, Chesson, and India VII), *P. cynomolgi* (Mulligan, PT1, Berok-PT2, RO, Ceylonensis, Gombok, and B strain), *P. knowlesi* (Hackeri, Malayan, Philippine and Nuri), *P. fieldi* (ABI) and *P. simiovale* provided by the Centers for Disease control (CDC, see Coatney et al., 1971), and field isolates of *P. vivax* from Thailand (n=6, hospital based study in Bangkok) and Peru (n=5, Peruvian Amazon region in the Loreto Department). Peru samples were provided by the Naval Medical Research Unit No. 6 in Lima. DNA was purified from 200µL of whole blood samples using the QIAamp DNA Blood Mini kit (Qiagen GmbH, Hilden, Germany).

2.2 PCR amplification, cloning, and sequencing

The gene encoding the MSP-9 for *P. vivax* and the closely related malarial species found in non-human primates (NHPs) were amplified by polymerase chain reaction (PCR). *PvMSP-9* and *PcynMSP-9* sequences were amplified using degenerated primers forward AE374 5'-ATG CG(A/C) (C/G)TG A(C/A/G)C (T/A)T CAT C-3' and reverse AE377 5'-CTA TGG (A/T)GT GAC ATC (G/C)GT G-3', while the *PkMSP-9* sequences were amplified using the primers forward AE374 and reverse AE378 5'-CTA AGG CTC TAC AGT GTT C-3'. *P*

fieldi and *P. simiovale* sequences were amplified using forward AE444 5'-TGG TGA (A/G)GG G(A/G)C AC(A/G) TAG-3' and reverse AE377. We also extracted putative MSP9 orthologs from locally assembled 454 (Roche, Applied Science, Basel, Switzerland) reads for *P. gonderi* (complete CDS) and *P. inui* (partial CDS). Those sequences are also reported.

PCR amplifications were carried out in a 50 µl volume reaction using 20 ng of total genomic DNA, 3 mM MgCl₂, 1X PCR buffer, 1.25 mM of each deoxynucleoside triphosphate, 0.4 mM of each primer, and 0.03 U/µl AmpliTaq Gold® DNA polymerase (Applied Biosystems, Roche-USA). Amplification was achieved using the following conditions: an initial denaturation step at 94°C for 3 minutes; 36 cycles of: denaturation at 94°C for 1 minute, annealing at 57°C for 1 minute and extension at 72°C for 2 minutes; followed by a final extension at 72°C for 10 min. PCR products were purified from a gel using the QIAquick Gel extraction kit (Qiagen GmbH, Hilden, Germany). In order to rule out PCR errors, at least two independent PCR products were cloned in the pGEM®-T Easy Vector Systems (Promega, Madison, WI, USA) in all cases, and both strands for three or four clones from each sample were sequenced using an Applied Biosystems 3730 capillary sequencer. The identity of the sequences was determined by comparing our sequences with those previously deposited in the GenBank database by using BLAST. The sequences reported in this study were deposited in Genbank under the accession numbers KC935389 to KC935421 and KF516232.

2.2 Genetic diversity and phylogenetic analyses

Nucleotide sequence and protein alignments of the conserved N-terminal were performed using ClustalW as implemented in MEGA v4.1 software (Tamura et al., 2007) with manual editing. We also included sequences available in GenBank and PlasmoDB for *P. falciparum* (AF213645, XM_001350647, FJ406812, FJ406815, FJ406822, FJ406825, FJ406816, FJ406813, FJ406809, FJ406814, FJ406821, FJ406820, FJ406810, FJ406818, FJ406817, FJ406811, FJ406823, FJ406819 and J03902), *P. reichenowi* (CDC-1 Oscar, FJ406808), *P. vivax* (XM_001617324, AF435853), *P. cynomolgi* (B-strain, PCYB_145430), *P. knowlesi* (Malaysia H, AF435855), *P. coatneyi* (AY135361), *Plasmodium yoelii yoelii* str. 17XNL (XM725739), *P. berghei* (isolate NK-65, AY302245; isolate ANKA, PBANKA_144330), and *P. chabaudi chabaudi* (PCHAS_144550).

Intra specific variation in regions that could be aligned (repeats were excluded) was characterized in terms of nucleotide diversity (π) and the effect of natural selection was evaluated by estimating the difference of synonymous and non-synonymous substitutions (D_s - D_n) between a pair of sequences using the Nei and Gojobori's method with the Jukes and Cantor correction. The null hypothesis of neutrality was rejected if $D_n/D_s > 1$. Standard error was calculated using bootstrap with 1000 replications and rates of D_s and D_n were compared by the Z-test of selection using MEGA v4.1 (Tamura et al., 2007). The assumption of neutrality was also tested using the McDonald and Kreitman test or MK test (McDonald and Kreitman, 1991) as implemented in DnaSP v5.0, which compares the intra and inter specific number of synonymous and non-synonymous sites; significance was assessed by using a Fisher's exact test for the 2×2 contingency table as implemented in the program DnaSP v5.0 (Librado and Rozas, 2009). In this analysis using the MK test, we

compared the following species: *P. vivax* with *P. cynomolgi* and *P. falciparum* with *P. reichenowi*.

Phylogenetic relationships of msp-9 orthologs were estimated on a nucleotide sequences alignment of the N-terminal (1294 bp) that lacks low complexity regions in most *Plasmodium* species but *P. falciparum* and *P. reichenowi*; thus, the alignment also excluded the R1 repetitive region found solely in these two species (see Fig 1). A phylogeny was estimated by using the Neighbor-Joining method (Saitou and Nei, 1987) with the Kimura 2-parameters model as implemented in MEGA v5.0 (Tamura et al., 2007). The reliability of the nodes in the NJ trees was assessed by bootstrap method with 1000 pseudo-replications (Nei and Kumar, 2000). In addition, a Bayesian phylogenetic analysis was performed as implemented in MrBayes v3.2.1 (Ronquist and Huelsenbeck, 2003). We used a general time reversible model + gamma (GTR+G); this model had the lower number of parameter that best fit the data as estimated by MEGA v5.0. Bayesian support for the nodes was inferred in MrBayes using 3×10^6 Markov Chain Monte Carlo (MCMC) steps, and after convergence was reached, we discarded the 50% of the sample as a burn-in. Sampling was performed every 100 generations.

The long term effect of selection on the N-terminal was explored using maximum likelihood approaches as implemented in HyPhy (Kosakovsky Pond et al., 2005). Specifically, the ratio of k_A to k_S ($=\omega$), in which k_A and k_S are the number of nonsynonymous and synonymous changes per nonsynonymous and synonymous sites, respectively, was estimated for branches in the phylogenetic trees using the branch model implemented in HyPhy. The ω ratio measures the direction and magnitude of selection on amino acid replacements, with values of $\omega < 1$ (purifying selection), $\omega = 1$ (neutral evolution), and $\omega > 1$ (positive selection). The heterogeneity of ω among branches was incorporated using the free-ratio model where each branch has its independent ω value. HyPhy was run using a codon partition, MG94 substitution model. We performed a likelihood ratio test comparing our null model – the *P. vivax* branch was constrained to having an omega of 1 – against our alternative model in which the branch was undergoing positive or negative selection. The models were compared using a likelihood ratio test (LRT). Since the tree topology affects the ω values, we explored the robustness of our results by performing a set of tests where our antigen was studied using the phylogeny estimated from mitochondrial genomes (Pacheco et al., 2011). Since selection might be affecting only a subset of lineages (episodic selection), we also used a branch- and branch-site type models, to recognize lineages where selection is episodic (positively selected sites) as implemented in HyPhy (Kosakovsky Pond et al., 2005; 2011). In all analyses we only considered the conserved regions within species and between species from the MSP-9 alignments. Moreover, the repetitive regions were not included in the phylogeny since they are not conserved in all orthologs.

2.3 Protein structure prediction

To predict important aspects of the protein such as its secondary structure (PROFsec) and its solvent accessibility (PROFacc), we used the PredictProtein automatic service (<http://www.predictprotein.org>) (Rost et al., 2004). In addition, we used SEG to divide amino acid sequences into protein segments of low and high-complexity as implemented in <http://>

mendel.imp.ac.at/METHODS/seg.server.html (Wootton, 1994). The parameters that best determine low complexity regions in MSP-9 were: trigger window length (W): 25, trigger complexity (K1): 3.0 and extension complexity (K2): 3.3. We also used the web-based I-TASSER server, available at <http://zhanglab.ccmb.med.umich.edu/I-TASSER/>, which is an on-line workbench for high-resolution modeling of protein structure. All predictions from this server are tagged with a confidence score which tells how accurate the predictions are without knowing the experimental data (Roy et al., 2010; 2012; Zhang 2008).

3. Results

3.1 Description of MSP-9 orthologs in *Plasmodium* spp

We obtained almost complete msp-9 sequences from 16 different isolates of *P. vivax*, seven from *P. knowlesi*, seven from *P. cynomolgi*, one from *P. fieldi*, one from *P. simiovale*, one from *P. gonderi*, and one partial from *P. inui*. The MSP-9 analysis was performed using 64 sequences in total, including the ones available in GenBank. The analysis of the MSP-9 sequences showed the signal peptide, four conserved cysteines in positions 49, 52, 69 and 76 (species alignment) and one or more repetitive regions in all species analyzed but in *P. chabaudi* (Fig 1). Although putative MSP-9 orthologs in *P. falciparum* and *P. vivax* are found in different chromosomes (chromosome 12 in *P. falciparum* and 14 in all the six species with genomic information); the surrounding proteins are also conserved strengthening the hypothesis that these are indeed orthologous genes.

The repeated domains of MSP-9 varied in length, motif composition and number of repetitions among the species studied (Table 1 and Fig. 1). *P. falciparum* and its relative in African Apes, *P. reichenowi*, showed two repetitive regions with a similar pattern of motifs. The first repetitive region (R1) in the N terminal was characterized by VNDE(E/D)D and TND(E/D)ED motifs in *P. falciparum* and by a (V/L)NDEED in *P. reichenowi*; while the second region in the C terminal had blocks of KE and KEE repeats in both parasites. On the other hand, a repetitive C-terminal end pattern was not maintained in the three rodent parasites since only *P. yoelii* displayed a repetitive region (R2) of NDKVEDESKPSEST, GDKIEDESETSKST and DDKVEDESKPSEST motifs, but *P. berghei* and *P. chabaudi* did not show such a complex pattern; however, *P. berghei* did have the R1 region similar to the one found in *P. yoelii* (Table 1 and Fig. 1).

In other Plasmodia such as *P. gonderi*, *P. simiovale* and *P. fieldi*, only one region of low complexity (repetitive region) was found in their MSP-9 orthologs. Two other non-human primate malarias, *P. coatneyi* and *P. knowlesi*, show two repetitive regions in their MSP-9 (R1 and R2) with a pattern of repetitive motifs in a relatively extended region, R2 (Fig. 1). In human *Plasmodium* species, different patterns of repetitions were observed for *P. falciparum* and *P. vivax* (Table 1). Whereas PfMSP-9 only had two low complexity regions, we identified four distinctive repetitive (low complexity) regions in PvMSP-9 that expand in the C terminal (Table 1 and Supplementary Table S1, Fig. 2 and Supplementary Fig. S1). A first repetitive region (R1) displayed a DET(V/M) motif with homologous regions in related species (Supplementary Fig. S1A). The second repetitive region (R2) had the motifs: VPENV(E/G)AVTETEEVVE, AAPENAEEVATEEVVEA(T/P)PENAEA or VVPEN(T/A)EVATEEVVEAVPEDAEA and the third repetitive region (R3) had repetitions of (A/

V)VPE(H/N)(T/A)EVATEVVE, A(V/A)PENA(E/D) or PVHENA(E/D). Moreover, *PvMSP-9* exhibit extensive polymorphism in terms of the number of repeats (Fig. 2). Homologous R2 and R3 regions were only found in *P. cynomolgi* which is the closest related species to *P. vivax* (Supplementary Fig. S1B) (Escalante et al., 2005). We also observe that there are two major R2 and R3 patterns within *P. cynomolgi* that separate the strains into two distinctive clades; however, none of the *P. cynomolgi* isolate studied has the last repetitive region (R4) that contains mainly heptapeptides (Supplementary Fig. S1B). This R4 forms coiled-coil structures in the protein (Vargas-Serrato et al., 2002). We also found two conserved blocks in *PvMSP-9* between R2 - R3 and R3 - R4 of 28 and 30 amino acids, respectively.

3.2 Polymorphism in putative T cell epitopes

One important step for malaria vaccine development is to identify T-cell epitopes that can interact promiscuously with a broad range of HLA class II molecules, as the ones described in *PvMSP-9* (Lima-Junior et al., 2008). Epitopes pE, pJ, pK, pH and pL were conserved (100% even at nucleotide level) in all *P. vivax* isolates analyzed in this study, those have been previously shown to induce the highest cellular response in natural exposed individuals (Lima-Junior et al., 2012). In contrast, we found high nucleotide diversity ($n=8$, $\pi=0.069$, $dS-dN=0.1$, $p>0.05$) but with an excess of synonymous over non-synonymous polymorphisms for all putative epitopes (pE, pJ, pH-pK-pL) in *P. cynomolgi*.

Moreover, when performing the species alignment (Fig. 3), we found high similarity between these peptides in *P. vivax* and in the closely related species. However, from all five peptides analyzed, the region of the alignment corresponding to peptide pJ had the highest similarity among all the species in the *P. vivax* clade included in this study (Fig. 3), ranging from 67% (*P. gonderi* with *P. vivax*) to 87% between *P. cynomolgi* and *P. vivax*. Even further, this peptide was conserved between *P. vivax* and *P. falciparum* (73%). Regarding the overlapping peptides pK, pH and pL (Fig. 3), we observed that their highly immunogenic core ASIDSMI (Lima-Junior et al., 2012) was not as conserved as the carboxy-terminal region (DEIDFYEK) of peptide pL, which was found in all non-human primate *Plasmodium* species (EEIDFYEK) and in *P. falciparum* and *P. reichenowi* with two amino acid replacements (DEIDFFEK).

On the other hand, the epitope ABRA#14 (DSNIMNSINNVMD EIDFFEK), located in positions 487-506 of *PfMSP-9* and previously identified as a HLA-DQ-binding peptide that also induced T-cell proliferation and Th1-associated cytokine production in DQ8+ transgenic mice (Pimthanotai et al., 2000), was only found in *P. reichenowi* but not in the other species included in this study.

3.3 Genetic diversity and phylogenetic analyses

Table 2 describes the genetic variation found in different domains of *msp-9* across *Plasmodium* spp. When we compared the genetic diversity found in *PvMSP-9* and *PfMSP-9* with the polymorphism observed in non-human primate malarias (*P. cynomolgi* and *P. knowlesi*), we found that *P. cynomolgi* had the highest nucleotide diversity ($\pi=0.069$), and *P. falciparum* the lowest ($\pi=0.001$). In addition, the difference between synonymous and non-

synonymous substitutions in msp-9 showed an excess of synonymous over non-synonymous polymorphisms in *P. vivax*, *P. cynomolgi* and *P. knowlesi*, a pattern consistent with purifying selection. In these species the evidence of purifying selection found in the overall gene seems to be driven by the C-terminal of the gene. When analyzing samples from Peru ($n=5$, $\pi=0.005$, $dS-dN=0.007$, $p=0.023$) and Thailand ($n=6$, $\pi=0.005$, $dS-dN=0.011$, $p=0$) we also found an excess of synonymous over non-synonymous substitutions in PvMSP-9. Instead, in *P. falciparum*, the signature of selection on the gene was positive (consistent with balancing selection) mainly at the polymorphism found on the region previously described as RBC binding (Fig. 1). We also analyzed the patterns of variation on MSP-9 using the McDonald and Kreitman test by comparing *P. vivax* with *P. cynomolgi* and also by comparing *P. falciparum* with *P. reichenowi*. However, we found no evidence of natural selection acting on the gene by making these comparisons.

The phylogeny of MSP-9 is depicted in Figure 4 using DNA sequences and it is consistent with previous published phylogenies that used house-keeping, mitochondrial, or antigen encoding genes (Escalante et al., 1998, 2004; Pacheco et al., 2007; 2010; 2012). As expected, we observed that *P. vivax* and related species in nonhuman primates are a monophyletic group with *P. cynomolgi* and *P. vivax*; *P. knowlesi* and *P. coatneyi*; and *P. simiovale* and *P. fieldi* as sister clades. The phylogenetic base methods as implemented in HyPhy gave an overall omega lower than one ($\omega < 1$) that is consistent with purifying selection acting on this gene. Indeed, the branch leading to *P. vivax* had a value of 0.165 (CI: 0.130-0.199). We also compared this omega value with the one derived from the mitochondrial genes phylogeny (Pacheco et al., 2011) and the results obtained showed a similar omega value of 0.162 (CI: 0.128–0.197). All results indicating negative (purifying) selection were significant as determined by the likelihood ratio test results. In addition, we found no indication of episodic selection on *P. vivax*.

3.4 Protein characteristics and predicted structure

According to the secondary structure predicted results, there are changes in the percentage of coil elements (loops) found in the protein across the phylogeny (Table 3). Whereas *P. falciparum* has a greater percentage of helices, *P. vivax* and related species have an increased amount of loops (Table 3). Regarding the solvent accessibility prediction, the percentage of exposed residues in MSP-9 ranged from 72-84% in different *Plasmodium* species with a predominance of glutamic acid, lysine, and valine in the amino acid composition.

We also obtained the predicted tertiary structure of PfMSP-9 (C-score=-2.53 and TM-score=0.42±0.14) and PvMSP-9 (C-score=-0.51 and TM-score=0.65±0.13) (Supplementary Fig. S2). The protein is depicted in rainbow coloring scheme Blue (N-terminus) to Red (C-terminus). In *P. falciparum*, we observed that the greatest percentage of alpha helices and beta strands were found at the N-terminal of the protein; while a long region of coil elements, indicating an unstructured or disordered region was found at the C-terminal, corresponding to the repetitive region of the protein (LCR). In addition, when analyzing the predicted solvent accessibility results, we also found that the N-terminal part was the one containing most of the buried residues (Supplementary Fig. S3) which delineates the core

region of the protein, whereas regions with solvent exposed and hydrophilic residues such as the ones found at the LCR are potential hydration/functional sites. To further explore protein function, we analyze the predicted ligand binding sites of PvMSP-9 reported by I-TASSER. The highest confidence of the binding site score (0.69) was observed when the template 1w36b was compared with the query protein. When the BS-score is greater than 0.5 and the predicted binding site residues are clustered together, then the confidence level of the prediction is usually high. Based on this score, we found four predicted binding site residues in *P. vivax*: 449D, 454Y, 455E and 456K; which correspond to the carboxy-terminal region (DEIDFYEK, Fig. 3) of the immunogenic peptide pL (Lima-Junior et al., 2008). These putative binding site residues were also found in *P. cynomolgi*, *P. inui*, *P. fieldi*, *P. simiovale*, *P. knowlesi*, *P. gonderi* and *P. coatneyi* with only one substitution in position 449 (E/D); it is worth noting that both amino acids are acid.

4. Discussion

Merozoite surface protein-9 is a promising vaccine candidate antigen found in rodents, non-human primates and human *Plasmodium* species. Like previous studies (Vargas- Serrato et al., 2002), the common features founded in all the species were that glutamic acid (E) and lysine (K) were the most predominant amino acid in this protein (Table 1), four cysteines were highly conserved at the N-terminus, and the C-terminus exhibited great polymorphism, with repetitive regions clearly defined in most species (Fig. 1). Neither the four conserved cysteines nor the tandem repeat region, appear to participate directly in the PfMSP-9-5ABC interaction (Li et al., 2004); however, they might confer specific structural characteristics to the protein. It has been suggested that the N-terminus interacts with red blood cells by binding the band 3 protein in association with PfMSP-1 (Kushwaha et al., 2002). Moreover, when we compared the five erythrocyte high activity binding peptides (HABPs) found on positions 121-240 in *P. falciparum* (Curtidor et al., 2001) against *P. reichenowi*, which are part of ABRA(P) and 1 peptides (Li et al., 2004; Kushwaha et al., 2002), they were highly conserved between both species (¹²¹LQSHKKLIK(A/G)LKKNIESIQN¹⁴⁰, ¹⁴¹KK(H/V)L(I/V)YKNKSYNPLLL(S/T)(C/F)VK¹⁶⁰, ¹⁶¹KMNMLKENVDYIQKNQNLFK¹⁸⁰, ²⁰¹LKSqGHKKETSqNQNEN(n/K)DN²²⁰ and ²²¹QKYQEVNDE(D/E)(D/E)VNDEE(D/E)(T/V)ND²⁴⁰) but not in the species conformed by *P. vivax* and other non-human primate malaria; this finding suggests that *P. falciparum* and *P. vivax* may have a different way of interacting with reticulocytes and that other peptides might be involved.

According to the protein predicted results, residues (454Y, 455E and 456K) found in *P. vivax* might be important ligand-binding sites. These residues were conserved not only in all *P. vivax* isolates but also in *P. cynomolgi*, *P. inui*, *P. fieldi*, *P. simiovale*, *P. knowlesi*, *P. gonderi* and *P. coatneyi*. Interestingly, these residues are part of the immunogenic peptide DEIDFYEK (Fig. 3). In addition, its predicted structure is mainly composed of coil elements and exposed residues; these characteristics make it a peptide to be considered in studies aiming to understand the immunogenicity of PvMSP-9 in naturally exposed populations. Although the peptide pJ is highly conserved in all malarial species (Fig. 1 and 3), including *P. falciparum*, its predictive solvent accessibility suggests that it is mostly “buried” and appear to be part of the core domain of the protein (Lima-Junior et al., 2008).

It is worth noting that MSP-9, in *P. vivax* and related species, has repetitive motifs or low complexity regions at the C-terminal of the gene, these regions are relatively large when compared with other *Plasmodium* including *P. falciparum*. The only other species with similar low complexity regions is *P. cynomolgi*, its closest known related species (Escalante et al., 2005), indicating that these regions originated before *P. vivax* became a human parasite. Like other repetitive proteins in *Plasmodium* spp. (Rodriguez et al., 2008), *PvMSP-9* has extensive polymorphism in the number of motifs; however, when the repetitive motifs are aligned among them (see Rich et al., 1997) the motifs themselves accumulate more synonymous than non-synonymous substitutions (Table 4). These repeated structures in proteins often occur in linear arrays or as a superhelix (Andrade et al., 2001) such as in MSP-9. The coil-coiled structures at the C-terminus end of *PfMSP-9* and *PvMSP-9* exhibit an extensive solvent-accessible surface that could be well suited to binding large substrates (Andrade et al., 2001). In malaria, it has been proposed that repeat regions may function as ligands for host proteins, or serve to suppress the development of immunity through a strategy of serological cross reactivity (Nussenzweig V and Nussenzweig R.S., 1989). Furthermore, certain domains might provide an activation signal to B cells, as the first step of a strategy to force them down an inappropriate (T-cell-independent) differentiation pathway (Schofield, 1991). Even though there is no direct evidence of this phenomenon in MSP-9, there are studies indicating that this protein is immunogenic.

For instance, experiments in rabbits found that peptides based on *PfMSP-9* residues covering the most hydrophilic regions and a peptide representing its repetitive sequence, elicited antibody responses and exerted a strong inhibitory effect on the merozoite invasion of erythrocytes (Sharma et al., 1998). Moreover, immunogenicity studies with recombinant *PfMSP-9* proteins in rabbits and mice indicated that T-cell epitopes were localized in the middle portion of the protein (Pimtanonthai et al., 2000; Kushwaha et al., 2001). Similar studies in *P. vivax* using the MSP-9 N-terminal and the second block of repeats (*PvMSP-9*-R11) generated an immunological response with high titers of specific antibodies that recognized the parasite's protein (Oliveira-Ferreira et al., 2004; Lima-Junior et al., 2012). The actual motifs of this LCR accumulate more synonymous than non-synonymous substitutions indicating some functional constrain. On the other hand, the N-terminal region of *PvMSP-9* induced IFN- γ and IL-4 cytokine responses, confirming the presence of T-cell epitopes in this region of the protein (Lima-Junior et al., 2010). The presence of these T cell epitopes (Lima-Junior et al., 2010) indicates that the response induced by the LCRs could be low grade and non-neutralizing (Schofield 1991). Overall, there is an elicited immune response which association with protection in *P. falciparum* and/or *P. vivax* remains to be determined.

Generally, there is relatively high genetic diversity on individual functional proteins expressed on the surface of the merozoite and that are involved in the invasion of the red blood cell (Escalante et al., 1998, 2004; VanBuskirk et al., 2004); this is not the case for MSP-9 (Table 2) where there was limited variation outside the low complexity regions. However, the fact that this protein could also be involved in the protein complex for red blood cell invasion may select against the accumulation of polymorphism in those non-repetitive regions. Studying the role that natural selection may play in the absence of

polymorphism is difficult simply because it could be the result of demographic processes such as population expansions after a bottleneck. This situation is not unique to MSP-9, other merozoite proteins such as merozoite surface protein-8 (MSP-8) and merozoite surface protein-10 (MSP-10) show very limited genetic polymorphism (Pacheco et al., 2012). Establishing whether such conserved parts of the protein are indeed under purifying selection is important since they may offer valuable targets for antimalarial vaccines if protective immune responses could be elicited against them (Baum et al., 2003; Cole-Tobian and King, 2003; Martinez et al., 2004; Pacheco et al., 2010; 2012). Based on our knowledge of MSP-9 from *P. falciparum*, we could speculate that in *PvMSP-9* the putative binding sites are conserved by purifying selection in order to interact with MSP-1₁₉ kDa fragment in a complex that binds the red blood cell. The analysis of genetic diversity of MSP-1₁₉ and MSP-1₃₃ regions in *P. vivax* and *P. falciparum* confirmed that MSP-1₁₉ is less polymorphic than MSP-1₃₃ in both human malarial parasites (Pacheco et al., 2007). Moreover, in *P. vivax* and related species, there were more synonymous over non-synonymous substitutions in MSP-9, which contrasts to what has been previously reported in other antigenic proteins (Escalante et al., 1998; Tetteh et al., 2009). This suggests that the repetitive parts of this protein are under strong functional constraints and implies a much more complex interaction between merozoite surface proteins that could result in a different host's immunological response.

5. Conclusion

We found evidence consistent with purifying selection in the gene encoding *PvMSP-9*. Our findings could indicate functional constraint for maintaining the protein's structure in *P. vivax* and related species due to a role in the invasion process. We also observed a different pattern in *P. falciparum*, where the observed polymorphism (particularly in the N-terminal) could be explained by balancing selection. Thus, the genetic polymorphism found in the MSP9 orthologs of the two major human parasites may be the result of different mechanisms of interaction with the host. It is worth noting that it has been hypothesized that natural selection is less efficient when acting at the blood stage of *P. vivax* than *P. falciparum* (Schneider and Escalante 2013). Such predictions, together with the contrasting patterns observed in orthologous merozoite proteins such as MSP-1, MSP-10 (Pacheco et al 2007; 2012) and the one studied here; further reinforce the notion that the differences between *P. falciparum* and *P. vivax* need to be considered so findings in one species are not simply translated to the other.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the US National Institutes of Health (R01 GM080586 to AAE). We thank the DNA laboratory at the School of Life Sciences for their technical support. We also thank Jane Carlton for providing access to 454 preliminary data from *P. inui* and *P. gonderi*. The content is solely the responsibility of the authors and does not represent the official views of the NIH, the Centers for Disease Control and Prevention, Department of the Navy, Department of Defense, or the US Government.

References

- Andrade MA, Perez-Iratxeta C, Ponting CP. Protein repeats: structures, functions, and evolution. *J Struct Biol.* 2001; 134:117–131. [PubMed: 11551174]
- Barnwell JW, Galinski MR, DeSimone SG, Perler F, Ingravallo P. *Plasmodium vivax*, *P. cynomolgi*, and *P. knowlesi*: identification of homologue proteins associated with the surface of merozoites. *Exp Parasitol.* 1999; 91:238–249. [PubMed: 10072326]
- Baum J, Thomas AW, Conway DJ. Evidence for diversifying selection on erythrocyte-binding antigens of *Plasmodium falciparum* and *P. vivax*. *Genetics.* 2003; 163:1327–1336. [PubMed: 12702678]
- Coatney, RG.; Collins, WE.; Warren, M.; Contacos, PG. The Primate Malaria. US Government Printing Office; Washington DC: 1971.
- Cole-Tobian J, King CL. Diversity and natural selection in *Plasmodium vivax* Duffy binding protein gene. *Mol Biochem Parasitol.* 2003; 127:121–132. [PubMed: 12672521]
- Cowman AF, Crabb BS. Invasion of red blood cells by malaria parasites. *Cell.* 2006; 124:755–766. [PubMed: 16497586]
- Curtidor H, Urquiza M, Suarez JE, Rodriguez LE, Ocampo M, Puentes A, Garcia JE, Vera R, Lopez R, Ramirez LE, Pinzon M, Patarroyo ME. *Plasmodium falciparum* acid basic repeat antigen (ABRA) peptides: erythrocyte binding and biological activity. *Vaccine.* 2001; 19:4496–4504. [PubMed: 11483276]
- Escalante AA, Lal AA, Ayala FJ. Genetic polymorphism and natural selection in the malaria parasite *Plasmodium falciparum*. *Genetics.* 1998; 149:189–202. [PubMed: 9584096]
- Escalante AA, Cornejo OE, Rojas A, Udhayakumar V, Lal AA. Assessing the effect of natural selection in malaria parasites. *Trends Parasitol.* 2004; 20:388–395. [PubMed: 15246323]
- Escalante AA, Cornejo OE, Freeland DE, Poe AC, Durrego E, Collins WE, Lal AA. A monkey's tale: the origin of *Plasmodium vivax* as a human malaria parasite. *Proc Natl Acad Sci U S A.* 2005; 102:1980–1985. [PubMed: 15684081]
- Kariuki MM, Li X, Yamodo I, Chishti AH, Oh SS. Two *Plasmodium falciparum* merozoite proteins binding to erythrocyte band 3 form a direct complex. *Biochem Biophys Res Commun.* 2005; 338:1690–1695. [PubMed: 16289042]
- Kosakovsky P, Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics.* 2005; 21:676–679. [PubMed: 15509596]
- Kosakovsky P, Murrell B, Fourment M, Frost SD, Delpont W, Scheffler K. A random effects branch-site model for detecting episodic diversifying selection. *Molecular Biology and Evolution.* 2011 first published online June 13, 2011. 10.1093/molbev/msr125
- Kushwaha A, Rao PP, Suresh RP, Chauhan VS. Immunogenicity of recombinant fragments of *Plasmodium falciparum* acidic basic repeat antigen produced in *Escherichia coli*. *Parasite Immunol.* 2001; 23:435–44. [PubMed: 11489167]
- Kushwaha A, Perween A, Mukund S, Majumdar S, Bhardwaj D, Chowdhury NR, Chauhan VS. Amino terminus of *Plasmodium falciparum* acidic basic repeat antigen interacts with the erythrocyte membrane through band 3 protein. *Mol Biochem Parasitol.* 2002; 122:45–54. [PubMed: 12076769]
- Li X, Chen H, Oo TH, Daly TM, Bergman LW, Liu SC, Chishti AH, Oh SS. A co-ligand complex anchors *Plasmodium falciparum* merozoites to the erythrocyte invasion receptor band 3. *J Biol Chem.* 2004; 279:5765–5771. [PubMed: 14630931]
- Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics.* 2009; 25:1451–1452. [PubMed: 19346325]
- Lima-Junior JC, Tran TM, Meyer EV, Singh B, De-Simone SG, Santos F, Daniel-Ribeiro CT, Moreno A, Barnwell JW, Galinski MR, Oliveira-Ferreira J. Naturally acquired humoral and cellular immune responses to *Plasmodium vivax* merozoite surface protein 9 in Northwestern Amazon individuals. *Vaccine.* 2008; 26:6645–6654. [PubMed: 18832003]
- Lima-Junior JC, Banic DM, Tran TM, Meyer VS, De-Simone SG, Santos F, Porto LC, Marques MT, Moreno A, Barnwell JW, Galinski MR, Oliveira-Ferreira J. Promiscuous T-cell epitopes of *Plasmodium* merozoite surface protein 9 (PvMSP-9) induces IFN-gamma and IL-4 responses in

- individuals naturally exposed to malaria in the Brazilian Amazon. *Vaccine*. 2010; 28:3185–3191. [PubMed: 20189487]
- Lima-Junior JC, Rodrigues-da-Silva RN, Banic DM, Jiang J, Singh B, Fabrício-Silva GM, Porto LC, Meyer EV, Moreno A, Rodrigues MM, Barnwell JW, Galinski MR, de Oliveira-Ferreira J. Influence of HLA-DRB1 and HLA-DQB1 alleles on IgG antibody response to the *P. vivax* MSP-1, MSP-3a and MSP-9 in individuals from Brazilian endemic area. *PLoS One*. 2012; 7:e36419. [PubMed: 22649493]
- Lopera-Mesa TM, Kushwaha A, Mohammed A, Chauhan VS. *Plasmodium berghei* merozoite surface protein-9: immunogenicity and protective efficacy using a homologous challenge model. *Vaccine*. 2008; 26:1335–1343. [PubMed: 18272263]
- Martinez P, Suarez CF, Cardenas PP, Patarroyo MA. *Plasmodium vivax* Duffy binding protein: a modular evolutionary proposal. *Parasitology*. 2004; 128:353–366. [PubMed: 15151140]
- McDonald J, Kreitman M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*. 1991; 351:652–654. [PubMed: 1904993]
- Nei, M.; Kumar, S. *Molecular Evolution and Phylogenetics*. Oxford University Press; New York: 2000.
- Nussenzweig V, Nussenzweig RS. Rationale for the development of an engineered sporozoite malaria vaccine. *Adv Immunol*. 1989; 45:283–334. [PubMed: 2665440]
- Nwagwu M, Haynes JD, Orlandi PA, Chulay JD. *Plasmodium falciparum*: chymotryptic-like proteolysis associated with a 101-kDa acidic-basic repeat antigen. *Exp Parasitol*. 1992; 75:399–414. [PubMed: 1493872]
- Oliveira-Ferreira J, Vargas-Serrato E, Barnwell JW, Moreno A, Galinski MR. Immunogenicity of *Plasmodium vivax* merozoite surface protein-9 recombinant proteins expressed in *E. coli*. *Vaccine*. 2004; 22:2023–2030. [PubMed: 15121316]
- Pacheco MA, Poe AC, Collins WE, Lal AA, Tanabe K, Kariuki SK, Udhayakumar V, Escalante AA. A comparative study of the genetic diversity of the 42kDa fragment of the merozoite surface protein 1 in *Plasmodium falciparum* and *P. vivax*. *Infect Genet Evol*. 2007; 7:180–187. [PubMed: 17010678]
- Pacheco MA, Ryan EM, Poe AC, Basco L, Udhayakumar V, Collins WE, Escalante AA. Evidence for negative selection on the gene encoding rhoptry-associated protein 1 (RAP-1) in *Plasmodium* spp. *Infect Genet Evol*. 2010; 10:655–661. [PubMed: 20363375]
- Pacheco MA, Battistuzzi FU, Junge RE, Cornejo OE, Williams CV, Landau I, Rabetafika L, Snounou G, Jones-Engel L, Escalante AA. Timing the origin of human malarias: the lemur puzzle. *BMC Evol Biol*. 2011; 11:299. [PubMed: 21992100]
- Pacheco MA, Elango AP, Rahman AA, Fisher D, Collins WE, Barnwell JW, Escalante AA. Evidence of purifying selection on merozoite surface protein 8 (MSP8) and 10 (MSP10) in *Plasmodium* spp. *Infect Genet Evol*. 2012; 12:978–86. [PubMed: 22414917]
- Pimtanonthai N, Parra M, Johnson AH, David CS, Katovich Hurley C. Assessing the binding of four *Plasmodium falciparum* T helper cell epitopes to HLA-DQ and induction of T-cell responses in HLA-DQ transgenic mice. *Infect Immun*. 2000; 68:1366–73. [PubMed: 10678949]
- Rayner JC, Tran TM, Corredor V, Huber CS, Barnwell JW, Galinski MR. Dramatic difference in diversity between *Plasmodium falciparum* and *Plasmodium vivax* reticulocyte binding-like genes. *Am J Trop Med Hyg*. 2005; 72:666–74. [PubMed: 15964948]
- Rich SM, Hudson RR, Ayala FJ. *Plasmodium falciparum* antigenic diversity: evidence of clonal population structure. *Proc Natl Acad Sci USA*. 1997; 94:13040–45. [PubMed: 9371796]
- Rodriguez LE, Curtidor H, Urquiza M, Cifuentes G, Reyes C, Patarroyo ME. Intimate molecular interactions of *P. falciparum* merozoite proteins involved in invasion of red blood cells and their implications for vaccine design. *Chem Rev*. 2008; 108:3656–705. [PubMed: 18710292]
- Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 2003; 19:1572–1574. [PubMed: 12912839]
- Rost B, Yachdav G, Jinfeng L. The PredictProtein server. *Nucleic Acids Res*. 2004; 32:W321–W326. [PubMed: 15215403]
- Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*. 2010; 5:725–738.

- Roy A, Yang J, Yang Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research*. 2012; 40:W471–W477. [PubMed: 22570420]
- Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987; 4:406–425. [PubMed: 3447015]
- Schofield L. On the function of repetitive domains in protein antigens of *Plasmodium* and other eukaryotic parasites. *Parasitol Today*. 1991; 5:99–105. [PubMed: 15463458]
- Sharma P, Kumar A, Singh B, Bharadwaj A, Sailaja VN, Adak T, Kushwaha A, Malhotra P, Chauhan VS. Characterization of protective epitopes in a highly conserved *Plasmodium falciparum* antigenic protein containing repeats of acidic and basic residues. *Infect Immun*. 1998; 6:2895–2904. [PubMed: 9596765]
- Schneider KA, Escalante AA. Fitness components and natural selection: why are there different patterns on the emergence of drug resistance in *Plasmodium falciparum* and *Plasmodium vivax*? *Malar J*. 2013; 11:12–15.
- Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol*. 2007; 24:1596–1599. [PubMed: 17488738]
- Tetteh KK, Stewart LB, Ochola LI, Amambua-Ngwa A, Thomas AW, Marsh K, Weedall GD, Conway DJ. Prospective identification of malaria parasite genes under balancing selection. *PLoS One*. 2009; 4:e5568. [PubMed: 19440377]
- VanBuskirk KM, Cole-Tobian JL, Baisor M, Sevova ES, Bockarie M, King CL, Adams JH. Antigenic drift in the ligand domain of *Plasmodium vivax* duffy binding protein confers resistance to inhibitory antibodies. *J Infect Dis*. 2004; 190:1556–1562. [PubMed: 15478059]
- Vargas-Serrato E, Barnwell JW, Ingravallo P, Perler FB, Galinski MR. Merozoite surface protein-9 of *Plasmodium vivax* and related simian malaria parasites is orthologous to p101/ABRA of *P. falciparum*. *Mol Biochem Parasitol*. 2002; 120:41–52. [PubMed: 11849704]
- Vargas-Serrato E, Corredor V, Galinski MR. Phylogenetic analysis of CSP and MSP-9 gene sequences demonstrates the close relationship of *Plasmodium coatneyi* to *Plasmodium knowlesi*. *Infect Genet Evol*. 2003; 3:67–73. [PubMed: 12797974]
- Weber JL, Lyon JA, Wolff RH, Hall T, Lowell GH, Chulay JD. Primary structure of a *Plasmodium falciparum* malaria antigen located at the merozoite surface and within the parasitophorous vacuole. *J Biol Chem*. 1988; 263:11421–11425. [PubMed: 3042768]
- World Health Organization (WHO). [accessed July 24, 2012] World malaria report 2011. 2011. http://www.who.int/malaria/world_malaria_report_2011/en
- Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem*. 1994; 18:269–285. [PubMed: 7952898]
- Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*. 2008; 9:40. [PubMed: 18215316]

Highlights

- We investigated the gene encoding the merozoite surface protein (MSP-9) in several *Plasmodium* spp.
- Repeated domains in malarial parasites species varied in length, motif composition and number of repetitions.
- Natural selection is operating differently in *P. falciparum* and *P. vivax*.
- We found evidence of purifying selection on *P. vivax* and related species.

Fig. 1.

Diagram of the MSP-9 *Plasmodium* spp protein alignment (partial sequence of *P. inui* was not included). The length of the protein is indicated for each species. The putative signal peptide, cysteines, RBC binding regions (Kushwaha et al., 2002, Li et al., 2004) and repetitive regions (R, R1, R2, R3 and R4) are shown. Blocks with different patterns represent different repetitive regions. Gaps in the protein alignment are represented by blank areas.

Isolate	R1	R2	R3	R4
Salvador I	[A1] ₅	B1C1B1C3	D1D2E1E2{F1E3} ₂ F1	G1G2H1G3H1
Brazil-Belem	A1A2{A1} ₃	B1C1B1C3	D1D2E1E2{F1E3} ₂ F1	[G2H1] ₂ G3H1
Vietnam PA	[A1] ₆	B1C1B1C3C5	D1D2E1E3F1E3F1E4F1E3{F1} ₂	G1[G2H1] ₂ G3H1
Sumatra	[A1] ₆	B1C1B1C3	D1D2E1E3F1	G2H1[G3H1] ₂ G2H1[G3H1] ₂
Chesson	[A1] ₄	B1C1B1C3	D1D2E1{E3F1} ₂	G2H1G2H1[G3H1] ₂
India VII	[A1] ₅	B1C1B1C3	D1D2E1E3F1E3{F1E2} ₂ F1	[G2H1] ₂ G3H2
Vietnam II	[A1] ₆	B2C1B1C3	D1D2E1E3F1E3F1E4F1	G1[G2H1] ₂ G3H1
Mauritania	[A1] ₆	B1C3	D1D2E1E3F1E3	G1[G2H1] ₂
Brazil I	[A1] ₄	B1C1B1C4	D1D2E1{E3} ₂ F1E3F1	G1[G2H1] ₂ G3H1
Perú PC-11	[A1] ₆	B1C1B1C3	D1D2E1E3F1E3{F1} ₂	G1G2H1G3H1
Perú PC-15	[A1] ₅	[B1C2] ₂ B1C4	D1D2E1{E3F1} ₂	G1[G2H1] ₂ G3H1
Perú PC-17	[A1] ₃	B1C1B1C3	D1D2E1E2{F1E3} ₂ F1	[G2H1] ₂ G4H2
Perú PC-19	[A1] ₆	B1C1B1C3	D1D2E1E2{F1E3} ₂ F1	[G2H1] ₂ G3H1
Perú PC-27	[A1] ₆	B1C1B1C3	D1D2E1E2{F1E3} ₂ F1	[G2H1] ₂ G3H1
Thai-2	[A1] ₆	B1C1B1C3	D1D3D2E1F1E3F1	G1[G2H1] ₂ G3H1
Thai-10	[A1] ₆	B1C1B1C3	D1D2E1{E3F1} ₂ E4F1E3{F1} ₂	[G2H1] ₂ G3H1
Thai-40	[A1] ₅	B1C3	D1D2E1E3F1E4F1E3F2F1E3{F1} ₂	G1[G2H1] ₂ G3H1
Thai-45	[A1] ₃	B1C1B1C3	D1D2E1E3F1E3F1E4F1E3{F1} ₂	[G2H1] ₂ H1G2H1G3H1
Thai-50	[A1] ₅	B1C1B1C3	D1D2E1F1E3F1E4F1E3F1	G1[G2H1] ₂ G3H1
Thai-67	[A1] ₃	B1C1B1C3	D1D2E1E3{F1} ₂	G1[G2H1] ₂ G3H1

Fig. 2.

Repetitive regions (R1, R2, R3 and R4) of *PvMSP-9*. Letters A-H represent repetitive motifs and numbers next to the letters represent motifs with non-synonymous substitutions (see Supplementary Table S1).

Fig. 3.

Alignment of putative T-cell epitopes of *PvMSP-9* (Lima-Junior et al., 2008), including different *Plasmodium* species. The positions underlined correspond to the carboxy-terminal region of peptide pL.

Fig. 4.

Bayesian phylogenetic tree of the gene encoding MSP-9 using MrBayes. The phylogeny was estimated using solely alignable sites on the N-terminal (1,294 bp). The numbers on nodes of the tree are posterior probabilities represented as percentage based on 3×10^6 Markov Chain Monte Carlo (MCMC) steps; the first 1,000,000 (50%) were discarded as a burn-in. Sampling was performed every 100 generations.

MSP-9 repetitive regions in all species analyzed (only complete sequences are included). The most abundant amino acids are shown in percentage.

Table 1

Species (Isolates)	aa composition in protein			LCR position	aa composition in LCR				Repetitive regions
<i>P. falciparum</i> (19)	E	K	N	209-278	D	E	N	T	R1: [VNDE(E/D)D] ₂ [TND(E/D)ED] ₆
	15%	15%	13%		30%	23%	23%	10%	
				637-737	E	K			R2: [KE/KEE] _[13-18]
					47%	21%			
<i>P. yoelii</i> (1)	E	K	S	404-431	D	V	Y		R1: [YD] ₄ [DD(I/D)(A/V)] ₃
	11%	10%	10%		57%	11%	14%		
				584-676	D	E	S		R2: [(N/G/D)DK(V/I)EDES(K/E)(P/T)S(E/K)ST] ₄
					13%	23%	18%		
<i>P. berghei</i> (2)	I	K	S	404-431	A	D			R1: [YD] ₂ [DD(D/T/V)A] ₄
	11%	10%	10%		18%	54%			
<i>P. coatneyi</i> (1)	E	K	T	519-760	E	T	Q		R1: [DETD] ₂
	13%	10%	8%		19%	15%	14%		R2: [TDADQKEVSETE(Q/P)NT] ₃ [AH(V/G)DQQEVSETE(Q/P)NT] ₈
<i>P. gonderi</i> (1)	E	K		475-653	E	V			R1: [QTEE(I/L)] ₃
	14%	12%			23%	16%			
<i>P. knowlesi</i> (8)	E	K		466-506	D	E			R1: [D(L/E)(E/N)D] ₂
	15%	12%			24%	20%			
				526-701	E	Q	T		R2: [VEPE(Q/P)(T/D)(E/Q)(E/T/A)Q(N/K)T] ₁₋₁₂
					30%	15%	21%		
<i>P. simiovale</i> (1)	E	K	V	450-743	E	V			R1: DTAE[DTVE] ₂ DTTD
	13%	9%	9%		19%	15%			
<i>P. fieldi</i> (1)	E	K	V	445-597	E	V			R1: EDTAGDTV[D(N/D)ATEDT(I/A)] ₃ DDAT
	13%	8%	8%		21%	13%			
<i>P. cynomolgi</i> (8)	A	E	V	477-818	A	E	V		R1 _{Bstrain} : [DAT(I/S)DA(V/T)N] ₂

Species (Isolates)	aa composition in protein			LCR position	aa composition in LCR			Repetitive regions
<i>P. vivax</i> (20)	9%	1	6%	11%	17%	25%	20%	R1 _{Berok} : [EEA(T/A)N(X/G/D)TA VDTV] ₁₂₋₃ [EDATNDTV] ₁₀₋₂
								R2 _{Bstrain} : [VGTEE(V/E)E(A/V)V PEN(A/V/G)E] ₁₄₋₅ [V(A/V/E)PEN(T/A)E] ₅₋₉
								R2 _{Berok} : [PE(N/H)AEV(A/V/T)] ₅₋₁₀ [PENDEVV] ₁₋₅
	A	E	V	466-916	A	E	V	R1: [DET(V/M)] ₃₋₆
	10%	17%	14%		18%	26%	21%	R2: [VPENV(E/G)AVTETEEVVE] ₁₋₃ [(A/V)(A/V)PEN(A/T)EVA TEEVVEA(T/P/V)PE(N/D)AEA] ₁₋₃
								R3: [(A/V)VPE(H/N)(T/A)EVA TEVVE] ₁₂₋₃ [A(V/A)PENA(E/D)] ₁₂₋₇ [PVHENA(E/D)] ₁₋₆
								R4: [AVPE(S/N)(V/A/T)E] ₃₋₇ [AVTEA(V/A)E] ₂₋₇

Table 2

Polymorphisms found in the gene encoding MSP-9. Statistical significance is indicated in bold.

Species	Length (bp)	II (S.E.)	Ds (S.E.)	Dn (S.E.)	Ds-Dn (S.D.)	p (Z-stat)
<i>P. falciparum</i> (n=18)						
Gene ⁺	2070	0.001(0.001)	0.001(0)	0.003(0.001)	-0.002(0.001)	0.013 (2.525)
RBC binding region	1266	0.001(0.001)	0(0)	0.002(0.001)	-0.002(0.001)	0.032 (2.164)
<i>P. vivax</i> (n=17)						
Gene ⁺	1821	0.005(0.001)	0.012(0.005)	0.003(0.001)	0.008(0.003)	0.01 (-2.610)
RBC binding region*	1317	0.003(0.001)	0.003(0.002)	0.003(0.001)	0.001(0.002)	0.730 (-0.34)
<i>P. cynomolgi</i> (n=8)						
Gene ⁺	2105	0.069(0.004)	0.199(0.018)	0.055(0.005)	0.145(0.019)	0 (-7.805)
RBC binding region*	427	0.043(0.003)	0.162(0.019)	0.023(0.003)	0.139(0.021)	0 (-7.261)
<i>P. knowlesi</i> (n=8)						
Gene ⁺	1710	0.008(0.001)	0.021(0.005)	0.005(0.002)	0.016(0.005)	0.002 (-3.122)
RBC binding region*	918	0.006(0.001)	0.013(0.050)	0.005(0.002)	0.009(0.005)	0.1 (-1.66)

⁺ Corresponds to alignable regions (within species) of the gene without repetitions.

^{*} Corresponds to the *P. falciparum* homologous region in the alignment.

Table 3

Results of MSP-9 secondary protein structure (H: helix, E: extended sheet, L: loop) and solvent accessibility prediction (e: exposed residues, b: buried residues). In bold are labeled the most common amino acids.

Species	Secondary structure (% in protein)				Solvent accessibility (% in protein)		aa composition (% of residues)																			
	H	E	L		b	E	A: 2.8	C: 0.8	D: 6.9	E: 14.6	F: 2.6	G: 3.0	H: 1.1	I: 5.5	K: 14.5	L: 8.7	M: 3.0	N: 12.9	P: 2.4	Q: 4.1	R: 0.5	S: 3.9	T: 4.2	V: 5.4	W: 0.1	Y: 3.0
<i>P. falciparum</i>	81.73	1.49	16.78		24.63	75.37																				
<i>P. yoelii</i>	31.27	6.64	62.09		28.17	71.83	A: 2.8	C: 0.9	D: 8.8	E: 10.8	F: 1.5	G: 4.6	H: 0.7	I: 9.4	K: 9.9	L: 9.0	M: 1.5	N: 8.7	P: 3.4	Q: 2.2	R: 1.9	S: 9.9	T: 5.6	V: 5.2	W: 0.0	Y: 3.2
<i>P. coatneyi</i>	20.29	2.36	77.36		15.71	84.29	A: 4.2	C: 0.5	D: 6.8	E: 12.7	R: 0.8	G: 5.5	H: 2.9	I: 6.4	K: 9.9	L: 7.3	M: 2.5	N: 6.2	P: 2.1	Q: 6.9		S: 5.1	T: 8.4	V: 8.0	W: 0.0	Y: 1.8
<i>P. knowlesi</i>	14.12	3.42	82.45		18.54	81.46	A: 2.4	C: 0.7	D: 5.3	E: 15.1	F: 2.3	G: 6.0	H: 2.1	I: 6.7	K: 12.3	L: 8.0	M: 3.0	N: 6.1	P: 3.0	Q: 5.6	R: 0.4	S: 3.3	T: 8.4	V: 7.1	W: 0.0	Y: 2.1
<i>P. cynomolgi</i>	14.71	2.7	82.6		15.44	84.56	A: 9.1	C: 0.5	D: 7.0	E: 15.9	F: 2.5	G: 5.0	H: 2.0	I: 5.4	K: 8.3	L: 6.1	M: 2.8	N: 7.0	P: 3.7	Q: 2.6	R: 0.9	S: 3.2	T: 5.5	V: 11	W: 0.0	Y: 1.6
<i>P. vivax</i>	23.14	5.57	71.29		22.6	77.4	A: 10.4	C: 0.4	D: 6.1	E: 17.1	F: 2.1	G: 4.3	H: 2.6	I: 4.2	K: 7.4	L: 5.6	M: 2.2	N: 5.8	P: 3.9	Q: 2.3	R: 0.6	S: 3.1	T: 6.9	V: 13.5	W: 0.0	Y: 1.6

Table 4

Polymorphisms found in the low complexity region of the gene encoding *PvMSP-9*. Descriptions of the repetitive regions are in Table 1 (R1, R2, R3 and R4). Statistical significance is indicated in bold.

Repetitive regions	Π (S.E.)	Ds-Dn (S.D.)	p (Z-stat)
RLC1	0.037±0.034	0.219±0.168	0.189 (-1.321)
RLC2	0.085±0.018	0.322±0.174	0.001 (-3.308)
RLC3A *	0.085±0.031	0.214±0.194	0.054 (-1.949)
RLC3B *	0.074±0.034	0.189±0.204	0.259 (-1.134)
RLC3C *	0.001±0.001	0.002±0.002	0.270 (1.108)
RLC4	0.056±0.020	0.138±0.115	0.096 (-1.675)
RLC1+RCL2+RCL3+RLC4	0.015±0.002	0.038±0.010	0.000 (-4.053)

* A, B and C correspond to each repetitive motif of R4 (Table 1)