Published in final edited form as:

Infect Genet Evol. 2017 June ; 50: 7–19. doi:10.1016/j.meegid.2017.01.024.

Evolution of the merozoite surface protein 7 (MSP7) family in *Plasmodium vivax* and *P. falciparum*: a comparative approach

Andreína I. Castillo¹, M. Andreína Pacheco², and Ananias A. Escalante^{2,*}

¹Arizona State University, School of Life Sciences, Tempe, AZ, USA

²Department of Biology/Institute for Genomics and Evolutionary Medicine (igem), Temple University, Philadelphia, PA, USA

Abstract

Malaria parasites (genus *Plasmodium*) are a diverse group found in many species of vertebrate hosts. These parasites invade red blood cells in a complex process comprising several proteins, many encoded by multigene families, one of which is merozoite surface protein-7 (*mSP7*). In the case of *Plasmodium vivax*, the most geographically widespread human-infecting species, differences in the number of paralogs within multigene families have been previously explained, at least in part, as potential adaptations to the human host. To investigate this in *mSP7*, we studied its orthologs in closely related nonhuman primate parasites; investigating both paralog evolutionary history and genetic polymorphism. The emerging patterns were then compared with the human parasite *Plasmodium falciparum*. We found that the evolution of the *mSP7* family is consistent with a birth-and-death model, where duplications, pseudogenizations, and gene loss events are common. However, all paralogs in *P. vivax* and *P. falciparum* had orthologs in their closely related species in non-human primates indicating that the ancestors of those paralogs precede the events leading to their origins as human parasites. Thus, the number of paralogs cannot be explained as an adaptation to human hosts. Although there is no functional information for *mSP7* in *P. vivax*, we found evidence for purifying selection in the genetic polymorphism of some of its paralogs as well as their orthologs in closely related non-human primate parasites. We also found evidence indicating that a few of *P. vivax*'s paralogs may have diverged from their orthologs in non-human primates by episodic positive selection. Hence, they may have been under selection when the lineage leading to *P. vivax* diverged from the Asian non-human primates and switched into Homininae. All these lines of evidence suggest that *mSP7* is functionally important in *P. vivax*.

Graphical abstract

*Correspondence author: Ananias.Escalante@temple.edu.

Conflict of interest statement

The authors have no conflict of interest to declare.

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

Merozoite surface proteins 7; Multigene family; *Plasmodium falciparum*; *Plasmodium vivax*; Episodic positive selection; Pseudogenization

1. Introduction

Malaria is a vector-borne disease caused by protozoa of the genus *Plasmodium*. These parasites are found associated with a broad range of vertebrate hosts including primates (Garnham, 1966). Out of the numerous *Plasmodium* species, there are four that typically infect humans. Two of these, *Plasmodium falciparum* and *Plasmodium vivax*, account for most of the malaria morbidity and mortality worldwide (WHO, 2015). These two species differ in many epidemiologic and biologic characteristics including features in their genomes. In particular, they markedly diverge in terms of G+C content in their exons, frequency of low complexity regions, and some distinctive multigene families (Battistuzzi et al., 2016; Frech and Chen, 2011).

Multigene families are found in all known genomes of *Plasmodium* species. However, some are shared only by species within particular clades (Wasmuth et al., 2009). They are involved in vital functions such as cytoadherence, host cell recognition and binding, antigenic variation, and antigenic diversity (Frech and Chen, 2011). Out of the multigene families found in all known primate and rodent malarias, two (*msp3* and *msp7*) are expressed on the surface of asexual stage (merozoite) parasites and are considered part of a group commonly known as the Merozoite Surface Proteins (MSPs). These MSPs, together with other proteins, are involved in the invasion of host red-blood cells (Boyle et al., 2014).

Interestingly, out of these two MSP families (*msp3* and *msp7*), only *msp7* is shared across all known primate and rodent malaria-causing parasites. In contrast, the *msp3* genes in *P. vivax* (*Pvmsp3*) and their orthologs from related species parasitizing non-human primates in Southeast Asia, are not orthologs to those identified as the *Pfmsp3* family in *P. falciparum* and its related species (Rice et al., 2014). This leaves *msp7* as the only MSP gene family,

with a role in erythrocyte invasion, found across all the known primate and rodent malarial parasites (Boyle et al., 2014). Although this pattern suggests its importance, there is still limited functional information on *msp7*.

Some paralogs of the *msp7* family in *P. falciparum* (*Pfmsp7*) encode proteins that may play an important role in host erythrocyte invasion. Specifically, *Pfmsp7* paralogs (e.g. PF3D7_1335100) are known to participate in one of many “complexes” with *Pfmsp1* that bind to the erythrocyte and appear to be fundamental during erythrocyte invasion (Lin et al., 2016). In addition, *msp7* paralogous genes may play redundant roles during this process (Kadekoppala et al., 2010). Knock-out experiments have shown that the correct processing of *Pfmsp1* is not affected by deletion of some *Pfmsp7* multigene family members (Kauth et al., 2006), likely due to the fact that there are multiple complexes involving *Pfmsp1* (Lin et al., 2016). However, the disruption of PF3D7_1335100, the *Pfmsp7* paralog described in the *msp1* complex, resulted in a reduction of 20% of the parasite’s ability to invade erythrocytes. Furthermore, the disruption of five *P. falciparum* *msp7* paralogs resulted in a null phenotype (Kadekoppala et al., 2008).

Interactions between *msp1* and some *msp7* paralogs have also been found in other *Plasmodium* species. Experimental evidence from the rodent parasite *Plasmodium yoelii* showed that at least one *Pymmsp7* paralog (PY17X_1354000) interacts with *Pymmsp1*. This suggests that, as in *P. falciparum*, *msp7* paralogs in *P. yoelii* can form similar complexes with *msp1* and have a role in invasion of red blood cells in this species. Despite this evidence, not all *msp7* paralogs have a significant role in erythrocyte invasion, as paralogous genes may differ in their functions. A recent report suggests that some *Pfmsp7* paralogs may instead act as immunomodulators by affecting disease severity (Perrin et al., 2015). Furthermore, *msp7* paralogs may affect parasite tropism as suggested by knock-out experiments with *Plasmodium berghei* showing an increase in parasite preference for reticulocytes (Tewari et al., 2005). All these lines of evidence indicate that *msp7* is a multigene family that plays diverse and critical roles in parasite-host interactions during the invasion of the red blood cell.

How the divergence observed among the *msp7* paralogs across *Plasmodium* spp. relates with their functional diversity remains unknown. A first step, however, is to improve our understanding of the evolutionary history and mechanisms maintaining the genetic diversity of the paralogous genes belonging to this multigene family. Previous studies have shown that there is large variation in the number and composition of paralogs of the *msp7* family among *Plasmodium* species (Garzón-Ospina et al., 2010; Kadekoppala and Holder, 2010). In particular, an expansion of the *msp7* multigene family was suggested to have occurred in *P. vivax* and its closest related species found in Southeast Asian macaques, *Plasmodium cynomolgi* (Mongui et al., 2006; Tachibana et al., 2012). However, limited information on other species within that clade did not allow for further exploration of this pattern. In this investigation, *msp7* multigene family members in Asian and Africa *Plasmodium* species were analyzed. Different evolutionary mechanisms were explored within the clade that includes *P. vivax* and related species in order to assess how those affected the proposed expansion of the *msp7* multigene family in the human parasite. Since molecular phylogenetic studies have shown that *P. vivax* and *P. falciparum* originated independently as

human parasites (Escalante et al., 1998), the emerging patterns from *P. vivax* were contrasted with the available information from *P. falciparum* and related species. In particular, changes in the number and composition of paralogs were compared between each of these human parasites and their respective closely related species in order to test whether they could be interpreted as an adaptation to the human host. Although this study found no evidence that the number of *msp7* paralogs originated as a parasite adaptation to infect humans, *Pvmsp7* paralogs appear to be under complex patterns of selection driven by ancient events that led to the introduction of this parasite lineage into Hominines.

2. Material and methods

Understanding the evolution of gene families involves characterization of the duplications/losses of their paralogs, as well as the processes driving the divergence/homogenization of those paralogs and their polymorphism (Innan and Kondrashov, 2010). Accordingly, we characterized all these levels of variation in the *msp7* gene family.

2.1 Sequence data

Each *Plasmodium* species has an *msp7* family with several paralogs. Multigene family size refers to the number of those paralogs in a given species. A specific *msp7* paralog in species “A” may have an ortholog in species “B”, thus, *msp7* paralogs may have orthologs whenever two species are compared. We used the PlasmoDB Gene IDs (nomenclature) assigned to *P. vivax* (Salvador I) (Aurrecoechea et al., 2009) to refer to specific paralogs.

Genetic diversity of each of the *msp7* paralogs within *P. vivax* (*Pvmsp7*) was ascertained first by using all complete or almost complete sequences available at PlasmoDB version 26 (Aurrecoechea et al., 2009) and the NCBI database (Benson et al., 2014). The data from PlasmoDB consisted of clinical isolates from diverse geographic regions and obtained via whole genome sequencing as part of the Hybrid Selection Initiative performed by the Broad Institute, where represented samples include NIAID-funded International Centers of Excellence for Malaria Research (ICEMRs), as well as non-ICEMR locations. In addition, the PlasmoDB database included the five sequenced *P. vivax* reference strains (Salvador I, North Korean, India VII, Mauritania I and Brazil I) publicly available (Neafsey et al., 2012). As a comparison, the genetic polymorphism observed in each of the *P. falciparum msp7* (*Pfmsp7*) paralogs was also analyzed. The data available in PlasmoDB for *P. falciparum* originated from the following sources: (1) whole genome sequencing of isolates collected from symptomatic malaria patients from Mali, generated through the 100 *Plasmodium* Genomes Whitepaper; (2) paired-end short-read sequences of clinical isolates from an endemic Gambian population from the Greater Banjul Area; and (3) genome sequences obtained from several Senegal isolates. In addition, informative *Pfmsp7* sequences available from the NCBI database were also included.

The study also included the *msp7* family from different non-human infecting *Plasmodium* species with publicly available genomes found in PlasmoDB and NCBI databases: *P. cynomolgi* (*Pcmsp7*, B-strain), *Plasmodium inui* (*Pimsp7*, San Antonio strain), *Plasmodium knowlesi* (*Pkmsp7*, H strain), *Plasmodium coatneyi* (*Pcomsp7*, Hackeri strain), *P. falciparum* (3D7), *Plasmodium reichenowi* (*Prmsp7*, Dennis strain), and the rodent malarias *P. yoelii*

(*Pymsp7*, YM strain), *P. berghei* (*PbmSP7*, ANKA strain), and *Plasmodium chabaudi chabaudi* (*Pchmsp7*, AS strain). Additionally, *msp7* sequences obtained from 454 reads (Roche, Applied Science, Basel, Switzerland) for a parasite from African primates, *Plasmodium gonderi* (*Pgmsp7*) were included.

Specific *msp7* multigene family paralogs were additionally sequenced using isolates provided by the Center for Disease Control (CDC): *P. vivax* (*Pvmsp7*: Indonesia I, Thailand III, Vietnam Palo Alto, and Sumatra strains), *P. cynomolgi* (*PcmSP7*: Berok, Cambodia, PT1, PT2, RO, *ceylonensis*, Gombak, and Mulligan strains), *P. inui* (*PimSP7*: Perlis, Perak, Philippine, Celebes II, Leaf Monkey I and II, Leucosphyrus, OS, Taiwan II, and N34 strains), *P. knowlesi* (*Pkmsp7*: Hackeri and Malayan strains), *Plasmodium fieldi* (*Pfmsp7*: N-3 strain), *Plasmodium simiovale*, and *Plasmodium hylobati*. Information on these species and strains can be found elsewhere (Coatney, 1971).

DNA extractions from blood samples were performed using QIAamp DNA blood mini kit (Qiagen, Hilden, Germany). Sets of degenerated primers were independently created for each *msp7* paralog using the publicly available genomes of *P. vivax*, *P. cynomolgi* and *P. knowlesi* (Table S1). Three independent polymerase chain reactions (PCR) for each *msp7* ortholog were performed using AmpliTaq polymerase (Applied Biosystems, Roche, USA), followed by purification of positive PCR products using QIAquick gel extraction kit (Qiagen, Hilden, Germany). Purified products were posteriorly cloned using pGEM-T Easy Vector Systems I (Promega, WI, USA). Two to three clones from each PCR were sequenced using an Applied Biosystems 3730 capillary sequencer. The orthology of newly sequenced *msp7* multigene family members was determined by reciprocal BLAST searches for all the species included here (Altschul et al., 1997). In particular, sequence similarity was assessed by evaluating the e-values compared to published *msp7* sequences (Table S2). The orthology of each *P. gonderi* *msp7* paralog was established with those found in *P. vivax* by using reciprocal BLAST searches (Table S2). All sequences obtained in this study were deposited in GenBank (KU307279-KU307446).

In addition to the BLAST e-values, newly obtained sequences were characterized as members of the *msp7* multigene family based on the identification of a signal peptide in the N-terminal region of the protein using SignalP 4.1 Server (Petersen et al., 2011) and the presence of the C-terminus domain commonly found conserved in members of this family (Marchler-Bauer et al., 2011), as well as their amino acid composition (Wilkins et al., 1999). The location and motive of repetitive tandem regions were identified using the Statistical Analysis of Protein Sequences (SAPS) online tool (Brendel et al., 1992).

2.2 Polymorphism and evolutionary genetic analyses

Interspecies nucleotide alignments were performed independently for the *msp7* multigene family and for individual *msp7* paralogs using the CLUSTALW algorithm incorporated into MEGA 6.06 (Tamura et al., 2013), followed by manual editing. Low complexity regions were not included because their alignment is unreliable (Battistuzzi et al., 2016) and cannot be studied with the type of genetic and phylogenetic analyses used in this investigation. Genetic diversity was calculated as among different isolates and for each *msp7* paralog of *P. falciparum*, *P. vivax*, *P. cynomolgi*, *P. inui* and *P. knowlesi* using MEGA 6.06 (Tamura et al.,

2013). Patterns consistent with natural selection acting on the observed polymorphism were assessed by calculating the differences of the average number of synonymous (dS) and non-synonymous substitutions (dN) between isolates using the Nei-Gojobori distance method and the Jukes and Cantor correction as implemented in MEGA 6.06. Given the number of sequences per paralog, the difference between dS and dN and its standard error was estimated by using bootstrap with 1,000 pseudo-replications, as well as a two tailed codon based Z-test on the difference between dS and dN (Nei and Kumar, 2000). Under the neutral model, synonymous substitutions are predicted to accumulate faster than non-synonymous (dS > dN) because synonymous substitutions have no effect on the fitness of the individuals carrying them whereas purifying selection is expected to act against non-synonymous substitutions since these are usually deleterious. Conversely, if positive selection is maintaining polymorphism, a higher rate of non-synonymous over synonymous substitutions is expected (dS < dN). We assumed as a null hypothesis that the observed polymorphism was not under selection (dS=dN).

Evidence of natural selection acting on the divergence among orthologs for each *msp7* paralog was evaluated in further detail. In particular, we tested for evidence indicating episodic selection for each *msp7* multigene family members using Hyphy's random effects Branch-Site REL model (Kosakovsky Pond et al., 2005; Kosakovsky Pond et al., 2011). Unlike other methods, this model does not require *a priori* partition between positively selected foreground branches and negatively selected or neutral background branches; assumptions that yield high rates of false positives and false negatives (Kosakovsky Pond et al., 2011). Instead, the Branch-Site REL model allows for the independent variation of three selective regimes over the branches in a phylogeny at any given site reducing the risk of false positives or negatives.

2.3 Phylogenetic analyses and recombination events

Phylogenies of the *msp7* paralogs among related species were estimated in order to explore their origin and diversification. However, many *msp7* paralogous genes do not have orthologs across all *Plasmodium* species (Garzón-Ospina et al., 2010; Kadekoppala and Holder et al., 2010). Thus, in the context of this investigation we found it more informative to perform independent phylogenetic analyses for the *msp7* paralogs within each major clade (Pacheco et al., 2011). In particular, we analyzed *P. vivax* and its closely related species found in Asian non-human primates using *P. gonderi* as outgroup. Separately, a phylogenetic analysis was carried out with the *msp7* paralogs found in *P. falciparum* and *P. reichenowi* (*Laverania* subgenus). For each *msp7* alignment, the most adequate substitution model was selected using the Akaike information criterion (AIC) method incorporated in Jmodeltest (Posada, 2008). The most fitted nucleotide substitution models were, for the most part, specific variants of the General Time Reversible (GTR) model. The nucleotide frequencies, fraction of invariable sites, and the alpha parameter of the gamma distribution were estimated in order to use them in Maximum Likelihood phylogenetic analysis.

Both Maximum Likelihood (ML) and Bayesian Inference (BI) methods were used to construct an *msp7* multigene family phylogenetic tree for both groups. In the case of the *P. vivax* clade phylogeny, paralogs PVX_082660 (531-567bp) and PVX_082690 (285-315bp)

and their respective orthologs were excluded due to their atypically short sequence lengths compared to other members of the *msp7* multigene family (828-1,608bp). Likewise, the *msp7* pseudogene PF3D7_1334900 (141bp) and its orthologous gene were excluded from the *P. falciparum*-*P. reichenowi* comparison. PhyML v3.0 (Guindon et al., 2010) with 200 bootstrap pseudo-replications was used to assess node significance on the ML phylogenetic tree construction. The BI analyses were performed using MrBayes v3.1.2 (Ronquist et al., 2012) with 2×10^7 Markov Chain Monte Carlo (MCMC) steps and sampling every 1,000 generations. A burn-in fraction of 50% was used. Convergence of the BI analysis was diagnosed by requiring a standard deviation between 0.01 and 0.05 among runs and a Potential Scale Reduction Factor (PSRF) between 1.00 and 1.02.

Finally, in order to explore whether recombination and/or gene conversion events affected the inferred phylogenies, these events were characterized among closely related paralogs, as estimated by the ML and BI phylogenetic trees (see Results and Discussion section) by using Recombination Detection Program (RDP3) with its default parameters and a cut-off value of 0.05 (Martin et al., 2010). This software combines several methods to detect and characterize recombination events on large sequence alignments and requires no user input regarding non-recombinant reference sequences. Nonetheless, the location of recombination breakpoints was also explored with the GARD tool available in <http://www.datamonkey.org/help/Recomb.php#GARD> (Kosakovsky Pond et al., 2006).

2.4 Conservation of High Activity Binding Peptides (HABP) in *P. falciparum*

In the particular case of *P. falciparum*, five peptides with significantly high binding activity to the red blood cell have been identified in one of its *msp7* paralogs (PF3D7_1335100) (García et al., 2007). Although attempts were made to locate these HABP in *P. vivax* and related species, no clear homology could be found. Nevertheless, given the proposed functional redundancy across *P. falciparum* *msp7* paralogs; we hypothesized that those motifs could be conserved among other *msp7* paralogs found in *P. falciparum* and other species of the *Laverania* clade such as *P. reichenowi* (*Prmsp7*). Thus, its conservation among *P. falciparum* paralogs and between closely related species is described. Furthermore, the putative ancestral sequences for all five peptides was estimated using MEGA 6.06 and plotted in the *Pfmsp7* and *Prmsp7* paralogs phylogeny.

3. RESULTS

3.1 Variation of *msp7* multigene family size

In all *Plasmodium* species, *msp7* paralogs are arranged in tandem inside syntenic blocks. This is the pattern expected when a multigene family is originated via gene duplication by unequal crossing over (Innan and Kondrashov, 2010). Specifically, the *msp7* family is found on the negative strand on chromosome 13 of *P. falciparum*, *P. reichenowi*, and rodent *Plasmodium* species, whereas it is in the positive strand on chromosome 12 of *P. vivax* and the Asian non-human primate *Plasmodium* species with publicly available genomes (Fig. 1). The *Pvmsp7* syntenic block is delimited by a flanking conserved hypothetical protein at the 5' end (PVX_082715) which has orthologs in all species in the Southeast Asian primate clade, and at the 3' end by a flanking conserved gene (chaperone binding putative protein,

PVX_082640) with its respective orthologs in all *Plasmodium* species with complete or partial genomes (Fig. 1). Orthologs for PVX_082640 and PVX_082715 are also found in rodent malarias and *P. falciparum*, but only the orthologs for PVX_082640 flank the syntenic block at the 5' end of the *msp7* block in those species. Additional open reading frames (ORF) unrelated to the *msp7* multigene family flank the 3' end of the syntenic block in *P. falciparum* (e.g. PF3D7_1335200) and in rodent *Plasmodium* species (PYYM_1351000).

Consistent with previous studies (Garzón-Ospina et al., 2010, Kadekoppala and Holder, 2010), the number of paralogs varied between the three major clades of mammalian parasites considered in this investigation which included: *Laverania* subgenus (*P. falciparum* - *P. reichenowi*- *P. gaboni*), rodent malarias, and *Plasmodium* species of Asian primates. Interestingly, when compared to the *P. falciparum* (*Laverania*) and *P. vivax* (Asian primates) clades, a reduction in the *msp7* family size was observed in the rodent *Plasmodium* species. The number of *msp7* paralogs (three) remained constant within the rodent clade and no pseudogenes were found. Importantly, these three rodent *msp7* paralogs are the ones with putative orthologs in all the primate parasites included in this investigation.

All *Pvmsp7* paralogs and their orthologs from closely related species share a similar amino acid composition independently of gene length; the similarity tends to be higher among orthologs than among paralogs (Fig. S1). Amino acids in high proportion throughout paralogs and orthologs are Lysine (K), Leucine (L), Glycine (G), Glutamic Acid (E), Asparagine (N) and Alanine (A) (Fig. S1). Paralog PVX_082710 and its corresponding *P. cynomolgi* ortholog (PCYB_122730) have been previously included as members of the *msp7* multigene family (e.g. Garzón-Ospina et al., 2010). Although both genes share a similar amino acid composition to that observed in *msp7* family members (Fig. S1), a detailed analysis showed that neither PVX_082710 nor PCYB_122730 had the conserved MSP7 C-terminus domain found in all other family members. Furthermore, they did not have the N-terminus signal peptide characteristic of the *msp7* multigene family (Kadekoppala and Holder, 2010) or share sequence similarity with any other *msp7* paralog (BLAST searches). Therefore, these two orthologs were excluded from further analyses. Even though PVX_082660 and its corresponding orthologs in *P. cynomolgi* also lacked the conserved MSP7 C-terminus domain, this gene and its correspondent orthologs showed sequence similarity with the N-terminus regions of other *msp7* multigene family members (e.g. PVX_082660 vs. PVX_082665, e-value = 5E-20, Table S2). Thus, we consider it to be a member of the MSP7 family. It is worth noting that in *P. cynomolgi*, the PCYB_122760 paralog reported by Tachibana et al., 2012 (see Fig. 1) has no orthologs in *P. vivax*. Thus, it is likely the result of a recent duplication event although it lacks the typical MSP7 N-terminus regions. This finding was corroborated for the Berok strain. Overall, the number of confirmed *msp7* paralogs shared by *P. vivax* and *P. cynomolgi* is 12 (Fig. 1).

An important observation is that each *P. vivax msp7* paralog has an ortholog in the closely related non-human primate parasite, *P. cynomolgi*. Furthermore, 10 of those 12 *msp7* paralogs shared by *P. vivax* and *P. cynomolgi* were also found in *P. fieldi* (Fig. 1), a species with a recent common ancestor shared with *P. vivax*, *P. cynomolgi*, and *P. inui*. (Muehlenbein et al., 2015; Pacheco et al., 2012). Although 10 paralogs found in *P. vivax* and

P. cynomolgi are shared with *P. fieldi*, it cannot be ascertained whether all 12 *msp7* paralogs are shared and syntenic between the three species due to the lack of genomic information on *P. fieldi*. Likewise, as in the case of *P. vivax* and *P. cynomolgi* (and possibly *P. fieldi*), all eight paralogs in the human parasite *P. falciparum* are shared with a non-human primate parasite found in African apes, *P. reichenowi* (Fig. 1). We found evidence indicating that seven out of the eight paralogs were also conserved in *P. gaboni*, a more distant species to *P. falciparum* within the *Laverania* subgenus (Ollomo et al., 2009; Otto et al., 2014; Pacheco et al., 2013). However, the sequences were not of good enough quality to be included in our analysis. Thus, these findings indicate that the origins of the *msp7* paralogs in *P. vivax* and *P. falciparum* prelate the most recent common ancestor for each of these human parasites and a respectively close related *Plasmodium* found in a non-human primate (see phylogenetic analyses below).

Unlike *P. vivax*, most of the proposed *P. cynomolgi* gene models for *msp7* include introns (Tachibana et al., 2012), for example, the paralogs PCYB_122740, PCYB_122750, PCYB_122770, PCYB_122780, and PCYB_122800. Yet, the sequences for those *P. cynomolgi* paralogs are in reading frame and can be aligned with their orthologs in *P. vivax*. This observation holds when different alleles of *P. cynomolgi* (obtained in this investigation) and orthologous genes from other closely related species are included in the alignments. Thus, the gene models proposed for *msp7* in *P. vivax* without introns (genome that has being curated further after its first release) seem to be appropriate for *P. cynomolgi*.

Finally, there is evidence consistent with pseudogenization on some *msp7* paralogs. In particular, *P. cynomolgi* has a pseudogene that is orthologs to PVX_082690, *P. inui* and *P. hylobati* have putative pseudogenes that are orthologs to PVX_082685, and *P. coatneyi* has a pseudogene that is orthologs to PVX_082660 (Fig. 1). Furthermore, *P. falciparum* and *P. reichenowi* share a pseudogene (PF3D7_1334900) in addition to their eight paralogs (Fig. 1).

3.2 Genetic diversity of the *msp7* paralogs in *P. vivax* and *P. falciparum*

No evidence was found indicating that *msp7* family size (number of paralogs) changes within each species (e.g.: *P. vivax*, *P. cynomolgi*, *P. inui*; see Fig. S2), thus *msp7* differs from *Pvmsp3* and its orthologs in non-human primates (Rice et al., 2014). However, as in *Pvmsp3* (Rice et al., 2014), there is sequence length polymorphism in many *msp7* paralogs in the *P. vivax* clade among different isolates within species (Fig. S2). This polymorphism is the result of low complexity regions (LCRs) with repetitive motifs located in the central regions of *msp7* multigene family members. While the size polymorphism and paralog composition suggest rapidly acting processes in the evolution of the *msp7* multigene family in some of the lineages inside the *P. vivax* clade, both features show little to no variation among species of the rodent clade and between *P. falciparum* and *P. reichenowi*.

Although most *P. vivax* paralogs seem to have higher average dS than dN substitutions, indicative of purifying selection, the null hypothesis of dN=dS was rejected in only six *Pvmsp7* paralogs. In particular, PVX_082650, PVX_082655, PVX_082665, PVX_082675, PVX_082680 and PVX_082685 showed significant dN<dS, a pattern expected when purifying selection is acting on the gene. In the case of *P. falciparum* (Table 1), the

polymorphism was markedly lower when compared to that found in *P. vivax*. Nevertheless, two *Pfmsp7* paralogs (PF3D7_1334500 and PF3D7_1335100) showed significant dN>dS, indicating that positive selection may be acting at those genes. As in *P. vivax*, the genetic polymorphism in orthologous genes from the non-human malarial parasites, *P. cynomolgi* and *P. inui*, showed evidence for purifying selection (Table 2). In particular, most *msp7* paralogs in *P. cynomolgi* have significantly higher average dS than dN. A similar pattern was observed in *P. inui*, however, it was not significant. Overall, the dS and dN patterns observed in many *Pvmsp7* paralogs and their orthologs in non-human primates are consistent with purifying selection.

3.3. Phylogenetic analysis of the *msp7* simian clade paralogs

Maximum likelihood (ML) and Bayesian inference (BI) phylogenetic trees were estimated to evaluate the relationships between *msp7* paralogs found on the *P. vivax* clade (Fig. 2). All pseudogenes and short length genes (PVX_082660 and PVX_082690 and their orthologs) were excluded from these analyses. Closely related paralogs that shared comparable amino acid composition (Fig. S1) and presented higher sequence similarity (Table S2) are labeled with identical or analogous color tones (Fig. 1 and Fig. 2). Orthologous genes formed well supported independent clades with the exception of PVX_082680 and PVX_082685.

Based on their phylogenetic relationships, *msp7* paralogs within the *P. vivax* clade were subdivided into three major groups (A-C). Groups A and C showed the highest posterior probabilities (Bayesian analysis) and bootstrap (ML analysis) support while group B was somewhat less supported (Fig. 2). Group A includes two of the three *msp7* paralogs (PVX_082645 and PVX_082695) that have orthologs in all the *Plasmodium* species considered in this study, including rodents. The phylogenetic relationships within ortholog PVX_082695 inside Group A paralogs clade are similar to those estimated for *Plasmodium* species using other loci and mtDNA genomes (Muehlenbein et al., 2015; Pacheco et al. 2011; Pacheco et al., 2012). Group B, considerably the largest in our phylogenetic analysis (Fig. 2), comprises paralogs that apparently originated during the radiation of the *P. vivax* clade. Group B also includes the *msp7* paralog PVX_082680 that has orthologs in all *Plasmodium* with genomic data available. The robustness of the group B branching pattern was confirmed by performing independent phylogenetic analyses excluding all *msp7* paralogs found outside group B (Fig. S3) and an additional analysis including only group B *msp7* paralogs with the highest sequence similarity (Fig. S4). In all these analyses no major changes were observed inside group B topological relations and the corresponding *P. gonderi msp7* paralogs formed a basal monophyletic group with low support. This is consistent with a scenario where duplication events may have occurred early on in *msp7* at the origin of the clade that includes *P. vivax*. Importantly, PVX_082685 appears to be undergoing a pseudogenization process in the lineage that includes *P. inui* and *P. hylobati*.

All lineages of paralogs included in group B have *P. vivax* and *P. cynomolgi* orthologs that are found adjacent in chromosome 12. These *P. vivax* and *P. cynomolgi* paralogs are shared with *P. fieldi* and less often by *P. simiolvale*. Although the lack of genomic data from *P. fieldi* and *P. simiovale* limit inferences based on how the phylogeny and paralog location in the chromosome relates across the clade, there are some clear duplications events that can be

observed. The PVX_082680 gene, for example, might have undergone a duplication event that resulted in PVX_082685 and its orthologs in *P. cynomolgi*, *P. fieldi*, and *P. simiovale*.

It is also worth noting that several paralogs with comparable amino acid composition and higher sequence similarity in group B are not adjacent in the *P. vivax* chromosome; this pattern suggests complex duplication/gene conversion events (Fig. 1) (Innan and Kondrashov, 2010). Although the lack of genomic data from some species did not allow us to properly reconstruct these processes, group B paralogs might have originated from a minimum of two major duplication events followed by additional gene duplications. Group C on the other hand, was formed by paralogs putatively originated by duplication events early in the radiation of the *P. vivax* clade, but their orthologs are not found in *P. knowlesi* and *P. coatneyi*, species with complete genome information. Furthermore, none of these *msp7* paralogs found in group C has orthologs in the rodent and *Laverania* clades.

Based on the phylogeny described above (Fig. 2), we performed a test for episodic selection on the divergence of the *msp7* paralogs among the species of the *P. vivax* clade. Table 3 shows the results of phylogenetic based test of selection as implemented in HyPhy. Four paralogs show evidence of episodic positive selection. Interestingly, these are the same paralogs whose polymorphisms show higher dS over dN indicating purifying selection (Table 1).

3.4 Recombination and selection patterns among group B *msp7* paralogs

Putative recombination events were evaluated only among group B paralogs (Fig. S3 and S5) based on their close phylogenetic relationship and the absence of distinct monophyly between PVX_082680 and PVX_082685 and their corresponding orthologs. An intricate pattern that encompassed all *msp7* paralogs of group B was detected between *P. vivax* and *P. cynomolgi*. These recombination events were observed among three major segments: one found in the N-terminus region of the protein and the other two on the C-terminus region (Fig. S5). Nonetheless, large variation in the location of the putative recombination breakpoints was also found. It is not clear how different paralog combinations were involved in each putative event detected, making it impossible to identify global recombination patterns among the group's paralogs. The amount of phylogenetic information found, when all species were included in the most generalized single recombination event detected (Fig. S5, it encompasses all group B paralogs), was lower than that observed in non-recombinant and highly conserved segments. Therefore, it was possible to conclude that recombination would not have a significant effect on our phylogenetic and selection analyses.

3.5 Conservation of HABP domains and phylogeny of *P. falciparum* paralogs

Plasmodium falciparum and *P. reichenowi* share all *msp7* paralogous genes indicating that they originated before these extant species evolved from a common ancestor (Fig. 1 and Fig. 3). Although the quality of the *P. gaboni* data did not allow its inclusion in this analysis, seven of the *Pfmsp7/Prmsp7* out of the eight genes were also found in *P. gaboni* (Fig. 1). The sequence conservation of the five HABPs described in PF3D7_1335100 (García et al., 2007) was evaluated among *msp7* paralogs in the *Laverania* subgenus (Fig. S6), no clear homology was detected in other species. A visual inspection of the *Laverania* alignments indicated that

these peptides were relatively conserved across *P. falciparum* paralogs and with their orthologous genes in *P. reichenowi* (Fig. S6). The number of HABPs with considerable sequence similarity among *Pfmsp7* paralogs varies from 1 to 4 (Table 4). Epitopes originally located in the C-terminus regions showed higher sequence conservation between *P. falciparum* paralogs. Particularly, epitope HA_26114 presented the fewest amino acid changes among *P. falciparum msp7* paralogs (e.g. 12/21, 60% conserved amino acids respect to PF3D7_1334800 and 15/21, 71% conserved amino acids respect to PF3D7_1334300; Fig S6). In addition to finding that all the HABP domains were shared by *P. falciparum* and *P. reichenowi* (Fig S6), the *msp7* paralogs without HABP domains were also found in *Pfmsp7* and *Prmsp7* families as can be observed in the phylogeny (Fig. 3). We also depicted the evolution of the HABPs sequences including their estimated ancestral sequences. In particular, this phylogeny shows that the putative HABPs were independently lost in two duplication events.

4. Discussion

Multigene families in *Plasmodium* harbor extraordinary variation to the extent that it is difficult to identify clear orthologs even in closely related species (Frech and Chen, 2011; Neasfy et al., 2012; Rice et al., 2014). In the case of *msp7*, orthologs are identifiable across species but they have diverged enough to hamper our ability to describe evolutionary processes at the genus level. This is worsened by the fact that genomic information is biased toward those *Plasmodium* species of biomedical importance, a sampling problem that is likely leading to an incorrect assessment of the number of duplication/loss events (Ness et al., 2011). Nevertheless, there are some clear patterns that emerged from our investigations.

This study found extensive variation in the numbers of paralogous genes for *msp7* across the genus *Plasmodium* (Garzón-Ospina et al., 2010; Kadakoppala and Holder, 2010), a pattern that has also been observed in other multigene families such as *msp3* (Rice et al., 2014) and SERA (Arisue et al., 2011). However, it is common to find variation in the actual number of paralogs within a *Plasmodium* species and yet, this is not the case with *msp7*. Indeed, we found no evidence of diversity in terms of the actual number of paralogs within the two human malarial parasites with multiple genomes, *P. falciparum* and *P. vivax*.

Plasmodium vivax and *P. cynomolgi* exhibited the largest numbers of *msp7* paralogs. This high number follows similar patterns reported for these two species in the SERA and *msp3* families. Nevertheless, there is no evidence of a recent expansion in the *P. vivax* lineage. Thus, the events leading to the high number of *msp7* paralogous genes in these two species likely occurred before their most recent common ancestor (2.36–5.27 Ma, Pacheco et al., 2011). Furthermore, some of these paralogs could be as old as the divergence of these two species from *P. ffieldi*, which we found to possess 10 out of the 12 *msp7* genes found in *P. vivax* and *P. cynomolgi* (Muehlenbein et al., 2015; Pacheco et al. 2012). However, we cannot confirm this older origin for all 12 extant *msp7* paralogs in *P. vivax* due to the absence of publicly available *P. ffieldi* genome sequences. Finally, as in *P. vivax*, the *Pfmsp7* paralogs are older than the origin of *P. falciparum* as evidenced by the complete conservation of *msp7* paralogs between *P. falciparum* (9) and *P. reichenowi* (9). Furthermore, most *Pfmsp7* paralogs (minimum 7) are found in *P. gaboni* (another parasite from African Apes). All these

observations indicate that the relatively high number of *msp7* paralogs is not an adaptation to the human host in *P. vivax* and *P. falciparum*.

In the case of *P. cynomolgi*, paralog PCYB_122760, previously reported for the genome of strain B, was also found in the Berok strain genome (Tachibana et al., 2012) but no orthologous gene was identified in any of the *P. vivax* isolates. Although previous studies suggested that PCYB_122760 and Pv082690 were orthologs (Garzón-Ospina et al., 2014), those paralogs cannot be reliably aligned. Thus, PCYB_122760 more likely originated from additional duplication events that took place in *P. cynomolgi* after its split from its common ancestor with *P. vivax*.

Tachinaba et al. (2010) automatic annotation proposed several introns for *msp7* paralogs in *P. cynomolgi*. Such a massive emergence of introns in a short period of time goes against the overall trends regarding the evolution of introns (Roy and Gilbert, 2005; Csuros et al., 2011) indicating a very interesting phenomenon. Alternatively, these introns could be an artifact of the automatic annotation (Sêton Bocco and Csuros, 2016), a hypothesis supported by visual examination of the alignments showing that the proposed introns are in reading frame and align well with the *msp7* gene models for *P. vivax*.

Evidence of an ancestral larger number of paralogous genes for the clade that includes *P. vivax* can be found in *P. gonderi*, a basal species to the Asian clade that is found in Africa (a parasite of white-eyelid mangabeys and mandrills) (Pacheco et al., 2012). Indeed, the larger family size in *Pgmsp7* (at least 9 paralogs) in relation to others in the *P. vivax* clade suggests that numerous paralogs may have had an early origin and perhaps many were present in the common ancestor of Asian primate malarias. However, lack of data from additional species does not allow to properly test this hypothesis. In particular, the *msp7* phylogeny (Fig. 2) showed that many *P. gonderi* paralogs may form a monophyletic group. Although this putative monophyletic group has low support in our phylogenetic analyses, it is still possible that some *Pgmsp7* paralogs originated independently via lineage specific duplication events in this African non-human primate parasite so the actual high number of *Pgmsp7* could be a convergent trait. Alternatively, gene conversion/recombination events within *Pgmsp7* could also result in an apparent monophyletic group in paralogs that originated earlier in the radiation of the species conforming this clade.

The phylogenetic gaps in the *msp7* sampling of paralogous genes likely led to an incorrect assessment of the number of duplication/loss events (Ness et al., 2011). Regardless of the obvious limitations in our sampling, gene duplications/losses and pseudogenization are evident in some clades (Fig. 1). This is a pattern consistent with a birth-and-death type of process where duplication events generate new paralogs with some duplications losing functionality (Nei et al., 1997; Nei and Rooney, 2005). We found a lower number of paralogs among some of the species that share a recent common ancestor with *P. vivax*. In particular, the number of paralogous genes found in the genomes of *P. knowlesi* (5), *P. coatneyi* (5), and *P. inui* (7) is lower than in *P. vivax* (12) and *P. cynomolgi* (12 plus a pseudogene) (Fig. 1 and 2). A similar pattern has been observed in families such as *msp3* (Rice et al., 2014). Unfortunately, we cannot confirm the actual sizes of *msp7* family in *P. simiovale* (minimum six paralogs) and *P. hylobati* (minimum four paralogs) due to the

absence of publicly available genomes. If a larger number of paralogs is ancestral in the group as suggested by *P. gonderi* data, the number of gene loss events found in *P. inui*, *P. knowlesi* and *P. coatneyi* is noteworthy as it has been documented in the same species for other multigene families expressed in the parasite's merozoitic stages (Arisue et al., 2007; Arisue et al., 2011; Rice et al., 2014). These parasites are remarkably different in terms of their life cycle and host range: *P. inui* (a quartan malaria), *P. knowlesi* (quotidian) and *P. coatneyi* (tertian) are parasites of macaques and surilis while *P. hylobati* is found in gibbons (Coatney et al., 1971). Another important element to consider is that these species diverged as part of a series of complex biogeographic processes involving multiple hosts (Muehlenbein et al., 2015; Pacheco et al., 2012). Thus, many of these loss-duplication events likely took place in a relatively short period of time accelerated by the interplay of selection and drift.

Gene duplications are expected to generate some functional redundancy (Innan and Kondrashov, 2010), however, having genes with identical functions may not be advantageous. Under this scenario, purifying selection is expected to relax and genes can start to accumulate deleterious mutations that lead to pseudogenization (Nei et al., 1997; Nei and Rooney, 2005; Innan and Kondrashov, 2010). Although we did not find evidence of pseudogenization taking place within *P. vivax*, we found indications of such a process in *P. cynomolgi* in our Sanger sequences (PVX_082690 ortholog). Furthermore, we observed putative pseudogenization events in *P. coatneyi* (corresponding to the PVX_082660 and PKH_121850 orthologs), as well as in *P. hylobati* and *P. inui* (corresponding to the PVX_082685 ortholog). In the case of *Laverania* subgenus, both *P. falciparum* and *P. reichenowi* also share a pseudogene (PF3D7_1334900 and PRCDC_1333900). It is also worth noticing that pseudogenization events were not detected in the genomes of rodent malarias.

These differences in family size and composition provide additional evidence for a birth-and-death type of process for the evolution of *msp7* in these species (Nei et al., 1997; Nei and Rooney, 2005). Under this model, paralogs can be produced via tandem or block duplications and the degeneration of paralogs via pseudogenization is common. In addition, multigene families evolving under the birth-and-death model are also prone to show subgroups with higher sequence similarity among divergent group of genes. This pattern can be observed in members of the *msp7* family (e.g. among paralogs PVX_082680, PVX_082685 and PVX_082655, Fig. 1 and Table S2).

The number of recombination events detected among closely divergent paralogs in the present study (Fig. S5) suggests that, as in other multigene families (Bethke et al., 2006), recombination played a role in the expansion of the *msp7* family via unequal crossing over. Furthermore, the increment of nucleotide diversity observed in recombinant regions indicates that this process might also have a role in generating genetic diversity as described in other *Plasmodium* multigene families (Nielsen et al., 2003; Garzón-Ospina et al., 2010; Pacheco et al., 2013; Garzón-Ospina et al., 2014; Rice et al., 2014). However, the conserved non-recombinant segments still allow reconstructing phylogenetic relationships among paralogs, especially if such study focuses on specific clades and low complexity regions are excluded.

Despite the numerous gene duplications and deletions observed in *msp7*, three paralogs (PVX_082645, PVX_082680 and PVX_082695) appear to have persisted in the three major *Plasmodium* clades evaluated in this study (Fig. 1). This might indicate functional importance. Indeed, PVX_082645 was reported to be under directional selection in *P. vivax* in a study that used a different approach (Cornejo et al., 2014). Although the specific functions of *Pvmsp7* is a matter that needs to be investigated, it has been reported that the proteolytic processing patterns of *msp7* multigene family members identified in *P. falciparum* and *P. vivax* showed certain similarities (Mongui et al., 2006).

A hypothesis emerging from all available data is that some of the *Pvmsp7* paralogs will likely be involved in the invasion of the red blood cell and they could form complexes with *msp1* gene as described in both rodent malaras and *P. falciparum* (Lin et al., 2016). Consistent with the functional importance of *Pvmsp7*, the limited transcriptomic data (Aurrecochea et al., 2009) together with the extensive evidence for purifying selection (Table 1), indicate that all *Pvmsp7* genes are expressed and likely still producing functional proteins. This pattern of extensive purifying selection is observed also in *P. cynomolgi*. In addition, some *Pvmsp7* paralogs also showed evidence of episodic (positive) selection in their divergence from *P. cynomolgi* by using phylogenetic methods (Table 3). These results show how natural selection may have differentially acted on *Pvmsp7* during its evolutionary history. First *Pvmsp7* paralogs diverged (episodic positive selection), but did not originate, driven by positive selection in an ancient parasite lineage that shifted from an Asian non-human primate (Cercopithecidae) to hominines. Then such paralogs that underwent an adaptive divergence in the *P. vivax* lineage, have been under purifying selection in the extant *P. vivax* populations indicating functional constraints. Interestingly, there is evidence that two of those genes (PVX_082675 and PVX_82680) are immunogenic and the latter is recognized by semi-immune individuals without fever after being challenged with sporozoites indicating that those antibodies could be associated with protection (Arévalo-Herrera et al., 2016; Hostetler et al., 2015).

Whereas the overall genetic polymorphism of *msp7* paralogs appears to be under functional constraint (purifying selection), positive selection may also interact with low complexity regions (Battistuzzi et al., 2016). However, this hypothesis cannot be tested with the evolutionary genetic analyses used in this investigation. Although previous studies have suggested that those LCRs in *Plasmodium* may have a role in immune evasion (Singh et al., 2004), their actual motifs could be maintained by purifying selection (Pacheco et al., 2012; Pacheco et al., 2013; Chenet et al., 2013). Whether this is the case in *msp7* needs to be addressed. The number and type of repetitive motifs of the LCRs varied between paralogs and orthologs in some species such as *P. inui* and *P. cynomolgi* (Table S3). For example, large strings of Glutamic Acid were observed in the PVX_082670, PVX_082675 and PVX_082680 orthologs in both *P. inui* and *P. cynomolgi*. This has been observed previously in other genes such as paralogs in the *msp3* multigene family (Rice et al., 2014), and the gene encoding the circumsporozoite protein (Pacheco et al., 2012; Pacheco et al., 2013).

The specific roles that *Pvmsp7* paralogs play in invasion remains unknown. We can only speculate that they may have a similar function to that in *Pfmsp7*. More than one *Pfmsp7* paralog may have the capacity to interact with proteins in the erythrocyte membrane (either

band 3 or ~52 kDa MSP1 protein), as observed for PF3D7_1335100 (García et al., 2007). Consistent with this notion, HABPs identified in PF3D7_1335100 originally located in the C-terminus were relatively conserved between *Pfmsp7* paralogs (Table 4). These peptides were also conserved in *P. reichenowi* and *P. gaboni* consistent with the observation made in *Pfmsp7* regarding their functional importance (Fig. S6). Interestingly, while it is clear that HABP epitopes located on the C-terminus region of *PMSP7* have a key role in erythrocyte interaction and invasion, the most conserved epitope (HA_26114) did not exhibit the highest peptide binding activity and resulted in the lowest percentages in invasion inhibition assays (24%) compared to other peptides, which displayed remarkably higher (50% or larger) inhibitory capacity (García et al., 2007). Thus, some of these motifs could actually be under other form of functional constraints and it is not necessarily an indication of their involvement in invasion of the red blood cell.

In summary, there is extraordinary diversity in *msp7* across *Plasmodium* species as evidenced by the number of paralogs and ongoing pseudogenization processes; a pattern consistent with a birth-and-death type of dynamic. The *msp7* diversity in the number of paralogs is particularly high in the clade that includes *P. vivax* and non-human primate parasites from Southeast Asia, such diversity may have been accelerated by the interplay of selection and drift (Muehlenbein et al., 2015; Pacheco et al., 2012) as suggested by the extraordinary phenotypic diversity and the complex biogeographic processes that affected the parasite hosts in the region. Whether there is comparable variation in the *Laverania* clade that includes *P. falciparum* is a matter that cannot be addressed at this time. However, the available evidence clearly indicates that the number of *msp7* paralogs in *P. vivax* and *P. falciparum* prelates the origin of these species as human parasites.

We also found evidence indicating that few *Pvmsp7* paralogs may have diverged from their orthologs in non-human primates by episodic selection. Thus, these paralogs may have been affected by the introduction of the lineage leading to *P. vivax* into Hominines from an ancestral host species that likely was a catarrhine. This observation, together with the conservation in the number of paralogs and the population data showing purifying selection acting on those paralogs, indicates that *Pvmsp7* is functionally important and may show some level of sub-functionalization as the divergence among paralogs seems to indicate. Finally, we observed conservation of HABP peptides across *Pfmsp7* paralogs and their orthologs in *P. reichenowi*. This pattern is consistent with experimental evidence pointing to functional redundancy among some of the *Pfmsp7* paralogs. All these findings from the evolutionary genetic analyses support the functional evidence emerging from *P. falciparum* indicating that *msp7* plays an important role in the invasion of the red blood cell and that such function may be shared across all *Plasmodium* species, including *P. vivax*.

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 and 1U19AI089702). We thank Amanda Poe and the DNA laboratory at the School of Life Sciences at Arizona State University for their technical

support. We thank the late William E. Collins from the Centers for Disease Control and Prevention for providing access to valuable specimens. We thank Monica Acosta and Judith Recht for editing the manuscript. The content is solely the responsibility of the authors and does not represent the official views of the US National Institute of Health.

References

- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997; 25:3389–33402. [PubMed: 9254694]
- Arévalo-Herrera M, Lopez-Perez M, Dotsey E, Jain A, Rubiano K, Felgner PL, Davies DH, Herrera S. Antibody profiling in naïve and semi-immune individuals experimentally challenged with *Plasmodium vivax* sporozoites. *PLoS Negl Trop Dis.* 2016; 10(3):e0004563. [PubMed: 27014875]
- Arisue N, Hirai M, Arai M, Matsuoka H, Horii T. Phylogeny and evolution of the SERA multigene family in the genus *Plasmodium*. *J Mol Evol.* 2007; 65:82–91. [PubMed: 17609844]
- Arisue N, Kawai S, Hirai M, Palacpac NM, Jia M, Kaneko A, Tanabe K, Horii T. Clues to evolution of the SERA multigene family in 18 *Plasmodium* species. *PLoS One.* 2011; 6:e17775. [PubMed: 21423628]
- Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Miller JA, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Stoeckert CJ Jr, Treatman C, Wang H. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.* 2009; 37:D539–D543. [PubMed: 18957442]
- Battistuzzi FU, Schneider KA, Spencer MK, Fisher D, Chaudhry S, Escalante AA. Profiles of low complexity regions in Apicomplexa. *BMC Evol Biol.* 2016; 29(16):47.
- Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res.* 2014; 42:D32–D37. [PubMed: 24217914]
- Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, Karlin S. Methods and algorithms for statistical analysis of protein sequences. *Proc Natl Acad Sci USA.* 1992; 89:2002–2006. [PubMed: 1549558]
- Bethke LL, Zilversmit M, Nielsen K, Daily J, Volkman SK, Ndiaye D, Lozovsky ER, Hartl DL, Wirth DF. Duplication, gene conversion, and genetic diversity in the species-specific acyl-CoA synthetase gene family of *Plasmodium falciparum*. *Mol Biochem Parasitol.* 2006; 150:10–24. [PubMed: 16860410]
- Boyle MJ, Langer C, Chan JA, Hodder AN, Coppel RL, Anders RF, Beeson JG. Sequential processing of merozoite surface proteins during and after erythrocyte invasion by *Plasmodium falciparum*. *Infect Immun.* 2014; 82:924–936. [PubMed: 24218484]
- Carlton JM, Escalante AA, Neafsey D, Volkman SK. Comparative evolutionary genomics of human malaria parasites. *Trends Parasitol.* 2008; 24:545–550. [PubMed: 18938107]
- Chenet SM, Pacheco MA, Bacon DJ, Collins WE, Barnwell JW, Escalante AA. The evolution and diversity of a low complexity vaccine candidate, merozoite surface protein 9 (MSP-9), in *Plasmodium vivax* and closely related species. *Infect Genet Evol.* 2013; 20:239–248. [PubMed: 24044894]
- Coatney, RG., Collins, WE., Warren, M., Contacos, PG. The Primate Malaras. US Government Printing Office; Washington: 1971.
- Cornejo OE, Fisher D, Escalante AA. Genome-wide patterns of genetic polymorphism and signatures of selection in *Plasmodium vivax*. *Genome Biol Evol.* 2014; 7:106–119. [PubMed: 25523904]
- Csuros M, Rogozin IB, Koonin EV. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol.* 2011; 7(9):e1002150. [PubMed: 21935348]
- Escalante AA, Freeland DE, Collins WE, Lal AA. The evolution of primate malaria parasites based on the gene encoding cytochrome b from the linear mitochondrial genome. *Proc Natl Acad Sci USA.* 1998; 95:8124–8129. [PubMed: 9653151]
- Frech C, Chen N. Genome comparison of human and non-human malaria parasites reveals species subset-specific genes potentially linked to human disease. *PLoS Comput Biol.* 2011; 7:e1002320. [PubMed: 22215999]

- García Y, Puentes A, Curtidor H, Cifuentes G, Reyes C, Barreto J, Moreno A, Patarroyo ME. Identifying merozoite surface protein 4 and merozoite surface protein 7 *Plasmodium falciparum* protein family members specifically binding to human erythrocytes suggests a new malarial parasite-redundant survival mechanism. *J Med Chem.* 2007; 50:5665–5675. [PubMed: 17944453]
- Garnham, PCC. *Malaria parasites and other haemosporidia.* Blackwell Scientific Publications; Oxford: 1966.
- Garzón-Ospina D, Forero-Rodríguez J, Patarroyo MA. Heterogeneous genetic diversity pattern in *Plasmodium vivax* genes encoding merozoite surface proteins (MSP) -7E, -7F and -7L. *Malar J.* 2014; 13:495. [PubMed: 25496322]
- Garzón-Ospina D, Cadavid LF, Patarroyo MA. Differential expansion of the merozoite surface protein (msp)-7 gene family in *Plasmodium* species under a birth-and-death model of evolution. *Mol Phylogenet and Evol.* 2010; 55:399–408.
- Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 2010; 59:307–321. [PubMed: 20525638]
- Hostetler JB, Sharma S, Bartholdson SJ, Wright GJ, Fairhurst RM, Rayner JC. A library of *Plasmodium vivax* recombinant merozoite proteins reveals new vaccine candidates and protein-protein interactions. *PLoS Negl Trop Dis.* 2015; 9(12):e0004264. [PubMed: 26701602]
- Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet.* 2010; 11:97–108. [PubMed: 20051986]
- Kadekoppala M, Holder A. Merozoite surface proteins of the malaria parasite: the MSP1 complex and the MSP7 family. *Int J for Parasitol.* 2010; 40:1155–1161.
- Kadekoppala M, O'Donnell RA, Grainger M, Crabb BS, Holder AA. Deletion of the *Plasmodium falciparum* merozoite surface protein 7 gene impairs parasite invasion of erythrocytes. *Eukaryot Cell.* 2008; 7:2123–2132. [PubMed: 18820076]
- Kadekoppala M, Ogun SA, Howell S, Gunaratne RS, Holder AA. Systematic genetic analysis of the *Plasmodium falciparum* MSP7-like family reveals differences in protein expression, location, and importance in asexual growth of the blood-stage parasite. *Eukaryot Cell.* 2010; 9:1064–1074. [PubMed: 20472690]
- Kauth CW, Woehlbier U, Kern M, Mekonnen Z, Lutz R, Mücke N, Langowski J, Bujard H. Interactions between merozoite surface proteins 1, 6, and 7 of the malaria parasite *Plasmodium falciparum*. *J Biol Chem.* 2006; 281:31517–31527. [PubMed: 16940297]
- Kosakovsky Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics.* 2005; 21:676–679. [PubMed: 15509596]
- Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delpont W, Scheffler K. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol.* 2011; 28:3033–3043. [PubMed: 21670087]
- Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. GARD: a genetic algorithm for recombination detection. *Bioinformatics.* 2006; 22:3096–3098. [PubMed: 17110367]
- Lin CS, Uboldi AD, Epp C, Bujard H, Tsuboi T, Czabotar PE, Cowman AF. Multiple *Plasmodium falciparum* merozoite surface protein 1 complexes mediate merozoite binding to human erythrocytes. *J Biol Chem.* 2016; 291:7703–7715. [PubMed: 26823464]
- Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* 2011; 2011(39):D225–D229.
- Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefevre P. RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics.* 2010; 26:2462–2463. [PubMed: 20798170]
- Mello K, Daly TM, Morrissey J, Vaidya AB, Long CA, Bergman LW. A multigene family that interacts with the amino terminus of *Plasmodium* MSP-1 identified using the yeast two-hybrid system. *Eukaryot Cell.* 2002; 1:915–25. [PubMed: 12477792]

- Mongui A, Perez-Leal O, Soto SC, Cortes J, Patarroyo MA. Cloning, expression, and characterization of a *Plasmodium vivax* MSP7 family merozoite surface protein. *Biochem Biophys Res Commun*. 2006; 351:639–644. [PubMed: 17081497]
- Muehlenbein MP, Pacheco MA, Taylor JE, Prall SP, Ambu L, Nathan S, Alsisto S, Ramirez D, Escalante AA. Accelerated diversification of nonhuman primate malarias in Southeast Asia: adaptive radiation or geographic speciation? *Mol Biol Evol*. 2015; 32:422–439. [PubMed: 25389206]
- Neafsey DE, Galinsky K, Jiang RH, Young L, Sykes SM, Saif S, Gujja S, Goldberg JM, Young S, Zeng Q, Chapman SB, Dash AP, Anvikar AR, Sutton PL, Birren BW, Escalante AA, Barnwell JW, Carlton JM. The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. *Nat Genet*. 2012; 44:1046–50. [PubMed: 22863733]
- Nei M, Gu X, Sitnikova T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc Natl Acad Sci USA*. 1997; 94:7799–806. [PubMed: 9223266]
- Nei M., Kumar, S. Molecular evolution and phylogenetics. Oxford University Press; NY: 2000.
- Nei M, Rooney AP. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet*. 2005; 39:121–152. [PubMed: 16285855]
- Ness RW, Graham SW, Barrett SC. Reconciling gene and genome duplication events: using multiple nuclear gene families to infer the phylogeny of the aquatic plant family Pontederiaceae. *Mol Biol Evol*. 2011; 28:3009–18. [PubMed: 21633114]
- Nielsen KM, Kasper J, Choi M, Bedford T, Kristiansen K, Wirth DF, Volkman SK, Lozovsky ER, Hartl DL. Gene conversion as a source of nucleotide diversity in *Plasmodium falciparum*. *Mol Biol Evol*. 2003; 20:726–734. [PubMed: 12679555]
- Ollomo B, Durand P, Prugnolle F, Douzery E, Arnathau C, Nkoghe D, Leroy E, Renaud F. A new malaria agent in African hominids. *PLoS Pathog*. 2009; 5(5):e1000446. [PubMed: 19478877]
- Otto TD, Rayner JC, Böhme U, Pain A, Spottiswoode N, Sanders M, Quail M, Ollomo B, Renaud F, Thomas AW, Prugnolle F, Conway DJ, Newbold C, Berriman M. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat Commun*. 2014; 5:4754. [PubMed: 25203297]
- 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, Cranfield M, Cameron K, Escalante AA. Malarial parasite diversity in chimpanzees: the value of comparative approaches to ascertain the evolution of *Plasmodium falciparum* antigens. *Malar J*. 2013; 12:328. [PubMed: 24044371]
- Pacheco MA, Reid MJ, Schillaci MA, Lowenberger CA, Galdikas BM, Jones-Engel L, Escalante AA. The origin of malarial parasites in orangutans. *PLoS One*. 2012; 7(4):e34990. [PubMed: 22536346]
- Perrin AJ, Bartholdson SJ, Wright GJ. P-selectin is a host receptor for *Plasmodium* MSP7 ligands. *Malar J*. 2015; 14:238. [PubMed: 26045295]
- Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*. 2011; 8:785–786. [PubMed: 21959131]
- Posada D. jModelTest: phylogenetic model averaging. *Mol Biol Evol*. 2008; 25:1253–1256. [PubMed: 18397919]
- Rice BL, Acosta MM, Pacheco MA, Carlton JM, Barnwell JW, Escalante AA. The origin and diversification of the merozoite surface protein 3 (*msh3*) multi-gene family in *Plasmodium vivax* and related parasites. *Mol Phylogenet Evol*. 2014; 78:172–84. [PubMed: 24862221]
- Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 2012; 61:539–542. [PubMed: 22357727]
- Roy SW, Gilbert W. Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc Natl Acad Sci USA*. 2005; 102:5773–5778. [PubMed: 15827119]
- Sêton Bocco S, Cs rös M. Splice Sites Seldom Slide: Intron Evolution in Oomycetes. *Genome Biol Evol*. 2016; 8:2340–2350. [PubMed: 27412607]

- Singh S, Soe S, Mejia JP, Roussilhon C, Theisen M, Corradin G, Druilhe P. Identification of a conserved region of *Plasmodium falciparum* MSP3 targeted by biologically active antibodies to improve vaccine design. *J Infect Dis.* 2004; 190:1010–1018. [PubMed: 15295710]
- Tachibana S, Sullivan SA, Kawai S, Nakamura S, Kim HR, Goto N, Arisue N, Palacpac NM, Honma H, Yagi M, Tougan T, Katakai Y, Kaneko O, Mita T, Kita K, Yasutomi Y, Sutton PL, Shakhbatyan R, Horii T, Yasunaga T, Barnwell JW, Escalante AA, Carlton JM, Tanabe K. *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. *Nat Genet.* 2012; 44:1051–1055. [PubMed: 22863735]
- Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol.* 2013; 30:2725–2729. [PubMed: 24132122]
- Tewari R, Ogun SA, Gunaratne RS, Crisanti A, Holder AA. Disruption of *Plasmodium berghei* merozoite surface protein 7 gene modulates parasite growth in vivo. *Blood.* 2005; 105:394–396. [PubMed: 15339842]
- Wasmuth J, Daub J, Peregrín-Alvarez JM, Finney CAM, Parkinson J. The origins of apicomplexan sequence innovation. *Genome Res.* 2009; 19:1202–1213. [PubMed: 19363216]
- Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF. Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol.* 1999; 112:531–552. [PubMed: 10027275]
- WHO. World Malaria Report 2015. World Health Organization; Geneva: 2015. p. 280

Abbreviations

<i>mSP7</i>	merozoite surface protein-7
SAPS	Statistical Analysis of Protein Sequences
MCMC	Markov Chain Monte Carlo
PSRF	Potential Scale Reduction Factor
ML	Maximum Likelihood
BI	Bayesian Inference
RDP3	Recombination Detection Program
HABP	High Activity Binding Peptides
LCRs	Low Complexity Regions

Highlights

- The array containing *msp7* genes is characterized among *Plasmodium* species.
- The number of *msp7* paralogs is particularly high in the *P. vivax* clade.
- The *msp7* paralogs in *P. vivax* and *P. falciparum* predates their origins as human parasites.
- Few *Pvmsp7* paralogs may have diverged from their orthologs in non-human primates by episodic positive selection.
- Some level of conservation of the HABP peptides across *Pfmsp7* and *Prmsp7* paralogs was observed.

Figure 1.

Msp7 multigene family organization. Data on the chromosome and position of the *msp7* family is limited to those species with complete annotated genomes. Syntenic *msp7* blocks are found in the *P. vivax* clade (Chromosome 12), and the rodent and *Laverania* clades (Chromosome 13). Ortholog genes are indicated by vertical lines connecting across different species, while paralogs are depicted in horizontal lines for each *Plasmodium* species with genomic data available. Most paralogs are represented by diamond shapes; PVX_082660 and PVX_082690 and their respective orthologs are represented by triangles showing their similarity only to the C-terminus or N-terminus regions, respectively. Same color paralogs are more closely related phylogenetically and share overall similar sequence patterns (e-values, Table S2). Pseudogenes are indicated by non-continuous lines. *Plasmodium* species

without annotated genomes are indicated by a red rectangle. Question marks indicate paralogs that were not obtained experimentally but are present in closely related *Plasmodium* species.

Figure 2.

Bayesian inference (BI) and Maximum Likelihood (ML) *msp7* multigene family phylogenetic tree. BI and ML trees showed almost identical topologies, so only BI topology is shown with asterisks (*) indicating conflicting branching patterns. Posterior probabilities (PP) and bootstrap values (BV) are shown next to the phylogenetic tree nodes (PP/BV). Same color branches indicate that paralogs share similar sequence patterns (e-values, Table S2). The tree was constructed from 233 *msp7* paralogs excluding those of short sequence length (PVX_082660 and PVX_082690 and respective orthologs). A total of 858 nucleotide positions were included in the analysis and the GTR+I+ Γ nucleotide model (inv. sites =

0.0060; $\alpha=1.6470$) was used. Paralogs were divided into three mayor groups (Group A, Group B and Group C).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 3.

Bayesian inference (BI) and Maximum Likelihood (ML) phylogenetic tree of MSP7 multigene family paralogs found in *Plasmodium* species from the *Laverania* clade. BI and ML trees showed similar topologies, so only BI topology is shown with asterisks (*) indicating conflicting branching patterns. Posterior probabilities (PP) and bootstrap values (BV) are shown next to the phylogenetic tree nodes (PP/BV). Branch which share the same coloration are more closely related and share overall similar sequence patterns. The tree was constructed from 16 *msp7* multigene family members. A total of 2,157 nucleotide positions were included in the analysis and the GTR+ Γ nucleotide model ($\alpha = 2.9310$) was used. Putative ancestral sequences and extant sequences of epitope HAP26114 were determined for each paralog. Amino acid changes are colored; changes between same types of amino acids share the same color.

Table 1

Polymorphism in *msp7 P.vivax* and *P. falciparum* paralogs. Analyses were carried out in the complete gene and carboxyl/amino termini for each gene. Identifiable central regions were analyzed whenever possible.

Sp.	Paralog ID	n	π [SE]	dS	dN	dS-dN [SE]	p (Z stat)
<i>P. vivax</i>	PVX_082645	110	0.001 [0.000]	0	0.001	0.0 [0.0]	dN=dS 0.314 (1.011)
				0	0	0.0 [0.0]	dN > dS 0.028 (2.226)
				0.001 [0.000]	0.001	0.000 [0.001]	dN=dS 0.502 (0.674)
	PVX_082650	16	0.014 [0.002]	0.039	0.008	0.031 [0.008]	dN < dS 0.0 (-3.839)
				0.007 [0.002]	0.027	0.001 [0.008]	dN < dS 0.030 (-3.063)
				0.022 [0.003]	0.052	0.037 [0.016]	dN < dS 0.027 (-2.241)
	PVX_082655	45	0.015 [0.003]	0.048	0.007	0.041 [0.010]	dN < dS 0.0 (-4.172)
				0.055 [0.004]	0.068	0.012 [0.011]	dN=dS* 0.265 (-1.119)
				0.019 [0.004]	0.078	0.007 [0.019]	dN < dS 0.0 (-3.635)
	PVX_082660	97	0.0184 [0.014]	0.12	0.28	-0.160 [0.037]	dN > dS* 0.0 (4.0)
				0.011 [0.003]	0.027	0.007 [0.011]	dN < dS 0.080 (-1.767)
				0.001 [0.001]	0.001	0.000 [0.002]	dN=dS 0.885 (-1.145)
<i>P. falciparum</i>	PVX_082665	42	0.027 [0.003]	0.094	0.012	0.082 [0.015]	dN < dS 0.0 (-5.372)
				0.051 [0.004]	0.066	0.002 [0.015]	dN=dS* 0.892 (-0.135)
				0.016 [0.003]	0.074	0.005 [0.022]	dN < dS 0.006 (-2.813)
	PVX_082670	94	0.066 [0.005]	0.062	0.142	-0.080 [0.021]	dN > dS* 0.0 (4.058)
				0.036 [0.005]	0.112	0.093 [0.022]	dN < dS 0.0 (-3.944)
				0.001 [0.001]	0	-0.001 [0.001]	dN=dS 0.145 (1.468)
	PVX_082675	23	0.000 [0.000]	0	0	0.000 [0.000]	dN > dS* 0.019 (2.368)
				0.002 [0.001]	0	-0.001 [0.001]	dN=dS 0.206 (1.273)
				0.013 [0.002]	0.027	0.019 [0.006]	dN < dS 0.001 (-3.382)
	PVX_082680	23	0.008 [0.002]	0.017	0.006	0.011 [0.006]	dN=dS 0.080 (-1.768)
				0.017 [0.003]	0.038	0.026 [0.009]	dN < dS 0.005 (-2.840)

Sp.	Paralog ID	n	π [SE]	dS	dN	dS-dN [SE]	d (Z stat)
	PVX_082680	63	0.023 [0.002]	0.076	0.012	0.065 [0.012]	dN < dS 0.0 (-5.635)
				0.157	0.016	0.140 [0.026]	dN < dS 0.0 (-5.187)
				0.015	0.007	0.009 [0.008]	dN = dS 0.280 (-1.086)
	PVX_082685	82	0.035 [0.003]	0.043	0.013	0.031 [0.010]	dN < dS 0.003 (-3.070)
				0.056	0.003	0.053 [0.020]	dN < dS 0.014 (-2.492)
				0.036	0.02	0.016 [0.011]	dN = dS 0.142 (-1.477)
	PVX_082690	106	0.002 [0.002]	0.001	0.003	-0.002 [0.002]	dN = dS 0.399 (0.847)
				0.003	0.003	0.000 [0.003]	dN = dS 0.891 (0.138)
				0	0.002	-0.002 [0.001]	dN = dS 0.197 (1.297)
	PVX_082695	107	0.005 [0.002]	0.005	0.005	0.000 [0.006]	dN = dS 0.938 (-0.078)
				0.002	0.001	0.001 [0.001]	dN = dS 0.546 (-0.606)
				0.002	0.001	0.000 [0.002]	dN = dS 0.961 (-0.050)
	PF3D7_1335100	196	0.006 [0.001]	0.001	0.007	0.010 [0.002]	dN > dS 0.546 (2.623)
				0	0	0.000 [0.002]	dN = dS 0.290 (-1.064)
				0.002	0.013	-0.011 [0.004]	dN > dS 0.007 (2.769)
	PF3D7_1335000	186	0.011 [0.003]	0	0	0.000 [0.000]	dN = dS 0.815 (-0.234)
				0	0	0.000 [0.000]	dN = dS 0.300 (1.040)
				0	0.004	-0.003 [0.003]	dN = dS 0.674 (0.422)
	PF3D7_1334900	196	0.003 [0.002]	0	0	0.000 [0.000]	dN = dS 0.134 (1.509)
				0	0	0.000 [0.000]	dN = dS 0.452 (-0.755)
				0	0	0.001 [0.001]	dN = dS 0.0 (3.645)
	PF3D7_1334800	178	0.002 [0.001]	0	0.003	-0.003 [0.001]	dN > dS 0.002 (3.208)
				0	0.005	-0.004 [0.001]	dN > dS 0.031 (2.179)
				0	0.001	-0.001 [0.000]	dN > dS 0.048 (1.991)
<i>P. falciparum</i>	PF3D7_1334700	210	0.004 [0.001]	0	0.002	-0.002 [0.001]	dN > dS 0.048 (1.991)
				0	0.002	-0.002 [0.001]	dN > dS 0.048 (1.991)
				0	0.002	-0.002 [0.001]	dN > dS 0.048 (1.991)
	PF3D7_1334600	195	0.002 [0.001]	0	0.002	-0.002 [0.001]	dN > dS 0.048 (1.991)
				0	0.002	-0.002 [0.001]	dN > dS 0.048 (1.991)
				0	0.002	-0.002 [0.001]	dN > dS 0.048 (1.991)
	PF3D7_1334500	110	0.002 [0.001]	0	0.003	-0.003 [0.001]	dN > dS 0.0 (3.645)
				0	0.005	-0.004 [0.001]	dN > dS 0.002 (3.208)
				0	0.001	-0.001 [0.000]	dN > dS 0.031 (2.179)
	PF3D7_1334400	195	0.002 [0.001]	0	0.002	-0.002 [0.001]	dN > dS 0.048 (1.991)
				0	0.002	-0.002 [0.001]	dN > dS 0.048 (1.991)
				0	0.002	-0.002 [0.001]	dN > dS 0.048 (1.991)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Sp.	Paralog ID	n	π	[SE]	dS	dN	dS-dN	[SE]	d	(Z stat)	dN=dS
	PF3D7_1334300	209	0.003	[0.001]	0.001	0.003	-0.001	[0.001]	0.482	(-0.705)	dN=dS
	CDS (1398bp)										
	N-terminal (1-699)				0.006	0.003	-	[0.003]	0.364	(-0.912)	dN=dS
	C-Terminal (700-1398)				0	0	0.000	[0.000]	0.390	(0.862)	dN=dS

Δ Low Complexity Region (LCR)
* Include the central region of the gene

Table 2

Polymorphism in *msp7* multigene family *P. vivax* clade paralogs.

Species	N	Paralog ID *	π [SD]	dS	dN	DN-dS [SD]	Z test	Neutrality
<i>P. cynomolgi</i>	8	PVX_082645	0.088 [0.007]	0.186	0.077	0.109 [0.032]	0.002 (-3.220)	dN < dS
	9	PVX_082650	0.101 [0.005]	0.189	0.091	0.097 [0.020]	0 (-4.773)	dN < dS
	8	PVX_082655	0.073 [0.005]	0.121	0.066	0.055 [0.016]	0.001 (-3.548)	dN < dS
		PVX_082655 ^I	0.060 [0.005]	0.149	0.042	0.106 [0.020]	0 (-5.009)	dN < dS
	9	PVX_082660	0.073 [0.008]	0.125	0.069	0.056 [0.028]	0.047 (-2.012)	dN < dS
	8	PVX_082665	0.083 [0.006]	0.117	0.067	0.111 [0.022]	0 (-5.159)	dN < dS
	8	PVX_082670	0.055 [0.004]	0.100	0.048	0.052 [0.014]	0 (-3.634)	dN < dS
	8	PVX_082675	0.108 [0.006]	0.204	0.098	0.106 [0.024]	0 (-4.628)	dN < dS
	7	PVX_082680	0.060 [0.005]	0.1	0.054	0.046 [0.016]	0.005 (-2.880)	dN < dS
	8	PVX_082685	0.079 [0.005]	0.163	0.066	0.096 [0.021]	0 (-4.465)	dN < dS
	8	PVX_082690	0.053 [0.009]	0.096	0.048	0.049 [0.033]	0.143 (-1.473)	dN = dS
	8	PVX_082695	0.037 [0.004]	0.067	0.031	0.036 [0.014]	0.010 (-2.628)	dN < dS
<i>P. inui</i>		PVX_082700	0.037 [0.004]	0.072	0.030	0.042 [0.012]	0 (-3.750)	dN < dS
		PVX_082700 ^I	0.037 [0.004]	0.070	0.031	0.040 [0.013]	0.003 (-3.016)	dN < dS
	8	PVX_082645	0.036 [0.003]	0.051	0.033	0.018 [0.008]	0.026 (-2.248)	dN < dS
	9	PVX_082670	0.020 [0.002]	0.027	0.018	0.009 [0.006]	0.157 (-1.424)	dN = dS
	10	PVX_082675	0.036 [0.003]	0.032	0.038	-0.006 [0.006]	0.340 (0.958)	dN = dS
	9	PVX_082680	0.036 [0.003]	0.059	0.034	0.016 [0.008]	0.051 (-1.972)	dN = dS
	6	PVX_082685	0.024 [0.003]	0.027	0.022	0.005 [0.007]	0.472 (-0.722)	dN = dS
	9	PVX_082695	0.030 [0.004]	0.044	0.027	0.017 [0.010]	0.102 (-1.649)	dN = dS
	9	PVX_082700	0.023 [0.002]	0.025	0.023	0.002 [0.006]	0.764 (-0.301)	dN = dS
	3	PVX_082660	0.009 [0.003]	0.029	0.004	0.026 [0.015]	0.081 (-1.758)	dN = dS
<i>P. knowlesi</i>	3	PVX_082675	0.045 [0.004]	0.026	0.053	-0.026 [0.010]	0.007 (2.725)	dN > dS
	3	PVX_082680	0.053 [0.006]	0.073	0.052	0.021 [0.015]	0.175 (-1.365)	dN = dS
	3	PVX_082695	0.010 [0.003]	0.013	0.01	0.003 [0.008]	0.683 (-0.410)	dN = dS

* *P. vivax* PlasmoDB nomenclature (Carlton et al., 2008).

No putative intron included (see PlasmomDB).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Branch and episodic selection in *P. vivax* clade *msp7* paralogs.

Branch	Paralogs*											
	PVX_082645	PVX_082650	PVX_082655	PVX_082660	PVX_082665	PVX_082670	PVX_082675	PVX_082680	PVX_082685	PVX_082695	PVX_082700	
<i>P. vivax</i>	0.29	0.5049 ^e	0.9424 ^e	0.382	0.7166	0	0.7002 ^e	0.3562 ^e	0.6558 ^e	0.3434	0.3261	
<i>P. cynomolgi</i> A	0.5686	0.3629	0.4563	1.1777	0.407	0.4941	0.6719	0.6854	0.408	0.5891	0.5792	
<i>P. cynomolgi</i> B	0.366	0.4991	0.6414	0.8418	0.4205	0.6893	0.7768	0.5279	0.6437	0.6125	0.4469	
<i>P. fieldi</i>	0.689	0.9202	0.449	2.4149	0.9976 ^e	-	0.9245	0.833	0.7149 ^e	0.5441	1.372	
<i>P. simiovale</i>	0.3321	0.4205	-	-	0.3877	-	-	0.6704 ^e	0.6179 ^e	0.8414	-	
<i>P. inui</i>	0.6568	-	-	-	-	0.8482	1.2523	0.6648	-	1.0493	0.8212	
<i>P. hyllobati</i>	0.939 ^e	-	-	-	-	-	-	0.7507	-	-	0.8425	
<i>P. knowlesi</i>	0.3012	-	-	0.1688	-	-	0.8071	0.6998	-	1.005	-	
<i>P. coatneyi</i>	0.3694 ^e	-	-	-	-	-	1.6571	1.0023	-	0.4085	-	
<i>P. gonderi</i>	0.13	0.0206	0.1644	-	-	0.4781	0.2704	0.1243	0.1524	0.1558	-	

* *P. vivax* PlasmoDB nomenclature (Carlton et al. 2008).

Positively selected branches (model 2, codeml) are indicated by bolded omega values.

^e Indicate branches with signature of episodic selection (branch-site model, HyPhy).

Table 4

Detection of HABPs among *P. falciparum* MSP7 paralogs

Location	Region	HABP	Sequences	Paralog [*]						
				PF3D7_1335100 [¶]	PF3D7_1335000	PF3D7_1334800	PF3D7_1334700	PF3D7_1334500	PF3D7_1334400	PF3D7_1334300
20 kDa fragment's N- and C-terminal	N-terminal	26101	IKNKKLEKLNIVSGDFVGNV	x	x	x		x		
	Central	26107	NLGLFGKNVLSKVKQAQSETDY	x						
		26114	EKDKEYHEQFKNYIYGVYSYA	x	x	x	x		x	
19 kDa fragment's C-terminal	C-terminal	26115	KQNSHLSEKKIKPEEEYKKF	x	x			x	x	
		26116	EKPEEEYKKFLEYSFNLLNTM	x	x			x		x

HABP: High Activity Binding Peptides.

Paralogs with similar HABP sequences in the corresponding genomic region are indicated by x.

^{*} *P. falciparum* PlasmoDB nomenclature (Mello et al., 2002).

P. falciparum HABPs (García et al., 2007).

[¶]HABP epitopes have been described in PF3D7_1335100 (García et al., 2007).