Evolution of Multigene Families and Single Copy Genes in *Plasmodium* spp.

by

Andreina I. Castillo Siri

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved May 2016 by the
Graduate Supervisory Committee:

Michael Rosenberg, Chair
Ananias Escalante
Jesse Taylor
James Collins

ARIZONA STATE UNIVERSITY

August 2016

ABSTRACT

The complex life cycle and widespread range of infection of *Plasmodium parasites,* the causal agent of malaria in humans, makes them the perfect organism for the study of various evolutionary mechanisms. In particular, multigene families are considered one of the main sources for genome adaptability and innovation. Within *Plasmodium*, numerous species- and clade-specific multigene families have major functions in the development and maintenance of infection. Nonetheless, while the evolutionary mechanisms predominant on many species- and clade-specific multigene families have been previously studied, there are far less studies dedicated to analyzing genus common multigene families (GCMFs). I studied the patterns of natural selection and recombination in 90 GCMFs with diverse numbers of gene gain/loss events. I found that the majority of GCMFs are formed by duplications events that predate speciation of mammal *Plasmodium* species, with many paralogs being neutrally maintained thereafter. In general, multigene families involved in immune evasion and host cell invasion commonly showed signs of positive selection and species-specific gain/loss events; particularly, on *Plasmodium* species is the simian and rodent clades. A particular multigene family: the merozoite surface protein-7 (*msp7*) family, is found in all *Plasmodium* species and has functions related to the erythrocyte invasion. Within *Plasmodium vivax*, differences in the number of paralogs in this multigene family has been previously explained, at least in part, as potential adaptations to the human host. To investigate this I studied *msp7* orthologs in closely related non-human primate parasites where homology was evident. I also estimated paralogs' evolutionary history and genetic

polymorphism. The emerging patterns where compared with those of *Plasmodium falciparum*. I found that the evolution of the *msp7* multigene family is consistent with a Birth-and-Death model where duplications, pseudogenization and gene lost events are common. In order to study additional aspects in the evolution of *Plasmodium*, I evaluated the trends of long term and short term evolution and the putative effects of vertebrate-host's immune pressure of gametocytes across various *Plasmodium* species. Gametocytes, represent the only sexual stage within the *Plasmodium* life cycle, and are also the transition stages from the vertebrate to the mosquito vector. I found that, while male and female gametocytes showed different levels of immunogenicity, signs of positive selection were not entirely related to the location and presence of immune epitope regions. Overall, these studies further highlight the complex evolutionary patterns observed in *Plasmodium*.

DEDICATION

To my family, my friends, and my mentors.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1. INTRODUCTION

This work focuses on understanding evolutionary mechanisms affecting the evolution of multigene families using *Plasmodium* spp., the causal agent of malaria in humans, as a model organism. The putative selective role of host-parasite interactions from a complex life cycle is evaluated. In addition, both short and long term evolutionary trends of transmissible *Plasmodium* spp. stages are explored.

1.1 *Plasmodium life cycle*

The *Plasmodium* life cycle begins when sporozoites are injected into the vertebrate host after an *Anopheles* mosquito feeds from an infected vertebrate host. Sporozoites travel to the host liver and within minutes invade the host's hepatocytes and replicate as hepatic schizonts. Eventually, merozoites are produced and released into the blood stream completing the exo-erythrocytic cycle. In certain *Plasmodium* species (*P. vivax*, *P. ovale* and *P. cynomolgi*), some of the liver parasites remain quiescent only to resume replication after several weeks or months. This life cycle stage, which is known as the hypnozoite (Dembélé et al., 2014; Siciliano and Alano, 2015), is thought to cause malaria relapses (Markus, 2015).

After entering the blood stream, parasites invade the host red blood cells and undergo several rounds of asexual replication (the erythrocytic cycle). Invasion of new red blood cells occurs when already infected ones are ruptured by the formation of mature schizonts and the release of new merozoites (Siciliano and Alano, 2015). Symptoms associated with malaria infection are caused by the cycle of parasite replication, red blood cell invasion, rupture and release of merozoites. Variation in the

periodicity of this cycle, as well as in the type of red blood cells infected by the parasite, are commonly observed among *Plasmodium* species (Carlton et al., 2008). A small proportion of parasites commit to the sexual pathway and differentiate into male and female gametocytes (Kuehn et al., 2010). This differentiation is highly flexible and is thought to be mediated by a variety of biological and environmental stressors (Alano, 2007; Talman et al., 2004). When ingested during the mosquito blood meal, various changes in parasite environment (drop of body temperature, presence of xanthurenic acid and increase of pH) result in gametocyte activation and the formation of gametes (Sinden, 2015).

Finally, gametes fuse in the *Anopheles* mosquito midgut and produce a zygote, which later develops into a motile ookinete capable of traversing the midgut epithelium and transforming into an oocyst. The thousands of sporozoites produced by the oocyst proceed to navigate towards the mosquito's salivary glands where they can be injected into another vertebrate host and complete the sporogonic cycle (Siciliano and Alano, 2015).

1.2 *Plasmodium and host interaction*

Parasites of the *Plasmodium* genus are capable of infecting a wide range of vertebrate hosts; however, each *Plasmodium* species can only infect certain host types (*e.g.*, reptiles, birds, rodents and primates). Several studies have shown that parasite adaptation can occur as a response to interaction with different types of vertebrates (*e.g.*, primates vs. rodents), and variable selective constraints (Assefa et al., 2015; Frech and Chen, 2011; Prugnolle et al., 2008). It is possible for host-switch events to occur when

closely related vertebrates share a common environment (Krief et al., 2010; Mu, 2005). The causal agents of major human malarias, *P. vivax* and *P. falciparum*, are thought to have originated from two independent host-switch events between humans, Southeast Asian macaques (Carlton et al., 2013; Escalante et al., 2005) and African gorillas (Liu et al., 2010), respectively. Thus, host-switch events are fundamental in the evolutionary history of *Plasmodium* (Duval and Ariey, 2012).

Plasmodium parasites also spend part of their life cycle in an *Anopheles* mosquito vector. Parasite-vector associations are thought to be species-specific to a certain extent, with some mosquito species having higher vectorial capacity than others (Kamali et al., 2012; Tainchum et al., 2015). Furthermore, numerous studies have found evidence that both *P. falciparum* and *P. vivax* are adapted to different *Anopheles* mosquito species worldwide (Sinka et al., 2012), and that selection leading to parasite adaptation to different vector species can occur locally (Joy et al., 2008; Molina-Cruz et al., 2012).

1.3 *Origins of Plasmodium spp. - host associations*

Vectorial capacity seems to have been gained independently after the divergence of various mosquito lineages (Kamali et al., 2012). Primary *P. falciparum* vectors, *Anopheles gambiae* and *An. funestus* diverged between 30-40 Mya, while the main Asian and South American vectors of *P. vivax* (*An. stephensi* and *An. darlingi*, respectively) are thought to have diverged approximately 100 Mya (Kamali et al., 2014; Neafsey et al., 2015). In contrast, when measured using protein-coding nuclear genes and the rate of pairwise amino acid sequence divergence, *Plasmodium* parasite associations with their respective vertebrate hosts are thought to be more recent. Specifically, the split of *P.*

falciparum and *P. reichenowi* is thought to have occurred about 3.0–5.5 Mya, a time that overlaps with the estimated divergence between humans and chimpanzees (4.9-6.8 Mya) (Kumar et al., 2005; Schrago and Voloch, 2013). On the other hand, parasites of the rodent clade are estimated to have diverged around 13-25 Mya, coinciding with the diversification of the family Muridae (Silva et al., 2015). Overall, mammal malarias are thought to have radiated approximately in the late Mesozoic (around 64 Mya), establishing an overlap with the divergence between the primate and rodent lineages (Silva et al., 2015).

1.4 *Immune response to Plasmodium spp. infection and parasite's evasion*

Innate immune mechanisms, thought to be triggered by conserved molecules among *Plasmodium* species, limit the initial density of blood-stage parasites irrespective of the *Plasmodium* species or strain (Stevenson and Riley, 2004). Immune mechanisms directed at blood-stage parasites involve inflammatory processes and antibody responses. Inflammation is triggered by pattern recognition receptors (PRRs) expressed by immune cells in response to *P. falciparum* infection (Crompton et al., 2014). In addition, immunoregulatory cytokines (IL-10 and TGF-ß), which contribute to the regulation of innate responses, are produced by the innate (macrophages) and adaptive (T cells) immune systems. Production of these cytokines activates the dendritic cells (DC), enhances the effect of parasite-derived maturation stimuli, and facilitates clonal expansion of antigen-specific CD4+T cells (Pouniotis et al., 2004; Stevenson and Riley, 2004). Furthermore, high antibody levels, CD4+ and CD8+ T cell responses have also been found against proteins expressed in the infective sporozoite stage (CSP, LSA-1 and

TRAP), where they are thought to reduce severity of disease in recently infected individuals (Offeddu et al., 2012).

Also, mechanisms such as cytoadherence, rosetting, antigenic variation and antigenic diversity, are used by the parasite to evade anti-malarial immunity within the vertebrate host (Deroost et al., 2016). Cytoadherence in *P. falciparum* allows infected erythrocytes to sequester in the microvasculature of multiple organs and evade host's immune responses by passage through the spleen (splenic entrapment). In certain species (*P. vivax* and *P. yoelii*), infection can be maintained in reticulocyte-rich environments such as the bone marrow (Malleret et al., 2015), or by creating reticulocyte-rich environment that enhances cell invasion (Deroost et al., 2016). Rosetting is thought to shield infected erythrocytes from opsonization and facilitate invasion of new erythrocytes by decreasing the distance between infected and non-infected cells (Lee et al., 2014; Niang et al., 2014). In addition, antigenic variation permits the evasion of vertebrate immune responses by altering the expression of surface proteins. Specifically, only one antigen type is expressed at a time during infections, while other loci are not transcribed (Scherf et al., 2008). This helps the parasite to evade the host immune response and also extends the parasite's survival within a single host (Abdi et al., 2016).

While the immune response of the vertebrate host involves adaptive and innate immune mechanisms, *Anopheles* mosquitoes combat *Plasmodium* infection via physical barriers (the peritrophic matrix and endothelium) and innate immune mechanisms (Crompton et al., 2014). Large bottlenecks are caused by parasite transversal of mosquito physical barriers (Saraiva et al., 2016). On the other hand, the innate immune response

consists of: 1) hemocytes becoming capable of phagocytosis in regions of high hemolymph flow, and 2) production of humoral factors leading to lysis and melanization of *Plasmodium* parasites (King and Hillyer, 2012).

1.5 *Plasmodium spp. multigene families*

Multigene families have a fundamental role as sources of adaptation and diversification among *Plasmodium* species (DeBarry and Kissinger, 2011; Kooij et al., 2005; Kuo and Kissinger, 2008; Weir et al., 2009). Furthermore, the largest differences among species' genomes have been found within their multigene families (Tachibana et al., 2012). Comparative studies performed within three of the major malaria clades have found common trends regarding clade-specific multigene family evolution. Repeated lineage-specific gene duplication and/or deletion events, evidenced by variation in the number of paralogs and paralog composition, have been described in the simian clade (Tachibana et al., 2012). A similar trend has been observed in the *Laveranian* subgenus, where certain multigene families (*Rifin* and *Stevor*) have variable size and composition, while others (*Phist*, *Fikk* and *var*) maintain a common family organization and share some easily identifiable orthologs (Otto et al., 2014b). Lineage-specific duplication and/or loss events have also been observed in multigene families unique to the rodent malaria clade (Otto et al., 2014a). Important functions, many of them related to parasite-host interaction and cell invasion, are known to be performed by subsets of species- or clade-specific multigene families (Frech and Chen, 2011).

Specifically within *Plasmodium*, clade- and species-specific multigene families have functions associated with cytoadherence and antigenic variation (*var*, SICA*var*,

Stevor, *Rifin*, *fikk*, *pir*), recognition and invasion of erythrocytes (*msp7*, *msp3*) and organelle formation (ETRAMP, *Phist*) (Reid, 2015; Tachibana et al., 2012). For the most part, clade- and species-specific multigene families have a tendency to show large numbers of paralogs and variable paralog composition among *Plasmodium* species, or even among strains of the same species, in comparison to multigene families shared by largely divergent *Plasmodium* species (Cheeseman et al., 2009; Iyer et al., 2006; Reid, 2015; Rice et al., 2014). Many of these multigene families are located in the highly recombinant sub-telomeric chromosome regions, or in internal regions with sub-telomeric-like repeats, a fact that is thought to contribute to facilitating recombination and generating variability (Taco W. A. Kooij et al., 2005; Kuo and Kissinger, 2008; Pain et al., 2008).

While many important clade- and species-specific multigene families are associated with sub-telomeric chromosome regions, subsets of species- and clade-specific genes potentially linked to virulence and transmission have also been found in internal regions of chromosomes (Frech and Chen, 2011; Taco W. A. Kooij et al., 2005). Internal chromosome regions tend to be highly syntenic among *Plasmodium* species, particularly those that are closely related (DeBarry and Kissinger, 2011). However, changes in synteny are usually associated with clade- and species-specific expansions and contractions of genus common multigene families (Tachibana et al., 2012).

The variation of the number of paralogs among *Plasmodium* species is thought to be related to host-specific adaptations (Bethke et al., 2006; Martens et al., 2008). This is a possibility since, in addition to the other functions seen in clade- and species-specific

multigene families (*e.g.*, immune evasion, cytoadherence, and mediation of cell invasion), genus-common multigene families tend to have functions related to metabolism (Ponsuwanna et al., 2016), development and maintenance of parasite structures (Taco W.A. Kooij et al., 2005), chaperones (Külzer et al., 2012), and life cycle regulation (Dorin-Semblat et al., 2011).

1.6 *Multigene family evolutionary models*

For decades, the prevalence and importance of gene duplication as a source of genomic innovation and diversification has been widely recognized (Ohno, 1970). Gene duplicates arise via two main mechanisms: unequal crossing over and retroposition, each can result in different patterns of organization and relationship among multigene family members (Walsh and Stephan, 2001; Zhang, 2003). The fixation and loss of multigene family members can be categorized under several evolutionary models depending on the predominant selective forces in action, divergence dynamics, and the potential effects that gene function has on these dynamics (Dittmar and Liberles, 2010; Innan and Kondrashov, 2010). These models have been categorized in the following groups: 1) models that consider the fixation of duplicated genes to be a neutral process (sub-functionalization and duplication–degeneration–complementation), 2) models in which the duplication itself is positively selected (increase dosage and neo-functionalization), and 3) models where duplication occurs in genes with genetic variation in the population (adaptive radiation and permanent heterozygote).

Alternatively, the evolution of multigene families can also be characterized in terms of the changes in the number, composition, and phylogenetic relationship among

paralogs. Specifically, under the Birth-and-Death model paralogs evolve separately and putative gain/loss and loss of function events occur independently. On the other hand, under the Concerted Evolution model, multigene family members evolve as a unit and tend to be highly homogeneous (Eirín-López et al., 2012; Nei et al., 1997; Nei and Rooney, 2005; Szöll\Hosi and Daubin, 2011). Within the *Plasmodium* genus, evolutionary patterns of several multigene families are thought to better reconcile with the Birth-and-Death model (Arisue et al., 2011, Garzón-Ospina et al., 2010; Nishimoto et al., 2008).

1.7 *This study*

Plasmodium parasites are characterized by a complex life cycle, and a capacity to infect a wide range of vertebrate hosts and *Anopheles* mosquitoes. Considering the importance of gene duplication in the development of organismal novelty and adaptability, multigene families represent a prime example to study genus-specific evolutionary patterns across *Plasmodium* species. Important roles in cell invasion, immune evasion, and other essential aspects of parasite's life have been attributed to many species and clade-specific multigene families. Nonetheless, while the study of multigene families found in a reduced number of *Plasmodium* species is primordial for understanding parasite-host interactions, the study of genus common multigene families (GCMFs) could aid in gaining insight into the evolutionary trends within the genus. I studied gain/loss events and the mechanism (recombination events and long term selective forces) shaping multigene family evolution among *Plasmodium* species with variable life history traits, geographic distributions, and host interactions with the

objective of gaining a better understanding of the capacity of GCMFs to shape organism's genomes.

I am particularly interested in evaluating genus-specific evolutionary patterns in the merozoite surface protein 7 multigene family (*msp7*). While many proteins are involved in the delicate processes which allow the interaction between merozoite and the erythrocyte, *msp7* is the only multigene family with members across largely divergent *Plasmodium* species involved is this process. By studying the evolutionary trends this multigene family, is possible to can gain a better understanding of the relationship between the development of functional divergence in critical points of the *Plasmodium* life cycle, and gain/loss events in multigene family evolution.

Finally, I have also focused on a specific *Plasmodium* life stage of potential interest in the development of malaria transmission blocking strategies. The reduced effectiveness of numerous policies intended for the treatment and eradication of malaria, and the complexity of *Plasmodium* parasite's immune evasion strategies, furthers the need to develop new treatment and eradication protocols. In this regard, besides their relevance in understanding genus common evolutionary trends, conserved genes shared across largely divergent *Plasmodium* species are of enormous significance in the development of universal malaria treatment and control strategies, especially in regions where more than one *Plasmodium* species co-occur and where transmission is low. I characterized the diversity and divergence of genes with gametocyte-biased expression in known human malarias and closely related *Plasmodium* species, and also evaluated the putative association between sex-biased expression and immunogenicity.

Overall, this dissertation covers a range of critical aspects in the evolution of the *Plasmodium* genome, highlights the importance of these elements in parasite-hosts interactions, and in the development of strategies of universal clinical interest for malaria.

CHAPTER 2. Evolution of the merozoite surface protein 7 (MSP7) in *Plasmodium vivax* and *P. falciparum*: a comparative approach.

2.1 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). Among the numerous *Plasmodium* species, there are four that typically infect humans. Two of those, *Plasmodium falciparum* and *P. vivax*, account for most of the malaria morbidity and mortality worldwide (WHO, 2015). These two species differ in many epidemiological and biological characteristics including divergent features in their genomes. In particular, they markedly differ in terms of their exonic G+C content, frequency of low complexity regions, and some distinctive multigene families (Battistuzzi et al. 2016; Carlton et al. 2008; 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 malarias, there are two (*msp3* and *msp7*) that are expressed on the surface of the asexual stage as part of a group commonly known as the Merozoite Surface Proteins (MSPs). These MSPs are involved in the invasion of the host red blood cell (Boyle et al. 2014).

Putative orthologs of *msp3* were originally described in many *Plasmodium* species, including *P. falciparum* (*Pfmsp3*) and *P. vivax* (*Pvmsp3*). However, recent studies have found that the *msp3* genes are not homologs among *Plasmodium* species

(Rice et al 2014). In particular, *Pvmsp3* genes and their orthologs from related species parasitizing non-human primates have evolved independently from those identified as the *Pfmsp3* family in *P. falciparum* and its related species (Rice et al. 2014). This leaves *msp7* genes as the only MSP family found across the known primate malarial parasites and one of the few that seems to be shared across all known *Plasmodium* lineages parasitic to mammals with a role in erythrocyte invasion (Boyle et al. 2014).

Indeed, evidence from the *msp7* family in *P. falciparum* (*Pfmsp7*) indicates that some paralogs encode proteins that may play an important role in the invasion of the host erythrocyte. Specifically, *Pfmsp7* paralogs (*e.g.*, PF3D7_1335100) are known to participate in one of many "complexes" with *Pfmsp1* (Kadekoppala and Holder 2010; Lin et al. 2016; Mello et al. 2002) that bind with the erythrocyte and that appear to be fundamental during the erythrocyte invasion (Lin et al. 2016). In addition, the available evidence indicates that *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 the deletion of some of the *Pfmsp7* multigene family members (Kauth et al. 2006), which is likely due to the fact that there are multiple complexes involving *Pfmsp1* (Lin et al. 2016). In particular, the disruption of PF3D7_1335100, the *Pfmsp7* paralog described in the *msp1* complex, resulted in a reduction of only 20% of the parasite's ability to invade erythrocytes. Nonetheless, the disruption of five of the *P. falciparum msp7* paralogs resulted in a null phenotype (Kadekoppala et al. 2008).

*Msp*1 also interacts with some *msp*7 paralogs in other *Plasmodium* species. Experimental evidence from *P. yoelii* showed that at least one *Pymsp7* paralog

(PY17X_1354000) definitively interacts with *Pymsp1*, while the other members of the family are expressed independently. This suggests that, as in *P. falciparum*, *msp7* paralogs are interacting with *msp1* during the invasion of the red blood cell. However, not all *msp7* paralogs might have an essential role (Mello et al. 2002; Mello et al. 2004). Indeed, recent investigations suggest that some of the *Pfmsp7* paralogs may actually play an immunomodulatory role leading to disease severity (Perrin et al. 2015). Furthermore, *msp7* paralogs may affect the parasite tropism as suggested by experiments with *P. berghei* where knock-out experiments showed an apparent increase in the parasite's use of reticulocytes (Tewari et al. 2005).

How the divergence observed among the *msp7* paralogs across *Plasmodium* spp. relates to their functional diversity remains unknown. A first step, however, is to improve our understanding of the evolutionary history and genetic diversity of the paralogous genes belonging to this multigene family. Previous studies have shown that there is a large variation in the number of paralogs and composition of the *msp7* family among *Plasmodium* species (Kadekoppala and Holder et al. 2010). In particular, an expansion of the *msp7* multigene family was suggested in *P. vivax* and its closest relative found in Southeast Asian macaques, *Plasmodium cynomolgi* (Mongui et al. 2006, Tachibana et al. 2012). However, the limited information on other species within that clade did not allow further exploration of this pattern. Here, I characterized *msp7* multigene family members in simian and *Laveranian Plasmodium* species with the intention to determine if the evolutionary mechanisms affecting the proposed expansion of the *msp7* multigene family in *P. vivax*, were different to those of *P. falciparum* and related species. In addition, I evaluate the hypothesis that complex selection patterns are related to ancient events

leading to the introduction of this parasite lineage into Hominines. Furthermore, given the proposed functional redundancy across *P. falciparum msp7* paralogs; I hypothesize that those motifs important for protein binding to the erythrocyte (Garcia et al., 2007) could be conserved among *msp7* paralogs found in *P. falciparum* and *P. reichenowi* (*Prmsp7*).

2.2 Material and methods

2.2.1 *Sequence data*

In this investigation, I will define the size of the *msp7* multigene family in each *Plasmodium* species to be the number of paralogs of that family in that species' genome. Furthermore, 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. I will use the PlasmoDB Gene IDs (nomenclature) assigned to *P. vivax* (Salvador I) (Aurrecoechea et al. 2009) to refer to specific paralogs.

First, I investigated the genetic diversity of each of the *msp7* paralogs within *P. vivax* (*Pvmsp7*) by using all sequences available at PlasmoDB version 26 (Aurrecoechea et al. 2009). The data consist 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 representative 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) publically available (Neafsey et al. 2012). As a comparison, I also analyzed the genetic polymorphism observed in each of the *P. falciparum msp7* (*Pfmsp7*) paralogs. The data available in PlasmoDB was obtained 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, I also included informative *Pfmsp7* sequences available in the NCBI database.

Second, I studied the *msp7* family from different *Plasmodium* species with publicly available genomes found in PlasmoDB and NCBI (Benson et al. 2014) databases: *P. cynomolgi* (*Pcmsp7*, B-strain), *P. inui* (*Pimsp7*, San Antonio strain), *P. knowlesi* (*Pkmsp7*, H strain), *P. coatneyi* (*Pcomsp7*, Hackeri strain), *P. falciparum* (3D7), *P. reichenowi* (*Prmsp7*, Dennis strain), and the rodent malarias *P. yoelii* (*Pymsp7*, YM strain), *P. berghei* (*Pbmsp7*, ANKA strain), and *P. chabaudi chabaudi* (*Pchmsp7*, AS strain). Additionally, I included sequences obtained from 454 reads (Roche, Applied Science, Basel, Switzerland) for a parasite from African primates, *P. gonderi* (*Pgmsp7*), in this study.

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

DNA extraction from blood samples was performed using QIAamp DNA blood mini kit (Qiagen, Hilden, Germany). I created sets of independent degenerated primers

for each *msp7* paralog using the publicly available genomes of *P. vivax*, *P. cynomolgi* and *P. knowlesi* (Table S1). Polymerase chain reactions (PCR) for each *msp7* ortholog was 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 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, I assessed sequence similarity by evaluating the e-values with respect to published *msp7* sequences. The orthology of each of the *P. gonderi msp7* paralogs with those found in *P. vivax* was established using reciprocal BLAST searches. All sequences obtained in this study were deposited in GenBank (KU307279-KU307446).

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

2.2.2 *Phylogenetic Analyses*

I performed interspecies alignments 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 of protein and nucleotide sequences. For each alignment, the most adequate substitution model was selected using the Akaike information criterion (AIC) method incorporated in Jmodeltest (Posada 2008). The most supported nucleotide substitution models were, for the most part, specific variants of the General Time Reversible (GTR) model. The nucleotide frequencies, fraction of invariant sites, and the shape parameter of the gamma distribution of substitution rates across sites were estimated in order to use them in Maximum Likelihood phylogenetic analysis.

I estimated phylogenies of the *msp7* paralogs among related species in order to explore their origin and diversification. However, many *msp7* paralogs do not have orthologs across all *Plasmodium* species (Garzón-Ospina et al. 2010). Thus, in the context of this investigation, I found it to be more informative to estimate phylogenetic relationships within specific groups. I estimated a phylogeny for the *P. vivax* clade with its closely related species using *P. gonderi* as an out-group, and then a separate phylogenetic analysis was carried out with the *msp7* paralogs found in *P. falciparum* and *P. reichenowi* (*Laverania* subgenus). 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, I excluded paralogs PVX_082660 (531-567bp) and PVX_082690 (285-315bp) and their respective orthologs due to their shorter sequence length relative to other members of the *msp7* multigene family (828-1,608bp). Likewise, I excluded the *msp7* pseudogene PF3D7_1334900 (1,143bp) and its ortholog gene 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 $2X10^7$ Markov Chain Monte Carlo (MCMC) steps; sampling performed every 1,000 generations and a burn-in fraction of 50%. 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.

2.2.3 *Polymorphism and Evolutionary Analyses*

I evaluated evidence of recombination and/or gene conversion events among closely related paralogs, as estimated by the ML and BI phylogenetic trees (see Results and Discussion section), by using Recombination Detection Program (RDP3) (Martin et al. 2010) with its default parameters and a cut-off value of 0.05. This software combines numerous 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).

I determined the genetic diversity (π) 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). Duplicated genes that are expressed and functional could be under purifying or positive selection. Patterns consistent with natural selection acting on the observed polymorphism were assessed by estimating the differences in 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. 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 accumulate faster than non-synonymous because they do not affect the parasite fitness and/or purifying selection is expected to act against non-synonymous substitutions ($dS \geq dN$). Conversely, if positive selection is maintaining polymorphism, a higher incidence of non-synonymous substitutions is expected ($dS < dN$). I 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, in order to evaluate if different family members showed variable levels of selection, I 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). This model does not require the *a priori* partition between positively selected foreground branches and negatively selected or neutral background branches, and instead, allows for the independent variation of three selective regimes to the branches in a phylogeny at any given site reducing the risk of false positives or negatives.

2.2.4 *Conservation of Highly Activity Binding Peptides (HABP)*

Five peptides with significantly high binding activity against red blood cells have been identified in one of the *P. falciparum msp7* paralogs (PF3D7_1335100) (García et al. 2007). Three of these high activity binding peptides (HABPs) were identified on the conserved C-terminus region of the *msp7* multigene family, while the other two were

found in the less conserved N-terminus and central regions of the protein. I attempted to locate these HABP in *P. vivax* and related species but there was not clear homology. Finally, I inferred the putative ancestral sequence for all five peptides using MEGA 6.06 and included in the *Pfmsp7* and *Prmsp7* paralogs phylogeny.

2.3 Results

2.3.1 *Variation on the 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). In particular, the *msp7* family is found on the antisense strand in the chromosome 13 of *P. falciparum*, *P. reichenowi*, and the rodent *Plasmodium* species, whereas it is on the sense strand of chromosome 12 of *P. vivax* and the simian non-human primate *Plasmodium* species with publically available genomes (Fig. 1). The *Pvmsp7* syntenic block is delimited by a flanking conserved hypothetical protein at the 5' end (PVX_082715) and its orthologs are found in all species in the Southeast Asian primate clade. Likewise, the 3' end is delimited 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 are flanking 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 is present at 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), we observed that the number of paralogs varied among the three major clades of mammalian parasites considered in this investigation: *Laverania* subgenus (*P. falciparum* - *P. reichenowi*- *P. gaboni*), rodent malarias, and *Plasmodium* species of Asian primates. All *Pvmsp7* paralogs and their orthologs from closely related species share a similar amino acid composition independent of gene length and the similarity tends to be higher among orthologs than among paralogs. Amino acids in higher proportions 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 of *msp7* family members (Fig. S1), a detailed analysis showed that neither PVX_082710 nor PCYB_122730 contained either the conserved MSP7 C-terminus domain found in all other family members nor do they have the N-terminus signal peptide characteristic of the *msp7* multigene family (Kadekoppala and Holder 2010). Furthermore, they did not have sequence similarity with any other *msp7* paralogs when BLAST searches were performed. 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 corresponding 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). Thus, I consider it a member of the MSP7 family.

It is worth noting that in *P. cynomolgi*, there is a paralog (PCYB_122760, see Fig. 1) that is likely the result of a recent duplication event; however, it lacks the typical MSP7 N-terminus regions. I corroborated this finding for the strain Berok. Overall, the number of confirmed *msp7* paralogs shared by *P. vivax* and *P. cynomolgi* is 12 (see Fig. 1). Interestingly, 10 of those paralogs were also found in *P. fieldi* (Fig. 1), a species that shares a recent common ancestor with *P. vivax, P. cynomolgi,* and *P. inui.* (Muehlenbein et al. 2015; Pacheco et al. 2012). Whether all the paralogs found in *P. vivax* and *P. cynomolgi* are shared by *P. fieldi* cannot be ascertained due to the lack of genomic information on this parasite; nevertheless, it is clear that many indeed are (at least 10 paralogs). Likewise, I cannot confirm the actual sizes of the *msp7* family in *P. simiovale* (minimum six paralogs) and *P. hylobati* (minimum four paralogs including the pseudogenization of the PVX_082685 ortholog) due to the absence of publicly available genomes. *P. gonderi,* a basal species to the simian clade that is found in Africa (a parasite of white-eyelid mangabeys and mandrills; Pacheco et al. 2012) shows a larger multigene family size (nine paralogs) in relation to other *Plasmodium* species such as *P. coatneyi, P. inui*, and *P. knowlesi.*

As in the case of *P. vivax* where all paralogs were shared with a non-human primate parasite (*P. cynomolgi* and this could be the case with *P. fieldi*), all eight paralogs in the human parasite *P. falciparum* are shared with a chimpanzee parasite, *P. reichenowi* (Fig. 1). I found evidence indicating that seven out of the eight paralogs were also conserved in *P. gaboni*, a more distant member of the *Laverania* subgenus (Ollomo et al. 2009; Otto et al. 2014; Pacheco et al. 2013), but the sequences were not of sufficiently high quality to be included in my analysis. When compared to the two clades that include

P. falciparum and *P. vivax*, what seems to be a reduction in the *msp7* family size is 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.

I found no evidence indicating that the *msp7* family size (number of paralogs) changes within each species (e.g.: *P. vivax*, *P. cynomolgi*, *P. inui*; see Fig. S2), so in that regard it 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*.

Finally, it is worth noting that I found evidence of ongoing pseudogenization processes on some *msp7* paralogs. In particular, *P. cynomolgi* has a pseudogene that is orthologous to PVX_082690, *P. inui* and *P. hylobati* have pseudogenes that are orthologous to PVX_082685, and *P. coatneyi* has a pseudogene that is orthologous to PVX_082660 (Fig. 1). Furthermore, *P. falciparum* and *P. reichenowi* share a pseudogene (PF3D7_1334900) in addition to their eight paralogs (Fig. 1).

Table 1 shows the nucleotide polymorphism of the different *Pvmsp7* paralogs; all of which exhibited relatively low levels of genetic polymorphism with π below 0.05. This measure, however, does not account for length polymorphisms in low complexity regions. Although most *P. vivax* paralogs seems to have higher average dS than dN indicative of purifying selection, the null hypothesis of dN=dS was rejected only in five *Pvmsp7* paralogs. In particular, PVX_082695 exhibited more non-synonymous than synonymous substitutions, a pattern consistent with positive selection. On the other hand, four paralogs: PVX_082650, 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 those 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.

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

Maximum likelihood (ML) and Bayesian inference (BI) phylogenetic trees were inferred to evaluate the relationships between *msp7* paralogs found in the *P. vivax* clade (Fig. 2). I excluded all pseudogenes and short length genes (PVX_082660 and PVX_082690 and their orthologs) from these analyses. I labeled closely related paralogs that shared comparable amino acid composition (Fig. S1) and presented higher sequence

similarity 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, suggesting that most *msp7* orthologs tend to be more closely related to each other than to paralogs.

I subdivided the *msp7* paralogs in the *P. vivax* clade into three major groups (A-C) based on their phylogenetic relationships. Groups A and C showed the highest posterior probability and bootstrap 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 (Muehlenbein et al. 2015; Pacheco et al. 2012). Group B, the largest in my phylogenetic analysis (Fig. 2), has 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 of these analyses, no major topological changes were observed inside the group B phylogeny and the corresponding *P. gonderi msp7* paralogs formed a basal monophyletic group with low support. This is consistent with the scenario that duplication events may have occurred early on *msp7* at the origin of the clade that includes *P. vivax*. 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 with *P. simiovale.* The PVX_082680 gene, for example, underwent a duplication event that gave rise to PVX_082685 and its orthologs in a monophyletic group that includes *P. vivax* and its most closely related species, *P. cynomolgi*, *P. fieldi*, and *P. simiovale* (Muehlenbein et al. 2015; Pacheco et al. 2012). Importantly, PVX_082685 is undergoing a pseudogenization process in the lineage that includes *P. inui* and *P. hylobati.*

It is worth noting that several paralogs with comparable amino acid composition and higher sequence similarity in group B did not show potentially adjacent location in the chromosome. Although the sampling problem (no genomic data from some species) does not allow me to properly describe all duplication events, this pattern implies that 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 simian 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 clade and *Laverania* subgenus.

Based on the phylogeny described above (Fig. 2), I performed a test for episodic selection. Table 3 shows the results of a phylogenetic-based test of selection used to detect patterns of episodic selection as implemented in HyPhy. Four paralogs show evidence of episodic selection. Interestingly, these are the same paralogs where their dS and dN patterns rejected neutrality (Table 1). All of them showed more synonymous than non-synonymous substitutions consistent with functional constrains.

2.3.3 *Recombination and selection patters among group B msp7 paralogs*

I evaluated putative recombination events 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 in the C-terminus region (Fig. S5). Nonetheless, I also found a large variation in the location of the putative recombinant break points. It is not clear how different paralog combinations were involved in each putative event detected, making it impossible to identify global recombination pattern among the group's paralogs. The amount of phylogenetic information found in recombinant segments in the most generalized single recombination event detected (Fig. S5), which encompassed all group B paralogs, was lower than those found in non-recombinant and highly conserved segments; therefore, it was possible to conclude that recombination should not have a significant effect hampering intended phylogenetic and selection analyses.

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

P. falciparum and *P. reichenowi* share all *msp7* paralogous genes so they originated before these extant species shared 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). I evaluated the sequence conservation of the five HABPs described in PF3D7_1335100 (García et al. 2007) 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 varied between 1 and 4 (Table 4). Epitopes originally located in the C-terminus regions showed higher sequence conservation between *P. falciparum* paralogs. In particular, 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 all the HABP domains in *P. falciparum* and *P. reichenowi* (Fig S6), the *msp7* paralogs without HABP domains were also shared between the *Pfmsp7* and *Prmsp7* families as can be observed in the phylogeny (Fig. 3). I also inferred the evolution of the HABPs sequences including their estimated ancestral sequences. The phylogeny shows that the putative HABPs were independently lost in two duplication events.

2.4 Discussion

There are several multigene families involved in immune evasion and host-parasite interaction making very difficult the development of effective treatment strategies against the *Plasmodium* parasite. These families harbor extraordinary variation to the extent that it is difficult to identify clear orthologs even in closely related species (Neasfy et al. 2012; Rice et al. 2014). In the case of *msp7,* orthologs are identifiable across species but they have still diverged enough to hamper our ability to describe evolutionary processes at the generic level in detail. This is worsened by the fact that the genomic information is biased toward those *Plasmodium* species of biomedical

importance. Thus, the existing data has a sampling problem in terms of taxa (Ness et al. 2011). These issues likely will be solved when more *Plasmodium* genomes become available. Nevertheless, there are some clear patterns that emerged from my investigations.

I observed extensive variation in the numbers of paralog genes for *msp7*, a pattern that has also been observed in the *msp3* (Rice et al. 2014) and SERA (Arisue et al. 2011) multigene families. However, in other multigene families it is common to find variation in the actual number of paralogs within species. This is not the case of *msp7*. Nevertheless, *msp7* exhibits complex patterns of gene gain/loss events that can be observed even among closely related *Plasmodium* species. Out of the *Plasmodium* species included in this study, *P. vivax* and *P. cynomolgi* contained the largest numbers of *msp7* paralogs. This high number of paralogs for *msp7* in the *P. vivax-P. cynomolgi* lineage 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 since the human parasite and *P. cynomolgi have* a comparable number of paralogs that are also orthologs between the two species. Thus, the events leading to the high number of *msp7* paralogs in these two species likely occurred before their most recent common ancestor (2.36–5.27 Mya, Pacheco et al. 2012). Furthermore, some of these paralogs could be as old as the divergence of these two species with *P. fieldi* (at least 10 out of 12) (Muehlenbein et al. 2015; Pacheco et al. 2012), however, I cannot confirm this older origin for all the extant *msp7* paralogs in *P. vivax* due to the absence of publicly available *P. fieldi* genomes. 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 ortholog gene had been identified in any of the *P. vivax* isolates. This suggests that an additional duplication event took place in this species after its split from its common ancestor with *P. vivax.*

Evidence of a larger ancestral number of paralogs in the clade that includes *P. vivax* can be found in *P. gonderi*, a basal species to the simian clade that is found in Africa (a parasite of white-eyelid mangabeys and mandrills) (Pacheco et al. 2012). In particular, 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 an early origin and might have even been present in the common ancestor of Asian primate malarias. However, lack of evidence from additional basal species does not allow a proper test this hypothesis. It is worth noting that 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 my 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 due to convergence. Alternatively, gene conversion/recombination events within *Pgmsp7* could also result in an apparent monophyletic group.

The gaps in the *msp7* sampling of paralogous genes likely lead to an incorrect assessment of the number of duplication/loss events (Ness et al. 2011). However, despite the obvious limitations in my 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 becoming no longer functional (Nei et al. 1997; Nei and Rooney 2005). I found that a reduction of the

multigene family may have taken place 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 less 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). If a larger number of paralogs is ancestral in the group as suggested by the *P. gonderi* data, the number of gene loss events found in *P. inui*, *P. knowlesi* and *P. coatneyi* is noteworthy due to the fact that similar reductions have been observed in the same species in 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; Cormier 2011). 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 events likely took place in a relatively short period of time accelerated by the interplay of selection and drift.

A similar pattern to the one found between *P. vivax* and *P. cynomolgi*, where all of the human parasite *msp7* paralogs predate the origin of the human parasite, can be observed in the case of *P. falciparum.* In particular, the *msp7* paralogs are not only conserved between *P. falciparum* (9) and *P. reichenowi* (9) but also many of them (minimum 7) are found in *P. gaboni* (another parasite from African Apes). The number of paralogs found in the primate parasites contrast with *Plasmodium* species from rodents

where only three paralogs constitute the *msp7* family. It is important to highlight that these three paralogs are the ones that have orthologs in all the species analyzed here (Fig. 1) a fact that may indicate their functional importance.

Gene duplication can generate redundancy; however, having genes with identical functions need not always be advantageous. Under this scenario, purifying selection is expected to relax and genes start to accumulate deleterious mutations that lead to pseudogenization. Although I did not find evidence of pseudogenization taking place within *P. vivax*, I have indications of such a process in *P. cynomolgi* using Sanger sequencing (PVX_082690 ortholog). Furthermore, I observed 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 the *Laverania* subgenus, both *P. falciparum* and *P. reichenowi* also have 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 groups 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).

The number of recombination events detected among closely related paralogs in the present study (Fig. S5) suggests that, as in other multigene families (Bethke et al. 2006), recombination might had been a pivotal force in the expansion of the *msp7* family via unequal crossing over or that segmental gene conversion has occurred among adjacent paralogs; furthermore, the elevated nucleotide diversity observed in recombinant regions indicates that recombination might also have a role in generating genetic diversity as described in other *Plasmodium* multigene families (Nielsen et al. 2003).

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 suggests functional importance. Although the specific function of *Pvmsp7* is a matter that needs to be investigated, it has been reported that the processing proteolytic 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 that they could form complexes with *msp1* gene as described in both rodent malarias and *P. falciparum* (Lin et al. 2016). Consistent with the functional importance of *Pvmsp7*, the limited transcriptomic data (Aurrecoechea et al. 2009) together with the evidence for purifying selection (Table 1), indicate that all *Pvmsp7* are expressed and likely still producing functional proteins. This pattern of strong 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 lines of evidence may look contradictory but they are not. A scenario

consistent with these two observations is that the paralogs diverged, but did not originate, as an adaptation in an ancient parasite lineage that shifted from an Asian non-human primate (Cercopithecidae) to hominins. Then such paralogs that underwent adaptive divergence from their orthologs in the ancestral lineage have been maintained by negative selection at the population level within the extant *P. vivax* populations. Interestingly, there is evidence that two of those genes (PVX_082675 and PVX_082680) are immunogenic and one in particular (PVX_082680) 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), 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 S2). 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. 2013). Previous studies have suggested that those LCRs may have a role in immune evasion (Chenet et al. 2013; Singh et al. 2004). Whether this is the case in *msp7* is a matter that needs to be investigated.

The specific role that *Pvmsp7* paralogs play in invasion remains unknown. We can only speculate that it may have a similar function as in *Pfmsp7.* In that regard, experimental evidence suggests that more than one *Pfmsp7* paralog have the capacity to

interact with proteins of the erythrocyte membrane (either band 3 or ~52 kDa MSP1 protein), as observed in PF3D7_1335100 (García et al. 2007). Consistent with this notion, HABP identified in PF3D7_1335100 originally located in the C-terminus regions 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 in the C-terminus region of *Pf*MSP7 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 inhibited essays (24%) with respect to the 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 forms of functional constraints and it is not necessarily an indication of their involvement in the 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* subgenus that includes *P. falciparum* is a matter that cannot be addressed at this time. However, what is certain is that the number of *msp7* paralogs in *P.*

vivax and *P. falciparum* predates their origin as human parasites. In particular, *P. vivax* has a conserved number of paralogs when compared to *P. cynomolgi*, a pattern that is also found between *P. falciparum* and *P. reichenowi.*

Although there is a paucity of functional information for *Pvmsp7*, I 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 Hominins from an ancestral host species that likely was a catarrhine. This observation, together with the conservation of the number of paralogs and the population data showing purifying selection acting on those paralogs, indicates that *Pvmsp7* is functionally important. In addition, while different paralogs show diverse levels of interaction with the erythrocyte, variable signs of selection across family members show positively maintained divergence, suggesting some level of sub-functionalization within the family. Finally, in the case of *Pfmsp7*, I observed some level of conservation of the 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 of these findings in the clades that include the two major malarial parasites support the evidence emerging from *P. falciparum* that *msp7* likely 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*. This hypothesis should be further assessed by performing additional knockout studies in *P. vivax* and other *Plasmodium* species.

Tables

Table 2-1. Polymorphism in *msp7 P. vivax* and *P. falciparum* paralogs.

Species	Paralog ID	N	π [SE]	dS	dN	dN-dS [SE]	Z test	Neutrality
P. vivax	PVX_082645	106	0.001 [0]	0	0.001	0.001 [0]	0.061 (1.892)	dN = dS
	PVX_082650	17	0.019 [0.002]	0.044	0.013	-0.031 [0.009]	0 (-3.721)	**dN < dS**
	PVX_082655	24	0.028 [0.003]	0.032	0.027	-0.005 [0.007]	0.431 (-0.791)	dN = dS
	PVX_082660	97	0.001 [0.001]	0.001	0.001	0 [0.002]	0.882 (-0.148)	dN = dS
	PVX_082665	37	0.016 [0.002]	0.02	0.015	-0.005 [0.004]	0.219 (-1.236)	dN = dS
	PVX_082670	89	0.001 [0]	0	0.001	0 [0]	0.337 (0.965)	dN = dS
	PVX_082675	31	0.016 [0.001]	0.026	0.013	-0.013 [0.005]	0.004 (-2.951)	**dN < dS**
	PVX_082680	17	0.036 [0.003]	0.073	0.026	-0.047 [0.011]	0 (-4.375)	**dN < dS**
	PVX_082685	53	0.012 [0.001]	0.022	0.009	-0.014 [0.005]	0.007 (-2.767)	**dN < dS**
	PVX_082690	106	0.002 [0.002]	0.001	0.003	0.002 [0.002]	0.393 (0.857)	dN = dS
	PVX_082695	96	0.003 [0.001]	0	0.003	0.003 [0.001]	0.004 (2.939)	**dN > dS**
	PVX_082700	89	0.001 [0]	0.002	0.001	-0.001 [0.001]	0.463 (0.737)	dN = dS
P. falciparum	PF3D7_1335100	154	0.002 [0.001]	0	0.002	0.002 [0.001]	0.014 (2.491)	**dN > dS**
	PF3D7_1335000	184	0 [0]	0	0	0 [0]	0.544 (0.608)	dN = dS
	PF3D7_1334900	198	0.001 [0]	0	0.001	0 [0.001]	0.742 (0.330)	dN = dS
	PF3D7_1334800	180	0 [0]	0	0	0 [0]	0.083 (1.748)	dN = dS
	PF3D7_1334700	175	0 [0]	0	0	0 [0]	0.125 (1.544)	dN = dS
	PF3D7_1334600	194	0 [0]	0.001	0	-0.001 [0.001]	0.433 (-0.787)	dN = dS
	PF3D7_1334500	103	0.001 [0]	0	0.002	0.002 [0.001]	0.003 (2.982)	**dN > dS**
	PF3D7_1334400	184	0.001 [0]	0	0.001	0 [0]	0.396 (0.852)	dN = dS
	PF3D7_1334300	196	0 [0]	0	0	0 [0]	0.615 (0.504)	dN = dS

Table 2-2. Polymorphism in *msp7* multigene family simian clade paralogs.

Species	N	Paralog ID*	π [SD]	dS	dN	DN-dS [SD]	Z test	Neutrality
P. cynomolgi	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**
	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**
	8	PVX_082700	0.039 [0.003]	0.076	0.031	0.045 [0.012]	0 (-3.729)	**dN < dS**
P. inui	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
P. knowlesi	3	PVX_082660	0.009 [0.003]	0.029	0.004	0.026 [0.015]	0.081 (-1.758)	dN = dS
	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).

Table 2-3. Branch and episodic selection in simian 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**
P. vivax	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
P. cynomolgi A	0.5686	0.3629	0.4563	**1.1777**	0.407	0.4941	0.6719	0.6854	0.408	0.5891	0.5792
P. cynomolgi B	0.366	0.4991	0.6414	0.8418	0.4205	0.6893	0.7768	0.5279	0.6437	0.6125	0.4469
P. fieldi	0.689	0.9202	0.449	**2.4149**	0.9976[e]	-	0.9245	0.833	0.7149[e]	0.5441	**1.372**
P. simiovale	0.3321	0.4205	-	-	0.3877	-	-	0.6704[e]	0.6179[e]	0.8414	-
P. inui	0.6568	-	-	-	-	0.8482	**1.2523**	0.6648	-	**1.0493**	0.8212
P. hylobati	0.939[e]	-	-	-	-	-	-	0.7507	-	-	0.8425
P. knowlesi	0.3012	-	-	0.1688	-	-	0.8071	0.6998	-	**1.005**	-
P. coatneyi	0.3694[e]	-	-	-	-	-	**1.6571**	1.0023	-	0.4085	-
P. gonderi	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 bolted omega values.

[e] Indicate branches with signature of episodic selection (branch-site model, Hyphy).

Table 2-4. Detection of HABPs among *P. falciparum* MSP7 paralogs.

Location	Region	HABPs	Sequences	Paralog*						
				PF3D7_1335100φ	PF3D7_1335000	PF3D7_1334800	PF3D7_1334700	PF3D7_1334500	PF3D7_1334400	PF3D7_1334300
20 kDa fragment's N- and C-terminal extremes	N-terminal	26101	IKNKKLEKLKNIVSGDFVGNY	x	x	x			x	
	Central	26107	NLGLFGKNVLSKVKAQSETDY	x						
		26114	EKDKEYHEQFKNYIYGVYSYA	x	x	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

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

**P. falciparum* PlasmoDB nomenclature (Mello et al. 2002).

ᵴ *P. falciparum* HABPs (García et al. 2007).

φ HABP epitopes have been described in PF3D7_1335100 (Garcia et al., 2007).

Figures

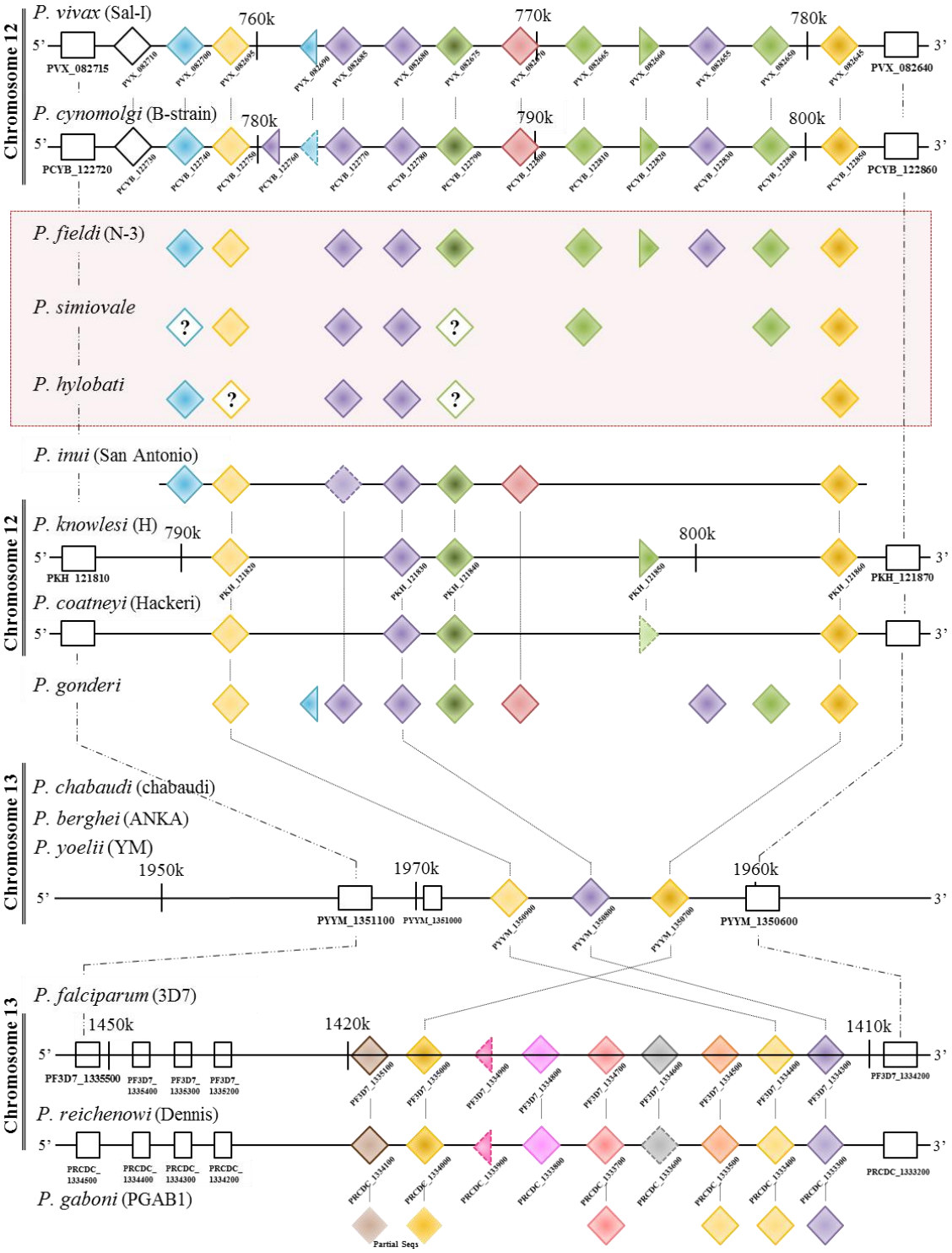

Chromosome 12
P. vivax (Sal-I)
760k
770k
780k
PVX_082715
PVX_082640
P. cynomolgi (B-strain)
780k
790k
800k
PCYB_122720
PCYB_122860
P. fieldi (N-3)
P. simiovale
P. hylobati
P. inui (San Antonio)
Chromosome 12
P. knowlesi (H)
790k
800k
PKH_121810
PKH_121870
P. coatneyi (Hackeri)
P. gonderi
Chromosome 13
P. chabaudi (chabaudi)
P. berghei (ANKA)
P. yoelii (YM)
1950k
1970k
1960k
PYYM_1351100
PYYM_1351000
PYYM_1350600
Chromosome 13
P. falciparum (3D7)
1450k
1420k
1410k
PF3D7_1335500
PF3D7_1334200
P. reichenowi (Dennis)
PRCDC_1334500
PRCDC_1333200
P. gaboni (PGAB1)
Partial Seqs

Figure 2-1. *Msp7* multigene family organization. Data on the chromosome and the 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 clade and *Laverania* subgenus (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. Paralogs with the same coloration are more closely related phylogenetically and share overall similar sequence patterns. 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.

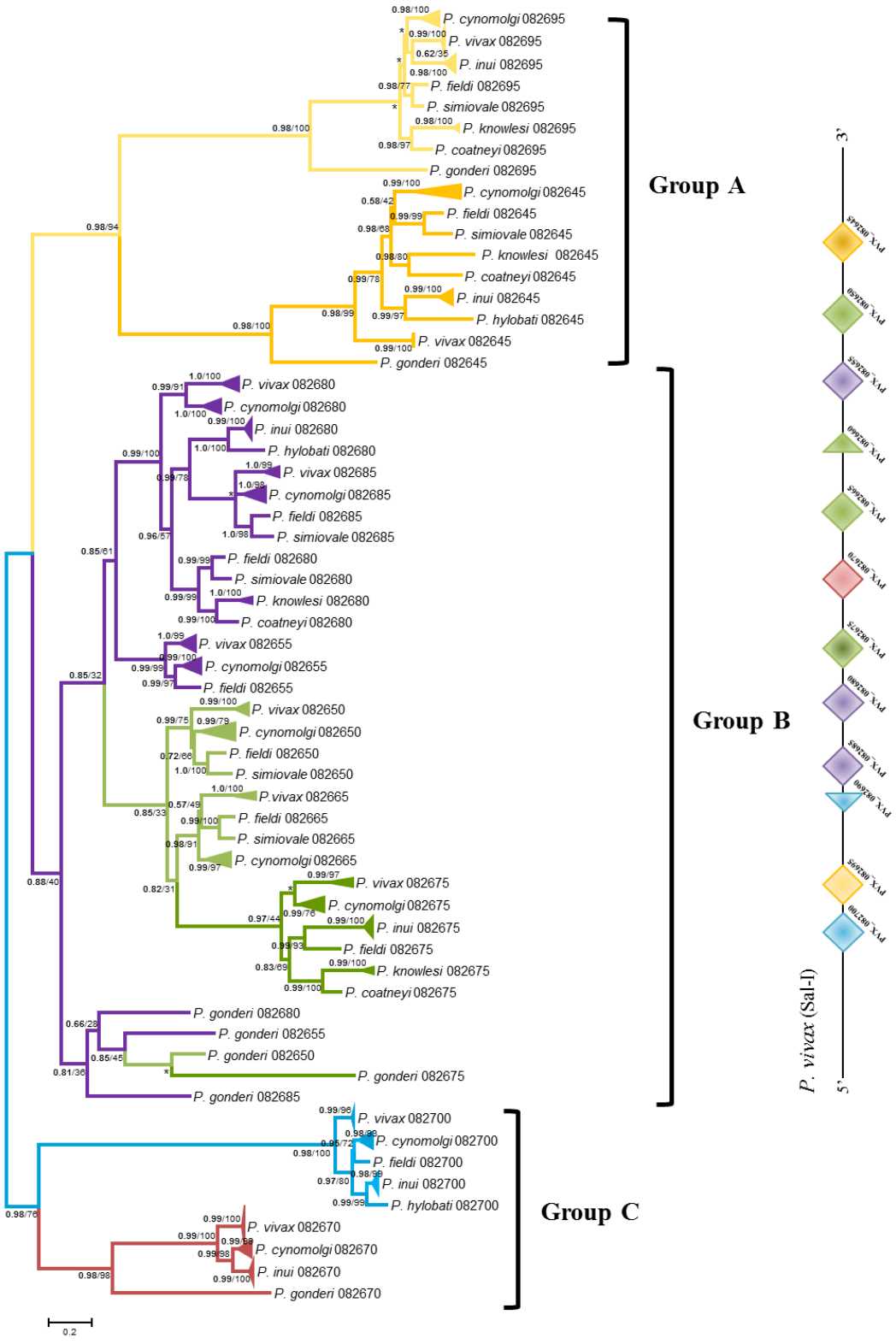

P. cynomolgi 082695
P. vivax 082695
P. inui 082695
P. fieldi 082695
P. simiovale 082695
P. knowlesi 082695
P. coatneyi 082695
P. gonderi 082695
P. cynomolgi 082645
P. fieldi 082645
P. simiovale 082645
P. knowlesi 082645
P. coatneyi 082645
P. inui 082645
P. hylobati 082645
P. vivax 082645
P. gonderi 082645
Group A
P. vivax 082680
P. cynomolgi 082680
P. inui 082680
P. hylobati 082680
P. vivax 082685
P. cynomolgi 082685
P. fieldi 082685
P. simiovale 082685
P. fieldi 082680
P. simiovale 082680
P. knowlesi 082680
P. coatneyi 082680
P. vivax 082655
P. cynomolgi 082655
P. fieldi 082655
P. vivax 082650
P. cynomolgi 082650
P. fieldi 082650
P. simiovale 082650
P.vivax 082665
P. fieldi 082665
P. simiovale 082665
P. cynomolgi 082665
P. vivax 082675
P. cynomolgi 082675
P. inui 082675
P. fieldi 082675
P. knowlesi 082675
P. coatneyi 082675
P. gonderi 082680
P. gonderi 082655
P. gonderi 082650
P. gonderi 082675
P. gonderi 082685
Group B
P. vivax 082700
P. cynomolgi 082700
P. fieldi 082700
P. inui 082700
P. hylobati 082700
P. vivax 082670
P. cynomolgi 082670
P. inui 082670
P. gonderi 082670
Group C
0.2
3'
5'
P. vivax (Sal-I)

Figure2-2. Bayesian inference (BI) and Maximum Likelihood (ML) multigene family phylogenetic tree for simian *msp7* paralogs. 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). Branch with the same coloration are not only closely related but share overall similar sequence patterns. The tree was constructed from 233 *msp7* paralogs excluding those 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; α=1.6470) was used. Paralogs were divided into three mayor groups (Group A, Group B and Group C).

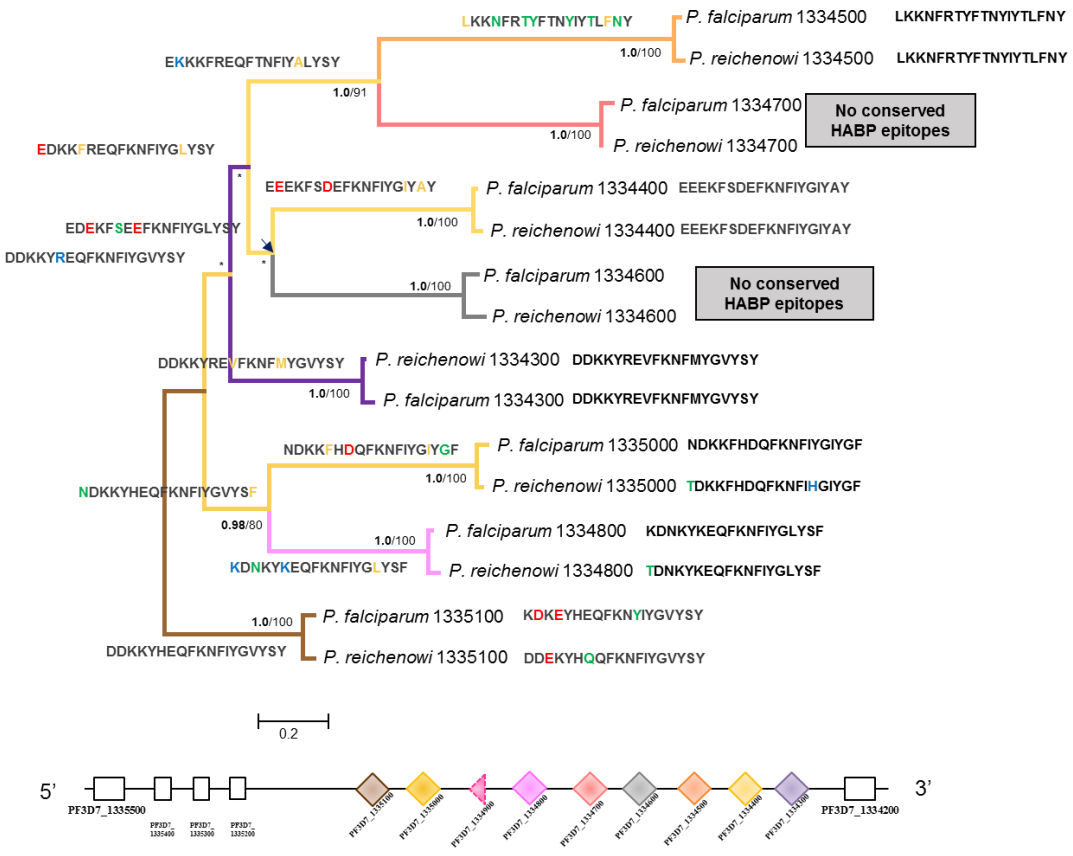


Figure 2-3. Bayesian inference (BI) and Maximum Likelihood (ML) phylogenetic tree of *msp7* multigene family paralogs found in *Plasmodium* species from the *Laverania* subgenus. BI and ML trees showed similar topologies, so only BI topology is shown with asterisks (*) indicating conflicting branching 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 (α= 2.9310) was used. Putative ancestral sequences and extant

sequences of epitope HABP26114 were determined for each paralog. Amino acid changes are colored with changes between same type amino acid sharing the same color.

CHAPTER 3. Evolutionary rates in gametocyte expressed genes and transmission blocking vaccine candidates in *Plasmodium* spp.

3.1 Introduction

By the end of 2015, an estimated 214 million malaria cases and approximately 438,000 malaria related deaths occurred worldwide with cases developing in the African (90%), South-East Asian (7%) and Eastern Mediterranean (2%) regions (WHO, 2015). The implementation of control and eradication strategies has resulted in an 18% decline of estimated malaria cases and 48% decline in worldwide mortality rates since the year 2000 (WHO, 2015; Zhou et al., 2014). While these policies have had a positive effect in reducing global malaria prevalence, *Plasmodium* resistance to antimarial drugs has become well-established in numerous malaria-endemic countries and is spreading into others (Fairhurst, 2015; Hastings et al., 2015; Mvumbi et al., 2015; WHO, 2015). Additional factors such as *Anopheles* resistance to insecticides (Riveron et al., 2015), complexity of transmission patterns, and the intricate array of antigens expressed during the parasite's life cycle (Kirkman and Deitsch, 2012), have hindered parasite eradication and negatively impact efforts to inhibit its expansion (Keitany et al., 2014). In order to address these issues, alternative malaria control and treatment strategies are being approached and the development of malaria vaccines has gained renewed interest.

Malaria vaccines can be classified into three groups depending on the parasite stage to which they are targeted: 1) pre-erythrocytic vaccines are aimed at the sporozoitic stage (*e.g.*, RTS,S currently in Phase 3 trial), they target the parasite when infection is

still asymptomatic and prevent blood stage infection (Agnandji et al., 2011); 2) blood stage vaccines are aimed at asexual blood stages, and negatively affect malaria transmission by preventing and controlling the progress of the disease in the vertebrate host (Carter, 2001); and 3) transmission blocking vaccines (TBVs), utilize antibodies produced by the immune response against *Plasmodium* gametocyte antigens (pre-zygotic) and antigens found in the mosquito (post-zygotic) to inhibit parasite development after ingestion via a blood meal (Miura et al., 2013; Nikolaeva et al., 2015). Pre- zygotic vaccine candidates are known to induce higher and more sustained antibody responses, which result in more efficient transmission blocking activity. On the other hand, post-zygotic candidates are not exposed to immune pressure in the vertebrate host, are subject to a lower antibody response, and tend to have lower polymorphism (Nikolaeva et al., 2015). Presently, there are approximately 24 suggested proteins described as potential transmission blocking antigens, the majority of which require further characterization before being considered as feasible TBV candidates (Sinden et al., 2012).

It has been proposed that high levels of nucleotide diversity may hamper the development of an effective malaria vaccine (Girard et al., 2007). Among the 6-cysteine family TBVs, limited levels of polymorphism have been found in *Pfs*16 and *Pfs*230, when compared to erythrocyte-stage antigens (Niederwieser et al., 2001), as well as in *Pfs*25, *Pfs*48/45 and their respective *Plasmodium vivax* orthologs (Da et al., 2013; Tachibana et al., 2015). Low levels of polymorphism have also been observed in *Pv*WARP, among samples from a temperature gradient (Gholizadeh et al., 2009; Miura et al., 2013).

In addition to low polymorphism, universal TBVs also require potential candidates that are conserved among different *Plasmodium* species, and that are capable of eliciting an immune response in more than one *Plasmodium* species (Alonso et al., 2011; Vaccines and others, 2011). This type of approach would be highly beneficial in regions where mixed infections by different parasite strains and/or species occur. The development of a single TBV effective against major human malarias is one of the proposed objectives of the PATH Malaria Vaccine Initiative and other research groups (Schwartz et al., 2012). Immunogenicity and effectiveness against both *P. falciparum* and *P. vivax* is of particular importance in locations where these species co-occur and where transmission is low; hence, evaluating evolutionary trends in these and closely related *Plasmodium* species can be vital in the development of vaccines that interrupt malaria transmission (Vaccines and others, 2011). Most of the preclinical development of sexual stage vaccines has been performed on *Plasmodium falciparum* with only two of them having currently reached Phase 1 trails (*Pfs*25-EPA/AS01 and *Pfs*230-EPA/AS01) (Schwartz et al., 2012). There are comparatively few studies conducted using other human malarias (Ouattara and Laurens, 2015), particularly *P. vivax*. So far, transmission-blocking activity against *P. falciparum* and *P. vivax* antibodies has been described for a single and highly conserved protective epitope of *An*APN1, a midgut molecule critical for ookinete invasion (Armistead et al., 2014). However, this finding could lead to the development of a future universal TBVs.

Another issue relevant to the development of TBV candidates is sex-dependent variation in responses to treatment and to the host's immune system. It has been suggested that the *Plasmodium* parasite's investment in the production of transmissible

stages (gametocytes) can be associated with the presence of numerous stressors such as antimalarial drugs, vaccines or changes of the in-host environment (Josling and Llinás, 2015). Under these circumstances, investment in asexual or sexual stages, or even among male or female gametocytes, appears to be adapted to maximize the parasite's fitness and transmission success (Carter et al., 2014). Specifically, investment in male or female gametocytes has been associated with sex-specific immune responses mounted during the course of infection (Bousema and Drakeley, 2011). Furthermore, there is evidence suggesting that male gametocytes are more vulnerable targets to drug treatment than female gametocytes (Delves et al., 2013). This collection of factors suggests that sex-specific responses to numerous stressors may result in divergent evolution of *Plasmodium* gametocytes, which in turn, could affect the effectiveness of pre- zygotic TBV candidates.

Previous studies have also suggested that differences in the number of *P. falciparum* immune epitopes in loci with male- or female-biased expression may be associated with different responses to the host immune system (Khan et al., 2012). In certain vaccine candidates, highly polymorphic domains, thought to be driven by host immune recognition, have been found within epitope regions of natural parasite populations. Also, although known TBV candidates have been described as showing low levels of polymorphism (Da et al., 2013; Tachibana et al., 2015), some variation in gamete-specific epitopes has been described in endemic *Plasmodium* strains (Stone et al., 2016). For example, significant geographical differentiation related to the presence of continent- and region-specific mutations has been found in *Pfs*48/45, even though this gene shows high sequence conservation worldwide (Feng et al., 2015). Thus, the putative

location of immune epitope regions could affect the distribution of positively selected sites in TBV candidate sequences. Furthermore, the association between the presence of immune epitopes and gametocyte sex could be of relevance in the selection and development of new TBV candidates.

In the present study, I characterized polymorphism and divergence of genes with gametocyte-biased expression in known human malarias and closely related non-primate *Plasmodium* species. I also evaluated the association between sex-biased expression, epitope distribution, and evolutionary signals of these genes.

3.2 Materials and Methods

3.2.1 *Sequence data.*

In a previous study, the male and female gametocyte proteomes of *P. berghei* were characterized using the partitioned *P. falciparum* gametocyte proteome as a baseline (Khan et al., 2012; Tao et al., 2014). I used the reanalyzed *P. berghei* proteome to select a sample of genes with gametocyte biased expression. Genes were selected based on their expression profile, orthology, and absence of paralogs. I also used the *P. falciparum* transcriptome of sexual and asexual life stages (López-Barragán et al., 2011), publicly available in PlasmoDB version 27.0 (Autorrecochea, 2009), to identify genes with pronounced gametocyte expression. Genes prominently expressed in other stages of the *Plasmodium* life cycle, such as members of the Merozoite Surface Protein family (*msp*1 and *msp*9) or circumsporozoite-related antigens, were excluded from the present analysis. Multigene family members were also excluded regardless of their expression profile (*e.g.*, SERA) with the exception of known TBV candidates (P28). Species-specific or clade-

specific genes were also excluded. These selection criteria lead to the inclusion of 381 orthologs from the original *P. berghei* gametocyte proteome re-analysis (784 male- and female-specific proteins). Expression for each gene was identified as male, female, or male-female common following previous categorization (Tao et al., 2014). In addition, I categorized genes as putative membrane and non-membrane located following the criteria used in the *P. berghei* female and male proteome (Khan et al., 2012).

I identified the 381 orthologs using reciprocal BLAST (Altschul et al., 1997) searches against 8 of the *Plasmodium* species with publicly available genomes. The PlasmoDB version 27.0 (Autorrecochea, 2009) database was used for a sequence search of all primate malarias included in this study: *P. vivax* (Salvador I strain), *P. cynomolgi* (B strain), *P. knowlesi* (H strain), *P. falciparum* (3D7 strain) and *P. reichenowi* (CDC strain); and the rodent malarias: *P. yoelii* (YM strain), *P. berghei* (ANKA strain) and *P. chabaudi chabaudi* (AS strain). Interspecies alignments were performed independently for each gene using the MUSCLE (Edgar, 2004) algorithm incorporated into SeaView version 4 (Gouy et al., 2010), followed by manual editing of protein and nucleotide sequences. I used an unrooted topology constructed using previously published *Plasmodium* species phylogenies (Pacheco et al., 2012) to perform each independent analysis.

3.2.2 *Synonymous and non-synonymous substitutions rates and MLEs of synonymous and non-synonymous branch lengths.*

I obtained estimates of synonymous and non-synonymous evolutionary rates over a specific time frame in order to identify variation of selective pressures among different

Plasmodium lineages. The absolute rates of synonymous and non-synonymous substitutions were estimated for each multisequence alignment using CodonRates v1.0 (Seo, 2004). Time constraints were fixed for all of the analyses using previously published time estimates for the split of *Plasmodium* species (Pacheco et al., 2012). The node estimating the split of rodent malarias was fixed to 12-14 Mya, and the *P. reichenowi- P. falciparum* split was fixed to 5-6 Mya. Significant differences in the overall absolute rate of synonymous and non-synonymous substitutions for each analyzed gene were evaluated with a two way ANOVA using the sex (Tao et al., 2014) and location categories (Khan et al., 2012) as factors. In addition, the maximum likelihood estimates (MLEs) of synonymous and non-synonymous branch lengths were also calculated. Significant differences among the branch MLEs were assessed with a three way ANOVA which included *Plasmodium* species, sex, and location as factors. Potential variation among sex (male, female, and male-female common) and location (membrane and non-membrane) categories was also assessed using box and whiskers plots. All statistical analyses were performed using R version 3.2.2.

3.2.3 *Evolutionary analyses.*

When developing effective elimination and control strategies, it is of interest to consider variations in the nature and strength of selective pressures acting on different *Plasmodium* species, particularly if the intention is for those strategies to be effective in regions where more than one *Plasmodium* species co-occurs. In the present study, I evaluated putative variation amidst the selective signals in different branches of the *Plasmodium* topology using Hyphy's random effects Branch-Site REL model (BSREL)

(Kosakovsky Pond et al., 2011; Pond and Muse, 2005). This analysis was executed independently for the 381 multisequence alignments and using the same topology. The number of classes in each branch was set equal to three and the alpha (α) and omega (ω) values were allowed to vary among branches and branch-site combinations. Internal and terminal branches with a corrected p-value <0.05 were considered to be under significant episodic selection. I recorded the strength of episodic selection and proportion of sites under this selective regime.

I further analyzed genes that showed significant signs of episodic selection in internal or terminal branches for signals of gene-wide positive selection using the branch-site unrestricted statistical test for episodic diversification (BUSTED) (Murrell et al., 2015). The test was performed using lineages in which signs of significant episodic selection were previously detected by BSREL as the *a priori* subset of foreground branches. This allowed me to test if branches with potential signals of episodic selection also showed signs of diversification, while also permitting flexible selection elsewhere in the phylogenetic tree.

3.2.4 *Codon selection analyses.*

3.2.4.1 *Genes with gametocyte biased expression.*

Previous studies have suggested that the trends of accelerated evolution observed in genes expressed in male gametocytes, may be associated with the presence of a significantly higher percentage of immune epitopes in comparison to genes with female-biased expression (Khan et al., 2012). In order to further test this hypothesis, I estimated the site-specific variation in the selective pressures along a multisequence alignment

using Hyphy's random effects likelihood analysis (REL) (Pond and Muse, 2005). The analysis was performed using publicly available *P. falciparum* (3D7, IT, 303.1, 7G8, BM-0008, CS2, Dd2, GB4, H209, HB3, M113-A, N011-A, RV3600, Santa Lucia, Senegal, UGK-396, O314, P164-C, T9-94 and TRIPS) and *P. reichenowi* (CDC) isolates. As previously suggested, I focused the analysis of *P. falciparum* immune epitopes due to their more detailed characterization (Khan et al., 2012). I obtained information regarding the location and length of immune epitopes reported in *P. falciparum* sequences from the Immune Epitope Database (IEDB) section found in PlasmoDB version 27.0 (Autorrecochea, 2009). The number of available sequences varied among loci due to differences in sequence quality among *P. falciparum* isolates, resulting in a minimum of 13 and a maximum of 20 *P. falciparum* sequences included in each REL test.

I determined the proportion of genes showing at least one significant positively selected site for the 381 genes. In addition, I determined the proportion of positively selected sites in relation to alignment length for all genes which showed at least one site under significant positive selection. In the case of genes with reported immune epitopes, I also estimated the proportion of these sites inside and outside putative epitope regions. In addition, I measured these values for each of the previously established sex (male, female, and male-female common) and location (membrane and non-membrane) categories, as well as in combinations of these categories (*e.g.*, male non-membrane genes).

3.2.4.2 *TBV candidates.*

I performed additional REL analyses for both *P. falciparum* and *P. vivax* for TBV candidates taking advantage of the large number of available worldwide isolates. *P. vivax* isolates were obtained from the database for clinical isolates representing diverse geographic regions, as part of the Hybrid Selection Initiative performed by the Broad Institute (Autorrecochea, 2009). In addition, the five sequenced *P. vivax* reference strains (Salvador I, North Korean, India VII, Mauritania I and Brazil I) publically available via the Malaria Research and Reference Reagent Resource Center were included. Alternatively, *P. falciparum* isolates were obtained 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, *P. falciparum* sequences available in the NCBI database (Benson et al., 2014) were included.

I assessed genetic diversity (π) and patterns consistent with natural selection acting on the observed polymorphism by calculating the differences of the average number of synonymous (dS) and non-synonymous substitutions (dN) between isolates using the Nei-Gojobori distance method (Nei and Gojobori, 1986), with the Jukes and Cantor correction implemented in MEGA 6.06 (Tamura et al., 2013). The difference between dS and dN and its standard error was estimated by using bootstrap with 500 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 accumulate faster than non-synonymous because they do not affect the parasite fitness and/or purifying selection is expected to act against non-synonymous substitutions ($dS \geq dN$). Conversely, if positive selection is maintaining polymorphism, a higher incidence of non-synonymous substitutions is expected ($dS < dN$). I assumed as a null hypothesis that the observed polymorphism was not under selection ($dS = dN$).

3.2 Results

3.3.1 *Synonymous and non-synonymous substitution rates and MLEs of synonymous and non-synonymous branch lengths.*

The differences in the overall absolute rates of synonymous substitutions among sex categories (male, female, and male-female common) were statistically significant (F= 3.20, p-value= 0.04). Also, a significant variation of the non-synonymous branch lengths MLEs was observed among location (F= 22.42, p-value= 2.342e-06), *Plasmodium* species (F= 104.0846, p-value= 2.2e-16) and the interaction of sex and location (F= 8.6625, p-value= 0.0001798).

Overall, synonymous branch length MLEs showed larger values than non-synonymous MLEs (Fig. 1). This trend was observed independently of the sex (male, female, and male-female common; Fig. 2 and Fig. 3) and location (membrane and non-membrane; Fig. 4 and Fig. 5) categories. Compared to other *Plasmodium* clades, species of the *Laveranian* subgenus had lower mean synonymous and non-synonymous branch length MLEs (Fig. 2-5). Notwithstanding, a large standard error and mean were observed in the *Laveranian* synonymous MLE branch length of the male-female common and non-

membrane categories (Fig. 2 and Fig. 4). Larger synonymous and non-synonymous branch length MLEs were observed in *P. vivax* and closely related species. Similar to the aforementioned case, the mean value and standard error of the synonymous MLEs of genes with male- and female-biased expression were noticeably larger in *P. knowlesi* (Fig. 2). Additionally, the means of synonymous and non-synonymous branch length MLEs were comparable in membrane genes; however, synonymous branch length MLEs were largely variable (Fig. 4 and Fig. 5). Genes with putative non-membrane expression showed slightly larger means of synonymous branch MLEs compared to non-synonymous branch MLEs (Fig. 4 and 5).

3.3.2 *Evolutionary analyses.*

Purifying selection was the dominant force for the majority of genes regardless of sex and location categories. Nevertheless, in 14 genes, I detected strong signals of diversifying selection in a small proportion of sites along certain branches of the phylogeny (Table S1). The most common annotated functions found were related to metal and nucleic acid binding activity (PVX_001080 and PVX_099105), as well as peptidase activity (PVX_082500 and PVX_098665). With the exception of gene PVX_098665, all sequences which showed signs of episodic selection in the BSREL analysis also presented evidence of gene-wide positive selection in the same branches of the *Plasmodium* topology (Table S2).

Both the *P. cynomolgi* and *P. berghei* terminal branches showed significant signs of episodic selection in the majority of the 14 genes, with *P. berghei* presenting the strongest selective signal. Five of these genes had characterized low confidence immune

epitopes in *P. falciparum*; nonetheless, no association between the strength of episodic selection and the presence of immune epitopes was found. Furthermore, no significant signals of episodic selection were found in *P. vivax* or *P. falciparum*, even though they were observed in closely related species.

3.3.3 *Codon selection analyses.*

3.3.3.1 *Genes with gametocyte biased expression.*

The number of genes with reported *P. falciparum* immune epitopes varied among the established sex and location categories. A larger proportion of genes with male-biased expression and genes with non-membrane location had reported immune epitopes. Furthermore, a significant proportion of male-female common expressed genes also harbored reported immune epitopes (Table S3). The proportion of genes with at least one site evolving under significant positive selection was larger in genes with reported epitopes from the male, male-female common and non-membrane categories. Alternatively in the female category, the proportion of genes with positively selected sites was lower in genes with reported epitopes. No variation in the proportion of genes with at least one site evolving under significant positive selection was observed between membrane-located genes regardless of the existence of epitopes (Table S3).

With the exception of the male-female common sex and the non-membrane location categories, the proportion of positively selected sites was marginally larger in genes with no reported epitopes than in those with epitopes. When the location of positively selected sites in genes with reported immune epitopes in relation to the locations of the epitopes was assessed, the proportion of positively selected sites was

slightly biased to the regions inside the reported immune epitopes in the non-membrane, male, and female categories. Thus, the presence of epitopes was associated with accelerated evolution in certain sex and location categories; however, there is little evidence of enrichment of positively selected sites within epitopes. The majority of the epitopes found in this study are of low confidence, and hence, sequence variation among isolates is to be expected. In order to address this, I evaluated the distribution of positively selected sites in genes where experimental epitopes have been characterized (high confidence); nonetheless, I found no association between the location of the epitope and the distribution of positively selected sites in the sequence.

3.3.3.2 *TBV candidates.*

Polymorphism levels were markedly low in all the evaluated TBV candidates with only *Pvs*230, *Pvs*ApiAP2 and *Pfs*47showing significant deviation from neutrality. A comparable proportion of sites under significant positive selection were found in both *P. falciparum* and *P. vivax* orthologs for some members of the 6-cyteine protein family: P230, P230p, P47 and P48/45, as well as ApiAP2. Evidence of positive selection was also detected in a single species in other TBV candidates (*Pfs*25 and *Pvs*28), with the proportion of sites being particularly large in *Pfs*25 (Table 1).

3.4 Discussion

Only *Plasmodium* gametocytes can infect the *Anopheles* mosquito vector and mediate the onward transmission of the disease. In *P. falciparum*, approximately 33% of the proteome is composed of gametocyte unique proteins (Florens et al., 2002; Lasonder et al., 2002). Furthermore, approximately 250 to 300 genes have shown specific mRNA

up-regulation in transcriptome analyses (Silvestrini et al., 2005; Young et al., 2005). Thus, it is possible that few critical expressed or up regulated genes during the gametocyte stages could be used to influence parasite transmission. However, although the gametocyte stage is common to all *Plasmodium* species, there are species-specific differences in gametocyte development. To begin with, the length of the sexual stage varies among *Plasmodium* species, with the longest cycle observed in *P. falciparum* (9-11 days) and the shortest in *P. vivax* and *P. berghei* (approximately 2 days). Also, *P. vivax* shows earlier onset of gametogenesis and a larger number of gametocytes in blood than *P. falciparum* (McKenzie et al., 2006). This type of inter-specific variation hinders the development of transmission-blocking strategies that can be effectively used in more than one species, a factor of particular relevance in areas where several *Plasmodium* species co-occur.

The effectiveness of transmission-blocking strategies can also be affected by genetic variation among *Plasmodium* parasites. *Plasmodium* species have markedly different evolutionary histories (Martinsen et al., 2008), and distinct evolutionary trends (Nikbakht et al., 2014). In some cases, these differences have been observed only in specific genes of human interest (Nikolaeva et al., 2015) or in the whole genome. Hence, the study of species-specific patterns is highly significant when making inferences about the long-term effects of transmission blocking strategies.

When comparing the rodent clade and *Laveranian* subgenus, previous studies have found no significant variation in non-synonymous substitutions rates in genes expressed in a single stage or those expressed in several life stages. Nonetheless, lower

selective constraints have been reported on gametocyte-expressed genes within the *Laverania* subgenus (Prugnolle et al., 2008). Variation of the synonymous and non-synonymous substitution rates in species from the simian and rodent clade, and the *Laveranian* subgenus suggest that genes with gametocyte-biased expression tend to evolve in a clade dependent manner (Fig. 1). While the mean synonymous substitutions MLEs are similar between species of the simian clade and *Laveranian* subgenus, the mean non-synonymous substitution MLEs is lower in the *Laveranian* subgenus (Fig. 1). This indicates that species of the simian clade could be evolving at an accelerated rate with respect to those of the *Laverania* subgenus. A possible explanation for this is that differences in the substitution rate could reflect the distinctive evolutionary history of *Plasmodium* species from both groups. The more recent divergence of *Plasmodium* species in the simian clade compared with those in the *Laveranian* subgenus (Pacheco et al., 2011), and their association with different vertebrate hosts (Mu, 2005; Prugnolle et al., 2013) may affect the strength of selective constrains. This could result in larger non-synonymous substitution rates in species of the simian clade as observed here.

Alternatively, the larger number of gametocytes in blood and an earlier onset of gametogenesis in *P. vivax* (McKenzie et al., 2006) is likely to result in a larger proportion of gametocytes exposed to the host immune system, which could also generate a pattern such as the one described here. Overall, these results suggest that, in the long term, immune selective pressures and the larger diversity in the *P. vivax* and *P. cynomolgi* genomes when compared to *P. falciparum* might be influencing a more rapid evolution of gametocyte expressed genes in certain species. Therefore, potential transmission blocking strategies could be less effective in *P. vivax* and closely related species. This is clinically

important because *P. vivax* is the most widespread human malaria species, and also because sporadic human infections caused by species closely related to *P. vivax* have been reported: *P. knowlesi* (White, 2008), and *P. cynomolgi* (Ta et al., 2014).

On the other hand, the large variation in the synonymous substitution rates estimated in *P. knowlesi*, *P. falciparum* and *P. reichenowi* indicates that certain genes with gametocyte-biased expression evolve more rapidly within these *Plasmodium* species (Fig. 1). The absence of a similar pattern in the non-synonymous substitution MLEs suggests that this variation might be either the result of relaxation of selective pressures or higher substitution rates in certain genes (Fig. 1). Overall, these results show that selective forces act differentially among *Plasmodium* species.

Parasite exposures to external stressors (*e.g.*, antimalarial drug treatments) or biological stressors (*e.g.*, high parasitemia) have been associated with increased commitment to gametogenesis *in vitro*; however, this association is less pronounced when gametogenesis is not completely regulated by environmental factors (Josling and Llinás, 2015) and instead is partially associated with naturally acquired immunity to asexual parasite stages (Bousema and Drakeley, 2011). It has been suggested that under stressful conditions, *Plasmodium* parasites may adjust gametocyte sex ratio in order to maximize reproduction by favoring a less female biased sex ratio when male parasites are too numerous (Reece et al., 2008; West et al., 2002), and that different gametocyte sexes have distinct responses to transmission blocking therapies (Delves et al., 2013). In addition, higher gametocyte density, particularly that of male gametocytes, positively affects transmission success, even when the majority of natural infections are female-

biased (Mitri et al., 2009). Differences between gametocyte sexes are not limited to life history but can also be traced to protein expression. particularly, a proportion of proteins expressed exclusively in either male or female gametocytes in *P. falciparum* and *P. berghei* (Khan et al., 2005; Tao et al., 2014). My results show little variation between the synonymous and non-synonymous substitution rates in genes with male, female, and male-female common expression among different *Plasmodium* species (Fig. 2 and Fig. 3) suggesting that long term evolution is not markedly different between sexes.

The limited variation observed in synonymous substitution rates within each sex category, with the exception of *P. knowlesi* (male and female categories; Fig. 2) and *P. falciparum* and *P. reichenowi* (male-female common category; Fig. 2), suggests that substitution rates are somehow constant within the specified time frame. The greater variation of synonymous substitution rates previously mentioned in the *Laveranian* subgenus, and in *P. knowlesi,* are restricted to specific sex categories. This could imply that accelerated evolution might be sex-dependent within certain species, particularly those with lengthy gametocyte stages such as *P. falciparum*, or it could be a result of accelerated evolution of specific genes. Alternatively, the common variation in non-synonymous substitution rates observed in each sex category for all *Plasmodium* species (Fig. 3) suggests that, while certain lineages might show signs of accelerated evolution, there is generally little variation in the selective pressures affecting genes within each lineage.

Non-membrane located genes and male-female common expressed genes from the *Laveranian* subgenus show a large variation of synonymous substitution rates,

indicating that both groups contain genes under accelerated evolution. Likewise, the larger variation of non-synonymous substitution rates observed in membrane-located genes compared with non-membrane genes (Fig. 5) suggest that certain membrane-located genes are exposed to higher selective pressures than others. This trend could be explained by the notion that exposure to the host's immune system results in accelerated evolution, which is something more likely to be observed in membrane than non-membrane located genes (Khan et al., 2012). Previous studies have found that, while non-membrane genes evolve more slowly than membrane genes with female and asexual expression, no significant variation is observed among membrane and non-membrane genes with male-biased expression. Furthermore, male non-membrane genes have been shown to evolve faster than female non-membrane genes (Khan et al., 2012).

In the present study, a small but significant variation in the non-synonymous branch lengths MLEs was observed among location categories. Male non-membrane genes showed signs of accelerated evolution relative to female, and male-female common non-membrane genes. The accelerated evolution of male, non-membrane genes could be driven by a parasite's life history traits. For instance, it has been found that low gametocyte density, as observed in some species (*e.g.*, *P. falciparum*) (McKenzie et al., 2006), favors a less female-biased sex ratio and affects local mate competition and fertilization success (Neal and Schall, 2014).

Genes showing signs of episodic selection were not found in *Plasmodium* species of human interest; however, episodic selection was commonly observed in *P. cynomolgi* and *P. berghei* (Table S1 and Table S2). This suggests that, while certain genes with

gametocyte biased expression could evolve under long-term diversifying selection, this trend does not seem to be observed in the main causal agents of human malarias. Nonetheless, given the cases of human infections by traditionally non-human malaria parasites (White, 2008; Ta et al., 2014), genes with gametocyte-biased expression should be evaluated in detail when presenting with this type of pattern, whether is in human malarias or not. The majority of genes with signs of episodic selection were hypothetical proteins with an unknown function (Table S1). It is likely that genes involved in functions related to host-parasite interaction or immune evasion are more prone to be positively selected (Kuo and Kissinger, 2008). Thus, it is possible that these genes could perform the aforementioned functions, but remain to be characterized. These results, however, should be taken with caution since, while efforts were made to include species representing three of the main malaria clades, there are only a few species representing each one. Therefore, it is possible that the power to detect significant estimates of adaptive evolution is reduced in our analyses. While larger power could be obtained by incorporating other species of the simian clade, no other rodent malaria genomes are currently available, and the *P. gaboni* genome, while available, is largely incomplete.

Even when using the reanalyzed *P. berghei* proteome (Tao et al., 2014), non-membrane located genes with male biased expression presented a larger proportion of immune epitopes relative to other categories. This result is in agreement with previous reports that show the proportion of genes with epitopes in the male, non-membrane category (0.292) to be larger than in the male membrane (0.107), female membrane (0.167), and female non-membrane (0.158) categories (Khan et al., 2012). In addition, the male-female common category, previously not included, also presented a similar level of

immunogenicity. Among genes where immune epitopes were reported, the proportion of genes with signs of positive selection was the largest in the male and non-membrane categories (Table 1). Nonetheless, when the distribution of positively selected sites along the sequences is taken into account, approximately less than half of the genes with reported immune epitopes show positively selected sites located inside the immune epitope region. This suggests that while selection caused by immune pressure does have an effect on gametocyte-expressed genes, particularly in those of the female-male common and non-membrane categories, selective forces do not seem to act disproportionately in the regions where immune epitopes are located.

The higher proportion of positively selected sites in genes with reported immune epitopes could indicate that evolution in a portion of gametocyte-expressed genes is indeed driven by the host's immune system. However, it is possible that alternative immunogenic regions remain to be described, or that additional forces shaping the evolution of these genes remain to be explored (*e.g.*, interaction with the mosquito vector). Furthermore, the distribution of positively selected sites was not significantly skewed to the putative location of immune epitope regions even when only high confidence immune epitopes were considered. This shows that the aforementioned trend can be observed regardless of the low confidence epitopes included in this analysis.

In the case of leading TBV candidates, lower levels of genetic variability have been observed in comparison to those of asexual and pre-erythrocytic vaccine candidates. It is possible that these differences are caused, in part, by the variable selective pressures generated by exposure to the vertebrate host and vector immune systems (Carter, 2001;

Da et al., 2013). Only a few studies have highlighted the importance of assessing the effect of long term evolution of known and potential TBV candidates, even when such studies are key for detecting novel TBV candidates (Sinden et al., 2012). Overall, leading TBV candidates included in the present study did not show significant signs of episodic selection with the exception of P28 (Table S1 and Table S2).

When I independently evaluated the presence of positively selected sites and the effects of natural selection in *P. vivax* and *P. falciparum* orthologs, I was found that positively selected sites were present in P47, P48/45, P230 and P230p in both species, with *Pfs*47 and *Pvs*230 respectively showing significant signs of positive and negative selection (Table 1). These results are in agreement with previously reported trends observed in both TBV candidates, where is has been observed that the distribution of non-synonymous polymorphisms is geographically skewed (Anthony et al., 2007; Doi et al., 2011). Antibodies that produce effective transmission blocking activity have been described in P48/45 (Roeffen et al., 2001; Tachibana et al., 2015) and P230 (Tachibana et al., 2012; Williamson, 2003); however, transmission blocking activity has been described in *Pv*s47 but not *Pf*s47 (van Schaijk et al., 2006). These differences highlight the complexity of developing effective antimalarial treatments, and further emphasize the need to explore additional TBV candidates with consistent selective signals and transmission blocking activity across *Plasmodium* species. Positively selected sites were also found in *Pvs*28 even though there was no deviation from neutrality. This suggests that P28 might be under diversifying selection in the former (Table 1).

Overall, my present results indicate that different patterns of accelerated evolution can be found across *Plasmodium* species, but are likely limited to a reduced number of genes. Different sex and location categories showed variable levels of immunogenicity; however, this variation did not seem to affect long term evolutionary trends among sex or location. Furthermore, short term evolutionary trends were not uniquely associated with the putative location of immune epitopes suggesting that immune pressures might not be the only factor shaping the evolution of gametocyte-expressed genes. Finally, while signs of episodic selection were not observed in known TBV candidates, with the exception of P28, *Plasmodium* species of human interest did show patterns consistent with positive selection acting on the observed polymorphism.

Tables

Table 3-1. Polymorphism and positively selected sites (REL) in putative TBV candidates.

Species	Name	Gene ID*	N	π [SD]	Ds	Dn	Dn-Ds [SD]	Z test	Neutrality	Prop. + selected sites (REL)
P. vivax	p230p	PVX_003900	80	0.001 [0]	0.001	0	0.001 [0.001]	0.233 (-1.200)	Dn = Ds	0.011410315
	p230	PVX_003905	111	0.001 [0]	0.002	0.001	-0.001 [0.001]	0.040 (-2.081)	**Dn < Ds**	0.010905125
	p48/45	PVX_083235	369	0.001 [0]	0.001	0.001	0.001 [0.001]	0.166 (1.392)	Dn = Ds	0.029748284
	p47	PVX_083240	166	0.003 [0.001]	0.002	0.003	0.001 [0.002]	0.543 (0.610)	Dn = Ds	0.043373494
	Hado	PVX_084290	101	0 [0]	0	0	0 [0]	1 (0)	Dn = Ds	0
	Soap	PVX_086220	108	0.001 [0.001]	0.002	0.001	-0.001 [0.002]	0.761 (-0.304)	Dn = Ds	0
	Gamer	PVX_093500	103	0.001 [0.001]	0.003	0	-0.003 [0.003]	0.283 (-1.078)	Dn = Ds	0
	Warp	PVX_093675	151	0.001 [0.001]	0.001	0.002	0.001 [0.001]	0.363 (0.913)	Dn = Ds	0
	p25	PVX_111175	315	0.003 [0.001]	0.003	0.003	0.001 [0.001]	0.538 (0.617)	Dn = Ds	0
	p28	PVX_111180	201	0.004 [0.001]	0.002	0.004	0.002 [0.001]	0.112 (1.601)	Dn = Ds	0.061674009
	ApiAP2	PVX_123760	56	0.001 [0]	0.002	0.001	-0.001 [0.001]	0.014 (-2.499)	**Dn < Ds**	0.012030516
P. falciparum	p230p	PVX_003900	113	0.001 [0]	0.001	0	-0.001 [0.001]	0.081 (-1.759)	Dn = Ds	0.008309688
	p230	PVX_003905	90	0.001 [0]	0.001	0.001	0 [0]	0.475 (0.719)	Dn = Ds	0.004163997
	p48/45	PVX_083235	244	0.002 [0.001]	0.001	0.002	0.001 [0.001]	0.123 (1.553)	Dn = Ds	0.015873016
	p47	PVX_083240	245	0.001 [0]	0	0.001	0.001 [0]	0.007 (2.762)	**Dn > Ds**	0.057831325
	Hado	PVX_084290	145	0 [0]	0	0	0 [0]	0.299 (-1.044)	Dn = Ds	0
	Soap	PVX_086220	197	0.001 [0.001]	0.001	0.001	0 [0.001]	0.531 (0.628)	Dn = Ds	0
	Gamer	PVX_093500	195	0 [0]	0	0	0 [0]	1 (0)	Dn = Ds	0
	Warp	PVX_093675	197	0 [0]	0	0	0 [0]	0.451 (0.756)	Dn = Ds	0
	p25	PVX_111175	215	0.001 [0]	0	0.001	0.001 [0.001]	0.125 (1.545)	Dn = Ds	0.870689655
	p28	PVX_111180	194	0 [0]	0	0	0 [0]	0.500 (0.676)	Dn = Ds	0
	ApiAP2	PVX_123760	87	0.001 [0]	0.001	0.001	0 [0]	0.747 (0.324)	Dn = Ds	0.008291874

* *P. vivax* nomenclature taken from Carlton et al., 2009

Figures

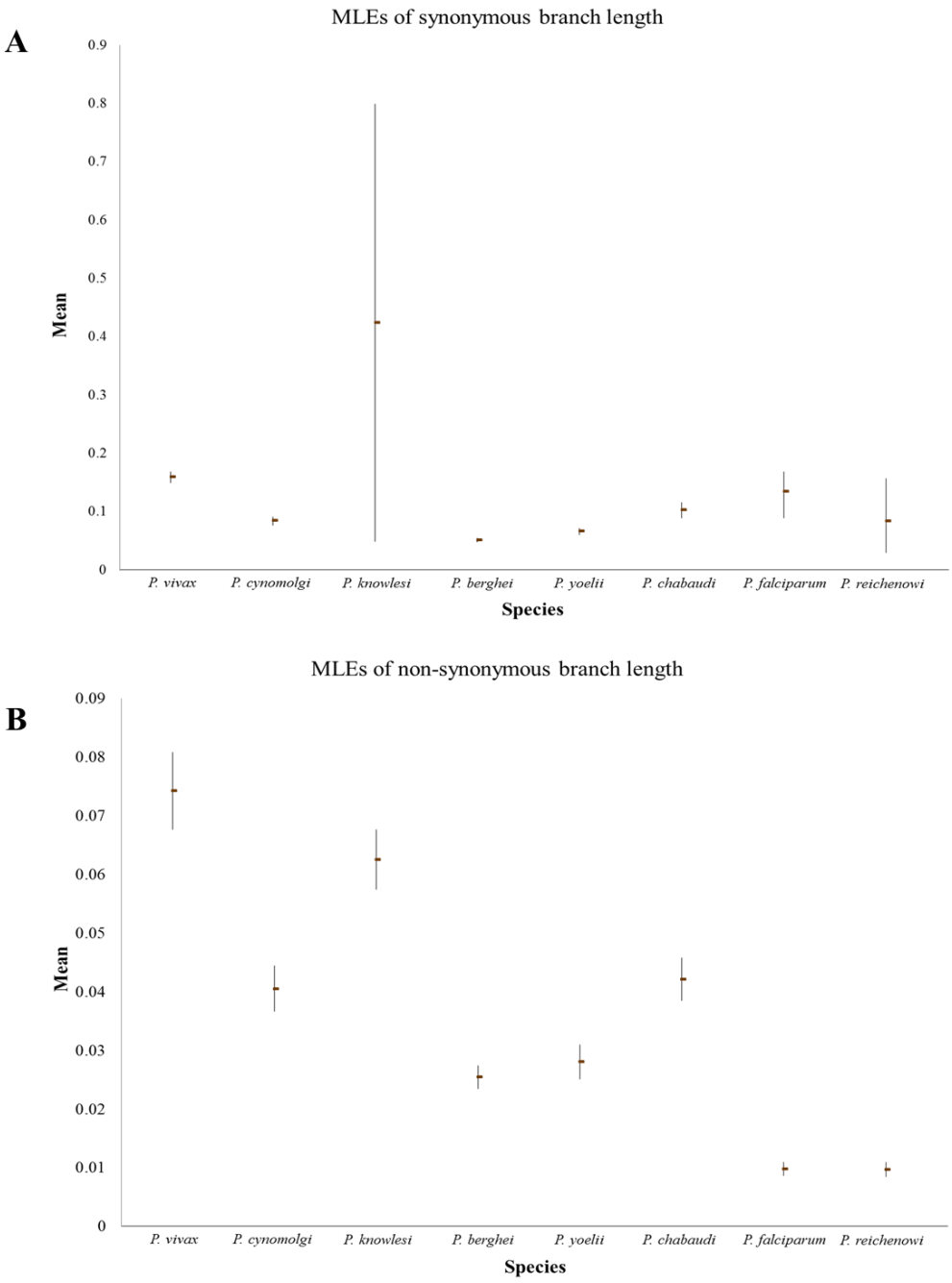


Figure 3-1. Species-specific synonymous and non-synonymous branch MLEs in genes with gametocyte biased expression. Mean values of synonymous (**A**) and non-synonymous (**B**) branch MLEs are marked as circles. Upper and lower confidence

intervals are indicated by error bars. Error bars larger than the graph scale are not included.

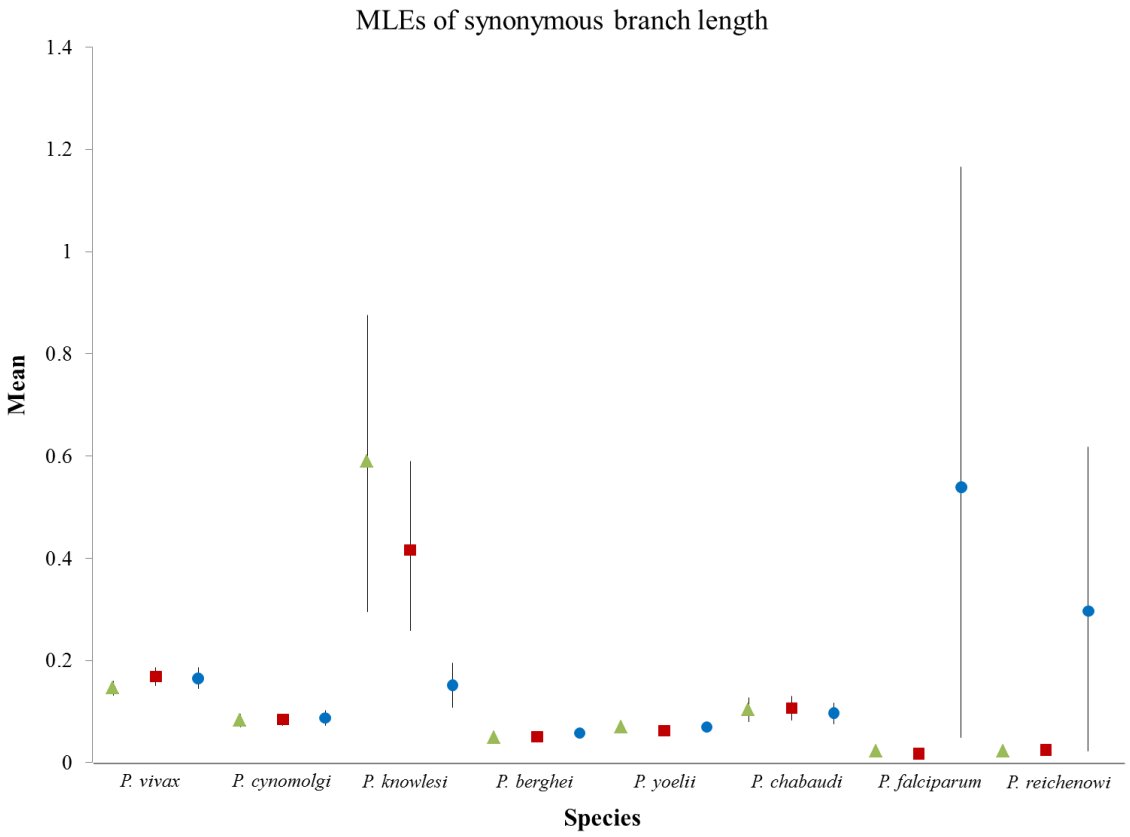


Figure 3-2. Species-specific synonymous branch MLEs in genes with gametocyte biased expression classified by sex categories. Mean values of synonymous branch MLEs of genes with female (**green triangle**), male (**red square**), and female-male common (**blue circle**) expression are marked as circles. Upper and lower confidence intervals are indicated by error bars.

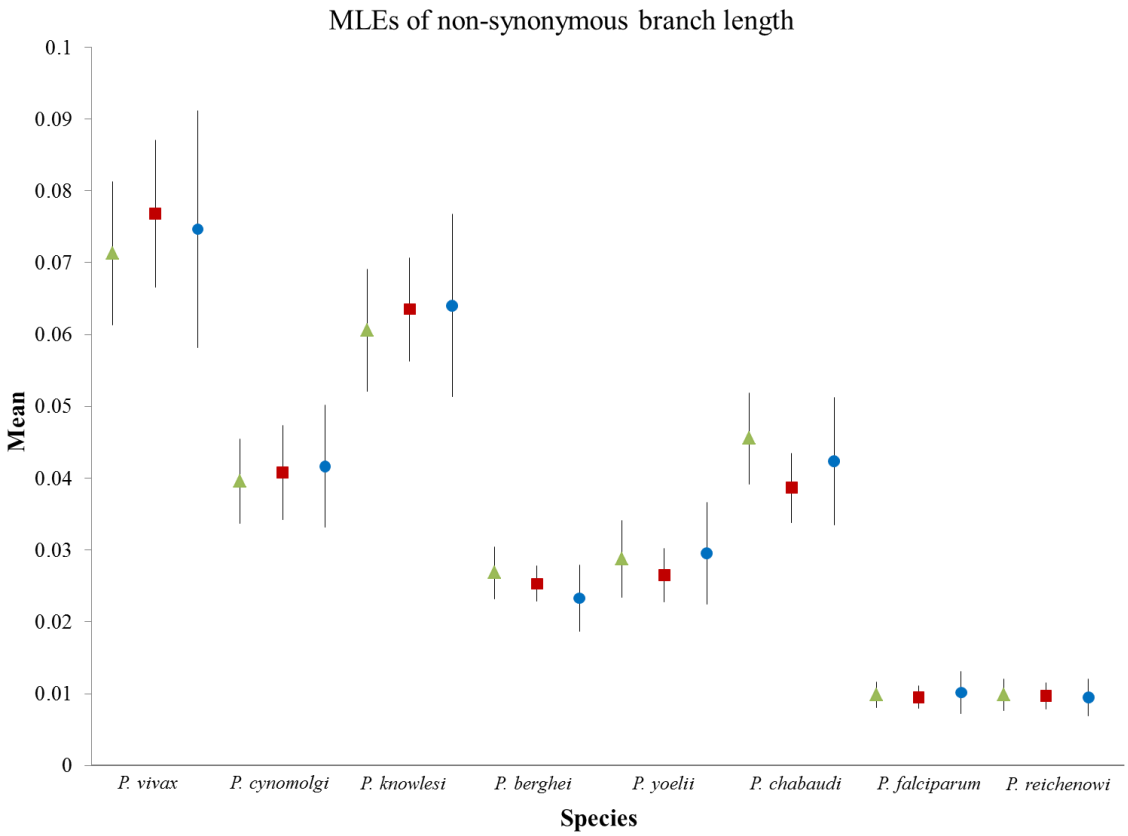


Figure 3-3. Species-specific non-synonymous branch MLEs in genes with gametocyte biased expression classified by sex categories. Mean values of non-synonymous branch MLEs of genes with female (**green triangle**), male (**red square**), and female-male common (**blue circle**) expression are marked as circles. Upper and lower confidence intervals are indicated by error bars.

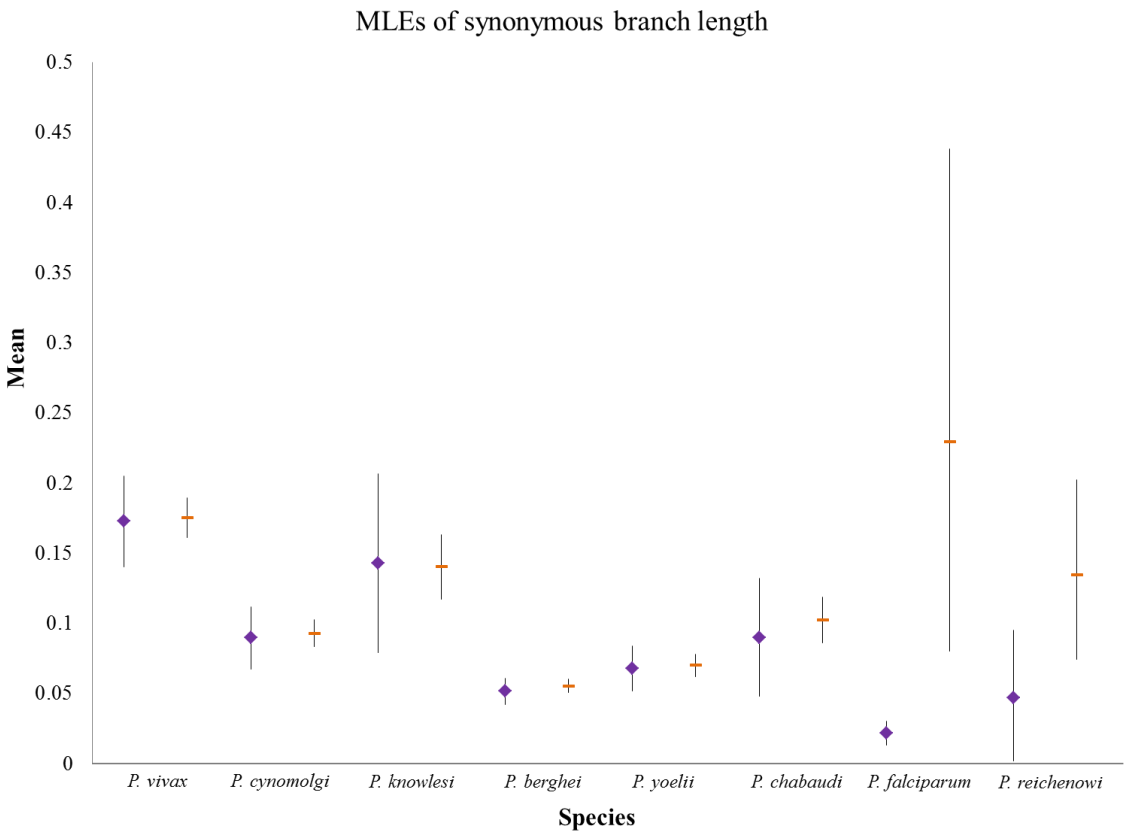


Figure 3-4. Species-specific synonymous branch MLEs in genes with gametocyte biased expression classified by location categories**.** Mean values of synonymous branch MLEs of genes with membrane (**purple diamonds**), and non-membrane (**orange lines**) location are marked as circles. Upper and lower confidence intervals are indicated by error bars.

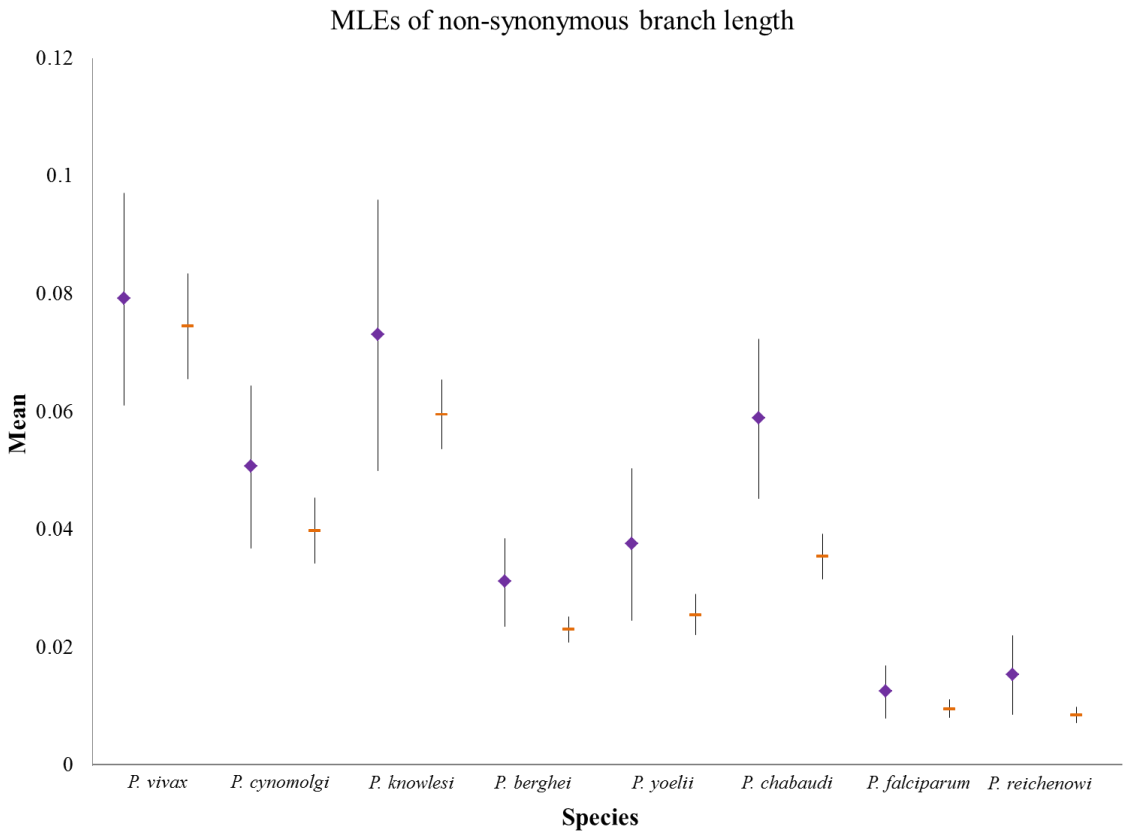


Figure 3-5. Species-specific non-synonymous branch MLEs in genes with gametocyte biased expression classified by location categories. Mean values of non-synonymous branch MLEs of genes with membrane (**purple diamonds**), and non-membrane (**orange lines**) location are marked as circles. Upper and lower confidence intervals are indicated by error bars in each mean value.

CHAPTER 4. Evolutionary trends in *Plasmodium* spp. genus common multigene families (GCMFs).

4.1 Introduction

Understanding the mechanisms of multigene family evolution and reconciling them with various evolutionary models has important implications for the study of organismal evolution and gene duplication dynamics (Okuda-Ashitaka et al., 1998; Rooney, 2004). Within parasitic protists (phylum *Apicomplexa*), several studies have pointed out the fundamental role of multigene families as a source of adaptation to biological niches and diversification among species (DeBarry and Kissinger, 2011; Kooij et al., 2005; Kuo and Kissinger, 2008; Weir et al., 2009). Specifically within the genus *Plasmodium* (the parasites which cause malaria), the largest differences among species' genomes have been found within their multigene families, with lineage-specific duplications and deletions being observed even among closely related species (Tachibana et al., 2012). Furthermore, differences in genome size and variation in selective pressures have been associated with adaptation to specific host types (Frech and Chen, 2011).

In their complex life cycle, parasites of the genus *Plasmodium* have to withstand a wide range of environments and selective pressures in order to successfully invade both *Anopheles* mosquitos and vertebrate hosts. The immune response of *Anopheles* mosquitoes acts as an important barrier to malaria transmission, and has led to specific evasive mechanisms by the parasite (Habtewold et al., 2008; Molina-Cruz and Barillas-Mury, 2014). *Anopheles* mosquitoes can combat *Plasmodium* infection using two defense mechanisms: 1), the presence of two physical barriers (the peritrophic matrix and midgut

epithelium) that hinder *Plasmodium* passage; and 2) the development of membrane-bound receptors that mark parasite cells for lysis or melanization (Saraiva et al., 2016). In humans, the humoral immune response targets both *Plasmodium* asexual blood stages and gametocyte-infected erythrocytes (Beeson et al., 2016; Stone et al., 2016). It also prevents the development of infections via the following mechanisms: blockage of erythrocyte invasion, opsonization and lysis of parasitized erythrocytes, and interference of infected erythrocyte adherence to the vascular endothelium (Ryg-Cornejo et al., 2016).

Within *Plasmodium*, certain multigene families are involved in immune evasion and mediation of host-parasite interactions. Clade-specific families found in species of the *Laveranian* subgenus: *P. falciparum* and *P. reichenowi* (*var*, *Rifin* and *Stevor*) are involved in functions related to immune evasion via antigenic variation, development of virulence, mediation of host-parasite interactions, and cytoadherence (Claessens et al., 2014; Kyes et al., 2007; Niang et al., 2009; Petter et al., 2008). Other multigene families are species-specific; for example, the SICAvar family (involved in antigenic variation) is unique to *P. knowlesi* (Lapp et al., 2009). Hypervariable multigene families, usually found in the sub-telomeric regions of *Plasmodium* chromosomes, are thought to be widely affected by recombination events (Claessens et al., 2014). It is believed that high recombination leads to rapid evolution and large gene turnover within these families (Kuo and Kissinger, 2008). Rapid turnover rates can also create species-specific arrays of paralogs in clade-specific families (Frech and Chen, 2013). The involvement of multigene families in functions related to both immune evasion and cell invasion shows that interactions with different types of host have the capacity to shape the evolution of *Plasmodium* parasites in their own unique ways. The complex life cycle and wide range

of vertebrate hosts within *Plasmodium*, provide an excellent opportunity to use comparative genomics as a mean to assess this hypothesis in genus common multigene families (GCMFs).

Variation in the number of paralogs and paralog composition within GCMFs is rarely as drastic as that observed in species-specific multigene families (Aoki et al., 2002; Gupta et al., 2015). However, many of these families also have highly important functions in cytoadherence, merozoite invasion, and immune evasion (Arisue et al., 2011; Gupta et al., 2015; Tachibana et al., 2012). On the other hand, many GCMFs perform a variety of housekeeping functions across different *Plasmodium* species, such as involvement in metabolic pathways, and participating in the formation and maintenance of several parasitic structures. While the mechanisms shaping the evolution of specific GCMFs have been examined, genome-wide trends have been characterized in far less detail. A more thorough description of GCMFs evolvability would be of great significance in understanding genus-specific evolutionary trends, and might also highlight additional families of clinical interest.

In the present study, I evaluate the evolutionary trends of GCMFs found in *Plasmodium* parasites. I also explore the putative variation of the composition and number of paralogs in each multigene family, and their relationship to recombination events and long term selective. Furthermore, I assess the putative role of host-parasite interactions in shaping these trends.

4.2 Methods

4.2.1 *Sequence data and classification of multigene family paralogs based on transcriptomic data.*

I identified multigene family members using BLAST (Altschul et al., 1997) searches against 11 *Plasmodium* species with current publicly available genomes. I used the PlasmoDB version 28.0 (Autorrecochea, 2009) and NCBI (Benson et al., 2015) databases for sequence searches of all primate malarias included in this study: *P. vivax* (Salvador I strain), *P. cynomolgi* (B-strain), *P. knowlesi* (H strain), *P. inui* (San Antonio strain), *P. coatneyi* (Hackeri strain), *P. falciparum* (3D7 strain) and *P. reichenowi* (CDC strain); and the rodent malarias: *P. yoelii* (YM strain), *P. berghei* (ANKA strain) and *P. chabaudi chabaudi* (AS strain). In addition, I obtained *P. gonderi* sequences from 454 reads (Roche, Applied Science, Basel, Switzerland) were also included.

I used the OrthoMCL database version 5.0. to select GCMFs. I established GCMFs as having at least one ortholog and a minimum of one paralog in any of the included *Plasmodium* species. This method assured the inclusion of multigene families with species- or clade-specific duplication events (referred henceforth as in-paralog families), and families with genus-specific duplication events (referred henceforth as out-paralog families). I found a total of 97 GCMFs. I performed interspecies alignments independently for each multigene family using the MUSCLE (Edgar, 2004) algorithm incorporated into SeaView version 4 (Gouy et al., 2010), followed by manual editing of protein and nucleotide sequences. Of the 97 selected families, I excluded seven (1-cys-glutaredoxin-like protein, Akrin repeat, cytosolic Fe-S cluster assembly factor, dynein

light chain type 2, tRNA pseudouridine synthase, leucine rich repeat antigen and alpha-beta hydrolase 2) from further analyses due to the presence of large unalignable regions and reduced numbers of informative sites. I also excluded paralogs with dramatically shorter sequence length from further analyses. Thus, I effectively obtained multisequence alignments for 90 multigene families.

I categorized paralogs as vector-specific, vertebrate-specific, or generalist based on the expression patterns found in the *P. falciparum* (López-Barragán et al., 2011) and *P. berghei* (Otto et al., 2014a) transcriptomes. I classified paralogs presenting twice the expression levels in stages associated to the vector host (gametocyte V and ookinete) in comparison to other life cycle stages as vector-specific. Similarly, I classified paralogs with twice the expression levels in stages associated to the vertebrate host (ring trophozoite, trophozoite and schizont) as vertebrate-specific. Finally, I classified paralogs with an expression profile less than double in either life cycle stage as generalist. Classification of paralogs was repeated independently using both transcriptomes.

4.2.2 *Phylogenetic tree construction.*

A Python pipeline was developed in order to automatize the multigene family tree building process. For each alignment, the most adequate substitution model was selected using the Akaike information criterion (AIC) method incorporated in Jmodeltest (Posada, 2008). Nucleotide frequencies, fraction of invariable sites, and the shape parameter (alpha) of the gamma distribution were specified for each analysis. Both Maximum Likelihood (ML) and Bayesian Inference (BI) methods were used to construct phylogenetic trees for each analyzed multigene family. PhyML v3.1 (Guindon et al.,

2010) with 1000 bootstrap pseudo-replications was used to assess node significance on 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. The prior parameters for each BI analysis were incorporated from the Jmodeltest results using the established Python pipeline. For each BI tree, the corresponding multigene family topology, stationary nucleotides frequencies, nucleotide substitution rates, proportion of invariable sites, and the shape parameter (alpha) of the gamma distribution were specified. Sampling was performed every 1000 generations with a burn-in fraction of 50%. Convergence of the BI analysis was diagnosed by requiring a standard deviation lower than 0.01 and a Potential Scale Reduction Factor (PSRF) close to 1.0.

4.2.3 *Number of multigene family paralogs.*

I used the software package Count (Csuos, 2010) to identify changes in the number of multigene family paralogs along the *Plasmodium* phylogeny and to make inferences about the evolutionary history of each multigene family. Count makes evolutionary inferences of family sizes along a phylogeny by using Wagner parsimony (Farris, 1970). Under Wagner parsimony loss and gain of individual family members are penalized independently, resulting in an inferred multigene family history that minimized penalty of gain/loss events (Csuos, 2010). I used the same tree topology, constructed using published *Plasmodium* species phylogenies (Pacheco et al., 2011), to infer lineage-specific gain/loss events across the 90 multigene families.

4.2.4 *Episodic Selection and recombination.*

I measured the putative effects of episodic selection acting on different branches of the multigene family tree using Hyphy's random effects Branch-Site REL model (BSREL) and its adaptive version (aBSREL) (Kosakovsky Pond et al., 2011; Pond et al., 2005; Smith et al., 2015). Neither the BSREL or aBSREL models require *a priori* partitions between positively selected foreground branches and negatively selected or neutral background branches. Nonetheless, the aBSREL model optimizes the number of selective regimes necessary to better assess evolution of each branch and reduces computational time. I executed the aBSREL model independently for the 90 multigene families using the ML phylogeny obtained from PhyML as input. The alpha (α) and omega (ω) values were allowed to vary among branches and branch-site combinations. Internal and terminal branches with a corrected *p*-value <0.01 were considered to be under significant episodic selection. I recorded the strength of episodic selection and proportion of sites under this selective regime on each case. Also, I evaluated significant differences in the distribution of the omega (ω) values and proportion of sites with an ANOVA test. I used the expression categories established by the *P. falciparum* and *P. berghei* transcriptomic data (vector-specific, vertebrate-specific, and generalist) as factors. All statistical analyses were performed using R version 3.2.2.

In addition, I performed Hyphy's test of Relaxed Selection (RELAX) (Wertheim et al., 2015) in multigene families showing significant signs of episodic selection. Given two subsets of branches in a phylogeny, RELAX can determine whether selective strength was relaxed or intensified in one subset relative to the other (Wertheim et al.,

2015). For each alignment, branches with signs of positive selection are included in a first subset while all other branches are included in a second subset. Then, using the RELAX test, the selective strength of branches on the first subset can be classified as relaxed or intensified in comparison to branches of the second subset.

I also assessed the role that recombination events might have in the evolution of GCMFs using the Recombination Detection Program, RDP4 (Martin et al., 2015) by performing independent analyses for each multigene family. I recorded the number of recombination events and multigene family members involved in each event. In addition, I evaluated the existence of potentially spurious signs of episodic selection due to recombination by repeating the aBSREL test using the FastNJ non-recombinant tree generated in RDP4.

4.2.5 *Polymorphism of larger GCMFs.*

I assessed the genetic diversity (π) among different isolates in the seven multigene families with largest variation in the number and composition of paralogs: Acyl-CoA synthase, Cytoadherence-linked asexual protein (CLAG), Lysophospholipase, NIMA related kinase (NEK), Papain, Plasmepsin, and Serine repeat antigen (SERA). I performed the analysis in both *P. vivax* and *P. falciparum* using the large number of worldwide isolates currently available for both species. I assessed patterns consistent with natural selection acting on the observed polymorphism by calculating the differences in the mean number of synonymous (dS) and non-synonymous substitutions (dN) using the Nei-Gojobori distance method (Nei and Gojobori, 1986), with the Jukes and Cantor correction implemented in MEGA 6.06 (Tamura et al., 2013). The difference between dS

and dN and its standard error was estimated by using bootstrap with 500 pseudo-replications, as well as a two-tailed codon based Z-test of the difference between dS and dN (Nei and Kumar 2000). Under the neutral model, synonymous substitutions accumulate faster than non-synonymous because they do not affect the parasite fitness and/or purifying selection is expected to act against non-synonymous substitutions ($dS \geq dN$). Conversely, if positive selection is maintaining polymorphism or driving divergence, then a higher incidence of non-synonymous substitutions is expected ($dS < dN$). I assumed as a null hypothesis that the observed polymorphism was not under selection ($dS = dN$).

I used worldwide *P. vivax* isolates from the database for clinical isolates representing diverse geographic regions as part of the Hybrid Selection Initiative performed by the Broad Institute available in PlasmoDB version 28.0 (Autorrecochea, 2009). In addition, the five sequenced *P. vivax* reference strains (Salvador I, North Korean, India VII, Mauritania I and Brazil I) publically available via the Malaria Research and Reference Reagent Resource Center were included. Alternatively, I obtained *P. falciparum* isolates, also available in PlasmoDB version 28.0 (Autorrecochea, 2009), from the following diverse 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.

4.3 Results

4.3.1 *Phylogenetic tree construction.*

Among the 90 analyzed multigene families, I found evidence that orthologs were more closely related than paralogs in 58. Tree topologies in these families showed paralogs clearly separated into two clades, with each clade including all species orthologs (Fig. S1). This is indicative of historical gene duplication events that predate speciation within the clade. The remaining gene family tree topologies indicated of species- and clade-specific duplication events, with putative evidence of lineage-dependent gain/loss events (Fig. S1). Tree topologies for six of the multigene families (conserved *Plasmodium* protein unknown function 6, conserved Rodent malaria protein unknown function, CLAG, Elongation factor1, Eukaryotic initiation factor 2a, and Glutathione synthetase) revealed higher sequence similarity among paralogs than among orthologs, suggesting predominance of different evolutionary mechanisms than those potentially acting on other GCMFs (Fig. S2).

4.3.2 *Number of multigene family paralogs.*

Of the 90 analyzed multigene families, 36 shared similarly biased *P. berghei* and *P. falciparum* expression profiles to the *Anopheles* mosquito (vector-specific = 14), vertebrate host (vertebrate-specific = 15), or both vector- and vertebrate hosts (generalist = 7). The expression profile varied in one paralog, but remained unchanged in other family members in 39 multigene families. The remaining 15 multigene families showed entirely different expression profiles (Table 1). Regardless of their expression profile, the majority of analyzed multigene families had just two paralogs in all included *Plasmodium*

species (Fig. 1, Fig. 2). Furthermore, with few notable exceptions (Table 2), I observed little variation in the number of multigene family paralogs across extant *Plasmodium* species or in the number of paralogs inferred for *Plasmodium* ancestors, independent of the paralogs' expression profile (vector specific, vertebrate specific, or generalist). Larger multigene families showed significant changes in the number of paralogs among *Plasmodium* species. This was particularly observed in the following families: CLAG, SERA, Plasmepsin, Papain, Acyl-CoA synthetase, and Lysophospholipase. Moreover, these families commonly had paralogs with vertebrate-biased expression and combinations of paralogs with either vector- or vertebrate-biased expression in both *P. falciparum* and *P. berghei* (Table 1). On the other hand, multigene families with species- or clade-specific duplication events tended to be more commonly associated with life stages expressed uniquely in either the mosquito vector or vertebrate host (Table 1).

4.3.3 *Episodic Selection and recombination.*

I found significant signs of episodic selection in 29 (Table S3) of the 90 multigene families evaluated. This pattern could be influenced by recombination events. When I used the FastNJ non-recombinant tree to confirm these results, signs of episodic selection were maintained in 23 of those families. The distribution, strength and proportion of sites under episodic selection remained relatively unchanged in the 23 families regardless of the use of non-recombinant tree. Signs of episodic selection were distributed in both internal and terminal branches of each family phylogeny. However, only 4 cases of episodic selection were found in terminal branches leading to *Laveranian* paralogs or their ancestors. The majority of branches in which signs of episodic selection were

detected, belonged to either the simian or rodent clades (Fig. S3). With the exception of two families (Asparagine tRNA ligase and a hypothetical protein), very strong positive selection was limited to a small number of sites (<10%). Furthermore, excluding the conserved *Plasmodium* protein unknown function 6 family, the proportion of sites under positive selection was lower in terminal branches of the phylogeny in all evaluated families. In contrast, the proportion of sites showing significant signs of positive selection was relatively larger in numerous ancestral branches leading to the split of family paralogs (Fig. S3).

I further tested branches that showed significant signs of episodic selection for signals of relaxed or intensified selection relative to other branches of the phylogeny (Table 3, Table S1). Selection was significantly intensified in 13 of the 23 analyzed multigene families, and significantly relaxed in only 3 of them: conserved *Plasmodium* protein, CLAG, and NEK. Intensification of selection was markedly larger in the Chaperonin (K=37.05) and hypothetical protein (K=50) multigene families. Nonetheless, only the hypothetical protein multigene family showed strong signs of episodic selection in the same evaluated branch (Fig. S3). On the other hand, positively selected branches in the Chaperonin multigene family tree showed similar strengths and proportions of sites under episodic selection as in other multigene families evaluated.

The intrinsically different nature of the immune response developed by the vector and vertebrate hosts against *Plasmodium* parasites can act as a selective mechanism driving the evolution of multigene families. However, I found no significant differences in the distribution of the omega (ω) values or their corresponding proportions of sites

(Table S4, Table S5) relative to the expression profiles of *P. falciparum* and *P. berghei* paralogs. This suggests that there are no differential selective patterns among paralogs with distinct expression profiles.

I found significant signals of recombination in 53 of the 90 analyzed multigene families (Table 4). Among these, recombination events were less frequently observed among ortholog than among paralog members of the same multigene family (Fig. 3). Furthermore, when recombination events occurred among paralog members of a multigene family, they were observed with a similar frequency among the same species paralogs than among paralog genes from different *Plasmodium* species (Fig. 3). Moreover, recombination events occurring among paralogs were more frequently observed in families showing duplication events predating species diversification (out-paralogs) than in families with putatively recent duplication events (in-paralogs). No significant signs of recombination were found among the intra-genomic paralogs of the previously described six families presenting higher sequence similarity among paralogs than orthologs (Fig. S2, Table S2).

I observed a median of two recombination events and a median length recombinant segment of 190 bp. among the analyzed multigene families. Most families had between one and two recombination events (Table 4). With the exception of Plasmepsin, large multigene families tended to show a larger number of recombination events among orthologs and same species paralogs. Multigene families showing two or more clearly defined out-paralogs tended to have single recombination events between paralogs (*e.g.*, Calcium dependent protein kinase (CDPK) and DHHC type zinc finger

protein). Alternatively, families with species- and clade-specific duplications had a tendency to show recombination events occurring between orthologs (*e.g.*, Adrenoxin reductase and Biotin acyl-CoA carboxylase).

4.3.4 *Polymorphism of larger GCMFs.*

Larger GCMFs showed reduced levels of polymorphism in both *P. falciparum* and *P. vivax* worldwide isolates. With the exception of a single *P. falciparum* (PF3D7_0215300) paralog of the Acyl-CoA synthetase multigene family, and two *P. falciparum* paralogs (PF3D7_0207400 and PF3D7_0207700) of the SERA multigene family, no significant signs of positive selection were found in either *P. vivax* or *P. falciparum*. On the other hand, excluding paralogs of the Lysophospholipase family, numerous *P. falciparum* paralogs showed significant signs of purifying selection in all multigene families evaluated (Table 5). Only some *P. vivax* paralogs of the CLAG, Acyl-CoA synthetase, NEK, and SERA families showed signs of purifying selection. Orthologs across the two species did not show similar selection patterns, with the exception of a single NEK paralog, which suggests that human *Plasmodium* species of the *Laveranian* subgenus and simian clade maintain different selective profiles. Furthermore, the observed signs of purifying selection on family-specific paralogs could suggest that they perform significant functions relevant to each multigene family.

4.4 Discussion

Repeated lineage-specific gene duplication and/or deletion events, have been described in the simian (Tachibana et al., 2012) and rodent clades (Otto et al., 2014a), and *Laveranian* subgenus (Otto et al., 2014b). Regardless of the clade in which they are

described, hyper-variable multigene families tend to be involved in functions related to immune evasion, cell invasion, sequestration, and virulence (Otto et al., 2014a, 2014b; Tachibana et al., 2012). However, while species- and clade-specific multigene families have been the source of much research, less is known about GCMFs even though they could represent an important source for the development of new malaria treatments and help understand the evolutionary forces acting within the *Plasmodium* genus.

In the present study, I found little variation in the number and composition of GCMFs paralogs. Family tree topologies showed two clearly and highly supported clades for each paralog in 58 of the multigene families evaluated (Fig. S1). This pattern suggests that a number of genus-common duplication events predate the divergence of three of the major *Plasmodium* clades. Moreover, conservation of the number and composition of paralogs observed in largely divergent *Plasmodium* species, and the reduced number of pseudogenization events (Table 2), suggests that gene duplicates are maintained despite marked differences in *Plasmodium* life cycle and in life history traits. Within the 58 described families, the Gene Ontology (GO) annotated functions included: energy and protein transport, metabolism, signaling, regulation of cell cycle processes, formation and maintenance of membrane structures, and DNA repair. Thus, it is possible that ancestral duplication events are beneficial for the parasite, leading to the preservation of duplicated copies after speciation events. Evolutionary models such as dosage balance, shielding against deleterious mutations, and positive dosage (Innan and Kondrashov, 2010) could be reconciled with the evolutionary patterns observed in these multigene families.

On the other hand, multigene families with species- or clade-specific duplication events showed variable gene tree topologies. In these cases, duplication events were closely related on the tree (Fig. S1). Associated GO functions in these families involved a variety of metabolic and regulatory processes (protein oxidation, hydrolases, ligases, etc.), but also included functions related to entry to host cell, immune evasion, phospholipid metabolism, and hemoglobin degradation. In this case, evolutionary models such as shielding against deleterious mutations, duplication degeneration complementation (DDC), positive dosage, neo-functionalization, and sub-functionalization could be reconciled with the gene duplication patterns observed (Innan and Kondrashov, 2010).

In general, GCMFs tended to have a lower number of paralogs than species- or clade-specific families. Little variation in the number of multigene family paralogs was observed independently of their expression profiles across extant *Plasmodium* species and in the inferred number of paralogs in *Plasmodium* ancestors (Fig. 1, Fig. 2), suggesting that host-parasite associations do not act as drivers for the occurrence and/or fixation of duplication events. Moreover, no significant differences in the distribution of the omega (ω) values or the corresponding proportion of sites under each selective regime, were observed in relation to *P. falciparum* and *P. berghei* expression. This is indicative that, while parasite-host interactions, and changes in parasite environment throughout *Plasmodium* life cycle are known drivers of adaptive evolution (Assefa et al., 2015; Molina-Cruz and Barillas-Mury, 2014; Prugnolle et al., 2008), they do not universally affect all multigene family members in the same manner. Nonetheless, these results should be assessed with caution given that changes in expression profiles were

observed between certain *P. berghei* and *P. falciparum* orthologs. Such variation shows that expression profiles are not entirely conserved across *Plasmodium* lineages, and should not be extrapolated casually into *Plasmodium* species without available expression data.

Multigene families involved in functions associated with host cell invasion, cytoadherence, immune evasion, and hemoglobin metabolism (CLAG, SERA, Plasmepsin, Papain and *msp*7) presented a larger number of family members, and were mostly expressed in *Plasmodium* stages associated with the vertebrate host (Table 1, Table 2). Furthermore, these families showed significant variation in the number and composition of paralogs among *Plasmodium* species. This pattern indicates predominance of lineage-specific duplication/loss events. The larger number of multigene family paralogs observed in primate malarias of both the simian clade and *Laveranian* subgenus suggests that changes in the number and composition of paralogs can be adaptive to specific host types.

Previous studies have proposed diverse hypotheses to explain the increase in the number of paralogs observed in primate malarias. Among the proposed hypotheses, it has been suggested that the variation could be a product of: repeated adaptation events (Ponsuwanna et al., 2016), functional redundancy among paralogs (Gupta et al., 2015), neo- or sub-functionalization of recently divergent paralogs (Bethke et al., 2006), or the result of an expansion in host range (Arisue et al., 2011). Overall, it is likely that essential family functions are performed by paralogs conserved among different *Plasmodium*

species, while species- and clade-specific gene duplicates may be involved in maintaining functional redundancy or in facilitating diversification and adaptability within the family.

Alternatively, some larger multigene families showed little variation in the number and composition of paralogs among the analyzed *Plasmodium* species (NEK, CDPK, Biotin acyl-CoA carboxylase, etc.). Expression profiles in these families' paralogs were found to be both vector- and vertebrate-specific (Table 1, Table 2), at least between the two *Plasmodium* species with available expression data (*P. berghei* and *P. falciparum*). Interestingly, they perform multiple functions throughout the parasite's life cycle, including: fatty acid synthesis (Chen et al., 2014), sexual and asexual development and commitment (Dorin-Semblat et al., 2011; Reininger et al., 2012, 2009), parasite differentiation and protein secretion (Moreno et al., 2011), and involvement in host-parasite interaction and development (Thompson et al., 2007). In this case, the conserved number of duplication events in different species indicates little involvement in family diversification and adaptability, and a more predominant role in sustaining parasite survival.

I found significant signs of episodic selection in 23 of the multigene families analyzed suggesting that the majority of GCMFs are evolving neutrally or under purifying selection. In families where signs of episodic selection were detected, species of the simian (23 branches) and rodent clades (13 branches) showed signs of episodic selection in internal and terminal branches more frequently than species of the *Laveranian* subgenus (4 branches) (Fig. S3). This trend shows that while most GCMFs do not appear to significantly deviate from neutrality, when adaptive signals are present,

they are more likely to be observed in species of the simian and rodent clades. This could be the product of: the different evolutionary histories among *Plasmodium* clades (Martinsen et al., 2008), their divergence times (Silva et al., 2015), the different associations with their respective vertebrate hosts (Mu, 2005; Prugnolle et al., 2013), or the result of lineage-specific processes essential for parasite survival.

On the other hand, significant episodic selection was also found in the branches leading to paralogs split in five multigene families (Fig. S3), implying that in these cases the duplicated genes have been positively selected. Previous studies have found that positive selection in the branch leading to a paralog split can be linked to emerging functions in the duplicated paralog (Hakes et al., 2007; Summers and Zhu, 2008; Van Zee et al., 2016). Nonetheless, multigene families where this pattern was observed shared the following GO term functions: ATPases, cell division, commitment to sexual and asexual stages, and recombination. This makes it difficult to establish what would the putative biological advantages of acquiring novel functions in newly duplicated paralogs.

Intensified selection was observed in all branches with significant signs of episodic selection with the exception of three families: CLAG, NEK and a conserved *Plasmodium* protein (Table 3, Table S1). These results support the hypothesis that positive selection is being driven by lineage-specific processes in the majority of multigene families. However, it also shows that relaxed selection can result in sequence diversification of certain paralogs. It is possible that this pattern is the product of a duplication event where the original paralogs maintained family function, leaving the

newly duplicated copy free of strong selective pressures (Innan and Kondrashov, 2010; Ohno, 1970).

The role of recombination in the divergence and emergence of novel paralogs has been previously established in many species- and clade-specific multigene families (Kuo and Kissinger, 2008). Here, I detected significant signs of recombination in 53 of the multigene families evaluated, indicating that recombination is an important evolutionary mechanism in GCMFs regardless of their location on the chromosome (Fig. 3). Recombination events were less likely to be detected in multigene families where duplication events predated *Plasmodium* speciation (Fig. 3). Potentially, this is the result of large sequence divergence between ancestral family paralogs. However, while less frequent, recombination was detected within some ancestral duplications. This could imply that recent recombination events have occurred or that highly conserved sequence structures are still maintained in ancestral paralogs. Furthermore, recombination events occurring in families with species- and clade-specific duplications shows that recombination is an important mechanism in the acquisition of novel paralogs in these families as well (Table 4).

Signs of episodic selection and recombination were found in 16 of the evaluated multigene families; nonetheless, the patterns of recombinant sequences and those showing significant selective signals did not overlap (Table 4, Fig. S3). This further supports the hypothesis that, within selected multigene families, paralogs are evolving independently and recombination likely acts as a mechanism for the creation of novel duplicated genes (Fawcett and Innan, 2011). Furthermore, the absence of comparable

selective patterns among family paralogs with clearly high sequence similarity, suggests that concerted evolution is not likely predominant among GCMFs (Fig. S2).

I also found low levels of polymorphism and signs of purifying selection in *P. falciparum* and *P. vivax* strains for members of the CLAG multigene family (Table 5). This result contrasts with previous reports of high polymorphism and positive selection on *P. falciparum* (Alexandre et al., 2011; Iriko et al., 2008). Sequence diversity was also markedly lower in comparison to previous studies performed on *P. falciparum* strains for the Acyl-CoA synthetase family (Bethke et al., 2006). Furthermore, while the previous studies using similar analysis methods found signs of positive selection in multiple paralogs of this family (Bethke et al., 2006), here only one paralog (PF3D7_0215300) showed significant signs of positive selection. The difference could be related to the larger number of samples included in this study in comparison to previous ones (5, 39, and 21 *P. falciparum* strains, respectively).

Also in contrast with previous reports (Ponsuwanna et al., 2016), I found no signs of positive selection in the Plasmepsin and Papain multigene families; however, I observed signs of purifying selection in *P. falciparum* paralogs in both families. Among other larger GCMFs, *P. falciparum* strains in the CLAG multigene family and *P. vivax* strains of the SERA multigene family showed the highest levels of polymorphism (Table 5). Furthermore, signs of positive selection were observed in two *P. falciparum* SERA paralogs (PF3D7_0207500 and PF3D7_0207800). There results further support the hypothesis that certain multigene family members are positively driven towards

diversification, while a majority of family paralogs are likely selected to maintain family function.

Overall, my results show that *Plasmodium* GCMFs are not subjected to rapid evolution or diversification. However, the conservation of numerous ancestral duplications across highly divergent *Plasmodium* species shows that gene duplication is an important aspect of parasite survival. Multigene family function appears as one of the most important factors in family evolvability, indicating that host-associated selective pressures can be highly relevant for some families and insignificant in others. Knock-out studies and characterization of multigene families in more divergent *Plasmodium* species, could help to determine the significance of putative ancestral duplication events in parasite survival, as well as their role in the colonization of the mammalian host.

Tables

Table 4-1. Expression category based on *P. falciparum* & *P. berghei* transcriptome.

Multigene family	*P. falciparum*	*P. berghei*
Actin	Generalist-Vertebrate	Generalist-Vertebrate
Acyl-CoA synthase	Generalist	Vector
Adrenodoxin reductase	Vector	Vertebrate
Alpha-beta hydrolase 2	Generalist-Vector	Vector
Asparagine tRNA ligase	Vector-Vertebrate	Vertebrate
ATP-dependent DNA helicase	Generalist-Vector	Vector-Vertebrate
Biotin acel-CoA carboxylase	Vector	Generalist-Vector
Calcium Dependent Protein Kinase (CDPK)	Generalist-Vertebrate	Vector-Vertebrate
Calcium-transporting ATPase (SERCA)	Vertebrate	Vertebrate
Casein kinase II beta chain	Vertebrate	Vector-Vertebrate
Cell division protein FtsH	Generalist-Vertebrate	Vertebrate
Chaperonin (CPN)	Vector	Vector-Vertebrate
Chromatin assembly factor 1 subunit	Generalist-Vertebrate	Vertebrate
ClpB protein	Generalist-Vertebrate	Vertebrate
Conserved and hypothetical *Plasmodium* protein	Generalist	Generalist-Vertebrate
Conserved *Plasmodium* protein	Vertebrate	Vertebrate
Conserved *Plasmodium* protein, unknown function	Vertebrate	Vertebrate
Conserved *Plasmodium* protein, unknown function 11	Vector	Vector
Conserved *Plasmodium* protein, unknown function 12	Vector	Vertebrate
Conserved *Plasmodium* protein, unknown function 2	Vector	Generalist-Vector
Conserved *Plasmodium* protein, unknown function 3	Vector	Vertebrate
Conserved *Plasmodium* protein, unknown function 4	Vector	Vector
Conserved *Plasmodium* protein, unknown function 6	Vertebrate	Vector
Conserved *Plasmodium* protein, unknown function 8	Vector	Vector
Conserved Rodent malaria protein, unknown function 2	Vertebrate	Vertebrate
Cysteine Repeat Modular Protein (CRMP)	Vector	Vector-Vertebrate
Cytoadherence-linked asexual protein (CLAG)	Vertebrate	Vertebrate
DEAD-DEAH box ATP-dependent RNA helicase	Generalist-Vertebrate	Vector-Vertebrate
DER1-like protein	Vector	Vertebrate
DHHC-type zinc finger protein	Vector-Vertebrate	Generalist-Vector
Dipeptidyl aminopeptidase (DPAP)	Vector-Vertebrate	Generalist-Vector
DNA mismatch repair protein MSH2	Vector-Vertebrate	Vector-Vertebrate
DNA-directed RNA polymerase II	Vertebrate	Vertebrate
DNA-directed RNA polymerase II second largest subunit	Vertebrate	Generalist-Vertebrate
DnaJ protein2	Vector-Vertebrate	Vector

Table 4-1. Expression category based on *P. falciparum* & *P. berghei* transcriptome (continued).

Multigene family	*P. falciparum*	*P. berghei*
Dynamin-like protein	Vector-Vertebrate	Vertebrate
Dynein heavy chain	Vector	Vector
Elongation factor 1 alpha (EF-1alpha)	Vertebrate	Vertebrate
Elongation factor G	Generalist-Vertebrate	Vector-Vertebrate
Eukaryotic initiation factor 2a	Vertebrate	Vertebrate
Exonuclease	Vector-Vertebrate	Vertebrate
Folate transporter (FT)	Vector	Generalist-Vertebrate
Glutathione reductase (GR)	Generalist-Vertebrate	Vector-Vertebrate
Glutathione synthetase (GS)	Vector	Vertebrate
Glycerol 3 phosphate dehydrogenase	Vertebrate	Vertebrate
Heat shock protein 40	Vector-Vertebrate	Vertebrate
Heat shock protein 70	Vertebrate	Vector-Vertebrate
Heat shock protein 90	Generalist-Vector	Vertebrate
High mobility group protein B1 (HMGB1)	Vector-Vertebrate	Vector-Vertebrate
Histidine tRNA ligase	Generalist-Vertebrate	Vector-Vertebrate
Histone H2B	Vector-Vertebrate	Vertebrate
Histone H3	Vertebrate	Vertebrate
Hypothetical protein	Vector	Vector
Inner membrane complex protein 1c (IMC1c)	Vector	Vector
Inorganic pyrophosphatase (VP)	Vertebrate	Vector-Vertebrate
Iron sulfur assembly protein (SufA)	Vector	Vector-Vertebrate
Kinesin-8	Generalist-Vector	Vector-Vertebrate
Lysophospholipase	Vector-Vertebrate	Vector-Vertebrate
Malate dehydrogenase (MDH)	Vector-Vertebrate	Vertebrate
Meiotic recombination protein DMC	Vector	Vector
Meiotic recombination protein SPO11	Vector	Vector
Merozoite Surface Protein 7 (MSP7)	Vertebrate	Vertebrate
Methionine aminopeptidase	Generalist-Vertebrate	Generalist-Vector
Methyltransferase	Vertebrate	Generalist-Vertebrate
NADP specific glutamate dehydrogenase (GDH)	Vertebrate	Generalist-Vertebrate
NIMA related kinase (NEK)	Vector	Vector
Novel putative transporter 1 (NPT1)	Vector	Vertebrate
Nucleotide binding protein	Generalist-Vertebrate	Vector-Vertebrate
P1-s1 nuclease	Vertebrate	Vertebrate
P28	Vector	Vector
Papain	Vertebrate	Vertebrate

Table 4-1. Expression category based on *P. falciparum* & *P. berghei* transcriptome (continued).

Multigene family	*P. falciparum*	*P. berghei*
Peptide release factor	Generalist-Vector	Generalist-Vertebrate
Peroxiredoxin thioredoxin peroxidase	Vector-Vertebrate	Generalist-Vertebrate
Phosducin like protein (PhLP)	Vector	Vertebrate
Phosphopantothenoylcysteine synthetase	Vector	Generalist
Plasmepsin	Vector-Vertebrate	Vector-Vertebrate
Pre mRNA splicing helicase BRR2	Generalist	Generalist
Pre-mRNA-splicing factor ATP-dependent RNA helicase	Generalist-Vertebrate	Generalist-Vertebrate
Protein phosphatase 2C	Vector	Vector
Rhoptry associated protein 2-3	Vertebrate	Vertebrate
SEL-1 protein	Generalist-Vertebrate	Vertebrate
Serine Repeat Antigen (SERA)	Vertebrate	Vertebrate
Serine tRNA ligase	Vector-Vertebrate	Vertebrate
Subpellicular microtubule protein 1 (SPM1)	Vector	Vector
Sun-family protein	Generalist-Vector	Vector-Vertebrate
Tetratricopeptide repeat protein, putative	Vector	Vector
Thioredoxin	Generalist	Vertebrate
Translation initiation factor IF-2	Generalist-Vertebrate	Vertebrate
Tubulin	Vertebrate	Vertebrate
Ubiquitin conjugating enzyme 2	Vector	Vertebrate
Ubiquitin-conjugating enzyme	Vector	Vertebrate

Table 4-2. Variation of the number multigene family of paralogs across *Plasmodium* species.

Family	*P. vivax*	*P. cynomolgi*	*P. inui*	*P. knowlesi*	*P. coatneyi*	*P. gonderi*	*P. chabaudi*	*P. berghei*	*P. yoelii*	*P. falciparum*	*P. reichenowi*
Actin	2	2	2	2	2	1	2	2	2	2	2
Acyl-CoA synthase	2	2	1	2	7	4	1	1	1	1	1
Adrenodoxin reductase	4	1	1	1	1	1	1	1	1	1	1
Alpha-beta hydrolase 2	2	2	2	2	2	2	2	2	2	2	2
Asparagine tRNA ligase	2	2	2	2	2	2	2	2	2	2	2
ATP dependent DNA helicase	1	1	1	1	1	1	2	2	2	2	2
Biotin acel-CoA carboxylase	2	2	2	2	2	2	2	2	2	2	2
Calcium Dependent Protein Kinase (CDPK)	5	5	5	5	5	5	5	5	5	7	7
Calcium-transporting ATPase (SERCA)	2	2	2	2	2	2	2	2	2	2	2
Casein kinase II beta chain	4	4	4	4	4	4	4	4	4	5	5
Cell division protein FtsH	2	2	2	2	2	2	2	2	2	2	2
Chaperonin (CPN)	3	3	3	3	3	3	3	3	3	3	3
Chromatin assembly factor 1 subunit	3	2	2	2	1	2	3	2	2	5	6
ClpB protein	2	2	2	2	2	2	2	2	2	2	2
Conserved and hypothetical *Plasmodium* protein	2	2	1	2	1	1	1	1	1	1	1
Conserved *Plasmodium* protein	2	2	1	2	2	2	1	1	1	2	2
Conserved *Plasmodium* protein, unknown function	2	1	1	1	1	1	1	1	1	1	1
Conserved *Plasmodium* protein, unknown function 11	1	1	1	2	2	1	1	1	1	1	1
Conserved *Plasmodium* protein, unknown function 12	2	2	2	2	2*	2	2	2	2	2	2
Conserved *Plasmodium* protein, unknown function 2	1	1	1	3	1	1	1	1	1	1	1
Conserved *Plasmodium* protein, unknown function 3	2	1	1	1	1	1	1	1	1	1	1
Conserved *Plasmodium* protein, unknown function 4	1	1	1	2	2	1	1	1	1	1	1
Conserved *Plasmodium* protein, unknown function 6	1	1	1	2	1	1	1	1	1	1	1
Conserved *Plasmodium* protein, unknown function 8	2	2	2	2	2	2	2	2	2	2	2
Conserved Rodent malaria protein, unknown function 2	2	2	2	2	2	2	2	2	2	2	2
Cysteine repeat modular protein (CRMP)	2	2	2	2	2	1	2	2	2	2	2
Cytoadherence-linked asexual protein (CLAG)	2	2	2	2	2	2	2	2	2	2	2
DEAD-DEAH box ATP-dependent RNA helicase	2	2	2	2	2	2	2	2	2	2	2
DER1-like protein	3	3	3	3	3	3	3	3	3	3	3
DHHC-type zinc finger protein	2	2	2	2	2	1	2	2	2	2	2
Dipeptidyl aminopeptidase putative DPAP	2	2	2	2	2	2	2	2	2	2	2
DNA mismatch repair protein MSH2	2	2	2	2	2	2	2	2	2	2	2
DNA-directed RNA polymerase II	2	2	2	2	2	2	2	2	2	2	2
DNA-directed RNA polymerase II second largest subunit	1	1	1	1	1	1	1	1	1	2	2
DnaJ protein2	2	2	2	2	2	2	2	2	2	2	2
Dynamin-like protein	4	4	3	4	3	3	4	4	4	4	4
Dynein heavy chain	4*	2*	3*	4*	3*	3*	4	4	4	4*	4*
Elongation factor 1 alpha (EF-1alpha)	2	2	2	2	2	2	2	2	2	2	2
Elongation factor G	1	1	1	2	2	1	1	1	1	1	1
Eukaryotic initiation factor 2a	1	1	2	2	2	2	2	2	2	2	2
Exonuclease	2	2	2*	2	2*	2*	2	2	2	2	2
Folate transporter (FT)	2	2	2	2	2	2	2	2	2	2	2
Glutathione reductase (GR)	1	1	1	2	2	1	1	1	1	1	1
Glutathione synthetase (GS)	2	2	2	2	2	2	2	2	2	2	2
Glycerol-3-phosphate dehydrogenase	2	2	2	2	2	2	1	1	1	2	2
Heat shock protein 40	2	2	2	2	2	2	2	2	2	3	3
Heat shock protein 70	2	2	2	2	2	2	1	1	2	1	1
Heat shock protein 90	2	2	2	1	2	2	2	2	2	2	2
High mobility group protein B1 (HMGB1)	2	2	2	2	2	2	2	2	3	2	2

Table 4-2. Variation of the number multigene family of paralogs across *Plasmodium* species (continued).

Family	*P. vivax*	*P. cynomolgi*	*P. inui*	*P. knowlesi*	*P. coatneyi*	*P. gonderi*	*P. chabaudi*	*P. berghei*	*P. yoelii*	*P. falciparum*	*P. reichenowi*
Histidine tRNA ligase	2	2	2	2	2	2	2	2	2	2	2
Histone H2B	2	2	2	2	2	2	2	2	2	2	2
Histone H3	2	1	1	1	1	1	1	1	1	1	1
Hypothetical protein	3	1	1	1	1	1	1	1	1	1	1
Inner membrane complex protein 1c (IMC1c)	2	2	2	2	2	2	2	2	2	2	2
Inorganic pyrophosphatase (VP)	3	3	3	3	3	3	3	3	3	3	3
Iron-sulfur assembly protein (SufA)	2	2	2	2	2	2	2	2	2	2	2
Kinesin-8	2	2	2	2	2	2	4	2	2	11	13
Lysophospholipase	3	3	3	3	3	3	3	3	3	3	2
Malate dehydrogenase (MDH)	2	2	2	2	2	2	2	2	2	2	2
Meiotic recombination protein DMC	2	2	2	2	2	1	2	2	2	2	2
Meiotic recombination protein SPO11	2	2	2	2	2	2	2	2	2	2	2
Methionine aminopeptidase	2	2	2	2	2	2	2	2	2	2	1
Methyltransferase	12	13	7	5	5	9	3	3	3	9	9
MSP7	2	2	2	2	2	2	2	2	2	2	2
NADP-specific glutamate dehydrogenase (GDH)	4	4	4	4	4	4	4	2	4	4	4
NIMA related kinase (NEK)	1	1	1	1	1	1	2	1	1	1	1
Novel putative transporter 1 (NPT1)	2	2	2	2	2	2	2	2	2	2	2
Nucleotide binding protein	3	2	2	2	2	1	1	1	1	2	2
P1-s1 nuclease	3	1	1	1	1	1	1	1	1	1	1
P28	2	2	3	3	3	3	1	1	1	3	3
Papain	2	2	2	2	2	2	2	2	1	2	2
Peptide release factor	3	2	3	2	3	3	3	3	3	3	3
Peroxiredoxin thioredoxin peroxidase	2	2	2	2	2	1	2	2	2	2	2
Phosducin-like protein (PhLP)	1	1	1	1	1	1	1	1	1	2	2
Phosphopantothenoylcysteine synthetase	7	7	7	7	7	3	7	7	7	10	10
Plasmepsin	3	3	3	3	3	3	3	3	3	3	3
Pre mRNA splicing factor ATP dependent RNA helicase	2	2	2	2	2	2	2	2	2	2	2
Pre mRNA splicing helicase	2	2	2	2	2	2	2	2	2	2	2
Protein phosphatase 2C	2	2	2	2	2	4	2	1	1	1	2
Rhoptry associated protein 2-3	2	2	2	2	2	2	2	2	2	2	2
SEL-1 protein	14	14	12	8	10	10	5	5	5	9	8
Serine Repeat Antigen (SERA)	2	2	2	2	2	2	2	2	2	2	2
Serine tRNA ligase	2	1	1	1	1	1	1	1	1	1	1
Subpellicular microtubule protein 1 (SPM1)	2	2	2	2	2	2	2	2	2	2	2
Sun family protein	1	1	2	1	0	2	1	1	1	2	2
Tetratricopeptide repeat protein	2	2	2	2	2	2	2	2	2	2	2
Thioredoxin	2	1	2	2	2	2	2	2	2	2	2
Translation initiation factor IF-2	2	2	2	2	2	2	2	2	2	2	2
Tubulin	2	2	2	2	2	2	2	2	2	2	2
Ubiquitin-conjugating enzyme	2	2	2	2	2	1	2	2	2	1	1
Ubiquitin-conjugating enzyme 2	2	1	2	2	2	1	2	2	2	2	2

* Represents pseudogenes.

Table 4-3. Significant RELAX test results for branches under episodic selection.

Test for selection	
Intensification	**Relaxation**
Acyl-CoA synthase	Conserved *Plasmodium* protein
Alpha beta hydrolase putative 2	Cytoadherence-linked asexual protein (CLAG)
Chaperonin putative	NIMA related protein kinase (NEK)
Conserved *Plasmodium* protein unknown function 6	
Conserved *Plasmodium* protein unknown function 12	
Conserved rodent malaria protein unknown function	
DEAD DEAH box ATP dependent RNA helicase putative	
DER1 like protein	
Dipeptidyl amino peptidase putative (DPAP)	
Hypothetical protein	
P28	
Papain	
Plasmepsin	

Table 4-4. Summary of recombination events and median recombinant length per multigene family.

Multigene family	Number of recombination events	Median length of recombinant segment
Actin	4	140.5
Acyl-CoA synthase	8	405.5
Adrenoxin reductase SV	1	121
Alpha beta hydrolase 2	5	77
Asparagine tRNA ligase	4	122
Biotin acyl-CoA carboxylase	1	468
Calcium transporting ATPase putative SERCA	2	104
Calcium Dependent Protein Kinase (CDPK)	2	270.5
Chromatin assembly factor 1 subunit	2	113
Cytoadherence-linked asexual protein (CLAG)	3	60
ClpBprotein	1	370.5
Conserved *Plasmodium* protein unknown function	1	60
Conserved *Plasmodium* protein unknown function 2	2	131
Conserved *Plasmodium* protein unknown function 4	1	257
Conserved *Plasmodium* protein unknown function 6	3	60
Conserved Rodent malaria protein unknown function	1	754
Cysteine Repeat Modular Protein (CRMP)	4	176
DEAD DEAH box ATP dependent RNA helicase	3	34
DHHC type zinc finger protein	2	445
Dipeptidyl amino peptidase putative (DPAP)	2	117
DNA directed RNA polymerase II	2	42.5
Dynein heavy chain	2	303.5
Elongation factor Tu putative tufA	1	402
Eukaryotic initiation factor 2a	2	299
Exonuclease	1	109
Glutathione reductase putative GR	1	46
Heatshock protein 40	2	478.5
Heatshock protein 90	3	44
Histone H2B	1	84
Histone H3	2	54
Hypothetical protein	1	368
Inorganic pyrophosphatase VP	1	45
Iron sulfur assembly protein SufA	1	100
Kinesin 8	1	1240
Lysophospholipase	6	65.5

Table 4-4. Summary of recombination events and median recombinant length per multigene family (continued).

Multigene family	Number of recombination events	Median length of recombinant segment
Meiotic recombination protein DMC	1	39
NADP specific glutamate dehydrogenase putative GDH	4	148
NIMA related protein kinase (NEK)	2	115
Novel putative transporter 1 NPT1	2	604
Nucleotide binding protein	2	604
P1s1 nuclease	2	418
P28	2	288.5
Papain	7	190
Phosducin like protein PhLP	1	52
Plasmepsin	1	108
Pre mRNA splicing factor ATP dependent RNA helicase	1	274
Pre mRNA splicing helicase	3	121
Protein phosphatase 2C	1	26
Rhoptry associated protein 23	1	233
Serine Repeat Antigen (SERA)	18	446.5
Tetratricopeptide repeat protein putative	1	113
Tubulin	7	426
Ubiquitin conjugating enzyme 2	1	84

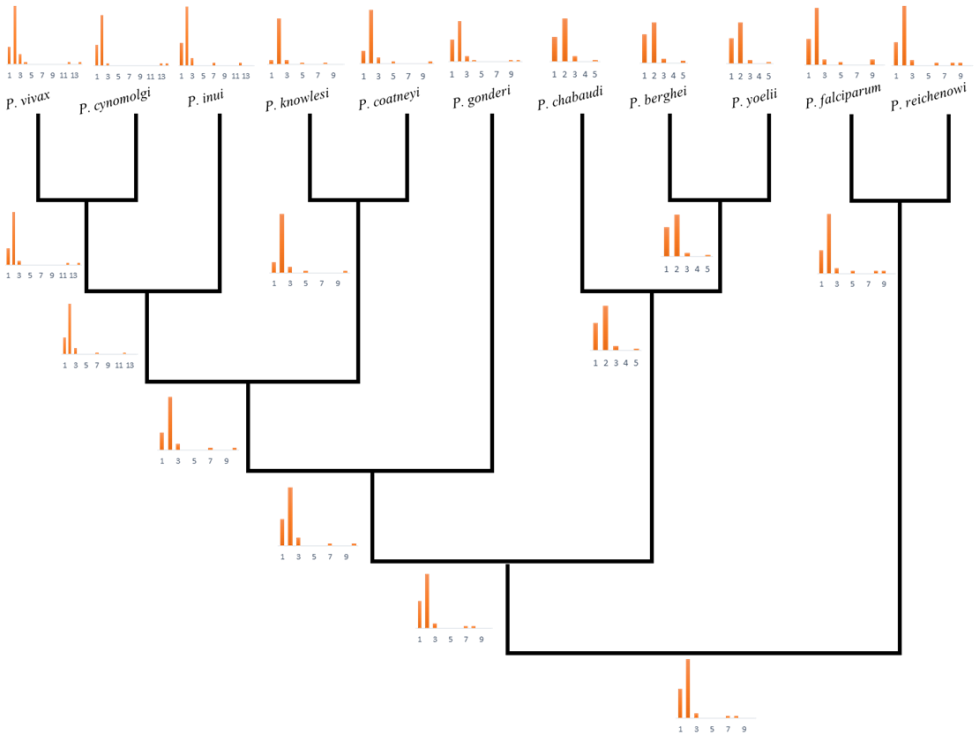
Table 4-5. Polymorphism and selection in paralogs with significant deviation from neutrality from larger *Plasmodium* GCMFs.

Family	Species	Gene ID*	N	π [SD]	Ds	Dn	Dn-Ds [SD]	Z test
Paralogs under significant purifying selection								
Acyl-CoA synthase	*P. falciparum*	PF3D7_1238800	151	0 [0]	0.001	0	0 [0]	0.024 (-2.284)
	P. vivax	PVX_002785	87	0.003 [0.001]	0.007	0.002	-0.005 [0.002]	0.005 (-2.861)
Cytoadherence-linked asexual protein (CLAG)	*P. falciparum*	PF3D7_0220800	55	0.004 [0.001]	0.012	0.002	-0.009 [0.002]	0 (-4.015)
		PF3D7_0302200	23	0.002 [0]	0.006	0.001	-0.005 [0.001]	0.001 (-3.533)
		PF3D7_0302500	47	0.002 [0]	0.003	0.001	-0.003 [0.001]	0.016 (-2.445)
	P. vivax	PVX_121885	35	0.005 [0.001]	0.009	0.003	-0.006 [0.002]	0.004 (-2.945)
NIMA related kinase (NEK)	*P. falciparum*	PF3D7_1228300	84	0 [0]	0.002	0	-0.002 [0.001]	0.002 (-3.195)
	P. vivax	PVX_079950	95	0 [0]	0.001	0	-0.001 [0]	0.049 (-1.986)
		PVX_124045	32	0.001 [0]	0.005	0	-0.004 [0.001]	0.002 (-3.129)
Papain	*P. falciparum*	PF3D7_1115300	104	0.004 [0.001]	0.009	0.003	-0.006 [0.003]	0.015 (-2.461)
		PF3D7_1115700	48	0.005 [0.001]	0.014	0.003	-0.011 [0.004]	0.011 (-2.574)
Plasmepsin	*P. falciparum*	PF3D7_0808200	191	0.001 [0]	0.003	0	-0.002 [0.001]	0.016 (-2.433)
		PF3D7_1408100	185	0 [0]	0	0	0 [0]	0.039 (-2.086)
		PF3D7_1465700	123	0 [0]	0.001	0	-0.001 [0]	0.042 (-2.060)
Serine repeat antigen (SERA)	*P. vivax*	PVX_003810	7	0.009 [0.001]	0.021	0.006	-0.015 [0.004]	0 (-3.706)
		PVX_003820	10	0.010 [0.001]	0.023	0.006	-0.017 [0.004]	0 (-5.106)
		PVX_003845	42	0.004 [0.001]	0.007	0.003	-0.004 [0.002]	0.019 (-2.378)
Paralogs under significant positive selection								
Acyl-CoA synthase	*P. falciparum*	PF3D7_0215300	157	0.001 [0]	0	0.002	0.001 [0]	0.001 (3.492)
Serine repeat antigen (SERA)	*P. falciparum*	PF3D7_0207400	84	0 [0]	0	0	0 [0]	0.025 (2.263)
		PF3D7_0207700	122	0.001 [0]	0	0.001	0.001 [0]	0.002 (3.106)

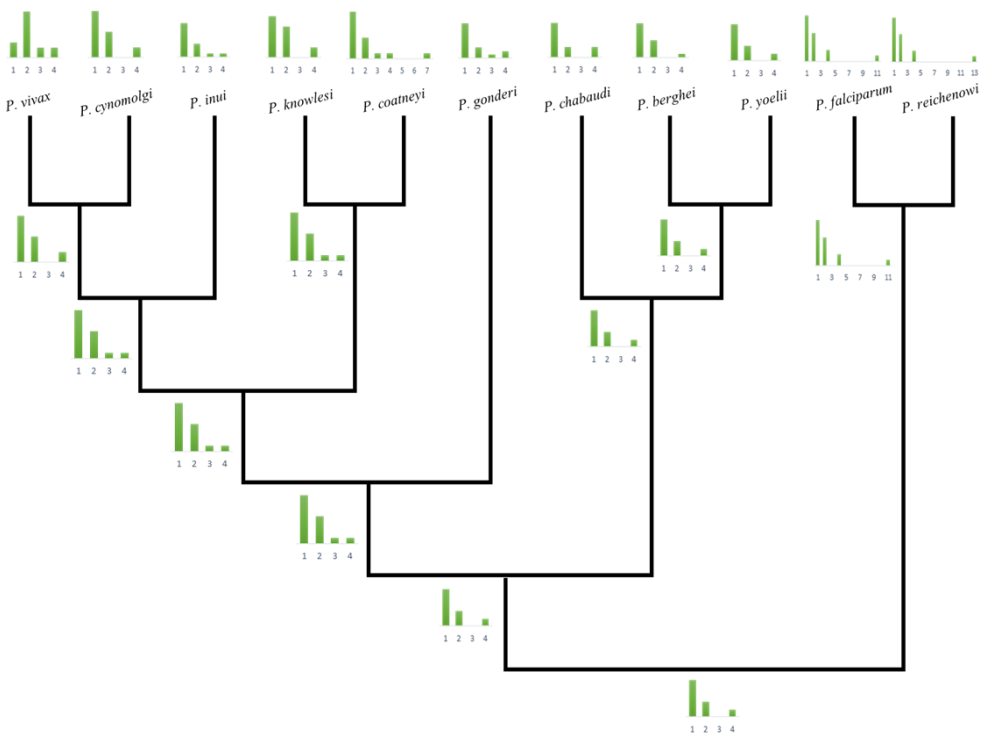
* *P. vivax* and *P. falciparum* PlasmoDB nomenclature (Mello et al. 2002; Carlton et al. 2008).

Figures

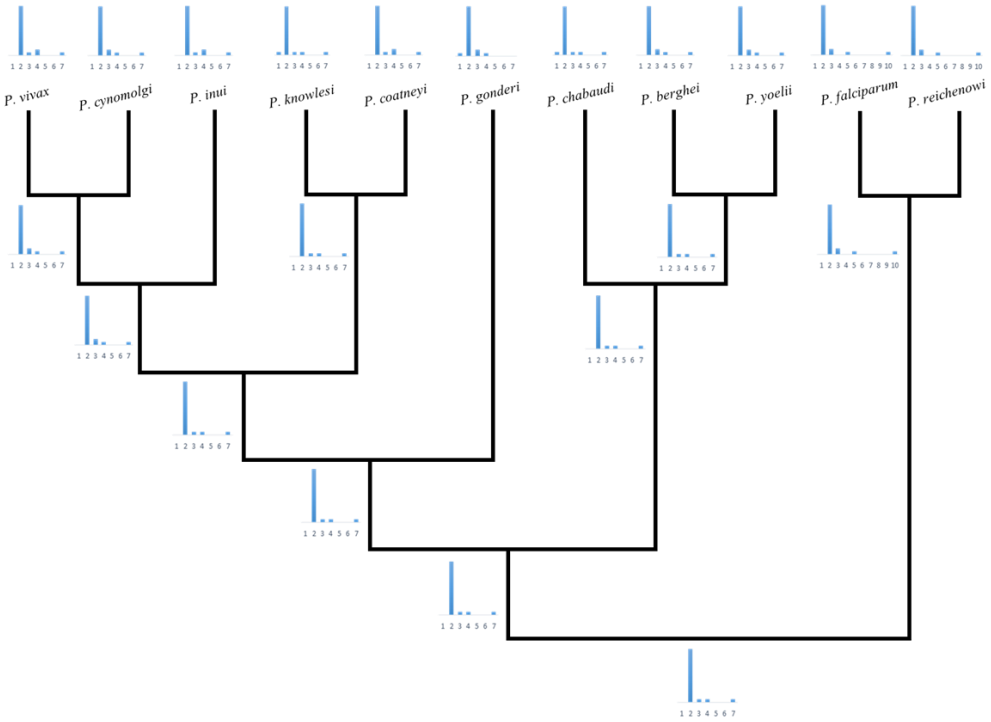
A


P. vivax
P. cynomolgi
P. inui
P. knowlesi
P. coatneyi
P. gonderi
P. chabaudi
P. berghei
P. yoelii
P. falciparum
P. reichenowi

B


P. vivax
P. cynomolgi
P. inui
P. knowlesi
P. coatneyi
P. gonderi
P. chabaudi
P. berghei
P. yoelii
P. falciparum
P. reichenowi

C


P. vivax
P. cynomolgi
P. inui
P. knowlesi
P. coatneyi
P. gonderi
P. chabaudi
P. berghei
P. yoelii
P. falciparum
P. reichenowi

D

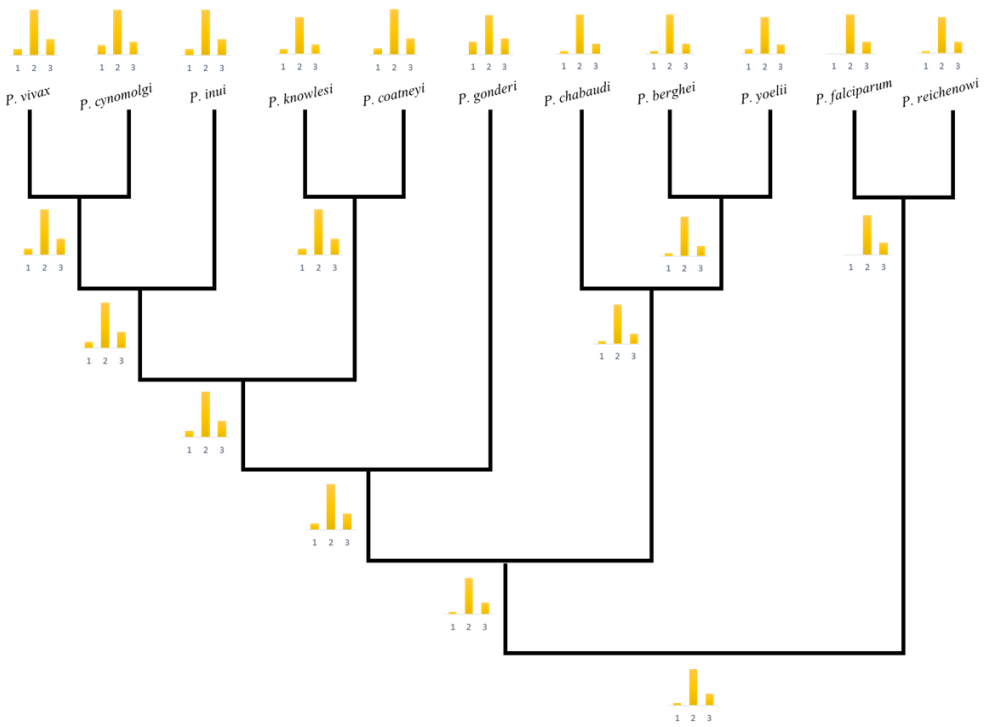
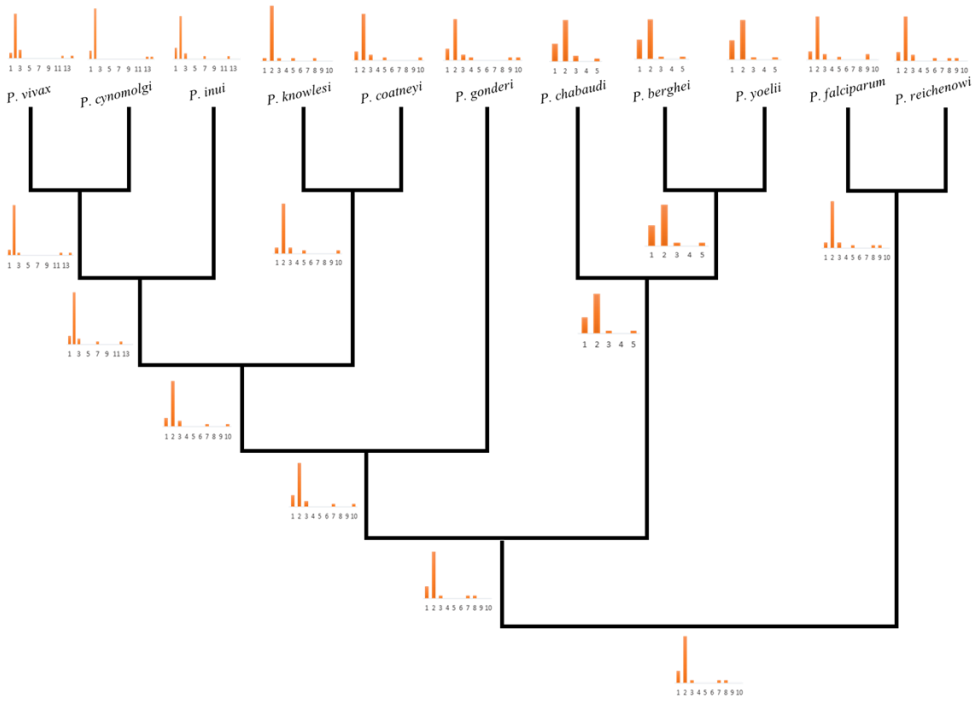
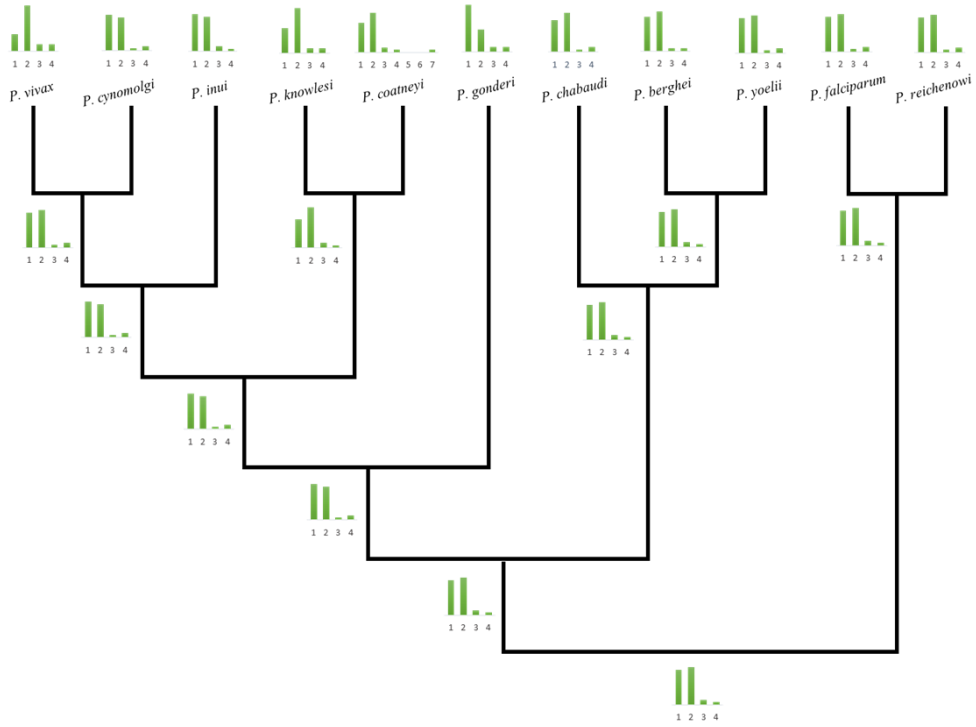


Figure 4-1. Number of paralogs vs. *P. berghei* expression patterns. Paralogs were classified as vector-specific, vertebrate-specific and generalist based on *P. berghei* transcriptome. Orange bars (**A**) indicate families composed exclusively of vertebrate-specific paralogs; green bars (**B**) indicate families composed exclusively of vector-specific paralogs; blue bars (**C**) indicate families composed of a combination of vector-specific and vertebrate-specific paralogs; yellow bars (**D** indicate families composed of generalist paralogs. The distribution of families with a given number of paralogs in each extant *Plasmodium* species is indicated by the height of the colored bars.

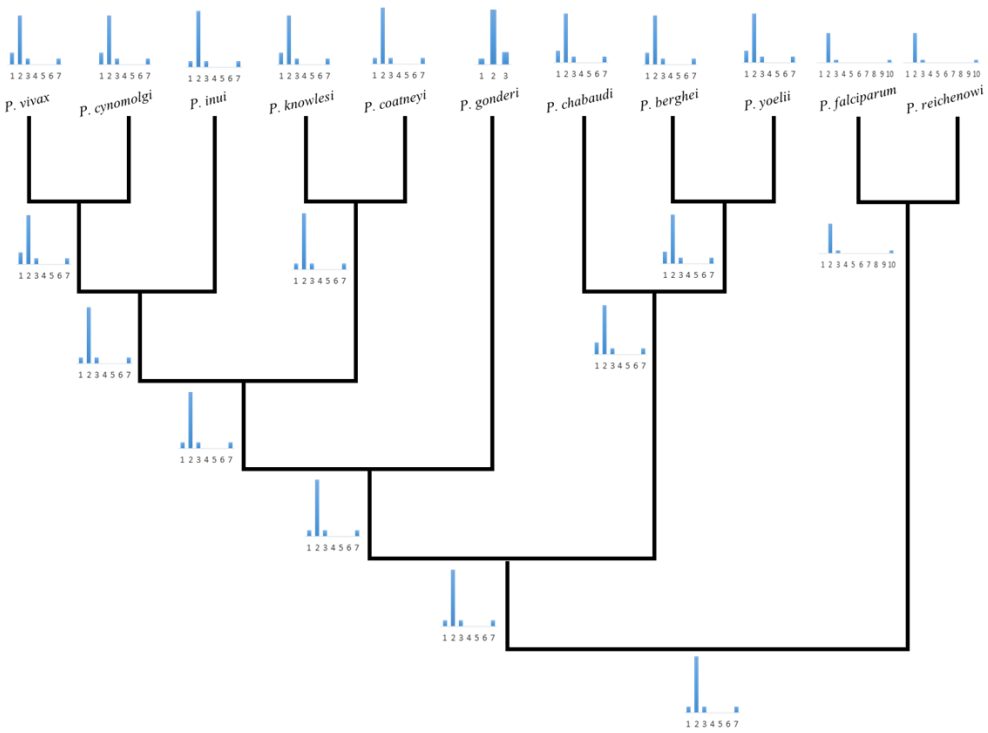
A


P. vivax
P. cynomolgi
P. inui
P. knowlesi
P. coatneyi
P. gonderi
P. chabaudi
P. berghei
P. yoelii
P. falciparum
P. reichenowi

B


P. vivax
P. cynomolgi
P. inui
P. knowlesi
P. coatneyi
P. gonderi
P. chabaudi
P. berghei
P. yoelii
P. falciparum
P. reichenowi

C


P. vivax
P. cynomolgi
P. inui
P. knowlesi
P. coatneyi
P. gonderi
P. chabaudi
P. berghei
P. yoelii
P. falciparum
P. reichenowi

D

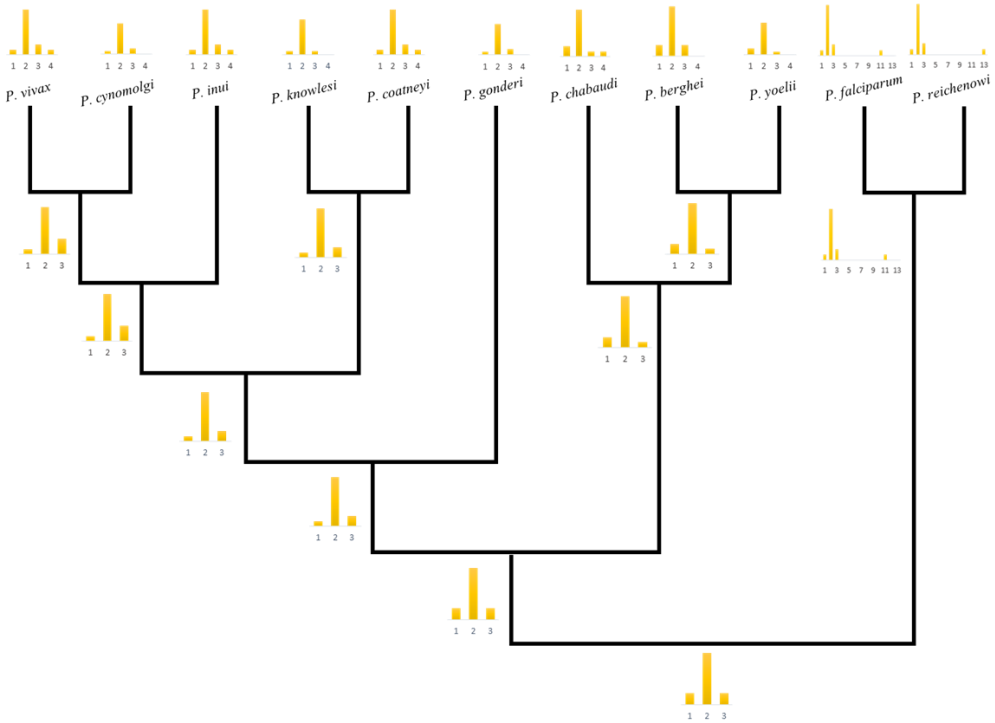

P. vivax
P. cynomolgi
P. inui
P. knowlesi
P. coatneyi
P. gonderi
P. chabaudi
P. berghei
P. yoelii
P. falciparum
P. reichenowi

Figure 4-2. . Number of paralogs vs. *P. falciparum* expression patterns. Paralogs were classified as vector-specific, vertebrate-specific and generalist based on *P. falciparum* transcriptome Orange bars (**A**) indicate families composed exclusively of vertebrate-specific paralogs; green bars (**B**) indicate families composed exclusively of vector-specific paralogs; blue bars (**C**) indicate families composed of a combination of vector-specific and vertebrate-specific paralogs; yellow bars (**D** indicate families composed of generalist paralogs. The distribution of families with a given number of paralogs in each extant *Plasmodium* species is indicated by the height of the colored bars.

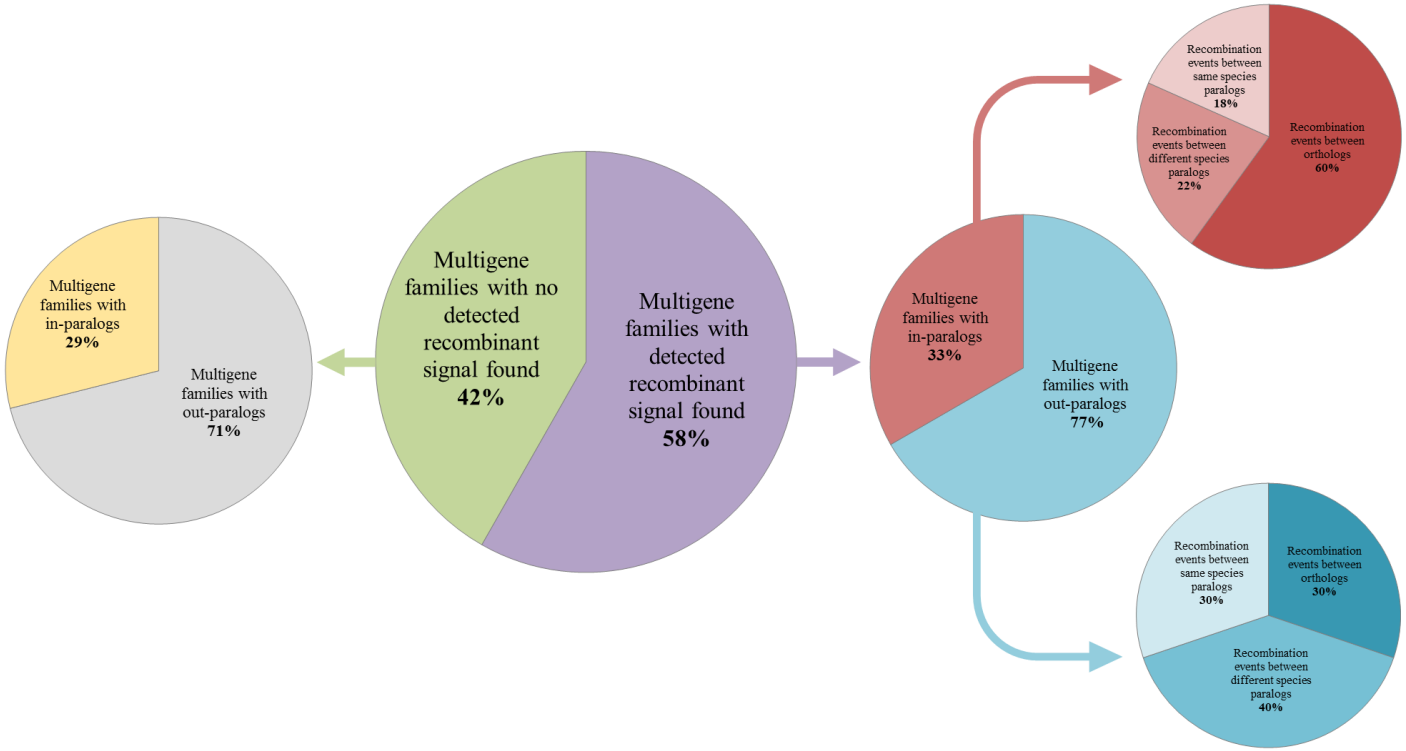


Figure 4-3. Recombination patters in GCMFs. The largest pie chart (green and purple) represents the total number of multigene families with informative multisequence alignments, families are split into those with significant recombination events and those without them. Smaller pie charts further divide groups with (red and blue) and without

(yellow and grey) recombination, on the basis of the evolutionary relationship among family members. Groups are further divided in the smallest pie charts to indicate the frequency of recombination events among orthologs (dark red and dark blue), same species paralogs (light red and light blue), and different species paralogs (medium red and medium blue).

CHAPTER 5. CONCLUSIONS.

The evolution of the merozoite surface family 7 (*msp7*) multigene family appears to be characterized by two independent expansion events in the simian clade and *Laveranian* subgenus. Furthermore, the lineage-dependent gain/loss events across closely related simian species show that the evolution of the *msp7* multigene family is consistent with the Birth-and-Death model. Within this clade, duplicated genes could have potentially originated in the simian ancestor and subsequently diversified by positive selection. The predominant signs of intra-specific purifying selection among *msp7* paralogs from the simian clade suggest that duplicated copies are preserved and are likely important for family function. On the other hand, the number of High Activity Binding Peptides (HABPs), identified for one *Pfmsp7* member of the multigene family (PF3D7_1353100), varied across *Laveranian* paralogs, with some paralogs showing a reduced number of HABPs. This suggest that members of the *msp7* family in the *Laveranian* subgenus might have differential significance or even variable roles during erythrocyte invasion. However, additional studies should be performed in order to determine if unique HABPs have independently evolved on other members of the *msp7* family.

Alternatively, long-term evolution of synonymous and non-synonymous substitutions shows a species-specific pattern among genes with a gametocyte biased expression. Overall, genes with male-biased, female-biased and male-female common expression showed similar rates of synonymous and non-synonymous substitutions with only a few genes showing significantly longer synonymous branch lengths. MLEs of

synonymous branch lengths were highly similar among genes coding for membrane and non-membrane located proteins; nonetheless, genes coding for membrane proteins showed slightly longer non-synonymous branch lengths. This is likely the result of selection imposed by the vertebrate's immune system.

On the other hand, genes with a male-biased, male-female common biased expression, and genes coding for non-membrane located proteins showed overall higher immunogenicity than other sex and location categories evaluated. Additionally, these categories also had a higher proportion of genes with specific sites of the coding sequence under significant positive selection. This indicates that the presence of immune epitopes might act as an indicator of rapid evolution in comparison to genes with no reported immune epitopes. Nonetheless, the similar proportion of positive selected sites inside and outside epitope regions suggests that these trends are not entirely related to immune pressure imposed by the host. An alternative hypothesis to be evaluated is the possibility that positive selected sites reflect the effects of interaction with the *Anopheles* vector.

To conclude, the majority of duplication events found in GCMFs likely predate speciation of mammalian *Plasmodium* species. It is possible that duplicated genes have been maintained neutrally across the genus *Plasmodium*; however, the reduced numbers of branches showing inter- and/or intra-specific signs of positive selection could indicate that the preservation of duplicated genes across the genus is beneficial for the parasite. Exploring the putative relationship between such preservation and multigene family function could help to better explain the role of multigene families in the evolution of *Plasmodium* genome. Contrary to expectations, multigene family size and composition

did not vary in a manner influenced by interaction with the mosquito vector or the vertebrate host. However, given the lineage-dependent variation previously discussed in *msp7* and gametocyte expressed single copy genes, it is unlikely that expression patterns will be universally maintained across the genus. Thus, this result should be re-evaluated in the future when additional transcriptomic data becomes available for additional *Plasmodium* species. Alternatively, inter-genic recombination appears to be a fundamental force in the development of sequence diversity in GCMFs as well as one of the potential mechanisms in duplication/loss events across the genus *Plasmodium*.

Overall, the present studies allowed a further exploration of evolutionary trends within the genus *Plasmodium*. A better understanding of the role that biological interactions, inter-, intra-specific selection, and recombination have as mechanisms in the development of functional novelty and adaptation within the genus has been obtained. Moreover, the present study also serves as a starting point for the evaluation of additional malaria treatment and prevention strategies, as well as a more detailed exploration of potential patterns involved in the evolution of the *Apicomplexan* genomes.

References

Agnandji ST, Lell B, Soulanoudjingar SS, Fernandes JF, Abossolo BP, Conzelmann C, Methogo BG, Doucka Y, Flamen A, Mordmüller B, Issifou S, Kremsner PG, Sacarlal J, Aide P, Lanaspa M, Aponte JJ, Nhamuave A, Quelhas D, Bassat Q, Mandjate S, Macete E, Alonso P, Abdulla S, Salim N, Juma O, Shomari M, Shubis K, Machera F, Hamad AS, Minja R, Mtoro A, Sykes A, Ahmed S, Urassa AM, Ali AM, Mwangoka G, Tanner M, Tinto H, D'Alessandro U, Sorgho H, Valea I, Tahita MC, Kaboré W, Ouédraogo S, Sandrine Y, Guiguemdé RT, Ouédraogo JB, Hamel MJ, Kariuki S, Odero C, Oneko M, Otieno K, Awino N, Omoto J, Williamson J, Muturi Kioi V, Laserson KF, Slutsker L, Otieno W, Otieno L, Nekoye O, Gondi S, Otieno A, Ogutu B, Wasuna R, Owira V, Jones D, Onyango AA, Njuguna P, Chilengi R, Akoo P, Kerubo C, Gitaka J, Maingi C, Lang T, Olotu A, Tsofa B, Bejon P, Peshu N, Marsh K, Owusu-Agyei S, Asante KP, Osei-Kwakye K, Boahen O, Ayamba S, Kayan K, Owusu-Ofori R, Dosoo D, Asante I, Adjei G, Adjei G, Chandramohan D, Greenwood B, Lusingu J, Gesase S, Malabeja A, Abdul O, Kilavo H, Mahende C, Liheluka E, Lemnge M, Theander T, Drakeley C, Ansong D, Agbenyega T, Adjei S, Boateng HO, Rettig T, Bawa J, Sylverken J, Sambian D, Agyekum A, Owusu L, Martinson F, Hoffman I, Mvalo T, Kamthunzi P, Nkomo R, Msika A, Jumbe A, Chome N, Nyakuipa D, Chintedza J, Ballou WR, Bruls M, Cohen J, Guerra Y, Jongert E, Lapierre D, Leach A, Lievens M, Ofori-Anyinam O, Vekemans J, Carter T, Leboulleux D, Loucq C, Radford A, Savarese B, Schellenberg D, Sillman M, Vansadia P; RTS,S Clinical Trials Partnership. 2011. A Phase 3 Trial of RTS,S/AS01 Malaria Vaccine in African Infants. N. Engl. J. Med. 367, 2284–2295. doi:10.1056/NEJMoa1208394.

Abdi, A.I., Warimwe, G.M., Muthui, M.K., Kivisi, C.A., Kiragu, E.W., Fegan, G.W., Bull, P.C., 2016. Global selection of *Plasmodium falciparum* virulence antigen expression by host antibodies. Sci. Rep. 6, 19882. doi:10.1038/srep19882

Alano, P., 2007. *Plasmodium falciparum* gametocytes: still many secrets of a hidden life. Mol. Microbiol. 66, 291–302. doi:10.1111/j.1365-2958.2007.05904.x

Alexandre, J.S., Kaewthamasorn, M., Yahata, K., Nakazawa, S., Kaneko, O., 2011. Positive selection on the *Plasmodium falciparum* clag2 gene encoding a component of the erythrocyte-binding rhoptry protein complex. Trop. Med. Health 39, 77–82. doi:10.2149/tmh.2011-12

Alonso, P.L., Brown, G., Arevalo-Herrera, M., Binka, F., Chitnis, C., Collins, F., Doumbo, O.K., Greenwood, B., Hall, B.F., Levine, M.M., Mendis, K., Newman, R.D., Plowe, C.V., Rodríguez, M.H., Sinden, R., Slutsker, L., Tanner, M., 2011. A Research Agenda to Underpin Malaria Eradication. PLoS Med. 8, e1000406. doi:10.1371/journal.pmed.1000406

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-33402.

Anthony, T.G., Polley, S.D., Vogler, A.P., Conway, D.J., 2007. Evidence Of Non-Neutral Polymorphism In *Plasmodium falciparum* Gamete Surface Protein Genes Pfs47 And Pfs48/45. Mol. Biochem. Parasitol. 156, 117–123. doi:10.1016/j.molbiopara.2007.07.008

Armistead, J.S., Morlais, I., Mathias, D.K., Jardim, J.G., Joy, J., Fridman, A., Finnefrock, A.C., Bagchi, A., Plebanski, M., Scorpio, D.G., Churcher, T.S., Borg, N.A., Sattabongkot, J., Dinglasan, R.R., Adams, J.H., 2014. Antibodies to a Single, Conserved Epitope in Anopheles APN1 Inhibit Universal Transmission of *Plasmodium falciparum* and *Plasmodium vivax*. Malaria. Infect. Immun. 82, 818–829. doi:10.1128/IAI.01222-13

Arévalo-Herrera, M., Lopez-Perez, M., Dotsey, E., Jain, A., Rubiano, K., Felgner, P.L., Davies, D.H., Herrera, S., 2016. Antibody profiling in naïve and semi-immune individuals experimentally challenged with *Plasmodium vivax* sporozoites. PLoS Negl. Trop. Dis. 10(3), e0004563.

Arisue, N., Hirai, M., Arai, M., Matsuoka, H., Horii, T., 2007. Phylogeny and evolution of the SERA multigene family in the genus *Plasmodium*. J. Mol. Evol. 65, 82-91.

Arisue, N., Kawai, S., Hirai, M., Palacpac, N.M.Q., Jia, M., Kaneko, A., Tanabe, K., Horii, T., 2011. Clues to Evolution of the SERA Multigene Family in 18 *Plasmodium* Species. PLoS ONE 6, e17775. doi:10.1371/journal.pone.0017775

Aoki, S., Li, J., Itagaki, S., Okech, B.A., Egwang, T.G., Matsuoka, H., Palacpac, N.M.Q., Mitamura, T., Horii, T., 2002. Serine Repeat Antigen (SERA5) Is Predominantly Expressed among the SERA Multigene Family of *Plasmodium falciparum*, and the Acquired Antibody Titers Correlate with Serum Inhibition of the Parasite Growth. J. Biol. Chem. 277, 47533–47540. doi:10.1074/jbc.M207145200

Assefa, S., Lim, C., Preston, M.D., Duffy, C.W., Nair, M.B., Adroub, S.A., Kadir, K.A., Goldberg, J.M., Neafsey, D.E., Divis, P., Clark, T.G., Duraisingh, M.T., Conway, D.J., Pain, A., Singh, B., 2015. Population genomic structure and adaptation in the zoonotic malaria parasite *Plasmodium knowlesi*. Proc. Natl. Acad. Sci. 112, 13027–13032. doi:10.1073/pnas.1509534112

Aurrecoechea, C., Brestelli, J., Brunk, B.P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J.C., Kraemer, E., Li, W., Miller, J.A., Nayak, V., Pennington, C., Pinney, D.F., Roos, D.S., Ross, C., Stoeckert, C.J. Jr., Treatman, C., Wang, H., 2009. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 37, D539-D543.

Battistuzzi, F.U., Schneider, K.A., Spencer, M.K., Fisher, D., Chaudhry, S., Escalante, A.A., 2016. Profiles of low complexity regions in *Apicomplexa*. BMC Evol. Biol. 29, 16:47.

Benson, D.A., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2014. GenBank. Nucleic Acids Res. 42, D32-D37.

Beeson, J.G., Drew, D.R., Boyle, M.J., Feng, G., Fowkes, F.J.I., Richards, J.S., 2016. Merozoite surface proteins in red blood cell invasion, immunity and vaccines against malaria. FEMS Microbiol. Rev. fuw001. doi:10.1093/femsre/fuw001

Bethke, L.L., Zilversmit, M., Nielsen, K., Daily, J., Volkman, S.K., Ndiaye, D., Lozovsky, E.R., Hartl, D.L., Wirth, D.F., 2006. Duplication, gene conversion, and genetic diversity in the species-specific acyl-CoA synthetase gene family of *Plasmodium falciparum*. Mol. Biochem. Parasitol. 150, 10–24. doi:10.1016/j.molbiopara.2006.06.004

Brendel V , Bucher P , Nourbakhsh IR , Blaisdell BE , Karlin S. 1992. Methods and algorithms for statistical analysis of protein sequences. Proc. Natl. Acad. Sci. USA. 89, 2002-6

Bethke, L.L., Zilversmit, M., Nielsen, K., Daily, J., Volkman, S.K., Ndiaye, D., Lozovsky, E.R., Hartl, D.L., Wirth, D.F., 2006. Duplication, gene conversion, and genetic diversity in the species-specific acyl-CoA synthetase gene family of *Plasmodium falciparum*. Mol. Biochem. Parasitol. 150, 10-24.

Bousema, T., Drakeley, C., 2011. Epidemiology and Infectivity of *Plasmodium falciparum* and *Plasmodium vivax* Gametocytes in Relation to Malaria Control and Elimination. Clin. Microbiol. Rev. 24, 377–410. doi:10.1128/CMR.00051-10

Boyle, M.J., Langer, C., Chan, J.A., Hodder, A.N., Coppel, R.L., Anders, R.F., Beeson, J.G., 2014. Sequential processing of merozoite surface proteins during and after erythrocyte invasion by *Plasmodium falciparum*. Infect. Immun. 82, 924-936.

Carlton, J.M., Escalante, A.A., Neafsey, D., Volkman, S.K., 2008. Comparative evolutionary genomics of human malaria parasites. Trends Parasitol. 24, 545-550.

Carlton, J.M., Das, A., Escalante, A.A., 2013. Genomics, Population Genetics and Evolutionary History of *Plasmodium vivax*. Advances in Parasitology. Elsevier, pp. 203–222.

Carter, L.M., Schneider, P., Reece, S.E., 2014. Information use and plasticity in the reproductive decisions of malaria parasites. Malar. J 13, 115.

Carter, R., 2001. Transmission blocking malaria vaccines. Vaccine 19, 2309–2314.

Cheeseman, I.H., Gomez-Escobar, N., Carret, C.K., Ivens, A., Stewart, L.B., Tetteh, K.K.,

Conway, D.J., 2009. Gene copy number variation throughout the *Plasmodium falciparum* genome. BMC Genomics 10, 353. doi:10.1186/1471-2164-10-353

Chen, N., LaCrue, A.N., Teuscher, F., Waters, N.C., Gatton, M.L., Kyle, D.E., Cheng, Q., 2014. Fatty Acid Synthesis and Pyruvate Metabolism Pathways Remain Active in Dihydroartemisinin-Induced Dormant Ring Stages of *Plasmodium falciparum*. Antimicrob. Agents Chemother. 58, 4773–4781. doi:10.1128/AAC.02647-14

Chenet, S.M., Pacheco, M.A., Bacon, D.J., Collins, W.E., Barnwell, J.W., Escalante, A.A., 2013. 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. 20, 239-248.

Claessens, A., Hamilton, W.L., Kekre, M., Otto, T.D., Faizullabhoy, A., Rayner, J.C., Kwiatkowski, D., 2014. Generation of Antigenic Diversity in *Plasmodium falciparum* by Structured Rearrangement of Var Genes During Mitosis. PLoS Genet. 10, e1004812. doi:10.1371/journal.pgen.1004812

Coatney, R.G., Collins, W.E., Warren, M., Contacos, P.G., 1971. The Primate Malarias. US Government Printing Office, Washington.

Cormier, L.A., 2011. New frontiers in historical ecology: ten-thousand year fever: rethinking human and wild-primate malarias. Left Coast Press, Walnut Creek.

Crompton, P.D., Moebius, J., Portugal, S., Waisberg, M., Hart, G., Garver, L.S., Miller, L.H., Barillas-Mury, C., Pierce, S.K., 2014. Malaria Immunity in Man and Mosquito: Insights into Unsolved Mysteries of a Deadly Infectious Disease. Annu. Rev. Immunol. 32, 157–187. doi:10.1146/annurev-immunol-032713-120220

Csuos, M., 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26, 1910–1912. doi:10.1093/bioinformatics/btq315

Cunningham, D., Lawton, J., Jarra, W., Preiser, P., Langhorne, J., 2010. The pir multigene family of *Plasmodium*: Antigenic variation and beyond. Mol. Biochem. Parasitol. 170, 65–73. doi:10.1016/j.molbiopara.2009.12.010

Da, D.F., Dixit, S., Sattabonkot, J., Mu, J., Abate, L., Ramineni, B., Ouedraogo, J.B., MacDonald, N.J., Fay, M.P., Su, X. -z., Cohuet, A., Wu, Y., 2013. Anti-Pfs25 Human Plasma Reduces Transmission of *Plasmodium falciparum* Isolates That Have Diverse Genetic Backgrounds. Infect. Immun. 81, 1984–1989. doi:10.1128/IAI.00016-13

DeBarry, J.D., Kissinger, J.C., 2011. Jumbled Genomes: Missing *Apicomplexan* Synteny. Mol. Biol. Evol. 28, 2855–2871. doi:10.1093/molbev/msr103

Dembélé, L., Franetich, J.-F., Lorthiois, A., Gego, A., Zeeman, A.-M., Kocken, C.H.M., Le Grand, R., Dereuddre-Bosquet, N., van Gemert, G.-J., Sauerwein, R., Vaillant, J.-C., Hannoun, L., Fuchter, M.J., Diagana, T.T., Malmquist, N.A., Scherf, A., Snounou, G., Mazier, D., 2014. Persistence and activation of malaria hypnozoites in long-term primary hepatocyte cultures. Nat. Med. 20, 307–312. doi:10.1038/nm.3461

Delves, M.J., Ruecker, A., Straschil, U., Lelievre, J., Marques, S., Lopez-Barragan, M.J., Herreros, E., Sinden, R.E., 2013. Male and Female *Plasmodium falciparum* Mature Gametocytes Show Different Responses to Antimalarial Drugs. Antimicrob. Agents Chemother. 57, 3268–3274. doi:10.1128/AAC.00325-13

Deroost, K., Pham, T.-T., Opdenakker, G., Van den Steen, P.E., 2016. The immunological balance between host and parasite in malaria. FEMS Microbiol. Rev. 40, 208–257. doi:10.1093/femsre/fuv046

Doi, M., Tanabe, K., Tachibana, S.-I., Hamai, M., Tachibana, M., Mita, T., Yagi, M., Zeyrek, F.Y., Ferreira, M.U., Ohmae, H., Kaneko, A., Randrianarivelojosia, M., Sattabongkot, J., Cao, Y.-M., Horii, T., Torii, M., Tsuboi, T., 2011. Worldwide sequence conservation of transmission-blocking vaccine candidate Pvs230 in *Plasmodium vivax*. Vaccine 29, 4308–4315. doi:10.1016/j.vaccine.2011.04.028

Dorin-Semblat, D., Schmitt, S., Semblat, J.-P., Sicard, A., Reininger, L., Goldring, D., Patterson, S., Quashie, N., Chakrabarti, D., Meijer, L., Doerig, C., 2011. *Plasmodium falciparum* NIMA-related kinase Pfnek-1: sex specificity and assessment of essentiality for the erythrocytic asexual cycle. Microbiology 157, 2785–2794. doi:10.1099/mic.0.049023-0

Duval, L., Ariey, F., 2012. Ape *Plasmodium* parasites as a source of human outbreaks. Clin. Microbiol. Infect. 18, 528–532. doi:10.1111/j.1469-0691.2012.03825.x

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. doi:10.1093/nar/gkh340

Escalante, A.A., Cornejo, O.E., Freeland, D.E., Poe, A.C., Durrego, E., Collins, W.E., Lal, A.A., 2005. A monkey's tale: the origin of *Plasmodium vivax* as a human malaria parasite. Proc. Natl. Acad. Sci. U. S. A. 102, 1980–1985.

Farris, J.S., 1970. Methods for Computing Wagner Trees. Syst. Zool. 19, 83. doi:10.2307/2412028

Fairhurst, R.M., 2015. Understanding artemisinin-resistant malaria: what a difference a year makes. Curr. Opin. Infect. Dis. 28, 417–425. doi:10.1097/QCO.0000000000000199

Fawcett, J., Innan, H., 2011. Neutral and Non-Neutral Evolution of Duplicated Genes with Gene Conversion. Genes 2, 191–209. doi:10.3390/genes2010191

Feng, H., Gupta, B., Wang, M., Zheng, W., Zheng, L., Zhu, X., Yang, Y., Fang, Q., Luo, E., Fan, Q., Tsuboi, T., Cao, Y., Cui, L., 2015. Genetic diversity of transmission-blocking vaccine candidate Pvs48/45 in *Plasmodium vivax* populations in China. Parasit. Vectors 8. doi:10.1186/s13071-015-1232-4

Frech, C., Chen, N., 2013. Variant surface antigens of malaria parasites: functional and evolutionary insights from comparative gene family classification and analysis. BMC Genomics 14, 1.

Frech, C., Chen, N., 2011. Genome Comparison of Human and Non-Human Malaria Parasites Reveals Species Subset-Specific Genes Potentially Linked to Human Disease. PLoS Comput. Biol. 7, e1002320. doi:10.1371/journal.pcbi.1002320

Florens, L., Washburn, M.P., Raine, J.D., Anthony, R.M., Grainger, M., Haynes, J.D., Moch, J.K., Muster, N., Sacci, J.B., Tabb, D.L., others, 2002. A proteomic view of the *Plasmodium falciparum* life cycle. Nature 419, 520–526.

García, Y., Puentes, A., Curtidor, H., Cifuentes, G., Reyes, C., Barreto, J., Moreno, A., Patarroyo, M.E., 2007. 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. 50, 5665-5675.

Garnham, P.C.C., 1966. Malaria parasites and other haemosporidia. Blackwell Scientific Publications, Oxford.

Garzón-Ospina, D., Cadavid, L.F., Patarroyo, M.A., 2010. Differential expansion of the merozoite surface protein (msp)-7 gene family in *Plasmodium* species under a birth-and-death model of evolution. Mol. Phylogenet. Evol. 55, 399–408. doi:10.1016/j.ympev.2010.02.017

Gholizadeh, S., Djadid, N., Basseri, H., Zakeri, S., Ladoni, H., 2009. Analysis of von Willebrand factor A domain-related protein (WARP) polymorphism in temperate and tropical *Plasmodium vivax* field isolates. Malar. J. 8, 137. doi:10.1186/1475-2875-8-137

Girard, M., Reed, Z., Friede, M., Kieny, M., 2007. A review of human vaccine research and development: Malaria. Vaccine 25, 1567–1580. doi:10.1016/j.vaccine.2006.09.074

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol. Biol. Evol. 27, 221–224. doi:10.1093/molbev/msp259

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307-321.

Gupta, A., Thiruvengadam, G., Desai, S.A., 2015. The conserved clag multigene family of malaria parasites: Essential roles in host–pathogen interaction. Drug Resist. Updat. 18, 47–54. doi:10.1016/j.drup.2014.10.004

Habtewold, T., Povelones, M., Blagborough, A.M., Christophides, G.K., 2008. Transmission Blocking Immunity in the Malaria Non-Vector Mosquito *Anopheles quadriannulatus* Species A. PLoS Pathog. 4, e1000070. doi:10.1371/journal.ppat.1000070

Hastings, I.M., Kay, K., Hodel, E.M., 2015. How Robust Are Malaria Parasite Clearance Rates as Indicators of Drug Effectiveness and Resistance? Antimicrob. Agents Chemother. 59, 6428–6436. doi:10.1128/AAC.00481-15

Hakes, L., Pinney, J.W., Lovell, S.C., Oliver, S.G., Robertson, D.L., 2007. All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol 8, R209. Hostetler, J.B., Sharma, S., Bartholdson, S.J., Wright, G.J., Fairhurst, R.M., Rayner, J.C., 2015. A library of *Plasmodium vivax* recombinant merozoite proteins reveals new vaccine candidates and protein-protein interactions. PLoS Negl. Trop. Dis. 9(12), e0004264.

Innan, H., Kondrashov, F., 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. 11, 4. doi:10.1038/nrg2689

Iriko, H., Kaneko, O., Otsuki, H., Tsuboi, T., Su, X., Tanabe, K., Torii, M., 2008. Diversity and evolution of the rhoph1/clag multigene family of *Plasmodium falciparum.* Mol. Biochem. Parasitol. 158, 11–21.

Iyer, J.K., Fuller, K., Preiser, P.R., 2006. Differences in the copy number of the py235 gene family in virulent and avirulent lines of *Plasmodium yoelii*. Mol. Biochem. Parasitol. 150, 186–191. doi:10.1016/j.molbiopara.2006.07.012

Janssen, C.S., 2004. *Plasmodium* interspersed repeats: the major multigene superfamily of malaria parasites. Nucleic Acids Res. 32, 5712–5720. doi:10.1093/nar/gkh907

Josling, G.A., Llinás, M., 2015. Sexual development in *Plasmodium* parasites: knowing when it's time to commit. Nat. Rev. Microbiol. 13, 573–587. doi:10.1038/nrmicro3519

Joy, D.A., Gonzalez-Ceron, L., Carlton, J.M., Gueye, A., Fay, M., McCutchan, T.F., Su, X. -z., 2008. Local Adaptation and Vector-Mediated Population Structure in *Plasmodium vivax* Malaria. Mol. Biol. Evol. 25, 1245–1252. doi:10.1093/molbev/msn073

Kadekoppala, M., Holder, A., 2010. Merozoite surface proteins of the malaria parasite: the MSP1 complex and the MSP7 family. Int. J. for Parasitol. 40, 1155-1161.

Kadekoppala, M., O'Donnell, R.A., Grainger, M., Crabb, B.S., Holder, A.A., 2008. Deletion of the *Plasmodium falciparum* merozoite surface protein 7 gene impairs parasite invasion of erythrocytes. Eukaryot. Cell 7, 2123-2132.

Kadekoppala, M., Ogun, S.A., Howell, S., Gunaratne, R.S., Holder, A.A., 2010. 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 9, 1064-1074.

Kamali, M., Marek, P.E., Peery, A., Antonio-Nkondjio, C., Ndo, C., Tu, Z., Simard, F., Sharakhov, I.V., 2014. Multigene Phylogenetics Reveals Temporal Diversification of Major African Malaria Vectors. PLoS ONE 9, e93580. doi:10.1371/journal.pone.0093580

Kamali, M., Xia, A., Tu, Z., Sharakhov, I.V., 2012. A New Chromosomal Phylogeny Supports the Repeated Origin of Vectorial Capacity in Malaria Mosquitoes of the *Anopheles gambiae* Complex. PLoS Pathog. 8, e1002960. doi:10.1371/journal.ppat.1002960

Kauth, C.W., Woehlbier, U., Kern, M., Mekonnen, Z., Lutz, R., Mücke, N., Langowski, J., Bujard, H., 2006. Interactions between merozoite surface proteins 1, 6, and 7 of the malaria parasite *Plasmodium falciparum*. J. Biol. Chem. 281, 31517-31527.

Keitany, G.J., Vignali, M., Wang, R., 2014. Live attenuated pre-erythrocytic malaria vaccines. Hum. Vaccines Immunother. 10, 2903–2909. doi:10.4161/21645515.2014.972764

Khan, S.M., Franke-Fayard, B., Mair, G.R., Lasonder, E., Janse, C.J., Mann, M., Waters, A.P., 2005. Proteome Analysis of Separated Male and Female Gametocytes Reveals Novel Sex-Specific *Plasmodium* Biology. Cell 121, 675–687. doi:10.1016/j.cell.2005.03.027

Khan, S.M., Reece, S.E., Waters, A.P., Janse, C.J., Kaczanowski, S., 2012. Why are male malaria parasites in such a rush?: Sex-specific evolution and host-parasite interactions. Evol. Med. Public Health 2013, 3–13. doi:10.1093/emph/eos003

King, J.G., Hillyer, J.F., 2012. Infection-Induced Interaction between the Mosquito Circulatory and Immune Systems. PLoS Pathog. 8, e1003058. doi:10.1371/journal.ppat.1003058

Kirkman, L.A., Deitsch, K.W., 2012. Antigenic variation and the generation of diversity in malaria parasites. Curr. Opin. Microbiol. 15, 456–462. doi:10.1016/j.mib.2012.03.003

Kooij, T.W.A., Carlton, J.M., Bidwell, S.L., Hall, N., Ramesar, J., Janse, C.J., Waters, A.P., 2005. A *Plasmodium* Whole-Genome Synteny Map: Indels and Synteny Breakpoints as Foci for Species-Specific Genes. PLoS Pathog. 1, e44. doi:10.1371/journal.ppat.0010044

Kooij, T.W.A., Franke-Fayard, B., Renz, J., Kroeze, H., van Dooren, M.W., Ramesar, J., Augustijn, K.D., Janse, C.J., Waters, A.P., 2005. *Plasmodium berghei* α-tubulin II: A role in both male gamete formation and asexual blood stages. Mol. Biochem. Parasitol. 144, 16–26. doi:10.1016/j.molbiopara.2005.07.003

Kosakovsky Pond, S.L., Frost, S.D., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676-679.

Kosakovsky Pond, S.L., Murrell, B., Fourment, M., Frost, S.D., Delport, W., Scheffler, K., 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol. Biol. Evol. 28, 3033-3043.

Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., Frost, S.D., 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics 22, 3096-3098.

Krief, S., Escalante, A.A., Pacheco, M.A., Mugisha, L., André, C., Halbwax, M., Fischer, A., Krief, J.-M., Kasenene, J.M., Crandfield, M., Cornejo, O.E., Chavatte, J.-M., Lin, C., Letourneur, F., Grüner, A.C., McCutchan, T.F., Rénia, L., Snounou, G., 2010. On the Diversity of Malaria Parasites in African Apes and the Origin of *Plasmodium falciparum* from Bonobos. PLoS Pathog. 6, e1000765. doi:10.1371/journal.ppat.1000765
Kuehn, A., Pradel, G., Kuehn, A., Pradel, G., 2010. The Coming-Out of Malaria Gametocytes, The Coming-Out of Malaria Gametocytes. BioMed Res. Int. BioMed Res. Int. 2010, 2010, e976827. doi:10.1155/2010/976827, 10.1155/2010/976827

Külzer, S., Charnaud, S., Dagan, T., Riedel, J., Mandal, P., Pesce, E.R., Blatch, G.L., Crabb, B.S., Gilson, P.R., Przyborski, J.M., 2012. *Plasmodium falciparum* -encoded exported hsp70/hsp40 chaperone/co-chaperone complexes within the host erythrocyte: Chaperones in the *P. falciparum* -infected host cell. Cell. Microbiol. 14, 1784–1795. doi:10.1111/j.1462-5822.2012.01840.x

Kumar S, Filipski A, Swarna V, Walker A, Hedges SB. 2005. Placing confidence limits on the molecular age of the human–chimpanzee divergence. Proceedings of the National Academy of Sciences of the United States of America.102:18842-18847. doi:10.1073/pnas.0509585102.

Kuo, C.-H., Kissinger, J.C., 2008. Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. BMC Evol. Biol. 8, 108. doi:10.1186/1471-2148-8-108

Kyes, S.A., Kraemer, S.M., Smith, J.D., 2007. Antigenic Variation in *Plasmodium falciparum*: Gene Organization and Regulation of the var Multigene Family. Eukaryot. Cell 6, 1511–1520. doi:10.1128/EC.00173-07

Lapp, S.A., Korir, C.C., Galinski, M.R., 2009. Redefining the expressed prototype SICAvar gene involved in *Plasmodium knowlesi* antigenic variation. Malar. J. 8, 181. doi:10.1186/1475-2875-8-181

Lasonder, E., Ishihama, Y., Andersen, J.S., Vermunt, A.M., Pain, A., Sauerwein, R.W., Eling, W.M., Hall, N., Waters, A.P., Stunnenberg, H.G., others, 2002. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. Nature 419, 537–542.

Lee, W.-C., Malleret, B., Lau, Y.-L., Mauduit, M., Fong, M.-Y., Cho, J.S., Suwanarusk, R., Zhang, R., Albrecht, L., Costa, F.T., others, 2014. Glycophorin C (CD236R) mediates vivax malaria parasite rosetting to normocytes. Blood 123, e100–e109.

Lin, C.S., Uboldi, A.D., Epp, C., Bujard, H., Tsuboi, T., Czabotar, P.E., Cowman, A.F., 2016. Multiple *Plasmodium falciparum* merozoite surface protein 1 complexes mediate merozoite binding to human erythrocytes. J Biol Chem. 291, 7703-7715.

Liu, W., Li, Y., Learn, G.H., Rudicell, R.S., Robertson, J.D., Keele, B.F., Ndjango, J.-B.N., Sanz, C.M., Morgan, D.B., Locatelli, S., Gonder, M.K., Kranzusch, P.J., Walsh, P.D., Delaporte, E., Mpoudi-Ngole, E., Georgiev, A.V., Muller, M.N., Shaw, G.M., Peeters, M., Sharp, P.M., Rayner, J.C., Hahn, B.H., 2010. Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. Nature 467, 420–425. doi:10.1038/nature09442

López-Barragán, M.J., Lemieux, J., Quiñones, M., Williamson, K.C., Molina-Cruz, A., Cui, K., Barillas-Mury, C., Zhao, K., Su, X., 2011. Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*. BMC Genomics 12, 587.

Malleret, B., Li, A., Zhang, R., Tan, K.S., Suwanarusk, R., Claser, C., Cho, J.S., Koh, E.G.L., Chu, C.S., Pukrittayakamee, S., others, 2015. *Plasmodium vivax*: restricted tropism and rapid remodeling of CD71-positive reticulocytes. Blood 125, 1314–1324. Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczycki, C.J., Lu, F., Marchler, G.H., Mullokandov, M., Omelchenko, M.V., Robertson, C.L., Song, J.S., Thanki, N., Yamashita, R.A., Zhang, D., Zhang, N., Zheng, C., Bryant, S.H., 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 2011, 39, D225-D229.

Markus, M.B., 2015. Do hypnozoites cause relapse in malaria? Trends Parasitol. 31, 239–245. doi:10.1016/j.pt.2015.02.003

Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., Lefeuvre, P., 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26, 2462-2463.

Martinsen, E.S., Perkins, S.L., Schall, J.J., 2008. A three-genome phylogeny of malaria parasites (Plasmodium and closely related genera): Evolution of life-history traits and host switches. Mol. Phylogenet. Evol. 47, 261–273. doi:10.1016/j.ympev.2007.11.012

Martens, C., Vandepoele, K., Van de Peer, Y., 2008. Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. Proc. Natl. Acad. Sci. 105, 3427–3432.

McKenzie, F.E., Wongsrichanalai, C., Magill, A.J., Forney, J.R., Permpanich, B., Lucas, C., Erhart, L.M., O'Meara, W.P., Smith, D.L., Sirichaisinthop, J., others, 2006. Gametocytemia in *Plasmodium vivax* and *Plasmodium falciparum* infections. J. Parasitol.

Mello, K., Daly, T.M., Long, C.A., Burns, J.M., Bergman, L.W., 2004. Members of the merozoite surface protein 7 family with similar expression patterns differ in ability to protect against *Plasmodium yoelii* malaria. Infect. Immun. 72, 1010-1018.

Mello, K., Daly, T.M., Morrisey, J., Vaidya, A.B., Long, C.A., Bergman, L.W., 2002. A multigene family that interacts with the amino terminus of *Plasmodium* MSP-1 identified using the yeast two-hybrid system. Eukaryot. Cell 1, 915-25.92, 1281–1285.

Mitri, C., Thiery, I., Bourgouin, C., Paul, R.E.L., 2009. Density-dependent impact of the human malaria parasite *Plasmodium falciparum* gametocyte sex ratio on mosquito infection rates. Proc. R. Soc. B Biol. Sci. 276, 3721–3726. doi:10.1098/rspb.2009.0962

Miura, K., Takashima, E., Deng, B., Tullo, G., Diouf, A., Moretz, S.E., Nikolaeva, D., Diakite, M., Fairhurst, R.M., Fay, M.P., Long, C.A., Tsuboi, T., 2013. Functional Comparison of *Plasmodium falciparum* Transmission-Blocking Vaccine Candidates by the Standard Membrane-Feeding Assay. Infect. Immun. 81, 4377–4382. doi:10.1128/IAI.01056-13

Mongui, A., Perez-Leal, O., Soto, S.C., Cortes, J., Patarroyo, M.A., 2006. Cloning, expression, and characterization of a *Plasmodium vivax* MSP7 family merozoite surface protein. Biochem. Biophys. Res. Commun. 351, 639-644.

Molina-Cruz, A., DeJong, R.J., Ortega, C., Haile, A., Abban, E., Rodrigues, J., Jaramillo-Gutierrez, G., Barillas-Mury, C., 2012. Some strains of *Plasmodium falciparum*, a human malaria parasite, evade the complement-like system of *Anopheles gambiae* mosquitoes. Proc. Natl. Acad. Sci. 109, E1957–E1962. doi:10.1073/pnas.1121183109

Moreno, S.N.J., Ayong, L., Pace, D.A., 2011. Calcium storage and function in *apicomplexan* parasites: Figure 1. Essays Biochem. 51, 97–110. doi:10.1042/bse0510097

Mu, J., 2005. Host Switch Leads to Emergence of *Plasmodium vivax* Malaria in Humans. Mol. Biol. Evol. 22, 1686–1693. doi:10.1093/molbev/msi160

Muehlenbein, M.P., Pacheco, M.A., Taylor, J.E., Prall, S.P., Ambu, L., Nathan, S., Alsisto, S., Ramirez, D., Escalante, A.A., 2015. Accelerated diversification of nonhuman primate malarias in Southeast Asia: adaptive radiation or geographic speciation? Mol. Biol. Evol. 32, 422-439.

Murrell, B., Weaver, S., Smith, M.D., Wertheim, J.O., Murrell, S., Aylward, A., Eren, K., Pollner, T., Martin, D.P., Smith, D.M., Scheffler, K., Kosakovsky Pond, S.L., 2015. Gene-Wide Identification of Episodic Selection. Mol. Biol. Evol. 32, 1365–1371. doi:10.1093/molbev/msv035

Mvumbi, D.M., Kayembe, J.-M., Situakibanza, H., Bobanga, T.L., Nsibu, C.N., Mvumbi, G.L., Melin, P., De Mol, P., Hayette, M.-P., 2015. *Falciparum* malaria molecular drug resistance in the Democratic Republic of Congo: a systematic review. Malar. J. 14. doi:10.1186/s12936-015-0892-z

Neafsey, D.E., Galinsky, K., Jiang, R.H., Young, L., Sykes, S.M., Saif, S., Gujja, S., Goldberg, J.M., Young, S., Zeng, Q., Chapman, S.B., Dash, A.P., Anvikar, A.R., Sutton, P.L., Birren, B.W., Escalante, A.A., Barnwell, J.W., Carlton, J.M., 2012. The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. Nat. Genet. 44, 1046-50.

Neafsey, D.E., Waterhouse, R.M., Abai, M.R., Aganezov, S.S., Alekseyev, M.A., Allen, J.E., Amon, J., Arca, B., Arensburger, P., Artemov, G., Assour, L.A., Basseri, H., Berlin, A., Birren, B.W., Blandin, S.A., Brockman, A.I., Burkot, T.R., Burt, A., Chan, C.S., Chauve, C., Chiu, J.C., Christensen, M., Costantini, C., Davidson, V.L.M., Deligianni, E., Dottorini, T., Dritsou, V., Gabriel, S.B., Guelbeogo, W.M., Hall, A.B., Han, M.V., Hlaing, T., Hughes, D.S.T., Jenkins, A.M., Jiang, X., Jungreis, I., Kakani, E.G., Kamali, M., Kemppainen, P., Kennedy, R.C., Kirmitzoglou, I.K., Koekemoer, L.L., Laban, N., Langridge, N., Lawniczak, M.K.N., Lirakis, M., Lobo, N.F., Lowy, E., MacCallum, R.M., Mao, C., Maslen, G., Mbogo, C., McCarthy, J., Michel, K., Mitchell, S.N., Moore, W., Murphy, K.A., Naumenko, A.N., Nolan, T., Novoa, E.M., O'Loughlin, S., Oringanje, C., Oshaghi, M.A., Pakpour, N., Papathanos, P.A., Peery, A.N., Povelones, M., Prakash, A., Price, D.P., Rajaraman, A., Reimer, L.J., Rinker, D.C., Rokas, A., Russell, T.L., Sagnon, N., Sharakhova, M.V., Shea, T., Simao, F.A., Simard, F., Slotman, M.A., Somboon, P., Stegniy, V., Struchiner, C.J., Thomas, G.W.C., Tojo, M., Topalis, P., Tubio, J.M.C., Unger, M.F., Vontas, J., Walton, C., Wilding, C.S., Willis, J.H., Wu, Y.-C., Yan, G., Zdobnov, E.M., Zhou, X., Catteruccia, F., Christophides, G.K., Collins, F.H., Cornman, R.S., Crisanti, A., Donnelly, M.J., Emrich, S.J., Fontaine, M.C., Gelbart, W., Hahn, M.W., Hansen, I.A., Howell, P.I., Kafatos, F.C., Kellis, M., Lawson, D., Louis, C., Luckhart, S., Muskavitch, M.A.T., Ribeiro, J.M., Riehle, M.A., Sharakhov,

I.V., Tu, Z., Zwiebel, L.J., Besansky, N.J., 2015. Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes. Science 347, 1258522–1258522. doi:10.1126/science.125852

Neal, A.T., Schall, J.J., 2014. Testing sex ratio theory with the malaria parasite *Plasmodium mexicanum* in natural and experimental infections: malaria sex ratio. Evolution 68, 1071–1081. doi:10.1111/evo.12334

Nei, M., Gu, X., Sitnikova, T., 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. USA. 94,7799-806.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.

Nei, M., Kumar, S., 2000. Molecular evolution and phylogenetics. Oxford University Press, NY. Nei, M., Rooney, A.P., Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. 39, 121-152.

Ness, R.W., Graham, S.W., Barrett, S.C., 2011. Reconciling gene and genome duplication events: using multiple nuclear gene families to infer the phylogeny of the aquatic plant family Pontederiaceae. Mol. Biol. Evol. 28, 3009-18.

Nielsen, K.M., Kasper, J., Choi, M., Bedford, T., Kristiansen, K., Wirth, D.F., Volkman, S.K., Lozovsky, E.R., Hartl, D.L., 2003. Gene conversion as a source of nucleotide diversity in *Plasmodium falciparum*. Mol. Biol. Evol. 20, 726-734.

Niang, M., Yan Yam, X., Preiser, P.R., 2009. The *Plasmodium falciparum* STEVOR Multigene Family Mediates Antigenic Variation of the Infected Erythrocyte. PLoS Pathog. 5, e1000307. doi:10.1371/journal.ppat.1000307

Niederwieser, I., Felger, I., Beck, H.-P., 2001. Limited polymorphism in *Plasmodium falciparum* sexual-stage antigens. Am. J. Trop. Med. Hyg. 64, 9–11.

Niang, M., Bei, A.K., Madnani, K.G., Pelly, S., Dankwa, S., Kanjee, U., Gunalan, K., Amaladoss, A., Yeo, K.P., Bob, N.S., Malleret, B., Duraisingh, M.T., Preiser, P.R., 2014. STEVOR Is a *Plasmodium falciparum* Erythrocyte Binding Protein that Mediates Merozoite Invasion and Rosetting. Cell Host Microbe 16, 81–93. doi:10.1016/j.chom.2014.06.004

Nikbakht, H., Xia, X., Hickey, D.A., Golding, B., 2014. The evolution of genomic GC content undergoes a rapid reversal within the genus *Plasmodium*. Genome 57, 507–511. doi:10.1139/gen-2014-0158

Nikolaeva, D., Draper, S.J., Biswas, S., 2015. Toward the development of effective transmission-blocking vaccines for malaria. Expert Rev. Vaccines 14, 653–680. doi:10.1586/14760584.2015.993383

Nishimoto, Y., Arisue, N., Kawai, S., Escalante, A.A., Horii, T., Tanabe, K., Hashimoto, T., 2008. Evolution and phylogeny of the heterogeneous cytosolic SSU rRNA genes in the genus *Plasmodium*. Mol. Phylogenet. Evol. 47, 45–53. doi:10.1016/j.ympev.2008.01.031

Offeddu, V., Thathy, V., Marsh, K., Matuschewski, K., 2012. Naturally acquired immune responses against *Plasmodium falciparum* sporozoites and liver infection. Int. J. Parasitol. 42, 535–548. doi:10.1016/j.ijpara.2012.03.011

Ohno, S., 1970. Evolution by Gene Duplication. Springer Berlin Heidelberg, Berlin, Heidelberg.

Okuda-Ashitaka, E., Minami, T., Tachibana, S., Yoshihara, Y., Nishiuchi, Y., Kimura, T., Ito, S., 1998. Nocistatin, a peptide that blocks nociceptin action in pain transmission. Nature 392, 286–289.

Ollomo, B., Durand, P. Prugnolle, F., Douzery, E., Arnathau, C., Nkoghe, D., Leroy, E., Renaud, F., 2009. A new malaria agent in African hominids. PLoS Pathog. 5(5), e1000446.

Otto, T.D., Böhme, U., Jackson, A.P., Hunt, M., Franke-Fayard, B., Hoeijmakers, W.A., Religa, A.A., Robertson, L., Sanders, M., Ogun, S.A., others, 2014a. A comprehensive evaluation of rodent malaria parasite genomes and gene expression. BMC Biol. 12, 1.

Otto, T.D., Rayner, J.C., Böhme, U., Pain, A., Spottiswoode, N., Sanders, M., Quail, M., Ollomo, B., Renaud, F., Thomas, A.W., Prugnolle, F., Conway, D.J., Newbold, C., Berriman, M., 2014b. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. Nat. Commun. 5, 4754. doi:10.1038/ncomms5754

Ouattara, A., Laurens, M.B., 2015. Vaccines Against Malaria. Clin. Infect. Dis. 60, 930–936. doi:10.1093/cid/ciu954

Pacheco, M.A., Battistuzzi, F.U., Junge, R.E., Cornejo, O.E., Williams, C.V., Landau, I., Rabetafika, L., Snounou, G., Jones-Engel, L., Escalante, A.A., 2011. Timing the origin of human malarias: the lemur puzzle. BMC Evol. Biol. 11, 299.

Pacheco, M.A., Cranfield, M., Cameron, K., Escalante, A.A., 2013. Malarial parasite diversity in chimpanzees: the value of comparative approaches to ascertain the evolution of *Plasmodium falciparum* antigens. Malar J. 17, 12:328.

Pacheco, M.A., Reid, M.J., Schillaci, M.A., Lowenberger, C.A., Galdikas, B.M., Jones-Engel, L., Escalante, A.A., 2012. The origin of malarial parasites in orangutans. PLoS One. 7(4), e34990.

Pain, A., Böhme, U., Berry, A.E., Mungall, K., Finn, R.D., Jackson, A.P., Mourier, T., Mistry, J., Pasini, E.M., Aslett, M.A., Balasubrammaniam, S., Borgwardt, K., Brooks, K., Carret, C., Carver, T.J., Cherevach, I., Chillingworth, T., Clark, T.G., Galinski, M.R., Hall, N., Harper, D., Harris, D., Hauser, H., Ivens, A., Janssen, C.S., Keane, T., Larke, N., Lapp, S., Marti, M., Moule, S., Meyer, I.M., Ormond, D., Peters, N., Sanders, M., Sanders, S., Sargeant, T.J., Simmonds, M., Smith, F., Squares, R., Thurston, S., Tivey, A.R., Walker, D., White, B., Zuiderwijk, E., Churcher, C., Quail, M.A., Cowman, A.F., Turner, C.M.R., Rajandream, M.A., Kocken, C.H.M., Thomas, A.W., Newbold, C.I., Barrell, B.G., Berriman, M., 2008. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. Nature 455, 799–803. doi:10.1038/nature07306

Perrin, A.J., Bartholdson, S.J., Wright, G.J., 2015. P-selectin is a host receptor for *Plasmodium* MSP7 ligands. Malar. J. 14, 238.

Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8, 785-786.

Petter, M., Bonow, I., Klinkert, M.-Q., 2008. Diverse Expression Patterns of Subgroups of the rif Multigene Family during *Plasmodium falciparum* Gametocytogenesis. PLoS ONE 3, e3779. doi:10.1371/journal.pone.0003779

Pond, S.L.K., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies, in: Statistical Methods in Molecular Evolution. Springer, pp. 125–181.

Ponsuwanna, P., Kochakarn, T., Bunditvorapoom, D., Kümpornsin, K., Otto, T.D., Ridenour, C., Chotivanich, K., Wilairat, P., White, N.J., Miotto, O., Chookajorn, T., 2016. Comparative genome-wide analysis and evolutionary history of haemoglobin-processing and haem detoxification enzymes in malarial parasites. Malar. J. 15. doi:10.1186/s12936-016-1097-9

Posada, D., 2008. jModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25, 1253-1256.

Pouniotis, D.S., Proudfoot, O., Minigo, G., Hanley, J.L., Plebanski, M., others, 2004. Malaria parasite interactions with the human host. J. Postgrad. Med. 50, 30.

Prugnolle, F., McGee, K., Keebler, J., Awadalla, P., 2008. Selection shapes malaria genomes and drives divergence between pathogens infecting hominids versus rodents. BMC Evol. Biol. 8, 223. doi:10.1186/1471-2148-8-223

Prugnolle, F., Rougeron, V., Becquart, P., Berry, A., Makanga, B., Rahola, N., Arnathau, C., Ngoubangoye, B., Menard, S., Willaume, E., Ayala, F.J., Fontenille, D., Ollomo, B.,

Durand, P., Paupy, C., Renaud, F., 2013. Diversity, host switching and evolution of *Plasmodium vivax* infecting African great apes. Proc. Natl. Acad. Sci. 110, 8123–8128. doi:10.1073/pnas.1306004110

Reece, S.E., Drew, D.R., Gardner, A., 2008. Sex ratio adjustment and kin discrimination in malaria parasites. Nature 453, 609–614. doi:10.1038/nature06954

Reid, A.J., 2015. Large, rapidly evolving gene families are at the forefront of host–parasite interactions in *Apicomplexa*. Parasitology 142, S57–S70. doi:10.1017/S0031182014001528

Reininger, L., Garcia, M., Tomlins, A., Muller, S., Doerig, C., 2012. The *Plasmodium falciparum*, Nima-related kinase Pfnek-4: a marker for asexual parasites committed to sexual differentiation. Malar J 11, 250.

Reininger, L., Tewari, R., Fennell, C., Holland, Z., Goldring, D., Ranford-Cartwright, L., Billker, O., Doerig, C., 2009. An Essential Role for the *Plasmodium* Nek-2 Nima-related Protein Kinase in the Sexual Development of Malaria Parasites. J. Biol. Chem. 284, 20858–20868. doi:10.1074/jbc.M109.017988

Rice, B.L., Acosta, M.M., Pacheco, M.A., Carlton, J.M., Barnwell, J.W., Escalante, A.A., 2014. The origin and diversification of the merozoite surface protein 3 (msp3) multi-gene family in *Plasmodium vivax* and related parasites. Mol. Phylogenet. Evol. 78, 172–184. doi:10.1016/j.ympev.2014.05.013

Riveron, J.M., Chiumia, M., Menze, B.D., Barnes, K.G., Irving, H., Ibrahim, S.S., Weedall, G.D., Mzilahowa, T., Wondji, C.S., 2015. Rise of multiple insecticide resistance in *Anopheles funestus* in Malawi: a major concern for malaria vector control. Malar. J. 14. doi:10.1186/s12936-015-0877-y

Roeffen, W., Teelen, K., van As, J., vd Vegte-Bolmer, M., Eling, W., Sauerwein, R., 2001. *Plasmodium falciparum*: Production and Characterization of Rat Monoclonal Antibodies Specific for the Sexual-Stage Pfs48/45 Antigen. Exp. Parasitol. 97, 45–49. doi:10.1006/expr.2000.4586

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539-542.

Rooney, A.P., 2004. Mechanisms Underlying the Evolution and Maintenance of Functionally Heterogeneous 18S rRNA Genes in *Apicomplexans*. Mol. Biol. Evol. 21, 1704–1711. doi:10.1093/molbev/msh178

Ryg-Cornejo, V., Ly, A., Hansen, D.S., 2016. Immunological processes underlying the slow acquisition of humoral immunity to malaria. Parasitology 1–9. doi:10.1017/S0031182015001705

Saraiva, R.G., Kang, S., Simões, M.L., Angleró-Rodríguez, Y.I., Dimopoulos, G., 2016. Mosquito gut antiparasitic and antiviral immunity. Dev. Comp. Immunol. doi:10.1016/j.dci.2016.01.015

Scherf, A., Lopez-Rubio, J.J., Riviere, L., 2008. Antigenic Variation in *Plasmodium falciparum*. Annu. Rev. Microbiol. 62, 445–470. doi:10.1146/annurev.micro.61.080706.093134

Schrago CG, Voloch CM. 2013. The precision of the hominid timescale estimated by relaxed clock methods. J Evol Biol. 26:746-55. doi: 10.1111/jeb.12076.

Schwartz, L., Brown, G.V., Genton, B., Moorthy, V.S., 2012. A review of malaria vaccine clinical projects based on the WHO rainbow table. Malar. J. 11, 1.

Seo, T.-K., 2004. Estimating Absolute Rates of Synonymous and Nonsynonymous Nucleotide Substitution in Order to Characterize Natural Selection and Date Species Divergences. Mol. Biol. Evol. 21, 1201–1213. doi:10.1093/molbev/msh088.

Siciliano, G., Alano, P., 2015. Enlightening the malaria parasite life cycle: bioluminescent *Plasmodium* in fundamental and applied research. Front. Microbiol. 6. doi:10.3389/fmicb.2015.00391

Silva, J.C., Egan, A., Arze, C., Spouge, J.L., Harris, D.G., 2015. A New Method for Estimating Species Age Supports the Coexistence of Malaria Parasites and Their Mammalian Hosts. Mol. Biol. Evol. 32, 1354–1364. doi:10.1093/molbev/msv005

Silvestrini, F., Bozdech, Z., Lanfrancotti, A., Giulio, E.D., Bultrini, E., Picci, L., deRisi, J.L., Pizzi, E., Alano, P., 2005. Genome-wide identification of genes upregulated at the onset of gametocytogenesis in *Plasmodium falciparum*. Mol. Biochem. Parasitol. 143, 100–110. doi:10.1016/j.molbiopara.2005.04.015

Smith, M.D., Wertheim, J.O., Weaver, S., Murrell, B., Scheffler, K., Kosakovsky Pond, S.L., 2015. Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. Mol. Biol. Evol. 32, 1342–1353. doi:10.1093/molbev/msv022

Sinden, R.E., 2015. The cell biology of malaria infection of mosquito: advances and opportunities: Malaria infection of the mosquito. Cell. Microbiol. 17, 451–466. doi:10.1111/cmi.12413

Singh, S., Soe, S., Mejia, J.P., Roussilhon, C., Theisen, M., Corradin, G., Druilhe, P., 2004. Identification of a conserved region of *Plasmodium falciparum* MSP3 targeted by biologically active antibodies to improve vaccine design. J. Infect. Dis. 190, 1010-1018.

Sinka, M.E., Bangs, M.J., Manguin, S., Rubio-Palis, Y., Chareonviriyaphap, T., Coetzee, M., Mbogo, C.M., Hemingway, J., Patil, A.P., Temperley, W.H., others, 2012. A global map of dominant malaria vectors. Parasit Vectors 5, 69.

Stevenson, M.M., Riley, E.M., 2004. Innate immunity to malaria. Nat. Rev. Immunol. 4, 169–180. doi:10.1038/nri1311

Stone, W.J.R., Dantzler, K.W., Nilsson, S.K., Drakeley, C.J., Marti, M., Bousema, T., Rijpma, S.R., 2016. Naturally acquired immunity to sexual stage *P. falciparum* parasites. Parasitology 1–12. doi:10.1017/S0031182015001341

Summers, K., Zhu, Y., 2008. Positive Selection on a Prolactin Paralog Following Gene Duplication in Cichlids: Adaptive Evolution in the Context of Parental Care. Copeia 2008, 872–876. doi:10.1643/CI-07-177

Ta, T.H., Hisam, S., Lanza, M., Jiram, A.I., Ismail, N., Rubio, J.M., 2014. First case of a naturally acquired human infection with *Plasmodium cynomolgi.* Malar J 13, 68.

Tachibana, M., Sato, C., Otsuki, H., Sattabongkot, J., Kaneko, O., Torii, M., Tsuboi, T., 2012. *Plasmodium vivax* gametocyte protein Pvs230 is a transmission-blocking vaccine candidate. Vaccine 30, 1807–1812. doi:10.1016/j.vaccine.2012.01.003

Tachibana, M., Suwanabun, N., Kaneko, O., Iriko, H., Otsuki, H., Sattabongkot, J., Kaneko, A., Herrera, S., Torii, M., Tsuboi, T., 2015. *Plasmodium vivax* gametocyte proteins, Pvs48/45 and Pvs47, induce transmission-reducing antibodies by DNA immunization. Vaccine 33, 1901–1908. doi:10.1016/j.vaccine.2015.03.008
Tachibana, S.-I., Sullivan, S.A., Kawai, S., Nakamura, S., Kim, H.R., Goto, N., Arisue, N., Palacpac, N.M.Q., Honma, H., Yagi, M., Tougan, T., Katakai, Y., Kaneko, O., Mita, T., Kita, K., Yasutomi, Y., Sutton, P.L., Shakhbatyan, R., Horii, T., Yasunaga, T., Barnwell, J.W., Escalante, A.A., Carlton, J.M., Tanabe, K., 2012. Plasmodium cynomolgi genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. Nat. Genet. 44, 1051–1055. doi:10.1038/ng.2375

Tainchum, K., Kongmee, M., Manguin, S., Bangs, M.J., Chareonviriyaphap, T., 2015. *Anopheles* species diversity and distribution of the malaria vectors of Thailand. Trends Parasitol. 31, 109–119. doi:10.1016/j.pt.2015.01.004

Talman, A.M., Domarle, O., McKenzie, F.E., Ariey, F., Robert, V., 2004. Gametocytogenesis: the puberty of *Plasmodium falciparum*. Malar. J. 3, 24.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725-2729.

Tao, D., Ubaida-Mohien, C., Mathias, D.K., King, J.G., Pastrana-Mena, R., Tripathi, A., Goldowitz, I., Graham, D.R., Moss, E., Marti, M., Dinglasan, R.R., 2014. Sex-partitioning of the Plasmodium falciparum Stage V Gametocyte Proteome Provides Insight into *falciparum*-specific Cell Biology. Mol. Cell. Proteomics 13, 2705–2724. doi:10.1074/mcp.M114.040956

Tewari, R., Ogun, S.A., Gunaratne, R.S., Crisanti, A., Holder, A.A., 2005. Disruption of *Plasmodium berghei* merozoite surface protein 7 gene modulates parasite growth in vivo. Blood 105, 394-396.

Thompson, J., Fernandez-Reyes, D., Sharling, L., Moore, S.G., Eling, W.M., Kyes, S.A., Newbold, C.I., Kafatos, F.C., Janse, C.J., Waters, A.P., 2007. *Plasmodium* cysteine repeat modular proteins 1?4: complex proteins with roles throughout the malaria parasite life cycle. Cell. Microbiol. 9, 1466–1480. doi:10.1111/j.1462-5822.2006.00885.x

Vaccines, malERA C.G. on, others, 2011. A research agenda for malaria eradication: vaccines. PLoS Med 8, e1000398.

van Schaijk, B.C.L., van Dijk, M.R., van de Vegte-Bolmer, M., van Gemert, G.-J., van Dooren, M.W., Eksi, S., Roeffen, W.F.G., Janse, C.J., Waters, A.P., Sauerwein, R.W., 2006. Pfs47, paralog of the male fertility factor Pfs48/45, is a female specific surface protein in *Plasmodium falciparum*. Mol. Biochem. Parasitol. 149, 216–222. doi:10.1016/j.molbiopara.2006.05.015

Van Zee, J.P., Schlueter, J.A., Schlueter, S., Dixon, P., Sierra, C.A.B., Hill, C.A., 2016. Paralog analyses reveal gene duplication events and genes under positive selection in Ixodes scapularis and other ixodid ticks. BMC Genomics 17. doi:10.1186/s12864-015-2350-2

Wasmuth, J., Daub, J., Peregrín-Alvarez, J.M., Finney, C.A.M., Parkinson, J., 2009. The origins of apicomplexan sequence innovation. Genome Res. 19, 1202-1213.
Weir, W., Sunter, J., Chaussepied, M., Skilton, R., Tait, A., de Villiers, E.P., Bishop, R., Shiels, B., Langsley, G., 2009. Highly syntenic and yet divergent: A tale of two *Theilerias*. Infect. Genet. Evol. 9, 453–461. doi:10.1016/j.meegid.2009.01.002

Wertheim, J.O., Murrell, B., Smith, M.D., Pond, S.L.K., Scheffler, K., 2015. RELAX: detecting relaxed selection in a phylogenetic framework. Mol. Biol. Evol. 32, 820–832.

White, N.J., 2008. Plasmodium knowlesi: The Fifth Human Malaria Parasite. Clin. Infect. Dis. 46, 172–173. doi:10.1086/524889

Williamson, K.C., 2003. Pfs230: from malaria transmission-blocking vaccine candidate toward function. Parasite Immunol. 25, 351–359.

Wilkins, M.R., Gasteiger, E., Bairoch, A., Sanchez, J.C., Williams, K.L., Appel, R.D., Hochstrasser, D.F., 1999. Protein identification and analysis tools in the ExPASy server. Methods Mol. Biol. 112, 531-552.

WHO., 2015. World Malaria Report 2015. World Health Organization, Geneva, 280.

Young, J.A., Fivelman, Q.L., Blair, P.L., de la Vega, P., Le Roch, K.G., Zhou, Y., Carucci, D.J., Baker, D.A., Winzeler, E.A., 2005. The *Plasmodium falciparum* sexual development transcriptome: A microarray analysis using ontology-based pattern identification. Mol. Biochem. Parasitol. 143, 67–79. doi:10.1016/j.molbiopara.2005.05.007

Zhou, S., Rietveld, A.E., Velarde-Rodriguez, M., Ramsay, A.R., Zhang, S., Zhou, X., Cibulskis, R.E., 2014. Operational research on malaria control and elimination: a review of projects published between 2008 and 2013. Malar. J. 13, 1–7.

APPENDIX A

SUPPLEMENTARY DATA FOR CHAPTER 2

Tables

Table S2-1. Primers used in the amplification of *msp7* simian clade paralogs.

Paralog	Forward sequence	Reverse sequence	bp	Temperature (°C)
PVX_082645	CTC CYT CAS CGC AAT GAA G		19	57.5-59
		CTC TTA AAS CTC AAG VGT G	19	53-55
	CTC CYT SAS CGC AAT GAA G		19	57-59.5
		CTC TTA AAS CWC AAG VGT G	19	53-55
PVX_082650	CAA CAA AAT GAG GAA AMA AAT TG		23	54-55.5
		CAC CTS AAG AGT GYT CAT C	19	55-57.5
PVX_082655	CAA AAW GAT GAA KAA AAC GAT CG		23	55-57.6
		CTA CAC CAC YTC AAK CGT G	19	55-59.5
PVX_082660	TCC TRC TGG GKT CCA TTT TG		20	56.4-60.5
		TAA CCC GCC ACT TTA CCA G	19	57.6
		AAT CTG CCA CTT CAC TGT CC	20	58.4
	TCC TRS TGK GBT CCA TTT TG		20	54.3-60.5
		CCT AAC CCG CCA CTT TAC	18	56.3
PVX_082665	TTG CAA AAA TGA RMR GAG TC		20	50.2-56.4
		TGT WCA TSA AGY TGA TGG C	19	53-55
PVX_082670	GGA AAA AAA DTG CWC TMT TCC		21	53.4-57.5
		TYG TGT TGA GGA AAC TTA GC	20	54.3-56.4
	GYT GGM AAA AAG GTG GAC		18	51.4-56.3
PVX_082675	GRT TAA TTT RTT TTY CTC CTC C		22	52.7-58.4
		GTG TTC ATC AAG YTR AWG GC	20	54.3-58.4
	CTC GTT GAA RAA AAA AYT GC		20	50.2-54.3
	AGA AGA RAA TGT AGA AGT GG		20	52.3-54.3
		AWG TAC TTA ATT TTG ACA TCG	21	51.7
PVX_082680	TGG GGG CGA CAA AAT GAA G		19	57.5
		CAC TTG CTC AGT TGG CTT C	19	57.5
	TYT TYG GTT CCC TCT TTG TG		20	54.3-58.4
		CTT YAY TTC AAT CGT GTT YAG C	22	54.7-60.1
	GGC YTT RAA GAA ACA GAT TG		20	52.3-56.4
		CCT TCT GTK GYT TTA AGT AG	20	52.3-56.4
PVX_082685	AWG TAY GTG ATG WTG TCT TCC		21	55.4-57.5
		TTC ATY CTC TSG YCC CTC	18	53.8-58.4
PVX_082690	AAT TTA GGA TGA WCG AGC GG		20	56.4
		CWA AAA TTG CTG TCC ASC TC	20	56.4-58.4
PVX_082695	TTA ACG CTC AYC ATG AAC GG		20	56.4-58.4
		TCT GAA ACA RCR TGW GGT AC	20	54.3-58.4
PVX_082700	AYA AAA RTA CTA TTC WTC TTG CC		23	53.9-57.6
		TGC AAC ATC CKV TTG ABC AAC	21	55.4-61.2

Table S2-2. Repetitive motifs found on *Plasmodium* species from the simian clade.

Paralog	Species (strain)	Repeat motive	Location	Length
PVX_082645	*P.cynomolgi* (RO)	[VNPTAN]$_2$	209-222	351
	P. hylobati	[PQNQA]2-PQNQ	245-258	504
	P.inui (Celebes II)	[DADNN/EE/NN]$_3$	35-55	507
		DADNN/EE/NN		
		[GASGG]$_4$	166-186	
		[APGPS]$_3$-APDPS-APGAS	293-318	
	P.inui (Leucosphyrus)	[ADNKNSH/D]$_2$	36-52	573
		[DADNN/EE/NN]$_3$	49-69	
		[SGE/GGT]$_3$	193-208	
		[APGPS]$_2$-APDPS-APGAS	313-333	
		[GSSVSSGSA/SG/VSA/S]$_3$	393-428	
	P.inui (Perlis)	[DADNN/EE/NN]$_3$	34-54	538
		[APGAS]$_2$	297-307	
		VSSESSGSSGSSGSA VSSGSSVSSGSSVSS GSSVSSGSAVSSESS GSSGPSGSA	343-396	
	P.inui (OS)	[DADNN/EE/NN]$_3$-DAD-NENN	42-69	429
		[SESSGSAVSSGSAVL]$_2$-SESSGS	245-280	
	P.inui (Taiwan II)	GTSGGGASGG GTSGGGT	182-198	458
		[SSESSG]$_3$- SAVSSGSAGSSG- SSGSSR- SAVSSGSAGSSG	267-316	
	P.inui (Perak)	[ENDADNN]$_2$ -NNDADNENN	47-69	555
		[ATDPS]$_3$	312-327	
		[SSGSAVSAV]$_2$ -SAVSSGSAV	370-397	
	P.inui (Leaf Monkey II)	[DADNN/EE/NN]$_3$	42-62	505
		[GASGA]$_2$	184-197	
	P.knowlesi (H)	STGSAA- [STGSTA]$_2$ -[STGSTG]$_2$ -STAST	204-238	383
PVX_082650	*P.vivax* (Sumatra)	ARGDPQSPA ARGDPQSPA A	279-297	470
PVX_082660	*P.cynomolgi* (B-Mulligan-PT1)	[AGGT]$_2$	154-163	188
	P. fieldi	AASKLVSK AASKSVSK A	145-161	196
PVX_082665	*P. fieldi*	VTPQPTERPA VTPQPTERPA VTPEPT	243-267	403
PVX_082670	*P.vivax* (6 strain)	[EADEGV]$_2$	197-211	411
	P.cynomolgi (Berok)	VEEEQGEEDLQGIFQLEEEPGEEDLQGIFQ LEDEPGEEYLQGIFQLEDEPGEEYLHGSFE SEEEAEQGE	241-304	457
	P.cynomolgi (Mulligan)	[AEEDEG]$_2$	184-200	469
		[LEEEQGEEDLQGIFQLEEEQGEEDLQGSFE]$_2$ LEEEAKQGE	248-316	
	P.cynomolgi (Gombok)	LEEEQGEEDLQGIFQLEDEPGEEYLQGSFG SEEEAGEEDLQGIFQLEEKPGEEYLQGSFE SEEEAEQGE	254-317	471
	P.cynomolgi (Ceylonensis)	KEEEQGEEDLQGIFQLEEEQGEEDLQGAFH LEEEQGEEDLQGVFHLEEE	238-287	442
	P.cynomolgi (Bstrain)	[AEEDEG]$_2$	194-210	479
		[LEEEQGEEDLQGIFQLEEEQGEEDLQGSFE]$_2$ -LEEEAKQGE	258-326	
	P.cynomolgi (PT1)	EEEEEK -EEEEK EKEKEK EEE	218-237	486
		[LEEEQGEEDLQGSFELEEEQGEEDLQGAFH]$_2$ -LEEEAKQGE	264-332	
	P.cynomolgi (PT2)	LEEEPGEEDLQGIFQLEDEPGEEYLQGIFQ LEDEPGEEYLQGIFQLEDEPGEEYLHGSFE SEEEAEQGE	251-304	447
	P.cynomolgi (RO)	[AEEDEG]$_2$	183-199	470
		[LEEEQGEEDLQGIFQLEEEQGEEDLQGSFE]$_2$ LEEEAKQGE	247-315	
PVX_082675	*P.cynomolgi* (Berok)	[EEQEEEQEGEQE]$_2$ -QEQEEEQEQEQE -EEQEQEQEEEQE -QEQEEEQEQEQE EEQ—EQEQ	87-151	492
	P.cynomolgi (PT2)	[EEQEEEQEGEQE]$_2$ -QEQEEEQEQEQE -EEQEQEQEEEQE -QEQEEEQEQEQE EEQ—EQEQ	87-151	481
	P.cynomolgi (Gombok)	EEQEEEQEEEHG EEQEEEQEEEL EEEQEEEQE	86-117	430
	P.cynomolgi (PT1)	EEEQEVEQEVEQGEEQGEEQ EEEQDEEQGEEQGEEQGEE	86-124	456
	P.cynomolgi (RO)	EEEQEEEQEEEQ EEEQEVEQEVEQ EEEQEEEQGEE	86-120	450
	P.cynomolgi (Ceylonensis)	EEKEEEQGEEQE VEQGEEQEEEQ GEEQEEEQEEE QGEEQEEEQEEE	83-128	465
	P.cynomolgi (Mulligan)	EEEQEEEQEEEQ EEEQEVEQEVEQ EEEQEEEQGEE	86-120	450
	P.cynomolgi (Bstrain)	EEEQEEEQEEEQ EEEQEVEQEVEQ EEEQEEEQGEE	86-120	461
	P. fieldi	[EEEQEQQ]$_2$ -[EGEQEQQ EEEQEQQ]$_2$ -[EEEQEQQ]3	86-168	488
	P.inui (Perlis)	EDQEEEQV EDQEEEPV EDQEEDQ EQQE	95-121	455
	P.inui (Leaf Monkey I)	EEQEEQQEEEHEEEHEE EQEEQQEEEQEE DQEEDQE	92-127	469
	P.inui (Perak)	EDQEEEQV EDQEEEPV EDQEEDQ EQQE	95-121	456
	P.inui (Philippines)	EDQEEEQK EEQKEDQE EEQVEDQE EEQVEDQE	97-128	469
	P.inui (Celebes II)	EQEEEQEE [DQEEEQKE]$_2$ -DQEEEQK EEQKEDQE	83-127	482
	P.inui (Leaf Monkey II)	EDQEEDQE EEQVEEQK EDQEEEQVE DQEEEQVE DQEEEQVE	89-121	485
	P.inui (N34)	[QEEEQGEE]$_2$ -QEEEQEED -QEEDQEEE QKEE	87-122	466
	P.inui (OS)	[EEQQ]$_2$ - [EEQQ]$_2$	84-103	485
PVX_082680	*P.inui* (Perak)	EKEEEEEETDK EKEEEEEEETDK EKEEEEETDK EKEEEEEETDK EKEEEEETDKE KEEEEEEETDK	140-217	509
	P.inui (Taiwan II)	EEEKKEEKEEEKKA EEEKKEEEEKKE EEEKEEEEEETDK [EKD/EEEEEETDK]2	132-237	536
	P.inui (OS)	[EEEKK/EEEKEEEK]$_2$ EEEKEEEKKE-- QEEKEEE-GEET DKEKEEEEEEET	130-197	481
	P.inui (Leaf Monkey II)	EKEEEEEETDK EKEEEEEEID KEKEEEEEETDK EKEEEEEETDK EK-EEEEATDK EKEEEEETDK	166-241	533
	P.inui (Philippines)	EEEEKKEEKEEEE EKEEEEEETDK EKDEEEEETGK EKEEEETDK EKEEEEEETDK EKEEEEEETDK	131-220	519
	P.inui (Perlis)	EKEEEKKEEEE EKEEEEEETDK EKDEEEEETDK EKEEEEEEETDK EKEEEEEETDK EKEEEETDK	131-215	514
	P.inui (N34)	[EKEEEEEETDK]$_6$	151-270	556
	P.inui (Celebes II)	[EKEEEEEEA/TDK]$_3$ -EK-EEEEETDK	144-186	500
		PSSTGEV PSSTGEV PSSTGEV PS	274-296	
	P. hylobati	[DPSHEET]$_5$- DP -[PQGTVPS]$_2$	238-329	384
PVX_082695	*P.vivax* (all strain)	[GGEE]$_2$ -SGEL TGE	129-143	311
	P.cynomolgi (all strain)	[QTGE]$_{3-4}$	128-139	307
	P.inui (Hawking)	QSDEKTAE QSDEKTAH QSDEKTA	124-146	311
	P.inui (Leaf Monkey II)	QSEEKTAE QSDEKTAH QSDEKTA	122-144	311
	P.inui (Celebes II)	[EQSDEKTA]$_2$ -QQSDEKTA -QQ -[QTTE]$_6$	129-178	340
	P. simiovale	QSGE [QTGE]$_2$	124-135	284
PVX_082700	*P.cynomolgi* (Berok)	DTHTDTVADTN TDTHTDTVADTN	228-251	448
	P. fieldi	EEQTPSESNEGKAEEEQTPSDSNEGKTE EEKTP	98-130	406

Figures

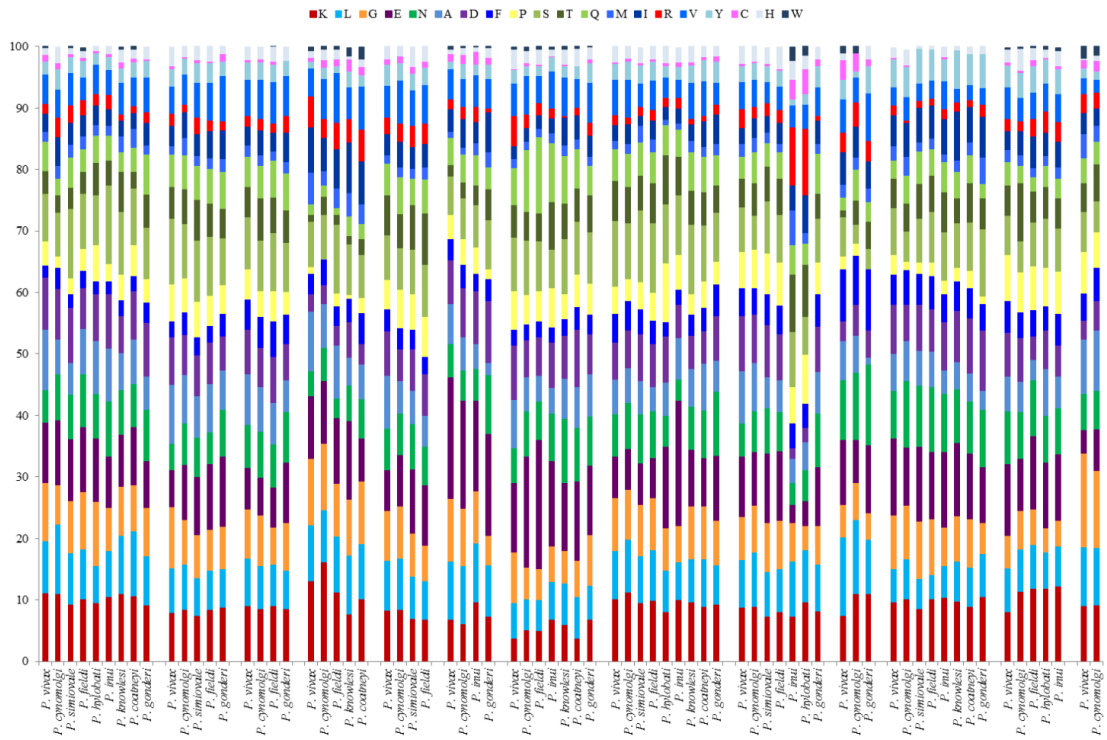
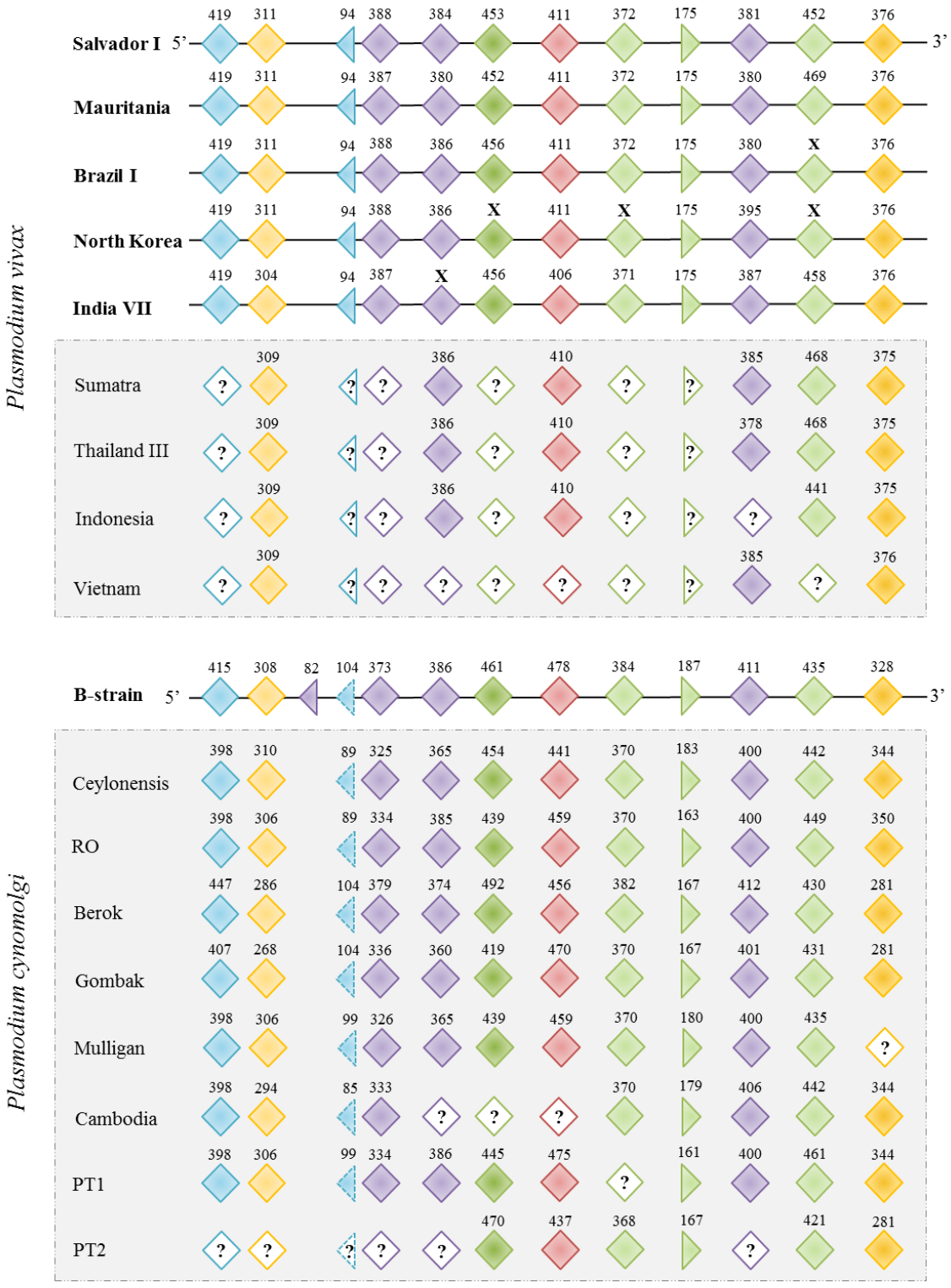


Figure S2-1. *Msp7* multigene family amino acid composition. The *msp7* paralogs share a similar amino acid composition independently of gene length. The similarity tends to be higher between orthologs than among paralogs within the same species. Amino acids in higher proportion are Lysine (K), Leucine (L), Glycine (G), Glutamic Acid (E), Asparagine (N) and Alanine (A).


Plasmodium vivax
Salvador I 5' 419 311 94 388 384 453 411 372 175 381 452 376 3'
Mauritania 419 311 94 387 380 452 411 372 175 380 469 376
Brazil I 419 311 94 388 386 456 411 372 175 380 X 376
North Korea 419 311 94 388 386 X 411 X 175 395 X 376
India VII 419 304 94 387 X 456 406 371 175 387 458 376
Sumatra ? 309 ? ? ? 386 ? 410 ? ? 385 468 375
Thailand III ? 309 ? ? ? 386 ? 410 ? ? 378 468 375
Indonesia ? 309 ? ? ? 386 ? 410 ? ? ? 441 375
Vietnam ? 309 ? ? ? ? ? ? ? ? 385 ? 376
Plasmodium cynomolgi
B-strain 5' 415 308 82 104 373 386 461 478 384 187 411 435 328 3'
Ceylonensis 398 310 89 325 365 454 441 370 183 400 442 344
RO 398 306 89 334 385 439 459 370 163 400 449 350
Berok 447 286 104 379 374 492 456 382 167 412 430 281
Gombak 407 268 104 336 360 419 470 370 167 401 431 281
Mulligan 398 306 99 326 365 439 459 370 180 400 435 ?
Cambodia 398 294 85 333 ? ? ? 370 179 406 442 344
PT1 398 306 99 334 386 445 475 ? 161 400 461 344
PT2 ? ? ? ? ? 470 437 368 167 ? 421 281

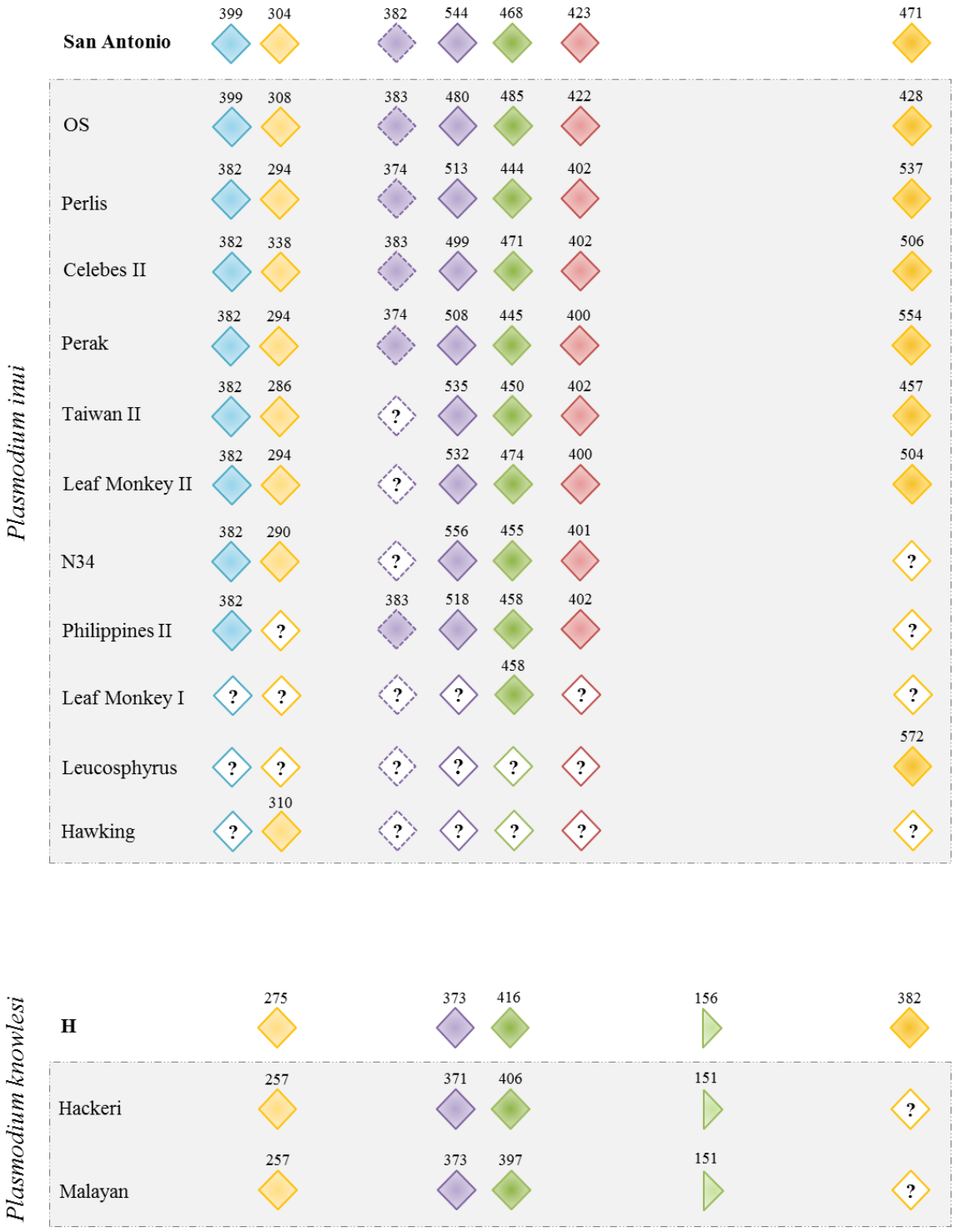


Figure S2-2. Sequences obtained for *P. vivax*, *P. cynomolgi*, *P. inui* and *P. knowlesi* in the laboratory. Putative location of paralogs on the chromosome follows the same order

of paralogs observed on the published *P. vivax* genome. Sequence length of each paralog and isolated is provided above the paralog by diamond and triangle shapes following Fig. 1. Paralogs marked with an X represent publicly available sequences which length could not be confidently measured due to partially missing data (Ns) or incomplete sequences found in a contig. Question marks indicate isolates for which laboratory amplification was not possible but are otherwise present in other *Plasmodium* species analyzed by different means.

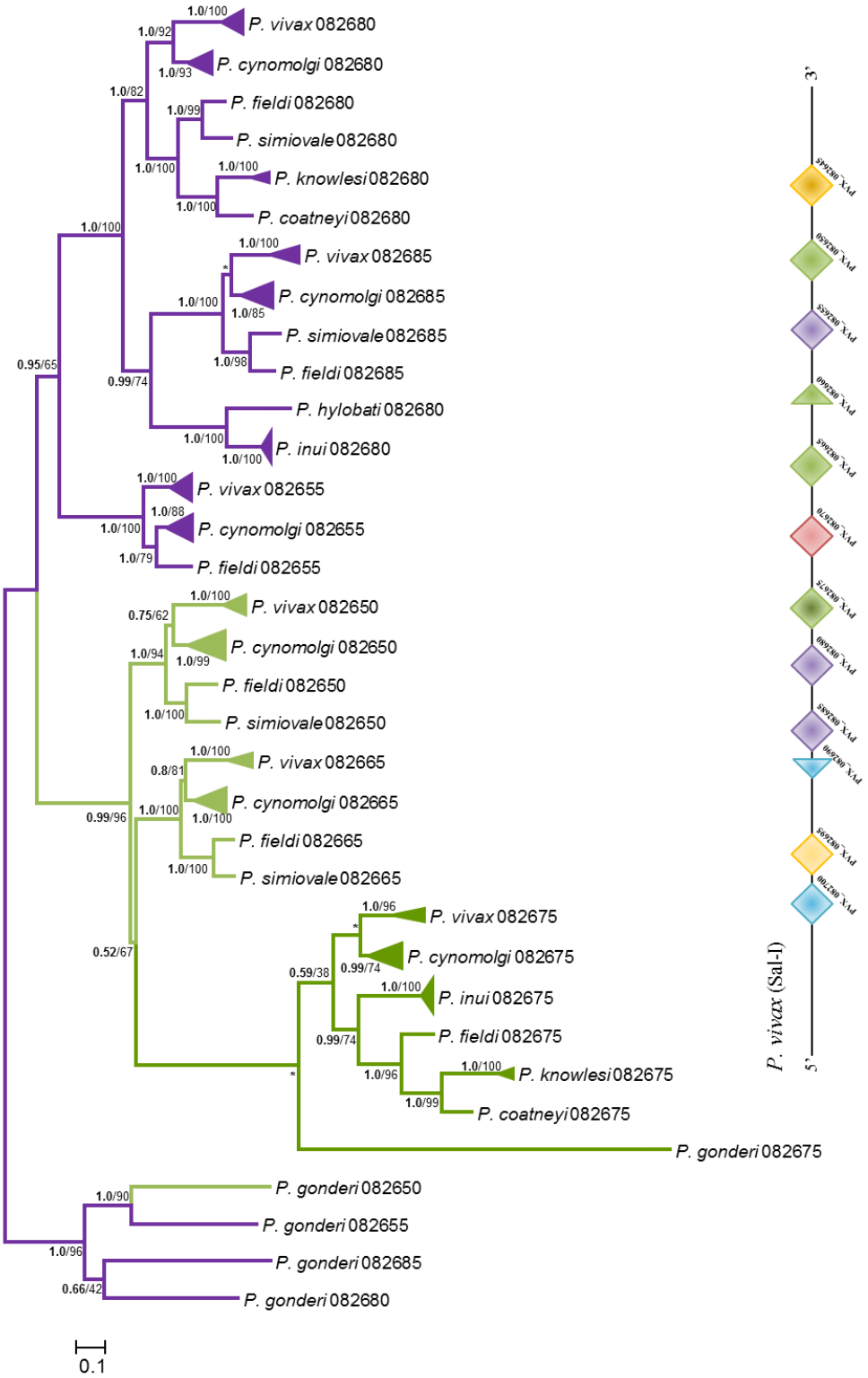

P. vivax 082680
P. cynomolgi 082680
P. fieldi 082680
P. simiovale 082680
P. knowlesi 082680
P. coatneyi 082680
P. vivax 082685
P. cynomolgi 082685
P. simiovale 082685
P. fieldi 082685
P. hylobati 082680
P. inui 082680
P. vivax 082655
P. cynomolgi 082655
P. fieldi 082655
P. vivax 082650
P. cynomolgi 082650
P. fieldi 082650
P. simiovale 082650
P. vivax 082665
P. cynomolgi 082665
P. fieldi 082665
P. simiovale 082665
P. vivax 082675
P. cynomolgi 082675
P. inui 082675
P. fieldi 082675
P. knowlesi 082675
P. coatneyi 082675
P. gonderi 082675
P. gonderi 082650
P. gonderi 082655
P. gonderi 082685
P. gonderi 082680
0.1
3'
5'
P. vivax (Sal-I)

Figure S2-3. Bayesian inference (BI) and Maximum Likelihood (ML) of six paralogs of *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). Branch which share the same coloration are more closely related and share overall similar sequence patterns. The tree was constructed from 123 sequences including only 6 paralogs (PVX_082650, PVX_082655, PVX_082665, PVX_082675, PVX_082680 and PVX_082685). A total of 1,158 nucleotide positions were included in the analysis and the GTR+I+Γ nucleotide model (inv. sites = 0.0390; α=1.4110) was used.

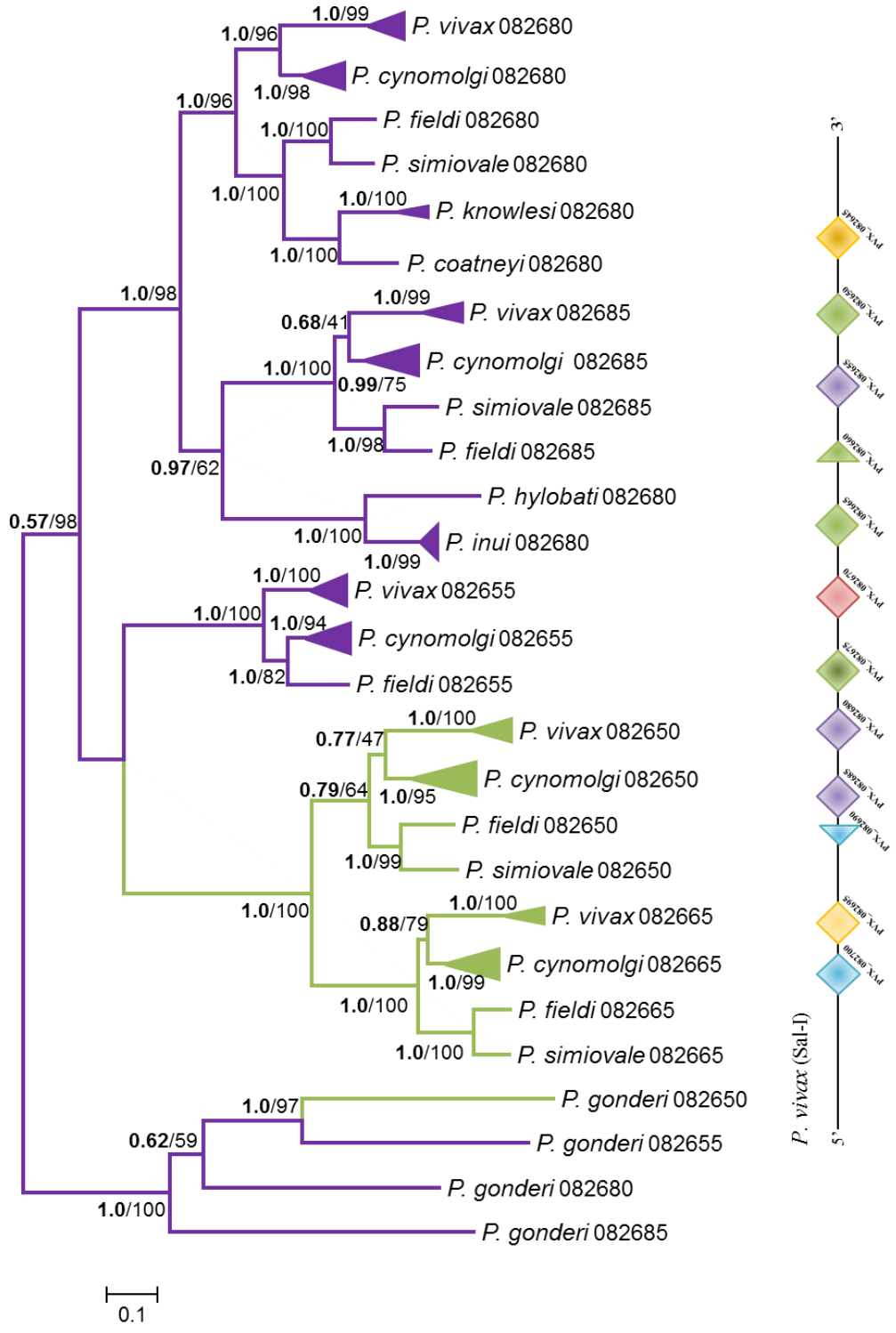


Figure S2-4. Bayesian inference (BI) and Maximum Likelihood (ML) of five paralogs MSP7 multigene family phylogenetic tree. BI and ML trees showed almost identical

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). Branch which share the same coloration are more closely related and share overall similar sequence patterns. The tree was constructed from 96 sequences including only 5 paralogs (PVX_082650, PVX_082655, PVX_082665, PVX_082680 and PVX_082685). A total of 1,158 nucleotide positions were included in the analysis and a special case of the GTR nucleotide model (012032) I+Γ (inv. sites = 0.0770; α=1.5510) was used.

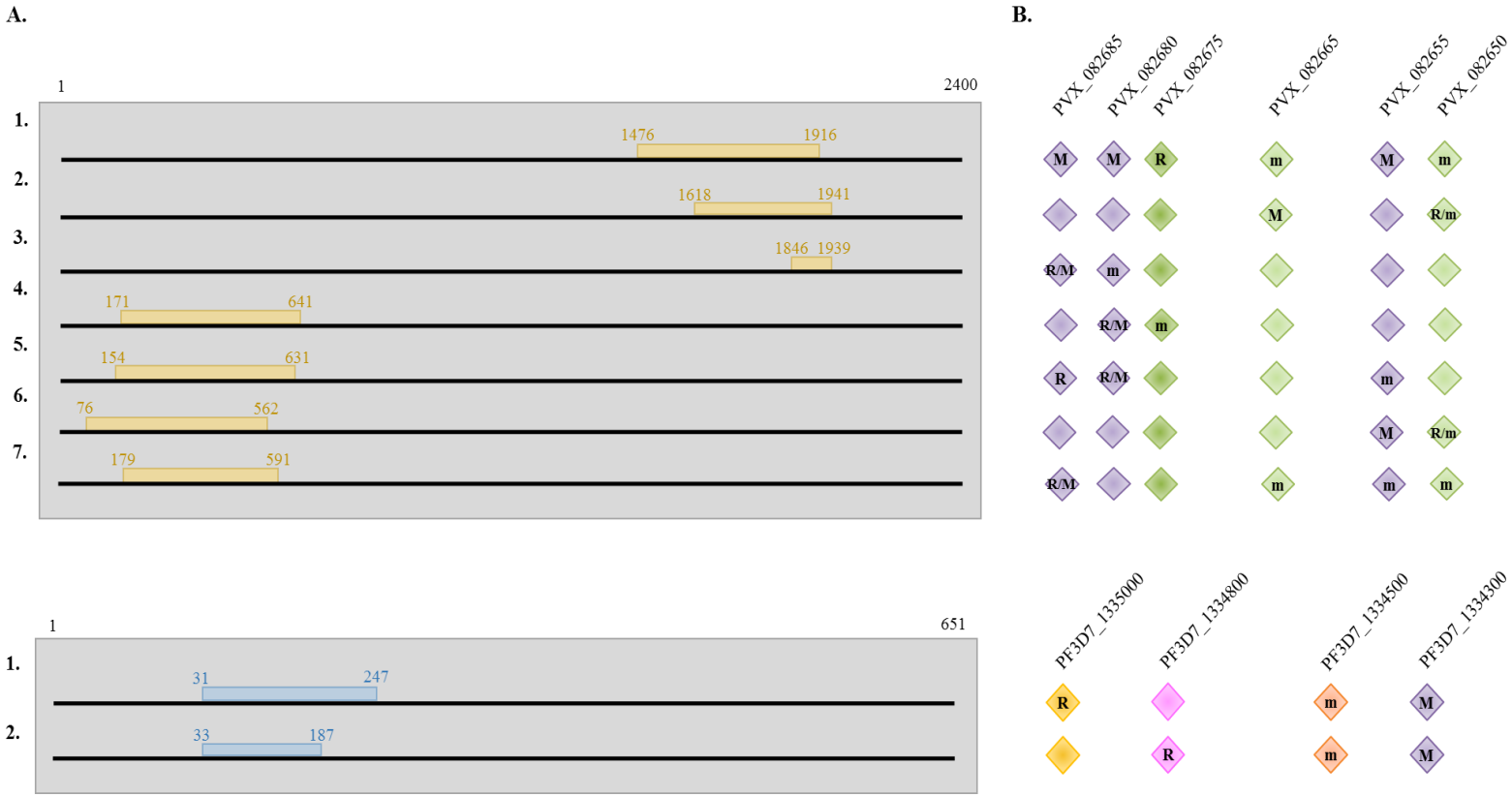


Figure S2-5. Recombination analysis of *msp7* multigene family paralogs. Six closely related *msp7* paralogs (PVX_082650, PVX_082655, PVX_082665, PVX_082675, PVX_082680 and PVX_082685) were analyzed using RDP3 with default parameters. **(A.)** Seven different recombination events were detected among the analyzed sequences. The nucleotide position of independent recombination brake points is indicated relatively to the alignment length (1-2,400bp). Recombination segments are indicated by the use of different color blocks (blue, orange and

yellow). **(B.)** Recombination patterns within *msp7* paralogs included in the analysis. The same coloration is shared by closely related paralogs with overall similar sequence patterns. Paralogs that were part of recombination events are indicated as: (R.) Recombinant, (M.) Mayor Parental, (m.) Minor Parental. Unmarked paralogs indicate that those were not part of a specific recombination event. (*e.g*., recombination event 7 was not detected on paralogs PVX_082675 and PVX_082680).

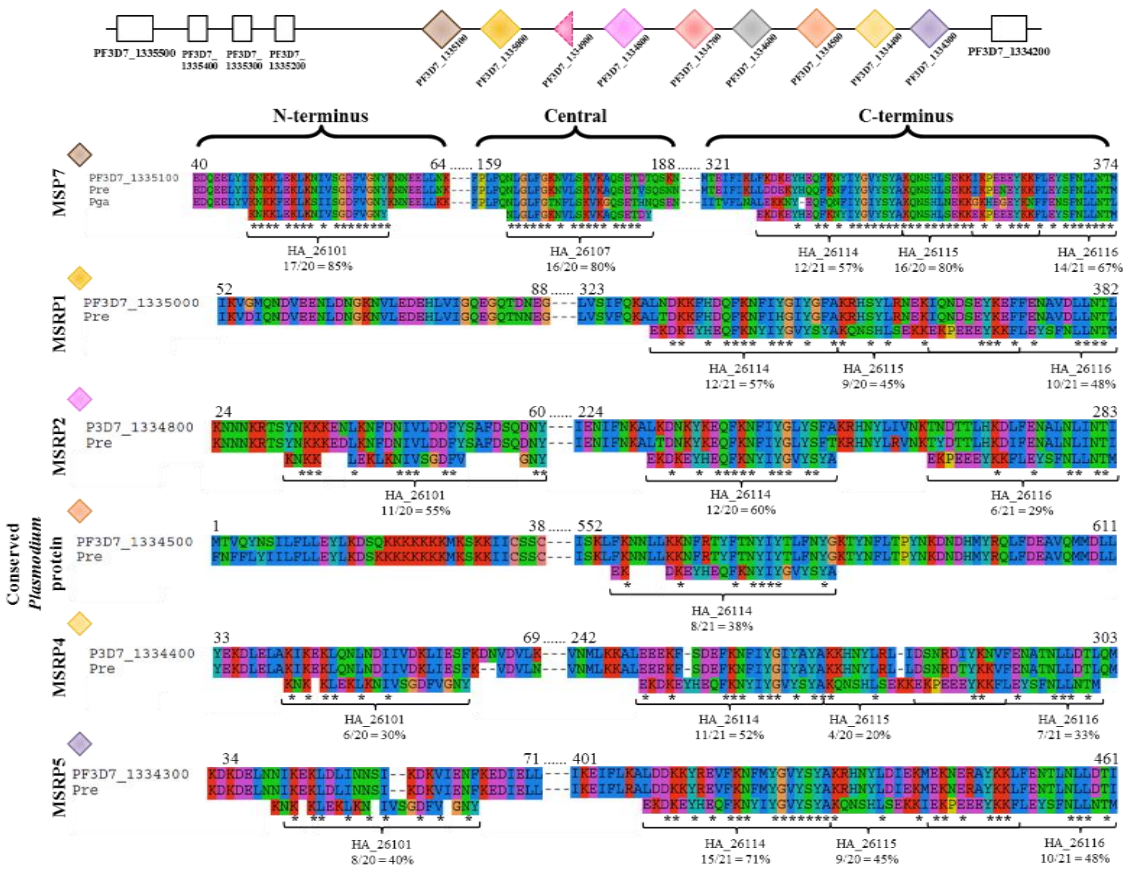


Figure S2-6. Conservation of HABP in *msp7 P. falciparum* paralogs. Alignment of HABPs (Garcia et al., 2007) and *Laverania* subgenus *msp*7 paralogs. Relative position on the alignment is indicated above each analyzed paralog. Asterisks (*) indicate conserved amino acid positions.

APPENDIX B

SUPPLEMENTARY DATA FOR CHAPTER 3

Tables

Table S3-1. Branches under significant episodic selection in gametocyte expressed genes.

Gene ID*	Sex	Loc.	Gene product	*P. cynomolgi*		*P. knowlesi*		*P. berghei*		*P. chabaudi*		*P. yoelii*		*P. reichenowi*		*P. berghei-P. yoelii*	
				ω+	**Prop. sites+**	**ω+**	**Prop. sites+**	**ω+**	**Prop. sites+**	**ω+**	**Prop. sites+**	**ω+**	**Prop. sites+**	**ω+**	**Prop. sites+**	**ω+**	**Prop. sites+**
PVX_001080	♀	-	hypothetical protein, conserved	2.19	12%	0.575	0.62%	14.1	0.44%	0	2.50%	0.248	11%	15.7	0.93%	**71.2**	**0.49%**
PVX_082500	♀	-	signal peptidase 21 kDa subunit, putative	0.2	33%	66.4	8.70%	**10000**	**1.20%**	2.21	7.60%	0.122	0.00%	0.126	0.00%	102	5.30%
PVX_084240	♂	-	hypothetical protein, conserved	**36.9**	**6.50%**	0.145	0.83%	12.2	7.90%	0.0762	0.00%	6.78	2.30%	0.118	6.40%	12.9	100%
PVX_086280	♂	NM	hypothetical protein, conserved	8.78	6.90%	**103**	**9.30%**	1.07	93%	0.776	14%	0.112	1.60%	0.122	0.00%	8.9	4.50%
PVX_088915	♂	NM	hypothetical protein, conserved	**51.4**	**1.40%**	1.75	30%	0.742	1.60%	1.21	12%	1470	0.24%	0.176	0.15%	0.0783	87%
PVX_089245	♂	NM	hypothetical protein, conserved	0.234	100%	3.13	21%	**3330**	**2.40%**	0.0964	100%	6.02	13%	0.137	2.30%	0.121	0.00%
PVX_094335	♀	NM	hypothetical protein, conserved	0.124	0.00%	4.6	8.60%	**10000**	**0.42%**	18.3	0.62%	0.206	2.60%	0.142	0.00%	8640	8.10%
PVX_096350	♀	-	hypothetical protein, conserved	1.97	13%	1.67	3.90%	2.87	14%	3.79	0.90%	3.64	2.60%	**3330**	**0.58%**	8.65	75%
PVX_098665	♀	-	signal peptidase complex subunit 3, putative	0.123	0.00%	0.0749	0.00%	0	0.12%	**64.4**	**13%**	0.0878	0.00%	5.49	53%	0.101	0.00%
PVX_099105	♀	NM	hypothetical protein, conserved	**10000**	**9.50%**	0.143	0.00%	136	10%	0.567	0.00%	5.11	8.70%	0.131	0.00%	2610	17%
PVX_099190	♂-♀	NM	ribonuclease H2 subunit C, putative	**73.8**	**3.70%**	2360	81%	0.124	0.00%	19.2	24%	0.0923	0.00%	0.133	0.00%	1210	0.66%
PVX_099520	♂-♀	NM	ubiquitin-like protein, putative	0.835	0.00%	0.734	26%	**136**	**0.65%**	4000	0.81%	0.17	48%	0.779	0.00%	3330	1.10%
PVX_111180	♀	-	28 kDa ookinete surface protein, putative (P28)	6.39	27%	**6.95**	**20%**	4.65	19%	**63.8**	**8.90%**	**15.6**	**16%**	15	5.00%	**10000**	**4.80%**
PVX_111535	♂	NM	hypothetical protein, conserved	**40**	**4.80%**	5.31	14%	7.72	4.50%	0.822	26%	43.9	0.54%	0	1.80%	0.52	0.00%

(♀) Female, (♂) Male, (**NM**) Non-membrane.* *P. vivax* nomenclature taken from Carlton et al., 2009 Branches with significant signature of episodic selection are shown in bolted values.

Table S3-2. Episodic diversifying selection on gametocyte expressed genes.

Gene ID*	Model	log L	AICc	Total tree length	ω1	ω2	ω3
PVX_001080	Unconstrained Model	-22195.53	44465.25	8.2	0.202 (93%)	0.218 (6.1%)	3330 (0.42%)
	(background branches)				0.00129 (55%)	0.347 (41%)	2.89 (3.7%)
	Constrained Model	-22201.34	44474.87	6.55	0.00 (24%)	0.00 (54%)	1.00 (22%)
	Evidence of episodic diversifying selection in the divergence between *P. berghei*-*P. yoelli*, with LRT p-value of 0.003						
PVX_082500	Unconstrained Model	-1912.79	3901.54	9.29	0.00 (95%)	0.00 (3.7%)	3460 (1.1%)
	(background branches)				0.0229 (88%)	0.0310 (8.5%)	3.01 (3.9%)
	Constrained Model	-1920.37	3914.61	4.05	0.00 (95%)	0.0672 (0.0%)	1.00 (5.1%)
	Evidence of episodic diversifying selection on *P. berghei*, with LRT p-value of 0.001						
PVX_084240	Unconstrained Model	-1751.55	3579.74	7.93	0.0437 (82%)	0.00 (12%)	40.6 (6.2%)
	(background branches)				0.0467 (90%)	0.971 (0.0%)	1.51 (10%)
	Constrained Model	-1760.13	3594.76	7.81	0.00 (31%)	0.00 (39%)	1.00 (29%)
	Evidence of episodic diversifying selection on *P. cynomolgi*, with LRT p-value of 0.000						
PVX_086280	Unconstrained Model	-1969.94	4016.25	6.22	0.753 (63%)	1.00 (28%)	84.6 (9.2%)
	(background branches)				0.00 (47%)	0.434 (48%)	0.405 (5.0%)
	Constrained Model	-1980.77	4035.79	5.64	1.00 (2.8%)	0.00 (54%)	1.00 (43%)
	Evidence of episodic diversifying selection on *P. knowlesi*, with LRT p-value of 0.000						
PVX_088915	Unconstrained Model	-10556.68	21187.78	7.43	0.137 (83%)	0.139 (15%)	51.3 (1.4%)
	(background branches)				0.0426 (77%)	0.454 (16%)	1.48 (7.0%)
	Constrained Model	-10569.96	21212.33	7.15	0.00 (81%)	0.0189 (0.0%)	1.00 (19%)
	Evidence of episodic diversifying selection on *P. cynomolgi*, with LRT p-value of 0.000						
PVX_089245	Unconstrained Model	-4939.45	9953.94	46.36	0.107 (86%)	0.762 (12%)	10000 (2.5%)
	(background branches)				0.0261 (63%)	0.481 (33%)	2.02 (3.2%)
	Constrained Model	-4945.73	9964.44	6.8	0.00 (77%)	0.921 (0.0%)	1.00 (23%)
	Evidence of episodic diversifying selection on *P. berghei*, with LRT p-value of 0.002						

Table S3-2. Episodic diversifying selection on gametocyte expressed genes (continued).

Gene ID*	Model	log L	AICc	Total tree length	ω1	ω2	ω3
PVX_094335	Unconstrained Model	-5993.48	12061.61	8.32	0.110 (92%)	0.177 (7.2%)	8210 (0.43%)
	(background branches)				0.0220 (82%)	0.366 (16%)	3.32 (2.2%)
	Constrained Model	-6000.18	12072.97	6.01	0.00 (86%)	0.931 (0.0%)	1.00 (14%)
	Evidence of episodic diversifying selection on *P. berghei*, with LRT p-value of 0.001						
PVX_096350	Unconstrained Model	-11453.59	22981.7	12.24	0.128 (91%)	0.130 (8.3%)	3330 (0.59%)
	(background branches)				0.00 (44%)	1.00 (19%)	0.262 (38%)
	Constrained Model	-11463.52	22999.53	8.8	0.0664 (0.0%)	0.00 (84%)	1.00 (16%)
	Evidence of episodic diversifying selection on *P. reichenowi*, with LRT p-value of 0.000						
PVX_098665	Unconstrained Model	-1878.25	3832.45	4.5	1.00 (73%)	1.00 (22%)	40.2 (5.1%)
	(background branches)				0.00 (74%)	0.210 (25%)	3.57 (1.3%)
	Constrained Model	-1881.22	3836.28	5.08	1.00 (4.6%)	0.00 (60%)	1.00 (36%)
	NO evidence of episodic diversifying selection on *P. chabaudi*, with LRT p-value of 0.052						
PVX_099105	Unconstrained Model	-666.22	1413.17	58.47	0.247 (83%)	1.00 (7.3%)	10000 (9.5%)
	(background branches)				0.00 (87%)	0.00777 (0.0%)	0.920 (13%)
	Constrained Model	-674.05	1426.45	6.68	0.00 (70%)	0.932 (0.0%)	1.00 (30%)
	Evidence of episodic diversifying selection on *P. cynomolgi*, with LRT p-value of 0.000						
PVX_099190	Unconstrained Model	-3307.38	6690.29	6.81	0.340 (81%)	0.358 (16%)	70.5 (3.7%)
	(background branches)				0.00 (60%)	0.581 (23%)	0.479 (17%)
	Constrained Model	-3313.3	6700.05	6.3	0.00 (4.8%)	0.00 (59%)	1.00 (37%)
	Evidence of episodic diversifying selection on *P. cynomolgi*, with LRT p-value of 0.003						
PVX_099520	Unconstrained Model	-9092.72	18259.96	10.83	0.375 (90%)	0.376 (9.5%)	129 (0.63%)
	(background branches)				0.0115 (70%)	0.496 (29%)	91.3 (0.27%)
	Constrained Model	-9098.63	18269.76	10.76	0.00 (21%)	0.00 (43%)	1.00 (35%)
	Evidence of episodic diversifying selection on *P. berghei*, with LRT p-value of 0.003						

Table S3-2. Episodic diversifying selection on gametocyte expressed genes (continued).

Gene ID*	Model	log L	AICc	Total tree length	ω1	ω2	ω3
PVX_111180	Unconstrained Model	-3184.61	6445.02	18.44	0.297 (60%)	0.413 (23%)	9.67 (16%)
	(background branches)				0.0330 (73%)	1.00 (17%)	11.4 (10%)
	Constrained Model	-3190.5	6454.71	17.67	0.00 (38%)	0.00 (2.6%)	1.00 (59%)
	Evidence of episodic diversifying selection on *P. knowlesi*, with LRT p-value of 0.003						
	Unconstrained Model	-3183.6	6443.01	17.06	0.579 (70%)	0.577 (20%)	72.3 (9.5%)
	(background branches)				0.0334 (74%)	1.00 (13%)	7.35 (14%)
	Constrained Model	-3195.37	6464.46	16.16	1.00 (3.3%)	0.00 (55%)	1.00 (42%)
	Evidence of episodic diversifying selection on *P. chabaudi*, with LRT p-value of 0.000						
	Unconstrained Model	-3186.06	6447.92	17.33	0.00 (75%)	0.00 (8.0%)	14.3 (17%)
	(background branches)				0.0318 (72%)	1.00 (16%)	9.05 (12%)
	Constrained Model	-3192.39	6458.49	16.66	0.00 (37%)	0.00 (4.1%)	1.00 (59%)
	Evidence of episodic diversifying selection on *P. yoelii*, with LRT p-value of 0.002						
	Unconstrained Model	-3182.22	6440.25	30.96	1.00 (75%)	1.00 (20%)	10000 (4.7%)
	(background branches)				0.0344 (73%)	1.00 (13%)	7.73 (14%)
	Constrained Model	-3189.24	6452.19	18.83	0.923 (0.0%)	0.922 (0.0%)	1.00 (100%)
	Evidence of episodic diversifying selectionin the divergence between *P. berghei*-*P. yoelli*, with LRT p-value of 0.001						
PVX_111535	Unconstrained Model	-4108.97	8293.02	7.22	0.393 (76%)	0.395 (19%)	52.2 (4.8%)
	(background branches)				0.0359 (70%)	0.243 (23%)	2.36 (6.9%)
	Constrained Model	-4118.55	8310.12	6.33	0.00 (4.9%)	0.00 (51%)	1.00 (44%)
	Evidence of episodic diversifying selection on *P. cynomolgi*, with LRT p-value of 0.000						

* *P. vivax* nomenclature taken from Carlton et al., 2009

Table S3-3. Distribution of positively selected sites (REL) in genes with and without known *P. falciparum* epitopes.

Category	# genes	# genes +	Prop. Genes +	# NE genes +	Prop. NE genes +	Prop. + sites in NE genes	# E genes	Prop. E genes	# E genes +	Prop. E genes +	Prop. E genes + from E	# E genes + inside	Prop. E genes + inside from E +	Prop. + sites in E genes	Prop. + sites in E genes inside	Prop. + sites in E genes outside
Female (♀)	146	32	0.219	20	0.625	0.034	25	0.171	12	0.375	0.480	5	0.417	0.025	0.034	0.042
Male (♂)	151	49	0.325	20	0.408	0.022	46	0.305	29	0.592	0.630	13	0.448	0.013	0.015	0.015
Male-female common	84	23	0.274	10	0.435	0.019	28	0.333	13	0.565	0.464	4	0.308	0.019	0.037	0.035
Membrane	35	6	0.171	3	0.500	0.035	6	0.171	3	0.500	0.500	1	0.333	0.019	0.022	0.017
Non-membrane	210	57	0.271	22	0.386	0.015	62	0.295	35	0.614	0.565	15	0.429	0.019	0.021	0.025
♀ Membrane	15	2	0.133	2	1.000	0.038	1	0.067	0	0.000	0.000	0	0.000	0.000	0.000	0.000
♀ Non-membrane	38	12	0.316	9	0.750	0.031	5	0.132	3	0.250	0.600	2	0.667	0.050	0.040	0.058
♀ N/A	93	18	0.194	9	0.500	0.036	19	0.204	9	0.500	0.474	3	0.333	0.014	0.028	0.018
♂ Membrane	9	2	0.222	0	0.000	0.000	2	0.222	2	1.000	1.000	1	0.500	0.027	0.022	0.017
♂ Non-membrane	104	25	0.240	5	0.200	0.022	32	0.308	20	0.800	0.625	9	0.450	0.013	0.015	0.016
♂ N/A	38	22	0.579	15	0.682	0.021	12	0.316	7	0.318	0.583	3	0.429	0.011	0.011	0.009
♂-♀ common Membrane	11	2	0.182	1	0.500	0.020	3	0.273	1	0.500	0.333	0	0.000	0.008	0.000	0.000
♂-♀ common Non-membrane	68	20	0.294	8	0.400	0.019	25	0.368	12	0.600	0.480	4	0.333	0.021	0.037	0.035
♂-♀ common N/A	5	1	0.200	1	1.000	0.007	0	0.000	0	0.000	0.000	0	0.000	0.000	0.000	0.000
Total	**381**	**104**	**0.273**	**50**	**0.481**	**0.025**	**99**	**0.260**	**54**	**0.519**	**0.545**	**22**	**0.407**	**0.017**	**0.021**	**0.022**

Legend. (♀) Female, (♂) Male, (**# genes**) Total number of genes, (**# genes +**) Total number of genes with positive selected sites, (**Prop. Genes +**) Proportion of genes with positive selected sites, (**# NE genes +**) Number of genes without immune epitopes with positive selected sites, (**Prop. NE genes +**) Proportion of genes without immune epitopes with positive selected sites, (**Prop. + sites in NE genes**) Proportion of positive selected sites in genes without immune epitopes, (**# E genes**) Total number of genes with immune epitopes, (**Prop. E genes**) Proportion of genes with immune epitopes, (**# E genes +**) Number of genes with immune epitopes with positive selected sites, (**Prop. E genes +**) Proportion of genes with immune epitopes and positive selected sites, (**Prop. E genes + from E**) Proportion of genes with immune epitopes and positive selected sites from genes with immune epitopes, (**# E genes + inside**) Number of genes with immune epitopes with positive selected sites inside putative immune epitope region, (**Prop. E genes + inside from E +**) Proportion of genes with immune epitopes and positive selected sites inside putative immune epitope region from genes with immune epitopes, (**Prop. + sites in E genes**) Proportion of positive selected sites in genes with immune epitopes, (**Prop. + sites in E genes inside**) Proportion of positive selected sites inside putative immune epitopes region, (**Prop. + sites in E genes outside**) Proportion of positive selected sites outside putative immune epitopes region

APPENDIX C

SUPPLEMENTARY DATA FOR CHAPTER 4

Tables

Table S4-1. RELAX results for GCMFs with positively selected branches.

Multigene family	Test for selection	Model	*log* L	# par	AIC_c	L_{tree}	Branch	ω_+	p_+	ω_n	p_n	ω_-	p-
Acyl-CoA synthase	**Intensification (K = 4.56) significant (p = 7.10E-09, LR = 33.51)**	Partitioned MG94xREV	-35696.7	93	71580.1	7.56	Reference	0.171	100%				
							Test	0.45	100%				
		General Descriptive	-34883.1	172	70112.6	308.9	All	0.0102	78%	0.943	21%	104	0.57%
		Null	-35101.7	96	70396.1	4662	Reference	9.86E-05	71%	0.443	25%	3.51	3.5%
							Test	9.86E-05	71%	0.443	25%	3.51	3.5%
		Alternative	-35084.9	97	70364.6	4486	Reference	9.00E-05	67%	0.287	26%	2.12	6.9%
							Test	3.5E-19	67%	3.36E-03	26%	30.8	6.9%
		Partitioned Exploratory	-35084.8	101	70372.5	4495	Reference	9.06E-05	68%	0.306	26%	2.27	6.2%
							Test	6.75E-15	67%	1.4E-12	26%	30.3	7%
Alpha beta hydrolase putative 2	**Intensification (K = 1.88), significant (p = 0.022, LR = 5.23)**	Partitioned MG94xREV	-7568.94	37	15212.5	2.44	Reference	0.18	100%				
							Test	0.36	100%				
		General Descriptive	-7403.53	60	14928.6	96.73	All	0.0145	79%	0.967	20%	71.4	0.97%
		Null	-7416.69	40	14914.1	22.02	Reference	0	46%	0.0869	45%	3.34	8.4%
							Test	0	46%	0.0869	45%	3.34	8.4%
		Alternative	-7414.08	41	14910.9	21.33	Reference	0	46%	0.0837	45%	2.86	9%
							Test	0	46%	9.34E-03	45%	7.25	9%
		Partitioned Exploratory	-7413.93	45	14918.7	21.19	Reference	0	46%	0.0865	45%	2.94	8.7%
							Test	1.79E-12	43%	2.91E-08	47%	6.62	10%
Asparagine tRNA ligase	Relaxation (K = 0.91), significant (p = 0.572, LR = 0.32)	Partitioned MG94xREV	-14897	55	29904.6	5.5	Reference	0.104	100%				
							Test	0.482	100%				
		General Descriptive	-14504.9	96	29203.6	137.9	All	1.88E-03	80%	0.9	20%	592	0.64%
		Null	-14603.2	58	29323.1	302.2	Reference	0.0205	90%	1	10%	1470	0.37%
							Test	0.0205	90%	1	10%	1470	0.37%
		Alternative	-14603	59	29324.8	306.1	Reference	0.02	90%	1	10%	3330	0.37%
							Test	0.0281	90%	1	10%	1640	0.37%
		Partitioned Exploratory	-14593.1	63	29313	221.4	Reference	0.0214	90%	0.998	9.80%	1470	0.27%
							Test	0	72%	1	23%	117	4.7%
Cell division protein FtsH	Relaxation (K = 1.00) significant (p = 0.986, LR = 0.00)	Partitioned MG94xREV	-14778.3	77	29711.5	10.35	Reference	0.0267	100%				
							Test	0.412	100%				
		General Descriptive	-14462	140	29206.7	73.05	All	2.24E-03	80%	0.477	20%	933	0.02%
		Null	-14525.9	80	29212.7	406.8	Reference	5.99E-04	78%	0.0572	21%	1	1.4%
							Test	5.99E-04	78%	0.0572	21%	1	1.4%
		Alternative	-14525.9	81	29214.8	406.9	Reference	5.99E-04	78%	0.0573	21%	1	1.4%
							Test	6.00E-04	78%	0.0573	21%	1	1.4%
		Partitioned Exploratory	-14524	85	29219	430.1	Reference	4.95E-04	75%	0.0482	24%	1	1.5%
							Test	3.45E-04	68%	6.25E-03	27%	1.09	5.2%

Table S4-1. RELAX results for GCMFs with positively selected branches (continued).

Multigene family	Test for selection	Model	*log* L	# par	AIC_c	L_{tree}	Branch set	ω_+	p_+	ω_n	p_n	ω_-	p-
Chaperonin	**Intensification (K = 37.05) significant (p = 1.07E-04, LR = 15.01)**	Partitioned MG94xREV	-1715.12	55	3544.13	7.7	Reference	0.0523	100%				
							Test	2.37	100%				
		General Descriptive	-1662.87	96	3529.82	24.54	All	1.29E-04	57%	0.778	43%	9980	0.06%
		Null	-1690.81	58	3501.96	41.23	Reference	0	22%	0.0401	76%	8.05	1.2%
							Test	0	22%	0.0401	76%	8.05	1.2%
		Alternative	-1683.31	59	3489.1	43.27	Reference	0	22%	0.0296	74%	1.14	4.2%
							Test	0	74%	0	22%	119	4.2%
		Partitioned Exploratory	-1680.85	63	3492.82	31.35	Reference	0	22%	0.046	77%	3.28	0.75%
							Test	0	1.8%	0	89%	143	8.7%
Conserved *Plasmodium* protein	**Relaxation (K = 0.36) significant (p = 6.08E-05, LR = 16.08)**	Partitioned MG94xREV	-7514.52	41	15111.9	3.13	Reference	0.274	100%				
							Test	1.27	100%				
		General Descriptive	-7397.57	68	14933.4	179.8	All	0.118	72%	0.704	27%	12.1	1%
		Null	-7444.29	44	14977.5	18.56	Reference	0.0668	75%	1	24%	52.4	0.54%
							Test	0.0668	75%	1	24%	52.4	0.54%
		Alternative	-7436.26	45	14963.5	101.6	Reference	0.0676	78%	1	22%	10000	0.4%
							Test	0.375	78%	1	22%	28.7	0.4%
		Partitioned Exploratory	-7428.97	49	14957.1	18.86	Reference	0.0641	78%	1	22%	63.9	0.37%
							Test	0.803	62%	0.897	34%	28.7	4.1%
Conserved *Plasmodium* protein unknown function 6	**Intensification (K = 1.41) significant (p = 0.0338, LR = 4.51)**	Partitioned MG94xREV	-6716.47	37	13507.8	2.62	Reference	0.684	100%				
							Test	0.808	100%				
		General Descriptive	-6629.17	60	13380.5	320.8	All	0.124	37%	0.744	57%	10.8	6%
		Null	-6649.03	40	13379	13.9	Reference	3.7E-15	34%	0.725	58%	7.94	8.6%
							Test	3.7E-15	34%	0.725	58%	7.94	8.6%
		Alternative	-6646.78	41	13376.6	13.28	Reference	0	32%	0.652	57%	5.58	11%
							Test	0	32%	0.546	57%	11.4	11%
		Partitioned Exploratory	-6644.99	45	13381.2	12.81	Reference	0	34%	0.846	60%	8.63	6%
							Test	0	43%	0.0000022	34%	5.83	23%
Conserved *Plasmodium* protein unknown function 12	**Intensification (K = 13.02) significant (p = 5.01E-07, LR = 25.26)**	Partitioned MG94xREV	-3970.18	37	8015.26	1.42	Reference	0.204	100%				
							Test	0.575	100%				
		General Descriptive	-3917.28	60	7956.92	6.25	All	0.101	90%	1	9.30%	9.89	0.89%
		Null	-3927.28	40	7935.61	6.5	Reference	0.065	85%	1	15%	62.1	0.26%
							Test	0.065	85%	1	15%	62.1	0.26%
		Alternative	-3914.65	41	7912.41	5.27	Reference	0.066	85%	1	12%	1.41	2.8%
							Test	4.33E-16	85%	1	12%	89.9	2.8%
		Partitioned Exploratory	-3914.42	45	7920.16	5.26	Reference	0.0656	85%	1	5.90%	1.06	9.3%
							Test	1.3E-16	61%	0.458	36%	103	2.9%

Table S4-1. RELAX results for GCMFs with positively selected branches (continued).

Multigene family	Test for selection	Model	*log* L	# par.	AIC_c	L_{tree}	Branch set	ω_+	p_+	ω_n	p_n	ω_-	p-
Conserved rodent malaria protein unknown function	**Intensification (K = 2.35) significant (p = 1.24E-07, LR = 27.96)**	Partitioned MG94xREV	-11340.5	51	22783.7	3.21	Reference	0.398	100%				
							Test	1.61	100%				
		General Descriptive	-11138.3	88	22454.7	419.9	All	0.144	71%	0.947	26%	7.34	2.9%
		Null	-11205.1	54	22519.1	32.26	Reference	0	25%	0.0983	55%	3.18	20%
							Test	0	25%	0.0983	55%	3.18	20%
		Alternative	-11191.2	55	22493.2	27.55	Reference	2.68E-16	25%	0.108	55%	2.63	20%
							Test	0	25%	5.31E-03	55%	9.75	20%
		Partitioned Exploratory	-11188.9	59	22496.7	30.15	Reference	0	24%	0.0931	54%	2.49	21%
							Test	1	11%	1	87%	132	2.2%
Cytoadherence-linked asexual protein (CLAG)	**Relaxation (K = 0.51), significant (p = 5.23E-06, LR = 20.75)**	Partitioned MG94xREV	-42634.4	63	85395.1	5.46	Reference	0.219	100%				
							Test	0.169	100%				
		General Descriptive	-41792.9	112	83810.7	85.97	All	2.15E-03	72%	0.958	27%	486	0.44%
		Null	-42007.8	66	84148	22.44	Reference	6.24E-03	62%	0.39	35%	4.79	3.20%
							Test	6.24E-03	62%	0.39	35%	4.79	3.2%
		Alternative	-41997.5	67	84129.3	20.04	Reference	1.07E-03	69%	0.49	28%	4.39	3.2%
							Test	0.0315	69%	0.697	28%	2.11	3.2%
		Partitioned Exploratory	-41975.8	71	84094.1	23.78	Reference	0	71%	0.501	25%	3.81	3.5%
							Test	8.06E-03	58%	0.605	38%	11.2	4.3%
DEAD DEAH box ATP dependent RNA helicase putative	**Intensification (K = 8.22) significant (p = 3.70E-09, LR = 34.78)**	Partitioned MG94xREV	-10738.5	57	21591.8	5.46	Reference	0.0671	100%				
							Test	0.171	100%				
		General Descriptive	-10478.8	100	21159.8	53.22	All	3.25E-04	82%	0.893	18%	3440	0.16%
		Null	-10532.4	60	21185.6	4151	Reference	3.80E-05	66%	0.0921	31%	2.08	3%
							Test	3.80E-05	66%	0.0921	31%	2.08	3%
		Alternative	-10515	61	21152.9	3995	Reference	3.88E-05	66%	0.0917	30%	1.41	3.6%
							Test	0	66%	2.98E-09	30%	16.5	3.6%
		Partitioned Exploratory	-10511.8	65	21154.7	3972	Reference	3.76E-05	65%	0.0661	29%	1	5.7%
							Test	0	78%	0.198	20%	54.3	2%
DER1 like protein	**Intensification (K = 8.03) significant (p = 1.22E-07, LR = 27.99)**	Partitioned MG94xREV	-4510.24	57	9135.98	12.41	Reference	0.0395	100%				
							Test	0.942	100%				
		General Descriptive	-4438.4	100	9081.44	44.01	All	0.0181	96%	0.905	4.10%	61.1	0.08%
		Null	-4466.22	60	9054.11	4953	Reference	9.47E-05	48%	0.0593	52%	83	0.33%
							Test	9.47E-05	48%	0.0593	52%	83	0.33%
		Alternative	-4452.22	61	9028.17	3327	Reference	8.18E-05	44%	0.0361	55%	2.4	1.8%
							Test	0	44%	2.64E-12	55%	1120	1.8%
		Partitioned Exploratory	-4451.77	65	9035.5	3309	Reference	8.24E-05	44%	0.0382	54%	2.34	1.6%
							Test	0	53%	4.78E-12	45%	2150	2.7%

Table S4-1. RELAX results for GCMFs with positively selected branches (continued).

Multigene family	Test for selection	Model	*log* L	# par.	AIC_c	L_{tree}	Branch set	ω_+	p_+	ω_n	p_n	ω_-	p-
Dipeptidyl amino peptidase putative (DPAP)	**Intensification (K = 3.47) significant (p = 2.85E-04, LR = 13.17)**	Partitioned MG94xREV	-14753.5	53	29613.6	4.95	Reference	0.138	100%				
							Test	0.259	100%				
		General Descriptive	-14434.6	92	29054.9	797.3	All	1.44E-04	63%	0.696	36%	9980	0.35%
		Null	-14520.7	56	29154.1	5132	Reference	1.97E-05	59%	0.21	38%	3.15	3.2%
							Test	1.97E-05	59%	0.21	38%	3.15	3.2%
		Alternative	-14514.1	57	29142.9	5276	Reference	1.96E-05	59%	0.201	37%	2.66	3.8%
							Test	4.71E-17	59%	3.82E-03	37%	29.8	3.8%
		Partitioned Exploratory	-14512.4	61	29147.7	6346	Reference	1.83E-05	58%	0.175	37%	2.12	4.7%
							Test	0	60%	4.25E-03	36%	229	3.9%
Heatshock protein 70	Intensification (K = 1.00) significant (p = 0.964, LR = 0.00)	Partitioned MG94xREV	-13700.1	61	27522.7	42.17	Reference	0.0131	100%				
							Test	2.17E-03	100%				
		General Descriptive	-13284.3	108	26786.2	10622	All	3.69E-04	78%	0.771	22%	3510	0.04%
		Null	-13387.1	64	26902.8	3537	Reference	1.89E-05	78%	0.028	21%	3.5	0.27%
							Test	1.89E-05	78%	0.028	21%	3.5	0.27%
		Alternative	-13387.1	65	26904.8	3552	Reference	1.87E-05	78%	0.028	22%	3.5	0.27%
							Test	1.87E-05	78%	0.028	22%	3.5	0.27%
		Partitioned Exploratory	-13385.9	69	26910.5	15785	Reference	1.82E-03	94%	0.0774	6.10%	3.99	0.23%
							Test	1.79E-05	77%	0.0288	0%	4.36	23%
Histone H3	Relaxation (K = 0.57) significant (p = 0.069, LR = 3.31)	Partitioned MG94xREV	-2076.83	57	4269.91	1.33	Reference	0.0151	100%				
							Test	0.349	100%				
		General Descriptive	-2034.58	100	4276.15	3.89	All	3.71E-03	92%	0.814	7.80%	331	0%
		Null	-2062.72	60	4247.93	22.14	Reference	0.0102	98%	0.0311	1.5%	1090	0.16%
							Test	0.0102	98%	0.0311	1.5%	1090	0.16%
		Alternative	-2061.06	61	4246.71	17.75	Reference	0.0106	98%	0.0107	1.5%	10000	0.12%
							Test	0.0764	98%	0.077	1.5%	182	0.12%
		Partitioned Exploratory	-2051.16	65	4235.25	4.51	Reference	8.91E-03	98%	0.0404	1.7%	1.45	0.29%
							Test	0.03	1.2%	0.0305	95%	182	3.6%
Hypothetical protein	**Intensification (K = 50.00) significant (p = 7.77E-16, LR = 65.06)**	Partitioned MG94xREV	-2280.61	37	4636.89	1.75	Reference	0.136	100%				
							Test	2.26	100%				
		General Descriptive	-2228.05	60	4580.52	11.17	All	0.0888	97%	1	0%	11.3	3.2%
		Null	-2256.83	40	4595.61	8.11	Reference	1.00E-04	0.44%	0.0891	97%	12.2	2.3%
							Test	1.00E-04	0.44%	0.0891	97%	12.2	2.3%
		Alternative	-2224.3	41	4532.65	6.66	Reference	0.0623	16%	0.0662	90%	1.15	8.7%
							Test	0	1.6%	0	90%	1000	8.7%
		Partitioned Exploratory	-2222.49	45	4537.46	6.92	Reference	0.0857	96%	0.49	1.5%	2.6	2.2%
							Test	0	1.7%	0	87%	11000000	12%

Table S4-1. RELAX results for GCMFs with positively selected branches (continued).

Multigene family	Test for selection	Model	*log* L	# par.	AIC_c	L_{tree}	Branch set	ω_+	p_+	ω_n	p_n	ω_-	p-
Meiotic recombination protein (DMC)	Relaxation(K = 0.70) significant (p = 0.109, LR = 2.57)	Partitioned MG94xREV	-7067.32	51	14237.5	7.76	Reference	0.0328	100%				
							Test	3.8	100%				
		General Descriptive	-6862.51	88	13903.6	75.87	All	1.47E-04	43%	0.679	57%	9990	0.27%
		Null	-6945.34	54	13999.6	742	Reference	7.67E-04	40%	0.032	60%	1460	0.53%
							Test	7.67E-04	40%	0.032	60%	1460	0.53%
		Alternative	-6944.06	55	13999.1	714.2	Reference	7.65E-04	39%	0.0321	60%	10000	0.51%
							Test	6.80E-03	39%	0.0914	60%	606	0.51%
		Partitioned Exploratory	-6911.27	59	13941.7	84.24	Reference	0	34%	0.0163	64%	5.31	1.2%
							Test	0.114	93%	0.515	2.9%	606	4.3%
Methyltransferase	Intensification (K = 1.80) significant (p = 0.077, LR = 3.12)	Partitioned MG94xREV	-4937.78	53	9983.02	6.28	Reference	0.12	100%				
							Test	0.592	100%				
		General Descriptive	-4832.79	92	9853.98	39.7	All	0.0129	77%	0.879	23%	88.2	0.13%
		Null	-4858.78	56	9831.18	574.1	Reference	1.23E-03	73%	0.437	27%	35.3	0.20%
							Test	1.23E-03	73%	0.437	27%	35.3	0.20%
		Alternative	-4857.22	57	9830.13	550.8	Reference	1.21E-03	73%	0.42	27%	8.21	0.57%
							Test	5.76E-06	73%	0.21	27%	43.9	0.57%
		Partitioned Exploratory	-4844.68	61	9813.29	75.5	Reference	0	20%	0.0313	68%	1	12%
							Test	0.286	66%	0.287	33%	1050	1.2%
NIMA related protein kinase (NEK)	**Relaxation (K = 0.63) significant (p = 0.024, LR = 5.12)**	Partitioned MG94xREV	-10324.9	91	20833.7	21.51	Reference	0.037	100%				
							Test	2.92	100%				
		General Descriptive	-9994.56	168	20331.4	298.4	All	4.15E-03	64%	0.42	36%	574	0.09%
		Null	-10168	94	20525.9	26667	Reference	4.14E-05	66%	0.101	34%	1470	0.12%
							Test	4.14E-05	66%	0.101	34%	1470	0.12%
		Alternative	-10165.4	95	20522.8	25156	Reference	4.14E-05	65%	0.099	35%	10000	0.11%
							Test	1.78E-03	65%	0.234	35%	323	0.11%
		Partitioned Exploratory	-10141.4	99	20483.1	6030	Reference	3.98E-05	64%	0.084	36%	3.16	0.4%
							Test	0.394	1.50%	0.395	95%	323	3.5%
P28	**Intensification (K = 2.00) significant (p = 0.005, LR = 8.01)**	Partitioned MG94xREV	-4455.56	37	8986.3	2.06	Reference	0.518	100%				
							Test	1.71	100%				
		General Descriptive	-4344.73	60	8812.58	347.9	All	0.0621	62%	0.973	29%	16.5	9.3%
		Null	-4362.3	40	8805.99	15.13	Reference	0.0341	68%	1	21%	8.28	11%
							Test	0.0341	68%	1	21%	8.28	11%
		Alternative	-4358.3	41	8800.05	14	Reference	0.0315	67%	1	20%	6.26	13%
							Test	9.91E-04	67%	1	20%	39.3	13%
		Partitioned Exploratory	-4356.93	45	8805.61	13.98	Reference	0.0391	70%	1	16%	5.6	14%
							Test	0.999	27%	1	63%	123	9.7%

Table S4-1. RELAX results for GCMFs with positively selected branches (continued).

Multigene family	Test for selection	Model	*log* L	# par.	AIC_c	L_{tree}	Branch set	ω_+	p_+	ω_n	p_n	ω_-	p-
Papain	**Intensification (K = 3.73) significant (p = 4.04E-06, LR = 21.24)**	Partitioned MG94xREV	-20869.5	83	41906.1	7.48	Reference	0.221	100%				
							Test	0.753	100%				
		General Descriptive	-20440.3	152	41188.2	216.9	All	1.06E-04	68%	0.947	31%	9980	0.78%
		Null	-20552.7	86	41278.5	96.74	Reference	2.33E-03	60%	0.329	34%	2.9	6.3%
							Test	2.33E-03	60%	0.329	34%	2.9	6.3%
		Alternative	-20542.1	87	41259.3	40	Reference	0	41%	0.12	46%	1.63	13%
							Test	0	41%	3.67E-04	46%	6.18	13%
		Partitioned Exploratory	-20540.4	91	41264.1	38.99	Reference	0	42%	0.142	47%	1.92	11%
							Test	0.164	28%	0.271	64%	7.69	7.7%
Plasmepsin	**Intensification (K = 1.33) significant (p = 7.316E-9, LR = 33.45)**	Partitioned MG94xREV	-21423.2	133	43114.5	16.23	Reference	0.11	100%				
							Test	0.0765	100%				
		General Descriptive	-20781.8	252	42075.3	2158	All	1.19E-04	78%	0.844	22%	9980	0.18%
		Null	-21026.1	136	42326.3	21851	Reference	3.33E-05	73%	0.371	27%	1090	0.12%
							Test	3.33E-05	73%	0.371	27%	1090	0.12%
		Alternative	-21009.3	137	42294.9	27151	Reference	4.74E-05	75%	0.436	25%	1040	0.1%
							Test	1.79E-06	75%	0.332	25%	10300	0.1%
		Partitioned Exploratory	-20996.5	141	42277.3	28952	Reference	4.77E-05	75%	0.437	25%	904	0.13%
							Test	2.35E-03	88%	0.869	12%	18.1	0%
Serine repeat antigen SERA	Relaxation (K = 0.52) significant (p = 0, LR = 92.78)	Partitioned MG94xREV	-79776.4	159	159872	13.89	Reference	0.244	100%				
							Test	0.629	100%				
		General Descriptive	-77836.3	304	156285	775.4	All	1.07E-04	60%	0.931	39%	10000	1.20%
		Null	-78517.7	162	157360	254.3	Reference	0.0221	73%	1	26%	585	0.49%
							Test	0.0221	73%	1	26%	585	0.49%
		Alternative	-78471.3	163	157270	381.6	Reference	0.0156	74%	1	26%	10000	0.52%
							Test	0.113	74%	1	26%	124	0.52%
		Partitioned Exploratory	-78350.4	167	157036	70.5	Reference	0.0256	76%	1	23%	46	0.58%
							Test	0	47%	1	45%	28.8	8.20%

Table S4-2. List of recombinant sequences, number of recombination events, and length of recombinant segments in GCMFs with significant recombination signal.

Multigene family	Length recombinant segment	Sequence code Recombinant 1	Sequence code Recombinant 2
Actin	91	PIN1 PVX_085830 PCYB_134420 PKH_133510 PCOA1	PGO2 PYYM_1463300
	88	PBANKA_145930 PCHAS_146160 PYYM_1463300	PCHAS_103090
	373	PBANKA_145930 PIN2 PGO2 PCHAS_146160 PYYM_1463300 PF3D7_1246200 PRCDC_1245600	PCHAS_103090 PBANKA_103010
	190	PYYM_1032200	PCHAS_103090
Acyl-CoA synthase	413	PF3D7_1479000 PF3D7_0301000 PRCDC_0935600	PF3D7_1253400 PRCDC_1370300
	410	PF3D7_1479000	PRCDC_1370300
	198	PRCDC_1370500 PF3D7_1200700 PRCDC_1200100	PRCDC_1370300
	1620	PCHAS_145560 PBANKA_145330 PYYM_1457300 PF3D7_1238800 PRCDC_1238000 PVX_100890 PKH_145350 PGO2	PF3D7_1479000 PGO1 PF3D7_1200700 PRCDC_1200100 PRCDC_1370500 PRCDC_0214200 PRCDC_0935400 PRCDC_0935600 PF3D7_1253400 PRCDC_1370300 PRCDC_1476900 PF3D7_1477900 PRCDC_0728400
	401	PRCDC_1370500 PF3D7_1200700	PF3D7_0301000
	302	PRCDC_0935400	PRCDC_1370300 PF3D7_1372400 PF3D7_1253400
	86	PRCDC_0935600	PF3D7_1253400 PF3D7_1372400
	1256	PF3D7_1200700	PRCDC_0728400
Adrenoxin reductase (SV)	121	PKH_093720	PVX_092585 PVXcontig7021 PVX_202290

Table S4-2. List of recombinant sequences, number of recombination events, and length of recombinant segments in GCMFs with significant recombination signal (continued).

Multigene family	Length recombinant segment	Sequence code Recombinant 1	Sequence code Recombinant 2
Alpha beta hydrolase putative 2	50	PRCDC_1400600 PF3D7_1401300	PIN1 PVX_089050 PKH_050460 PCOA1
	946	PF3D7_0826200 PRCDC_0825500	PF3D7_1401300 PRCDC_1400600
	128	PYYM_0704600 PCHAS_093360	PRCDC_1400600 PF3D7_1401300
	75	PGO1	PKH_050460 PVX_089050 PIN1 PCOA1
	77	PF3D7_1401300	PIN1
Asparagine tRNA ligase	38	PBANKA_030860 PCHAS_031080	PCHAS_110890
	350	PVX_002940 PIN1	PRCDC_0508800 PF3D7_0509600
	27	PGO1	PGO2
	206	PKH_102330	PVX_098040
Biotin acetylCoA carboxylase	468	PRCDC_1459300	PBANKA_132360
Calcium transporting ATPase putative SERCA	104	PVX_081455	PGO1
Calcium Dependent Portein Kinase (CDPK)	252	PRCDC_1336800 PCYB_122510 PGO4 PBANKA_135150 PCHAS_135610 PYYM_1353200 PF3D7_1337800	PF3D7_0717500 PVX_000555 PCYB_032120 PKH_030080 PIN1 PGO1 PYYM_0617000 PBANKA_061520 PCHAS_061690 PRCDC_0714800
	289	PCYB_032120 PVX_000555 PIN1 PCOA1	PBANKA_135150 PCHAS_135610
Chromatin assembly factor 1 subunit	26	PRCDC_1432600 PBANKA_101160 PCHAS_101240 PF3D7_1433300 PYYM_1013100 PCYB_132560 PCOA3 PGO3	PF3D7_1329300 PRCDC_1328300
	200	PCYB_123320	PF3D7_1433300 PRCDC_1432600
ClpBprotein	52	PGO1	PYYM_0714300 PBANKA_071420 PCHAS_072330
	484	PIN2	PF3D7_1116800

Table S4-2. List of recombinant sequences, number of recombination events, and length of recombinant segments in GCMFs with significant recombination signal (continued).

Multigene family	Length recombinant segment	Sequence code Recombinant 1	Sequence code Recombinant 2
Conserved *Plasmodium* protein unknown function 4	257	PIN1 PCYB_126530 PVX_231290 PVX_117995 PKH_125650	PF3D7_1449800 PRCDC_1449100
Conserved *Plasmodium* protein unknown function 2	90	PGO2 PVX_101220	PCHAS_131760 PBANKA_131430 PYYM_1315100
	64	PF3D7_1246500 PRCDC_1245900	PCHAS_131760 PBANKA_131430
Conserved *Plasmodium* protein unknown function 6	198	PCYB_113750 PVX_114125	PKH_112930
	316	PCYB_113750	PCOA1
	31	PF3D7_0620000 PRCDC_0618400	PCOA A PCOA B PKH_112930 PKH_112990
Conserved *Plasmodium* protein unknown function	60	PRCDC_1104200 PVX_090965 PCYB_091270 PKH_090290 PIN1 PCOA1 PGO1 PBANKA_094120 PCHAS_090310 PF3D7_1105700	PGO2
Conserved Rodent malaria protein unknown function	754	PBANKA_080030	PYYM_1001800
Cysteine Repeat Modular Protein (CRMP)	6554	PBANKA_061590 PCHAS_061760 PYYM_0617700	PIN1 PVX_099005 PCYB_071910 PKH_070870
	160	PVX_096410	PYYM_0815400
	103	PBANKA_081240	PGO1
	368	PCHAS_061760 PYYM_0617700	PF3D7_0718300 PRCDC_0715500
Cytoadherence-linked asexual protein (CLAG)	192	PF3D7_0935800 PRCDC_0933900	PKH_073370 PGO1 PCOA1 PIN1 PVX_086930
	99	PRCDC_0830800 PF3D7_0831600	PYYM_0839400 PBANKA_083630 PCHAS_083660
	46	PF3D7_0935800	PRCDC_0219700 PF3D7_0220800

Table S4-2. List of recombinant sequences, number of recombination events, and length of recombinant segments in GCMFs with significant recombination signal (continued).

Multigene family	Length recombinant segment	Sequence code Recombinant 1	Sequence code Recombinant 2
DEAD DEAH box ATP dependent RNA helicase putative	34	PCHAS_131290 PBANKA_130970 PYYM_1310500 PF3D7_1445900 PRCDC_1445200	PIN1 PVX_118190 PKH_126040 PCOA1
	12	PCOA2 PVX_123985 PCYB_145280 PKH_144390 PIN2	PYYM_1310500
	1186	PRCDC_1445200 PGO1 PF3D7_1445900	PCHAS_144390 PBANKA_144190 PYYM_1446000
DHHC type zinc finger protein	378 512	PBANKA_051200 PCYB_114680	PKH_113920 PCHAS_010890
Dipeptidyl amino peptidase putative DPAP	51	PCOA1 PVX_091465 PCYB_092280 PKH_091410 PIN1 PBANKA_093130 PCHAS_091300 PYYM_0932700 PF3D7_1116700 PRCDC_1115100	PVX_101280 PCYB_147220 PKH_146510
	183	PIN1 PVX_091465 PCYB_092280	PCHAS_146300 PBANKA_146070 PYYM_1464700
DNA directed RNA polymerase II	38	PCOA2 PVX_082395 PCYB_123350 PKH_122370 PIN2	PBANKA_080700 PCHAS_080730
	47	PCOA1	PRCDC_1328000 PF3D7_1329000
Dynein heavy chain	30	PRCDC_0726700 PKH_021460 PIN1 PF3D7_0729900	PGO2
	577	PF3D7_0729900	PGO1
Elongation factor Tu putative tufA	402	PCOA1	PIN1
Eukaryotic initiation factor 2a	570	PBANKA_061080 PCHAS_061250	PKH_131130 PCOA1
	28	PYYM_0612400 PBANKA_061080 PCHAS_061250	PIN1
Exonuclease	109	PCHAS_081090	PCYB_071720
Glutathione reductase putative GR	46	PGO2	PKH_072100 PCYB_073140

Table S4-2. List of recombinant sequences, number of recombination events, and length of recombinant segments in GCMFs with significant recombination signal (continued).

Multigene family	Length recombinant segment	Sequence code Recombinant 1	Sequence code Recombinant 2
Heatshock protein 40	593	PYYM_0310800	PKH_127060
	364	PYYM_0310800	PIN2
Heatshock protein 90	26	PRCDC_1221600 PVX_123745 PCYB_144800 PKH_143880 PIN2 PCOA2 PGO2 PBANKA_143730 PCHAS_143930 PYYM_1441400 PF3D7_1222300	PCOA1 PVX_087950 PIN1 PCYB_011550 PBANKA_080570 PCHAS_080600 PYYM_0808700 PF3D7_0708400 PRCDC_0706600
	1386	PYYM_1441400 PBANKA_143730	PF3D7_1222300 PRCDC_1221600
	44	PBANKA_080570 PCHAS_080600 PYYM_0808700	PIN2 PCYB_144800
Histone H2B	84	PYYM_0943400 PVX_090935 PCOA1 PCHAS_090250 PRCDC_1103600	PVX_122930
Histone H3	19	PGO1 PCYB_113960 PBANKA_111710 PCHAS_111660 PYYM_1119100 PF3D7_0617900 PRCDC_0616300	PVX_113665 PGO2 PKH_113870 PIN2
	89	PVX_114020	PGO2 PYYM_0109800
Hypothetical protein	368	PBANKA_131130 PCHAS_131460 PYYM_1312100	PCOA1 PVX_118120 PCYB_126760 PKH_125890 PIN1 PGO1
Inorganic pyrophosphatase VP	45	PVX_100710	PKH_145370
Iron sulfur assembly protein SufA	100	PKH_041280	PVX_080115
Kinesin 8	1240	PVX_081250 PIN1 PCOA1 PKH_020210	PYYM_0205600

Table S4-2. List of recombinant sequences, number of recombination events, and length of recombinant segments in GCMFs with significant recombination signal (continued).

Multigene family	Length recombinant segment	Sequence code Recombinant 1	Sequence code Recombinant 2
Lysophospholipase	236	PCOA3 PGO1 PVX_112700 PCYB_002250	PCYB_053800 PVX_090280 PKH_052800 PCOA5 PCOA4
	68	PKH_010790 PVX_088015 PIN1 PCOA1	PCOA2 PIN2
	80	PIN3	PCOA2 PVX_090280 PCYB_053800 PKH_052800
	48	PCOA1	PCHAS_122100 PBANKA_122030 PYYM_1223000
	63	PIN3	PCOA3 PCOA6 PCOA2
	51	PIN3	PVX_112700
Meiotic recombination protein DMC	39	PVX_091045 PIN2 PKH_090470 PCOA2 PGO2 PBANKA_093950 PCHAS_090480 PYYM_0941100 PF3D7_1107400 PRCDC_1105900	PKH_051570 PVX_089570 PBANKA_071400 PCHAS_072310 PYYM_0714100 PF3D7_0816800 PRCDC_0816100
NADP specific glutamate dehydrogenase putative GDH	53	PCOA1 PCYB_132800 PKH_131960	PIN2 PVX_085625 PCYB_134020 PGO1
	64	PKH_131960 PCOA1	PIN2 PVX_085625 PCYB_134020
	571	PKH_131960	PCYB_132800
	232	PIN2 PVX_085625 PCYB_134020	PGO1

Table S4-2. List of recombinant sequences, number of recombination events, and length of recombinant segments in GCMFs with significant recombination signal (continued).

Multigene family	Length recombinant segment	Sequence code Recombinant 1	Sequence code Recombinant 2
NIMA related protein kinase (NEK)	17	PBANKA_061670 PVX_096360 PCYB_032330 PKH_031300 PIN3 PCOA3 PCHAS_061840 PYYM_0618500 PF3D7_0719200 PRCDC_0716400	PVX_079950 PKH_100620 PGO4
	213	PRCDC_1201000	PIN3
Novel putative transporter 1 NPT1	1036	PIN1 PKH_020840 PCOA1	PYYM_0211200 PBANKA_020830 PCHAS_114680 PCHAS_020670
	172	PF3D7_0104800 PRCDC_0102700	PKH_020840 PCYB_021900
Nucleotide binding protein	38	PVX_092045 PCYB_093480 PKH_092640 PCOA2 PF3D7_1128500	PIN1 PVX_098980
	97	PRCDC_1127000 PVX_092045 PCHAS_092470 PF3D7_1128500	PF3D7_0910800
P1s1 nuclease	723	PBANKA_103060 PCHAS_103140 PYYM_1032700	PGO3
	113	PKH_133560 PCOA1	PCOA3
P28	406	PCYB_007100 PVX_111180 PKNH_0615600 PIN	PF3D7_1030900
	171	PCYB_007100 PVX_111180	PCHAS_0515000

Table S4-2. List of recombinant sequences, number of recombination events, and length of recombinant segments in GCMFs with significant recombination signal (continued).

Multigene family	Length recombinant segment	Sequence code Recombinant 1	Sequence code Recombinant 2
Papain	308	PKH_091260 PCOA3	PVX_091405
	114	PCOA3 PCOA2 PKH_091260	PVX_091405
	216	PKH_091260 PCOA3	PVX_091415 PIN3
	190	PCYB_092160 PVX_091410 PIN2	PGO3
	44	PKH_091240	PCOA3
	130	PVX_091405	PVX_091410 PIN2
	633	PCYB_125710 PKH_124810	PBANKA_132170
Phosducin like protein PhLP	52	PBANKA_120480 PCHAS_120550 PYYM_1207400	PYYM_0520200 PBANKA_051970 PCHAS_051990
Plasmepsin	108	PRCDC_1407400	PCHAS_101530
Pre mRNA splicing factor ATP dependent RNA helicase PRP22	274	PKH_071490 PIN2 PCOA2	PYYM_1142600 PBANKA_114020 PCHAS_113970
Pre mRNA splicing helicase	121	PRCDC_0419700 PF3D7_0422500	PGO1 PKH_052600
	131	PKH_052600 PVX_090165	PGO1
	100	PF3D7_1439100 PBANKA_130300 PCHAS_130620 PYYM_1303800 PRCDC_1438400	PCYB_053590 PVX_090165
Protein phosphatase 2C	26	PCYB_141870 PVX_122290 PKH_140810 PIN1 PCOA1	PRCDC_1308200
Rhoptry associated protein 23	233	PBANKA_110140 PYYM_1103600	PKH_103190 PGO1

Table S4-2. List of recombinant sequences, number of recombination events, and length of recombinant segments in GCMFs with significant recombination signal (continued).

Multigene family	Length recombinant segment	Sequence code Recombinant 1	Sequence code Recombinant 2
SERA	746	Pi18610-18643 PCYB_042210 PCYB_042230 PCYB_042280 PVX_003840 PVX_003830 PVX_003805 Pc12410-12443	PYYM_0305900 PBANKA_030490 PBANKA_030500 PCHAS_030720 PF3D7_0207600 PF3D7_0207700 PF3D7_0207800 PF3D7_0208000 PRCDC_0206600 PRCDC_0206700 PRCDC_0206800 PRCDC_0206900 PYYM_0305800 PYYM_0306000
	834	PKH_041210 PCOA1	PVX_003795 PCYB_042300 PKH_041260 Pi41952-41985 PCOA2
	697	PCOA1	PVX_003845 PCYB_042200
	337	PCOA3	PKH_041200
	236	PVX_003810 PCYB_042270 PKH_041230 PIN3 PCOA3	PVX_003840 PCYB_042210 PCYB_042220 PCYB_042230 PKH_041210 PVX_003830 PIN1 PCOA1 PGO1
	678	PGO2 PGO3 PGO4	PCOA4 PCYB_042190 PVX_003850 PIN4
	326	PRCDC_0206700 PF3D7_0207700 PF3D7_0207900 PRCDC_0206800	PRCDC_0206600 PF3D7_0207600
	891	PVX_003820	PVX_003840
	816	PCYB_042220 PVX_003835	PCYB_042200 PVX_003845
	94	PKH_041260 PCYB_042300	PCOA4 PVX_003850
	255	PCOA5 PKH_041210	PVX_003840 PCYB_042230
	920	PVX_003820	PVX_003830 PCYB_042210 PVX_003840
	791	PVX_003820	PVX_003830
	498	PCYB_042220	PVX_003835
	395	PVX_003830	PVX_003805
	198	PKH_041250 PCHAS_030710 PCYB_042290 PVX_003800 PIN6 PCOA6 PGO5	PVX_003850 PCYB_042190 PIN5
	269	PCYB_042220	PVX_003845
	74	PCYB_042200 PVX_003845 PIN7 PCOA7	PVX_003850 PIN5 PCOA4

Table S4-2. List of recombinant sequences, number of recombination events, and length of recombinant segments in GCMFs with significant recombination signal (continued).

Multigene family	Length recombinant segment	Sequence code Recombinant 1	Sequence code Recombinant 2
Tetratricopeptide repeat protein	113	PKH_111790 PVX_114650 PCYB_112700	PF3D7_0601600 PF3D7_0631000 PRCDC_0629400
Tubulin	724	PVX_098630 PCYB_071150 PIN2 PKH_070090 PCOA2	PBANKA_041770 PCHAS_041860 PYYM_0420500
	802	PCOA1 PCYB_053570	PIN2 PVX_098630 PCYB_071150
	426	PCYB_053570 PVX_090155 PIN1 PCOA1	PCHAS_041860 PBANKA_041770 PYYM_0420500
	772	PGO2	PRCDC_0901800 PBANKA_041770 PF3D7_0903700
	182	PVX_090155 PKH_052580 PCOA1	PF3D7_0422300 PRCDC_0419500
	147	PKH_070090 PVX_098630 PCYB_071150 PIN2 PCOA2 PGO1	PVX_090155 PCYB_053570 PIN1
	196	PVX_090155	PGO1
Ubiquitin conjugating enzyme 2	84	PYYM_1031800 PBANKA_102970 PF3D7_1412900. PRCDC_1412200.	PYYM_1360500 PBANKA_135840 PCHAS_136300

Table S4-3. List of multigene families with branches under significant episodic selection.

Multigene family
Acyl-CoA synthase
Alpha beta hydrolase 2
Asparagine tRNA ligase
Calcium Dependent Protein Kinase (CDPK)
Cell division protein FtsH
Chaperonin
Conserved *Plasmodium* protein
Conserved *Plasmodium* protein unknown function 6
Conserved *Plasmodium* protein unknown function 12
Conserved rodent malaria protein unknown function
Cytoadherence-linked asexual protein (CLAG)
DEAD DEAH box ATP dependent RNA helicase
DER1 like protein
DHHC type zinc finger protein
Dipeptidyl amino peptidase putative (DPAP)
DNA directed RNA polymerase II
Eukaryotic initiation factor 2a
Heat Shock protein 70
Heat Shock protein 90
Histone H3
Hypothetical protein
Lysophospholipase
Meiotic recombination protein DMC
Methyltransferase
NIMA-related protein kinase (NEK)
P28
Papain
Plasmepsin
Serine repeat antigen (SERA)

Table S4-4. Distribution of strength and proportion on sites under three different selective regimes in *P. falciparum* paralogs.

Name	B	Test p-value	ω-	Prop. Sites -	ωN	Prop. Sites N	ω+	Prop. Sites +	Expression
PF3D7_1246200	0	1	0.000	1.00	0.202	0.00	0.144	0.00	Vertebrate
PF3D7_0211800	0.003	1	0.198	0.63	0.358	0.00	0.315	0.37	Vertebrate
PF3D7_0106300	0.009	1	0.046	0.47	0.496	0.00	0.051	0.53	Vertebrate
PF3D7_1211900	8E-04	1	1.000	0.51	0.984	0.00	29.900	0.49	Vertebrate
PF3D7_1103700	0.007	1	0.000	0.88	0.000	0.11	0.000	0.01	Vertebrate
PF3D7_1342400	0.025	1	0.016	0.79	0.058	0.00	0.050	0.21	Vertebrate
PF3D7_0717500	0.015	1	0.016	0.77	0.017	0.23	0.139	0.00	Vertebrate
PF3D7_0217500	0.011	1	0.000	0.74	0.000	0.22	0.000	0.04	Vertebrate
PF3D7_0310100	0.021	1	0.041	1.00	0.080	0.00	0.138	0.00	Vertebrate
PF3D7_1119600	0.011	1	0.000	1.00	0.206	0.00	0.133	0.00	Vertebrate
PF3D7_1464900	0.006	1	0.000	1.00	0.000	0.00	0.116	0.00	Vertebrate
PF3D7_1329300	0.006	1	0.057	0.96	0.435	0.00	0.057	0.04	Vertebrate
PF3D7_1433300	0.004	1	0.000	0.71	0.000	0.24	0.000	0.05	Vertebrate
PF3D7_0816600	0.017	1	0.000	0.84	0.000	0.16	0.120	0.00	Vertebrate
PF3D7_1409600	0.01	1	1.000	0.58	1.000	0.41	382.000	0.01	Vertebrate
PF3D7_0620000	0.014	1	0.889	0.00	0.859	0.00	56.500	1.00	Vertebrate
PF3D7_1105700	0.026	1	0.010	0.86	0.120	0.00	0.084	0.14	Vertebrate
PF3D7_1222000	0.025	1	1.000	0.52	1.000	0.46	9350.000	0.02	Vertebrate
PF3D7_0803600	0.022	1	0.254	1.00	0.219	0.00	0.120	0.00	Vertebrate
PF3D7_0935800	0.021	1	1.000	0.48	0.414	0.23	5.600	0.29	Vertebrate
PF3D7_0831600	0.02	0.828	0.434	0.70	0.000	0.07	14.300	0.23	Vertebrate
PF3D7_0302200	0.011	1	0.000	0.94	0.000	0.00	15.900	0.06	Vertebrate
PF3D7_0220800	0.005	1	0.811	1.00	0.735	0.00	0.589	0.00	Vertebrate
PF3D7_0302500	0.008	1	0.818	1.00	0.816	0.00	0.620	0.00	Vertebrate
PF3D7_1445900	9E-04	1	0.000	0.83	0.000	0.15	0.000	0.02	Vertebrate
PF3D7_0609800	0.006	1	0.102	0.50	0.418	0.00	0.174	0.50	Vertebrate
PF3D7_1116700	0.018	1	0.000	0.99	0.000	0.00	12.400	0.01	Vertebrate
PF3D7_0318200	0.012	1	0.000	0.86	0.000	0.14	0.120	0.00	Vertebrate
PF3D7_1329000	0.019	1	0.013	1.00	0.117	0.00	0.091	0.00	Vertebrate
PF3D7_0215700	0.006	1	0.000	1.00	0.001	0.00	0.112	0.00	Vertebrate
PF3D7_1206600	0.008	1	0.000	1.00	0.182	0.00	0.130	0.00	Vertebrate
PF3D7_1149600	0.03	1	1.000	0.87	1.000	0.01	39.200	0.12	Vertebrate
PF3D7_1427500	0.001	1	0.901	0.00	1.000	0.03	14.200	0.97	Vertebrate
PF3D7_1145400	0.028	1	0.032	0.83	0.076	0.00	0.036	0.17	Vertebrate
PF3D7_1357100	0.005	1	0.000	1.00	0.196	0.00	0.146	0.00	Vertebrate
PF3D7_1357000	0.003	1	0.000	1.00	0.001	0.00	0.126	0.00	Vertebrate
PF3D7_0602400	0.011	1	0.000	1.00	0.157	0.00	0.114	0.00	Vertebrate
PF3D7_1438000	0.42	0.717	0.000	1.00	0.468	0.00	3330.000	0.00	Vertebrate
PF3D7_1106300	0	1	0.001	0.79	0.204	0.16	0.133	0.04	Vertebrate
PF3D7_1419800	0.002	1	0.957	0.00	1.000	0.43	23.400	0.57	Vertebrate
PF3D7_0501100	0.003	1	0.390	0.00	0.464	0.92	0.609	0.08	Vertebrate
PF3D7_0201800	0.021	1	0.000	0.93	0.000	0.00	4.490	0.07	Vertebrate
PF3D7_0818900	0.014	1	0.000	1.00	0.001	0.00	0.120	0.00	Vertebrate
PF3D7_0917900	0.003	1	0.000	0.84	0.000	0.13	0.000	0.03	Vertebrate
PF3D7_0831700	0.021	1	0.075	0.87	0.080	1.00	0.078	0.00	Vertebrate
PF3D7_0708400	0.014	1	0.000	1.00	0.190	0.00	0.142	0.00	Vertebrate
PF3D7_1202900	0.019	1	0.000	1.00	0.203	0.00	0.135	0.00	Vertebrate
PF3D7_1445100	0.021	1	0.034	1.00	0.567	0.00	0.142	0.00	Vertebrate
PF3D7_0714000	0	1	0.000	0.92	0.222	0.07	0.146	0.01	Vertebrate
PF3D7_0610400	0	1	0.001	0.92	0.221	0.07	0.145	0.01	Vertebrate
PF3D7_0617900	0	1	0.001	0.80	0.221	0.19	0.160	0.02	Vertebrate
PF3D7_1456800	0.017	1	0.030	0.60	0.038	0.00	0.033	0.40	Vertebrate
PF3D7_1235200	0.016	1	0.043	0.33	0.152	0.00	0.042	0.67	Vertebrate
PF3D7_1372400	0.084	0.0004	0.000	0.93	0.000	0.01	28.300	0.06	Vertebrate
PF3D7_1200700	0.033	0.6272	0.000	0.91	0.000	0.01	7.340	0.08	Vertebrate
PF3D7_0731600	0.032	1	0.000	0.85	0.000	0.03	7.330	0.13	Vertebrate
PF3D7_1479000	0.1	1	0.094	0.91	0.119	0.02	2.490	0.07	Vertebrate
PF3D7_0215300	0.016	1	0.390	0.09	0.388	0.13	0.385	0.78	Vertebrate
PF3D7_0301000	0.086	1	0.089	0.94	1.000	0.06	0.089	0.00	Vertebrate
PF3D7_1324900	0.01	1	0.000	1.00	0.198	0.00	0.143	0.00	Vertebrate
PF3D7_1015300	0.002	1	0.916	0.19	1.000	0.71	1.580	0.10	Vertebrate
PF3D7_0409300	0.032	1	1.000	0.98	0.548	0.00	225.000	0.02	Vertebrate
PF3D7_1455200	0.002	1	0.000	0.31	0.001	0.00	67.400	0.69	Vertebrate
PF3D7_1430700	0.004	1	0.434	0.59	0.470	0.18	0.449	0.22	Vertebrate
PF3D7_1416500	0.001	1	1.000	0.37	1.000	0.29	23.400	0.34	Vertebrate
PF3D7_0910800	0.021	1	0.039	1.00	0.000	0.00	0.116	0.00	Vertebrate
PF3D7_1412000	0.018	1	0.155	0.90	0.155	0.09	0.155	0.01	Vertebrate
PF3D7_1411900	0.009	1	0.756	1.00	0.966	0.00	0.575	0.00	Vertebrate
PF3D7_1030900	0	1	0.001	0.88	0.213	0.10	0.140	0.01	Vertebrate

Table S4-4. Distribution of strength and proportion on sites under three different selective regimes in *P. falciparum* paralogs (continued).

Name	B	Test p-value	ω-	Prop. Sites -	ωN	Prop. Sites N	ω+	Prop. Sites +	Expression
PF3D7_1115300	0.009	1	0.000	0.93	0.000	0.02	10000.000	0.06	Vertebrate
PF3D7_1115400	0.004	1	0.774	0.00	0.458	0.00	65.000	1.00	Vertebrate
PF3D7_1115700	0.01	1	0.168	1.00	0.192	0.00	0.141	0.00	Vertebrate
PF3D7_1458000	0.032	1	0.050	1.00	0.072	0.00	0.069	0.00	Vertebrate
PF3D7_1438900	0.008	1	0.000	0.85	0.000	0.15	0.120	0.00	Vertebrate
PF3D7_0808200	0.003	1	0.047	0.00	0.701	0.00	1.570	1.00	Vertebrate
PF3D7_1033800	0.015	1	0.250	0.00	0.297	0.00	0.295	1.00	Vertebrate
PF3D7_0917600	0.008	1	0.000	0.89	0.000	0.10	0.000	0.01	Vertebrate
PF3D7_1364300	0.016	1	0.000	0.87	0.000	0.13	0.123	0.00	Vertebrate
PF3D7_0702200	0.04	1	0.187	0.68	0.217	0.32	6.620	0.00	Vertebrate
PF3D7_0501500	0.013	1	0.000	0.94	1.000	0.07	0.051	0.00	Vertebrate
PF3D7_0501600	0.025	1	0.000	0.83	0.921	0.00	2.980	0.17	Vertebrate
PF3D7_1448400	0.003	1	0.881	0.00	0.942	0.00	30.100	1.00	Vertebrate
PF3D7_0207400	0.004	1	0.680	0.00	0.736	0.00	1.230	1.00	Vertebrate
PF3D7_0207500	0.008	1	0.510	0.00	0.469	0.12	0.472	0.88	Vertebrate
PF3D7_0207600	0.018	1	0.349	0.13	0.190	0.00	0.349	0.87	Vertebrate
PF3D7_0207700	0.013	1	0.994	0.84	1.000	0.06	15.800	0.10	Vertebrate
PF3D7_0207800	0.302	1	0.188	0.83	0.986	0.13	15.600	0.04	Vertebrate
PF3D7_0207900	0.017	1	0.591	0.00	0.695	0.00	1.510	1.00	Vertebrate
PF3D7_0208000	0.012	1	0.089	0.59	0.089	0.41	0.052	0.00	Vertebrate
PF3D7_0902800	0.008	1	0.383	0.00	0.425	0.00	0.357	1.00	Vertebrate
PF3D7_0717700	0.013	1	0.005	0.82	0.450	0.00	0.017	0.18	Vertebrate
PF3D7_0516600	0.007	1	0.526	1.00	0.224	0.00	0.399	0.00	Vertebrate
PF3D7_0903700	0.006	1	0.000	1.00	0.202	0.00	0.146	0.00	Vertebrate
PF3D7_0422300	0.023	1	0.000	1.00	0.202	0.00	0.146	0.00	Vertebrate
PF3D7_1139700	0.005	1	0.966	0.00	1.000	0.08	10000.000	0.92	Vector
PF3D7_0826200	0.012	1	0.418	0.00	0.534	0.00	0.638	1.00	Vector
PF3D7_0509600	0.019	1	0.026	0.78	0.015	0.00	0.031	0.22	Vector
PF3D7_1429900	0.002	1	0.458	0.98	0.620	0.00	0.726	0.02	Vector
PF3D7_1026900	0.035	1	0.153	1.00	0.125	0.00	0.137	0.00	Vector
PF3D7_1460000	0.038	0.073	0.167	0.98	0.152	0.00	78.700	0.01	Vector
PF3D7_1333000	0.005	1	0.901	0.00	1.000	0.06	66.200	0.94	Vector
PF3D7_1215300	0	1	0.001	0.91	0.219	0.08	0.145	0.01	Vector
PF3D7_1450600	0.007	1	0.155	0.00	0.193	0.81	0.192	0.19	Vector
PF3D7_1246500	0.012	1	0.054	0.89	0.130	0.11	0.055	0.01	Vector
PF3D7_0519000	0.007	1	0.987	0.00	0.808	0.00	6.890	1.00	Vector
PF3D7_1449800	0.019	1	0.000	1.00	0.191	0.00	0.137	0.00	Vector
PF3D7_1469100	0	1	0.000	0.94	0.184	0.05	0.139	0.01	Vector
PF3D7_1213400	0.008	1	0.134	1.00	0.126	0.00	0.244	0.00	Vector
PF3D7_1213200	0.014	1	0.349	0.15	0.351	0.85	0.271	0.00	Vector
PF3D7_0911300	0.305	1	0.525	1.00	1.000	0.00	10000.000	0.00	Vector
PF3D7_0718300	0.01	1	0.081	0.53	0.781	0.00	0.090	0.47	Vector
PF3D7_1032500	0.002	1	1.000	0.27	1.000	0.12	31.200	0.60	Vector
PF3D7_1468500	0	1	0.001	0.83	0.211	0.14	0.139	0.03	Vector
PF3D7_1027900	0.002	1	0.659	0.00	0.648	0.00	14.100	1.00	Vector
PF3D7_0528400	0	1	0.001	0.74	0.199	0.19	0.129	0.07	Vector
PF3D7_1247800	0.003	1	0.982	0.00	0.993	0.00	53.800	1.00	Vector
PF3D7_0523400	0.012	1	0.000	1.00	0.193	0.00	0.139	0.00	Vector
PF3D7_0706700	0.031	1	0.067	0.94	0.094	0.04	8.700	0.01	Vector
PF3D7_1037500	0.001	1	0.938	0.61	1.000	0.07	1.730	0.32	Vector
PF3D7_0729900	0.014	1	0.028	1.00	1.000	0.00	199.000	0.00	Vector
PF3D7_1023100	0.004	1	0.086	0.64	0.105	0.36	0.477	0.00	Vector
PF3D7_1122900	0.004	1	0.057	0.86	0.033	0.00	0.130	0.14	Vector
PF3D7_0905300	0.003	1	0.053	0.89	0.104	0.11	0.113	0.00	Vector
PF3D7_1330600	0.038	1	0.013	1.00	0.077	0.00	0.128	0.00	Vector
PFC10_API0028	0.011	1	0.000	0.88	0.534	0.00	0.000	0.12	Vector
PF3D7_0909400	0.013	1	0.061	1.00	0.156	0.00	0.326	0.00	Vector
PF3D7_0828600	0.014	1	0.103	1.00	0.188	0.00	0.176	0.00	Vector
PF3D7_1116500	0.045	1	0.009	1.00	0.012	0.00	0.140	0.00	Vector
PF3D7_0512200	0.013	1	0.095	0.67	0.072	0.21	5.640	0.12	Vector
PF3D7_1114800	0.001	1	1.000	0.50	0.909	0.00	11.700	0.50	Vector
PF3D7_1216200	0.003	1	1.000	0.80	0.999	0.00	658.000	0.20	Vector
PF3D7_0213100	0.013	1	0.000	1.00	0.001	0.00	0.116	0.00	Vector
PF3D7_1222300	0.007	1	0.000	1.00	0.001	0.00	0.128	0.00	Vector
PF3D7_0817900	0.004	1	0.000	0.86	0.000	0.13	0.000	0.02	Vector
PF3D7_1105100	0	1	0.001	0.92	0.222	0.08	0.146	0.01	Vector
PF3D7_1447500	0.023	1	0.174	0.00	0.212	0.30	0.233	0.70	Vector
PF3D7_1003600	0.006	1	0.000	0.88	0.000	0.11	0.000	0.01	Vector
PF3D7_0111000	0.021	1	0.032	1.00	0.026	0.00	0.118	0.00	Vector

Table S4-4. Distribution of strength and proportion on sites under three different selective regimes in *P. falciparum* paralogs (continued).

Name	B	Test p-value	ω-	Prop. Sites -	ωN	Prop. Sites N	ω+	Prop. Sites +	Expression
PF3D7_0215000	0.008	1	0.171	0.00	0.935	1.00	0.620	0.00	Vector
PF3D7_1477900	0.02	1	0.085	0.00	0.129	0.83	0.129	0.17	Vector
PF3D7_0709700	0.019	1	0.094	1.00	0.049	0.00	0.057	0.00	Vector
PF3D7_0618500	0.004	1	1.000	0.33	1.000	0.14	56.500	0.53	Vector
PF3D7_0816800	0.013	1	0.000	1.00	0.001	0.00	0.119	0.00	Vector
PF3D7_1107400	0.007	1	0.000	0.84	0.000	0.16	0.119	0.00	Vector
PF3D7_1027600	0.012	1	0.038	1.00	0.493	0.00	0.575	0.00	Vector
PF3D7_1217100	0.02	1	0.094	0.74	0.097	0.26	0.121	0.00	Vector
PF3D7_1201600	0.016	1	0.137	1.00	0.200	0.00	0.098	0.00	Vector
PF3D7_1228300	0.007	1	0.000	1.00	0.001	0.00	0.124	0.00	Vector
PF3D7_0719200	0	1	0.001	0.89	0.219	0.10	0.144	0.01	Vector
PF3D7_0525900	0	1	0.001	0.78	0.220	0.20	0.144	0.03	Vector
PF3D7_0104800	0.004	1	0.976	0.00	1.000	0.54	3.190	0.46	Vector
PF3D7_1428600	0.002	1	0.997	0.00	1.000	0.73	943.000	0.27	Vector
PF3D7_1215000	0.002	1	1.000	0.28	0.995	0.00	66.000	0.72	Vector
PF3D7_1036700	0	1	0.001	0.88	0.219	0.10	0.139	0.02	Vector
PF3D7_1006600	0.024	1	0.095	0.80	0.087	0.20	0.105	0.00	Vector
PF3D7_1102400	0.012	1	0.000	1.00	0.228	0.00	0.107	0.00	Vector
PF3D7_0412300	0.006	1	0.797	0.00	0.910	0.00	4.340	1.00	Vector
PF3D7_1465700	0.021	1	0.000	1.00	0.363	0.00	80.100	0.00	Vector
PF3D7_0311700	0.005	1	0.000	0.98	0.000	0.00	3000.000	0.02	Vector
PF3D7_1430200	0.014	1	0.000	0.60	0.000	0.25	0.000	0.14	Vector
PF3D7_1407900	0.006	1	0.001	0.00	0.384	0.00	72.100	1.00	Vector
PF3D7_1408100	0.033	1	0.055	0.28	0.100	0.00	0.055	0.72	Vector
PF3D7_1407800	0.002	1	0.000	0.24	0.402	0.00	0.000	0.76	Vector
PF3D7_1408000	0.01	1	0.068	0.22	0.070	0.00	0.068	0.78	Vector
PF3D7_1309200	0.02	1	0.000	0.74	0.000	0.26	0.122	0.00	Vector
PF3D7_0810300	0	1	0.001	0.85	0.212	0.12	0.140	0.02	Vector
PF3D7_1216000	0.016	1	0.000	0.98	0.000	0.00	47.500	0.02	Vector
PF3D7_0909500	0.007	1	0.852	0.00	0.528	0.00	67.100	1.00	Vector
PF3D7_1230600	0.005	1	0.000	0.77	0.000	0.20	0.000	0.04	Vector
PF3D7_0601600	0.002	1	0.000	1.00	0.001	0.00	0.124	0.00	Vector
PF3D7_0631000	0.001	1	1.000	0.69	0.479	0.00	3050.000	0.31	Vector
PF3D7_0319300	0.012	1	0.183	1.00	0.192	0.00	0.137	0.00	Vector
PF3D7_0812600	0	1	0.000	0.82	0.000	0.17	0.118	0.01	Vector
PF3D7_1345500	0.006	1	0.000	0.79	0.000	0.18	0.000	0.03	Vector
PF3D7_1412900	0	1	0.001	0.78	0.194	0.17	0.119	0.05	Vector
PF3D7_1412500	0.01	1	0.000	1.00	0.001	0.00	0.124	0.00	Generalist
PF3D7_1401300	0.012	1	0.753	0.00	0.823	0.00	16.800	1.00	Generalist
PF3D7_0918600	0.03	1	0.678	1.00	0.874	0.00	10000.000	0.00	Generalist
PF3D7_1337800	0.002	1	1.000	0.45	1.000	0.30	45.600	0.25	Generalist
PF3D7_0610600	0.005	1	0.258	0.00	0.126	0.01	0.126	0.99	Generalist
PF3D7_1239700	0	1	0.000	0.87	0.000	0.10	0.116	0.02	Generalist
PF3D7_0110700	0.016	1	0.066	0.93	0.066	0.07	0.066	0.00	Generalist
PF3D7_1116800	0.003	1	0.184	0.66	0.200	0.34	0.074	0.00	Generalist
PF3D7_1143800	0.003	1	0.961	0.00	0.910	0.00	10.800	1.00	Generalist
PF3D7_1416300	0.006	1	0.450	1.00	0.326	0.00	0.347	0.00	Generalist
PF3D7_1227100	0.005	1	0.082	0.00	1.000	1.00	0.154	0.00	Generalist
PF3D7_1233000	0.024	1	0.000	0.98	0.000	0.00	3.310	0.02	Generalist
PF3D7_0923800	0.002	1	0.000	1.00	0.001	0.00	0.114	0.00	Generalist
PF3D7_0934000	0.036	1	0.131	0.00	0.158	0.69	0.158	0.31	Generalist
PF3D7_0319400	0.001	1	1.000	0.59	0.979	0.00	25.300	0.41	Generalist
PF3D7_1253400	0.018	1	0.000	0.98	0.000	0.01	34.500	0.01	Generalist
PF3D7_1238800	0.03	1	0.047	0.97	0.087	0.00	0.047	0.03	Generalist
PF3D7_0527300	0.012	1	0.390	0.00	0.332	1.00	1.030	0.00	Generalist
PF3D7_1128500	0.03	1	0.060	0.00	0.060	0.03	0.060	0.97	Generalist
PF3D7_0932400	0.007	1	0.847	0.00	0.846	0.00	8.620	1.00	Generalist
PF3D7_0802200	0.002	1	0.940	0.00	0.981	0.00	18.400	1.00	Generalist
PF3D7_1030100	0	1	0.001	0.76	0.217	0.21	0.142	0.03	Generalist
PF3D7_0422500	0.021	1	0.045	1.00	0.039	0.00	0.050	0.00	Generalist
PF3D7_1439100	0.013	1	0.058	0.22	0.057	0.78	0.088	0.00	Generalist
PF3D7_0313100	0.007	1	0.026	0.47	0.229	0.33	0.268	0.20	Generalist
PF3D7_1129400	0.028	1	0.016	0.00	0.015	1.00	0.011	0.00	Generalist
PF3D7_0925500	0.011	1	0.000	1.00	0.197	0.00	0.143	0.00	Generalist
PF3D7_1457200	0	1	0.001	0.77	0.218	0.21	0.143	0.03	Generalist
PF3D7_0827100	0.008	1	0.734	0.00	0.827	0.00	1.460	1.00	Generalist

Table S4-5. Distribution of strength and proportion on sites under three different selective regimes in *P. berghei* paralogs.

Gene ID	B	P-value	ω-	Prop. Sites -	ωN	Prop. Sites N	ω+	Prop. Sites +	Expression
PBANKA_145930	0.03	1	0.0000	0.900	0.0000	0.100	0.1240	0.000	Vertebrate
PBANKA_090930	1.336	0.0447	0.2540	0.990	1.0000	0.001	3330.0000	0.010	Vertebrate
PBANKA_030860	3.858	0.0578	0.0130	0.980	0.0000	0.003	1530.0000	0.018	Vertebrate
PBANKA_110920	0.027	1	0.1380	0.870	0.1400	0.000	0.1380	0.130	Vertebrate
PBANKA_081950	0.023	1	0.0587	0.500	0.0590	0.500	0.0590	0.000	Vertebrate
PBANKA_020700	0.027	1	0.0346	1.000	0.4360	0.000	0.1150	0.000	Vertebrate
PBANKA_061040	0.029	1	0.0635	1.000	0.0522	0.000	0.0645	0.000	Vertebrate
PBANKA_135550	0.025	1	0.0230	1.000	0.0255	0.000	0.1250	0.000	Vertebrate
PBANKA_031420	0.018	1	0.0000	0.720	0.0000	0.280	0.1200	0.000	Vertebrate
PBANKA_061520	0.013	1	0.0000	1.000	0.1920	0.000	0.1200	0.000	Vertebrate
PBANKA_135150	0.028	1	0.0501	0.490	0.0513	0.510	0.1200	0.000	Vertebrate
PBANKA_092850	0.018	1	0.0389	1.000	0.2840	0.000	0.1330	0.000	Vertebrate
PBANKA_132830	0.02	1	0.0000	1.000	0.1870	0.000	0.1330	0.000	Vertebrate
PBANKA_145410	0.019	1	0.0652	0.900	0.0701	0.000	0.0650	0.100	Vertebrate
PBANKA_143105	0.015	1	0.0359	0.540	0.2790	0.000	0.2580	0.460	Vertebrate
PBANKA_020300	0.018	1	0.0346	0.790	0.0356	0.210	0.1360	0.000	Vertebrate
PBANKA_134430	0.016	1	0.0657	0.720	0.0661	0.000	0.0656	0.280	Vertebrate
PBANKA_071420	0.03	1	0.0164	1.000	0.0921	0.000	0.1200	0.000	Vertebrate
PBANKA_093120	0.449	1	0.0665	1.000	0.0658	0.000	1470.0000	0.005	Vertebrate
PBANKA_102640	0.028	1	0.0167	1.000	0.5020	0.000	0.1360	0.000	Vertebrate
PBANKA_103290	0.068	1	0.0990	0.780	0.0000	0.000	1.5600	0.220	Vertebrate
PBANKA_123380	0.018	1	0.1310	0.730	0.1250	0.000	0.1270	0.270	Vertebrate
PBANKA_142900	0.02	1	0.1950	0.690	0.1940	0.310	0.1220	0.000	Vertebrate
PBANKA_094120	0.006	1	0.0472	0.930	0.5400	0.000	0.0832	0.067	Vertebrate
PBANKA_080030	0.105	0.9895	0.4540	0.980	0.4530	0.001	23.9000	0.019	Vertebrate
PBANKA_100050	0.045	1	0.0000	0.690	0.3710	0.000	2.0800	0.310	Vertebrate
PBANKA_081240	0.045	0.0375	0.4330	0.930	0.1980	0.000	7.9300	0.068	Vertebrate
PBANKA_083630	0.047	1	0.8830	0.000	0.8830	0.000	1.1200	1.000	Vertebrate
PBANKA_140060	0.036	1	1.0000	1.000	0.4820	0.000	1.0600	0.000	Vertebrate
PBANKA_144190	0.028	1	0.0307	0.900	0.0310	0.096	0.1170	0.000	Vertebrate
PBANKA_051630	0.006	1	0.1530	0.860	0.2060	0.110	0.1790	0.031	Vertebrate
PBANKA_133180	0.017	1	0.0000	0.830	0.0000	0.140	0.0000	0.030	Vertebrate
PBANKA_080700	0.017	1	0.0263	0.780	0.0316	0.220	0.0294	0.000	Vertebrate
PBANKA_134400	0.032	1	0.0193	0.970	0.0895	0.000	0.0193	0.025	Vertebrate
PBANKA_060520	0.016	1	0.0373	1.000	0.3710	0.000	0.1240	0.000	Vertebrate
PBANKA_101710	0.007	1	0.0662	0.550	0.0843	0.380	0.0673	0.062	Vertebrate
PBANKA_052040	0.013	1	0.0066	0.930	0.1740	0.000	0.0936	0.071	Vertebrate
PBANKA_090360	0.064	1	0.0177	0.950	0.0177	0.051	0.1310	0.000	Vertebrate
PBANKA_113330	0	1	0.0008	0.920	0.2220	0.071	0.1460	0.007	Vertebrate
PBANKA_113340	0	1	0.0008	0.920	0.2220	0.071	0.1460	0.007	Vertebrate
PBANKA_144770	0.028	1	0.0324	1.000	0.0261	0.000	0.0692	0.000	Vertebrate
PBANKA_134560	0.014	1	0.0571	0.920	0.0590	0.000	0.0575	0.079	Vertebrate
PBANKA_061080	0.039	1	0.1300	0.820	0.2370	0.000	0.1300	0.180	Vertebrate
PBANKA_081060	0.017	1	0.0565	0.140	0.0576	0.440	0.0564	0.410	Vertebrate
PBANKA_094060	0.029	1	0.0520	0.900	0.0737	0.000	0.0521	0.100	Vertebrate
PBANKA_082470	0.029	1	0.0593	0.880	0.0592	0.120	0.0504	0.000	Vertebrate
PBANKA_111180	0.572	1	0.2470	1.000	0.2470	0.000	3330.0000	0.003	Vertebrate
PBANKA_093290	0.02	1	0.1310	1.000	0.2130	0.000	0.1200	0.000	Vertebrate
PBANKA_143190	0.015	1	0.0000	0.980	0.0000	0.003	11.5000	0.013	Vertebrate
PBANKA_031000	0.107	1	0.0094	1.000	0.0004	0.000	0.1360	0.000	Vertebrate
PBANKA_081890	0.01	1	0.0975	0.180	0.1160	0.820	0.1830	0.000	Vertebrate
PBANKA_080570	0.012	1	0.0178	0.830	0.2610	0.000	0.0965	0.170	Vertebrate
PBANKA_143730	0.011	1	0.0514	0.990	0.2940	0.000	0.0522	0.007	Vertebrate
PBANKA_060190	0.024	1	0.0000	1.000	0.0008	0.000	0.1150	0.000	Vertebrate
PBANKA_130890	0.023	1	0.0802	0.770	0.0821	0.029	0.0836	0.200	Vertebrate
PBANKA_094180	0.008	1	0.0508	0.390	0.2250	0.540	0.3070	0.064	Vertebrate
PBANKA_142060	0.009	1	0.0000	0.920	0.0000	0.075	0.0000	0.007	Vertebrate
PBANKA_010880	0.004	1	0.0000	1.000	0.0014	0.000	0.1380	0.000	Vertebrate
PBANKA_111710	0.006	1	0.0000	1.000	0.0008	0.000	0.1450	0.000	Vertebrate
PBANKA_144980	0.027	1	0.0651	0.270	0.0659	0.730	0.0685	0.000	Vertebrate
PBANKA_080590	0.01	1	0.3140	0.100	0.3120	0.900	0.2500	0.000	Vertebrate
PBANKA_031180	0.067	1	0.0000	0.670	0.0000	0.030	1.7400	0.300	Vertebrate
PBANKA_111770	0.027	1	0.2660	1.000	0.2620	0.000	0.2610	0.000	Vertebrate
PBANKA_134010	0.008	1	0.0000	1.000	0.1920	0.000	0.1430	0.000	Vertebrate
PBANKA_100690	0.018	1	0.0397	0.950	0.0333	0.021	8.5400	0.033	Vertebrate
PBANKA_101400	0.036	1	0.0713	0.590	0.0728	0.410	0.1410	0.000	Vertebrate
PBANKA_020830	0.024	1	0.1690	0.000	0.1630	1.000	0.1430	0.000	Vertebrate
PBANKA_091970	0.025	1	0.0325	0.300	0.0132	0.000	0.0327	0.700	Vertebrate
PBANKA_103060	0.039	1	0.2020	0.000	0.2280	1.000	0.2060	0.000	Vertebrate

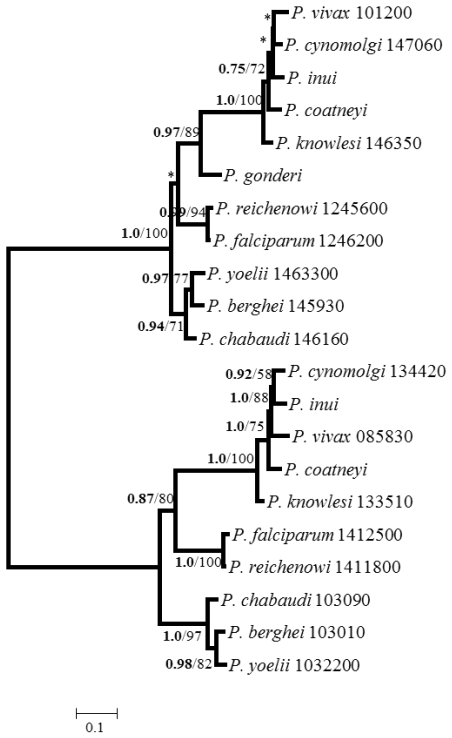
Table S4-5. Distribution of strength and proportion on sites under three different selective regimes in *P. berghei* paralogs (continued).

Gene ID	B	P-value	ω-	Prop. Sites -	ωN	Prop. Sites N	ω+	Prop. Sites +	Expression
PBANKA_0514900	0.083	0.9299	0.0000	0.800	0.0000	0.007	4.2400	0.200	Vertebrate
PBANKA_093240	0.039	1	0.2650	0.830	0.2660	0.160	146.0000	0.002	Vertebrate
PBANKA_132170	0.04	1	0.0000	0.660	0.6990	0.300	0.7760	0.036	Vertebrate
PBANKA_101610	0.113	1	0.0000	0.780	0.0000	0.001	3.2600	0.220	Vertebrate
PBANKA_130280	0.012	1	0.1370	0.510	0.1390	0.150	0.1370	0.340	Vertebrate
PBANKA_143080	0.031	1	0.0197	1.000	0.1230	0.000	0.1390	0.000	Vertebrate
PBANKA_051970	0.013	1	0.5020	0.530	0.5010	0.000	0.5010	0.470	Vertebrate
PBANKA_120480	0.02	1	0.0574	0.000	0.0581	0.910	0.0586	0.090	Vertebrate
PBANKA_101450	0.02	1	0.1660	0.000	0.2580	1.000	0.1580	0.000	Vertebrate
PBANKA_103440	0.033	1	1.0000	0.880	0.9030	0.000	122.0000	0.120	Vertebrate
PBANKA_081860	0.032	1	0.0159	1.000	0.1330	0.000	0.1420	0.000	Vertebrate
PBANKA_114020	0.017	1	0.0250	1.000	0.0243	0.000	15.9000	0.004	Vertebrate
PBANKA_052290	0.041	1	0.0942	0.670	0.2310	0.330	13.1000	0.003	Vertebrate
PBANKA_130300	0.149	1	0.1330	1.000	0.1330	0.003	3330.0000	0.001	Vertebrate
PBANKA_122020	0.047	1	0.2540	0.000	0.2860	1.000	0.1600	0.000	Vertebrate
PBANKA_110140	0.086	0.1805	1.0000	0.890	1.0000	0.042	24.7000	0.071	Vertebrate
PBANKA_041080	0.031	1	0.0628	0.840	1.0000	0.045	1.7700	0.110	Vertebrate
PBANKA_131220	0.038	1	0.0764	0.000	0.0955	1.000	0.0952	0.000	Vertebrate
PBANKA_030480	0.044	1	0.0000	0.680	1.0000	0.062	1.0500	0.260	Vertebrate
PBANKA_030490	0.035	1	0.1840	1.000	0.1670	0.000	0.1020	0.000	Vertebrate
PBANKA_030500	0.052	1	0.0000	0.420	0.8260	0.440	0.7930	0.140	Vertebrate
PBANKA_030510	0.041	1	0.0058	0.770	0.7890	0.000	2.4000	0.230	Vertebrate
PBANKA_061540	0.022	1	0.1070	1.000	0.0987	0.000	0.1060	0.000	Vertebrate
PBANKA_143170	0.013	1	0.3390	0.000	0.3560	0.000	0.3390	1.000	Vertebrate
PBANKA_091880	0.024	1	0.0000	0.960	0.0000	0.011	4.6800	0.027	Vertebrate
PBANKA_082630	0.027	1	0.1050	0.000	0.1160	0.530	0.1180	0.470	Vertebrate
PBANKA_132090	0.011	1	0.1170	0.280	0.1640	0.630	0.2780	0.089	Vertebrate
PBANKA_070360	0.042	1	0.0696	0.000	0.0783	0.000	0.0748	1.000	Vertebrate
PBANKA_123140	0.041	1	0.1800	0.000	0.1650	0.000	0.1650	1.000	Vertebrate
PBANKA_041770	0.016	1	0.0000	1.000	0.0008	0.000	0.1260	0.000	Vertebrate
PBANKA_052270	0.027	1	0.0000	0.910	0.0000	0.089	0.1260	0.000	Vertebrate
PBANKA_080600	0.027	1	0.0631	0.360	0.0643	0.640	0.0687	0.000	Vertebrate
PBANKA_142490	0.016	1	0.0000	0.830	0.0000	0.170	0.1180	0.000	Vertebrate
PBANKA_135840	0.014	1	0.0000	0.950	0.0008	0.000	8.7200	0.047	Vertebrate
PBANKA_102970	0.034	1	0.0000	0.980	1.0000	0.019	0.1030	0.000	Vertebrate
PBANKA_070440	0.095	0.0106	0.0000	0.900	0.8350	0.000	6.3500	0.100	Vector
PBANKA_101480	0.049	1	0.0824	1.000	0.0900	0.000	0.0504	0.000	Vector
PBANKA_051100	0.021	1	0.2550	0.320	0.2550	0.680	0.1780	0.000	Vector
PBANKA_094320	0.032	1	0.0875	1.000	0.0675	0.000	0.1110	0.000	Vector
PBANKA_040820	0.012	1	0.0205	0.750	0.0227	0.250	0.1180	0.000	Vector
PBANKA_134780	0.009	1	0.8910	0.000	1.0000	0.120	10000.0000	0.880	Vector
PBANKA_131430	0.017	1	0.2450	0.530	0.2610	0.000	0.2460	0.470	Vector
PBANKA_131350	0.039	1	0.0945	0.700	0.0999	0.047	0.1050	0.250	Vector
PBANKA_111920	0.038	1	0.9810	0.000	0.2130	0.000	1.1000	1.000	Vector
PBANKA_133230	0.025	1	0.1820	1.000	0.1790	0.000	0.1250	0.000	Vector
PBANKA_142920	0.02	1	0.6200	0.920	0.5950	0.000	0.6250	0.083	Vector
PBANKA_061590	0.021	1	0.2260	0.920	0.3560	0.020	1.2000	0.063	Vector
PBANKA_130970	0.017	1	0.0476	0.520	0.0231	0.000	0.0482	0.480	Vector
PBANKA_051200	0.009	1	0.2270	0.400	0.2270	0.600	0.0871	0.000	Vector
PBANKA_124300	0.016	1	0.1230	1.000	0.0926	0.000	0.0536	0.000	Vector
PBANKA_146070	0.042	1	0.1760	1.000	0.1560	0.000	0.1230	0.000	Vector
PBANKA_123820	0.024	1	0.1210	1.000	0.1160	0.000	0.7630	0.000	Vector
PBANKA_080430	0.018	1	0.1290	1.000	0.1190	0.000	0.1090	0.000	Vector
PBANKA_021400	0.016	1	0.0912	1.000	0.1410	0.000	0.1050	0.000	Vector
PBANKA_050730	0.011	1	0.0318	0.910	0.6040	0.058	0.5660	0.032	Vector
PBANKA_092540	0.015	1	0.0883	0.840	0.0888	0.160	0.1820	0.000	Vector
PBANKA_041610	0.018	1	0.0204	1.000	0.0186	0.000	0.1190	0.000	Vector
PBANKA_010120	0.021	1	0.0523	0.890	0.0553	0.000	0.0543	0.110	Vector
PBANKA_093150	0.016	1	0.0385	1.000	0.5050	0.000	0.2600	0.000	Vector
PBANKA_102340	0.022	1	0.0489	0.650	0.0510	0.350	0.1230	0.000	Vector
PBANKA_071190	0.023	1	0.0000	1.000	0.0008	0.000	0.1220	0.000	Vector
PBANKA_071290	0.021	1	0.0000	1.000	0.2210	0.000	0.1350	0.000	Vector
PBANKA_083480	0.039	1	0.2150	0.410	0.2170	0.590	0.2910	0.000	Vector
PBANKA_131130	0.041	1	0.2710	0.000	0.3090	0.740	0.3060	0.260	Vector
PBANKA_120200	0.021	1	0.1720	1.000	0.1700	0.000	0.1240	0.000	Vector
PBANKA_132050	0.017	1	0.0974	0.200	0.1270	0.000	0.1020	0.800	Vector
PBANKA_020270	0.025	1	0.1710	0.250	0.1400	0.000	0.1710	0.750	Vector
PBANKA_145330	0.03	1	0.0539	0.970	0.9120	0.029	0.2630	0.000	Vector
PBANKA_122030	0.032	1	0.0000	0.930	0.0000	0.015	2.7700	0.057	Vector

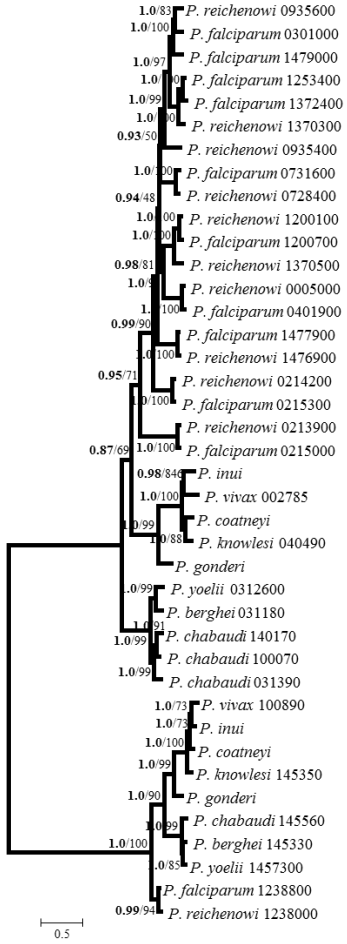
Table S4-5. Distribution of strength and proportion on sites under three different selective regimes in *P. berghei* paralogs (continued).

Gene ID	B	P-value	ω-	Prop. Sites -	ωN	Prop. Sites N	ω+	Prop. Sites +	Expression
PBANKA_071400	0.027	1	0.0430	0.610	0.0430	0.390	0.0411	0.000	Vector
PBANKA_093950	0.018	1	0.0191	0.920	0.0192	0.084	0.1180	0.000	Vector
PBANKA_051170	0.014	1	0.2670	1.000	0.2230	0.000	0.2410	0.000	Vector
PBANKA_143270	0.028	1	0.2540	0.980	0.2540	0.013	25.8000	0.007	Vector
PBANKA_124210	0.032	1	0.1650	1.000	0.1790	0.000	0.1230	0.000	Vector
PBANKA_061670	0	1	0.0008	0.770	0.2190	0.200	0.1440	0.028	Vector
PBANKA_124070	0.012	1	0.0000	1.000	0.2000	0.000	0.1250	0.000	Vector
PBANKA_081190	0.013	1	0.0277	0.950	0.1180	0.008	0.0794	0.046	Vector
PBANKA_122250	0.031	1	0.0377	1.000	0.0923	0.000	0.0912	0.000	Vector
PBANKA_051760	0.015	1	0.1050	0.310	0.1050	0.690	0.0681	0.000	Vector
PBANKA_132910	0.051	1	0.0668	1.000	0.0669	0.000	253.0000	0.004	Vector
PBANKA_040970	0.021	1	0.0274	0.350	0.0282	0.520	0.0279	0.130	Vector
PBANKA_081070	0.067	1	0.0614	1.000	0.1980	0.000	24.5000	0.000	Vector
PBANKA_144530	0.02	1	0.0000	0.940	0.0000	0.043	4.0600	0.013	Vector
PBANKA_112980	0.032	1	0.0322	0.900	0.8940	0.028	2.0600	0.075	Vector
PBANKA_103010	0.019	1	0.0168	1.000	0.1440	0.000	1.3100	0.000	Generalist
PBANKA_101160	0.023	1	0.0000	0.890	0.0000	0.041	3.5900	0.073	Generalist
PBANKA_090520	0.025	1	0.0607	1.000	0.0224	0.000	0.0114	0.000	Generalist
PBANKA_145960	0.021	1	0.1560	0.820	0.1560	0.000	0.1540	0.180	Generalist
PBANKA_010830	0.035	1	0.0000	0.980	0.0000	0.006	4.9100	0.014	Generalist
PBANKA_093130	0.029	1	0.0425	0.980	0.0452	0.009	12.1000	0.011	Generalist
PBANKA_031240	0.034	1	0.0054	0.980	0.0836	0.009	2.1600	0.006	Generalist
PBANKA_API0028	0.019	1	0.0000	0.980	0.0000	0.003	6.8000	0.012	Generalist
PBANKA_070210	0.034	1	0.1190	0.770	0.1210	0.230	0.1110	0.000	Generalist
PBANKA_121370	0.017	1	0.1390	0.520	0.1410	0.480	0.1160	0.000	Generalist
PBANKA_131890	4E-04	1	0.0000	1.000	0.1050	0.000	0.0903	0.000	Generalist
PBANKA_102620	0.031	1	0.0843	0.510	0.0846	0.490	0.1210	0.000	Generalist
PBANKA_083320	0.029	1	0.0924	0.860	0.0929	0.000	0.0929	0.140	Generalist
PBANKA_122800	0.026	1	0.2180	1.000	0.1870	0.000	0.1920	0.000	Generalist
PBANKA_061360	0.013	1	0.4680	0.150	0.4680	0.140	0.4680	0.700	Generalist
PBANKA_051410	0.037	1	0.0129	1.000	0.1700	0.000	0.1420	0.000	Generalist
PBANKA_140770	0.017	1	0.0827	1.000	0.0896	0.000	0.0847	0.000	Generalist
PBANKA_142720	0.005	1	0.1130	0.890	0.7290	0.000	0.1140	0.110	Generalist

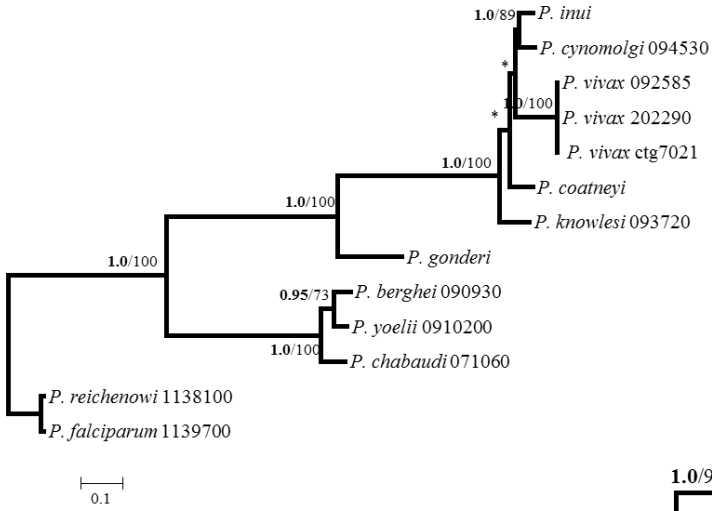
Supplementary Figures



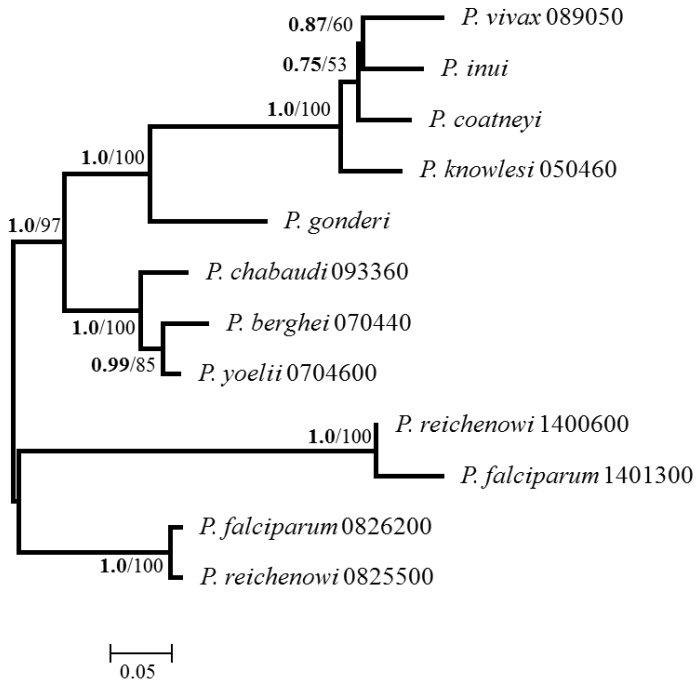
Actin



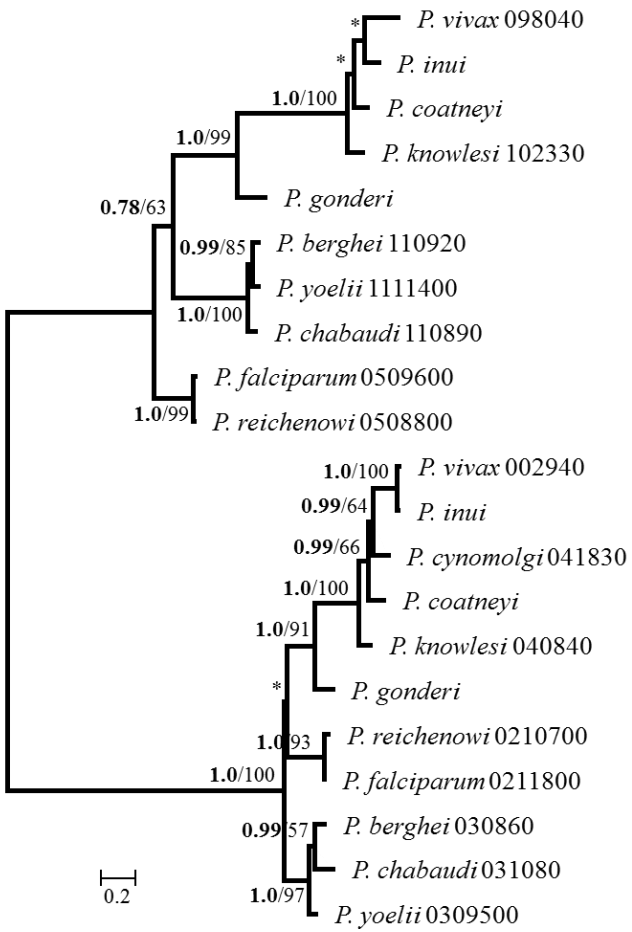
Acyl-coa synthase



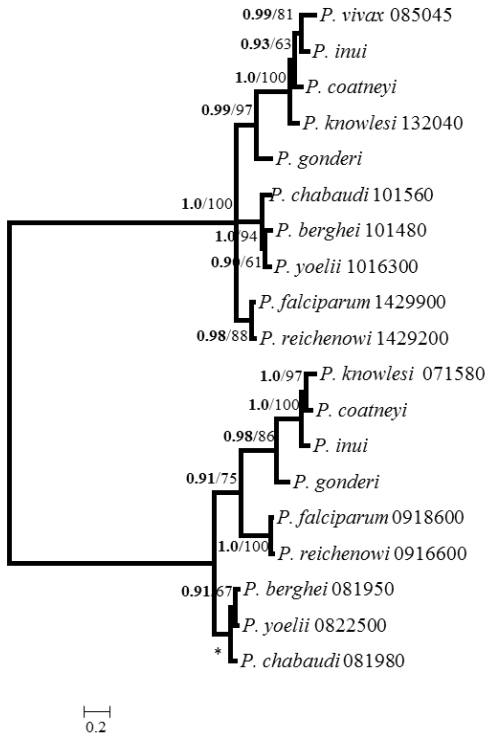
Adrenodoxin reductase



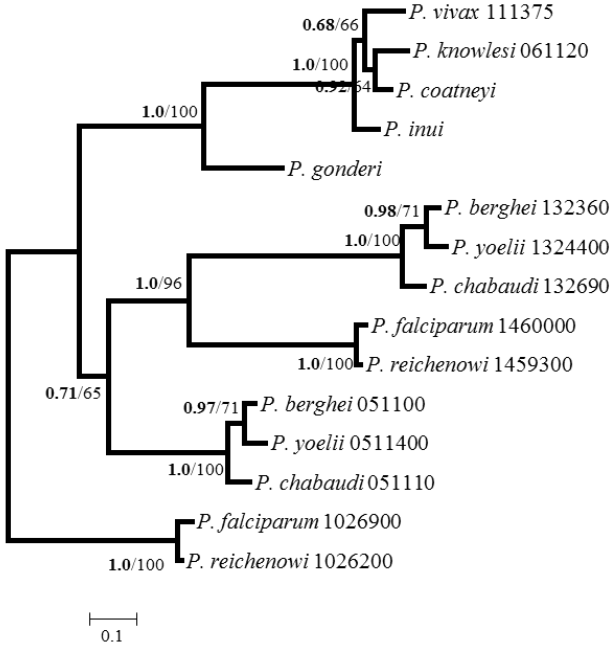
Alpha beta hydrolase



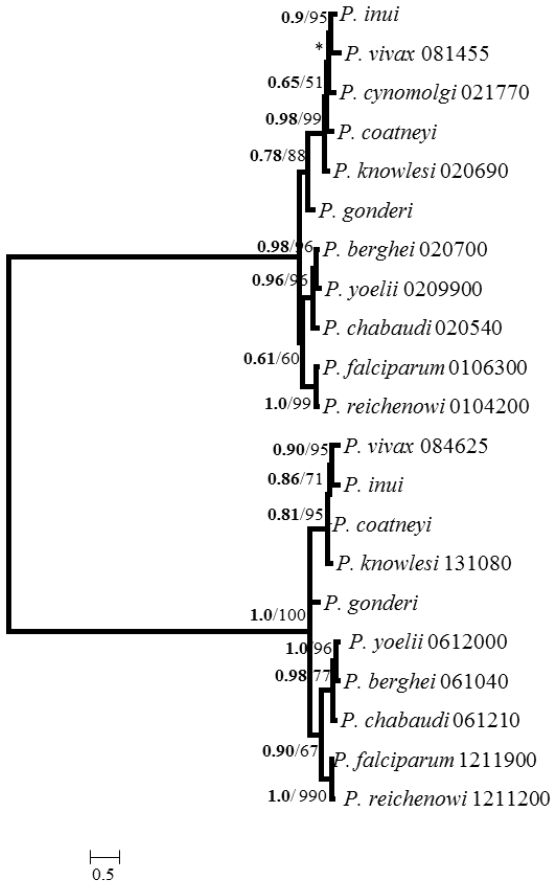
Asparagine tRNA ligase



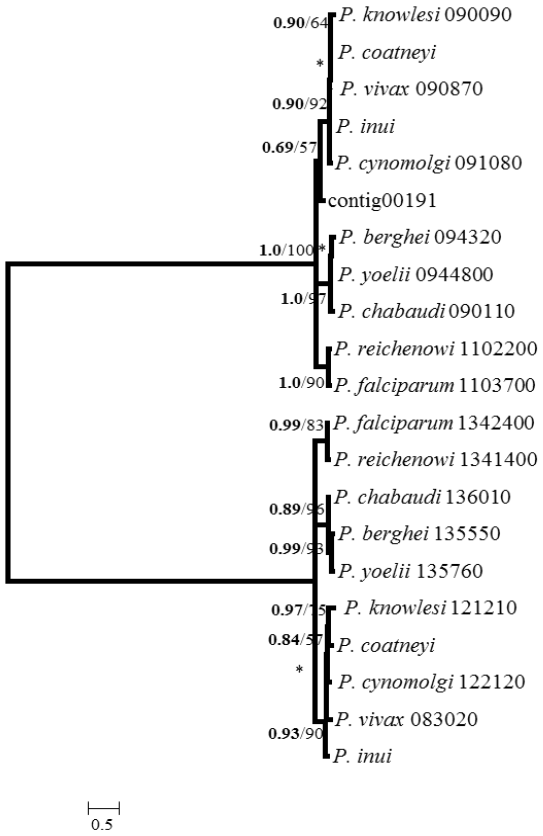
ATP dependent DNA helicase



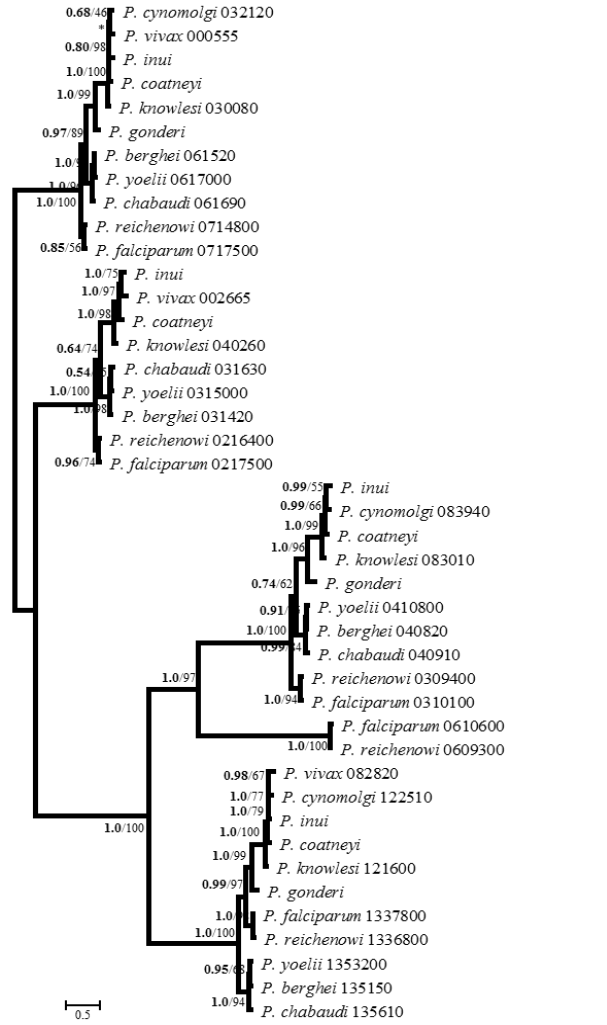
Biotin acyl-CoA carboxylase



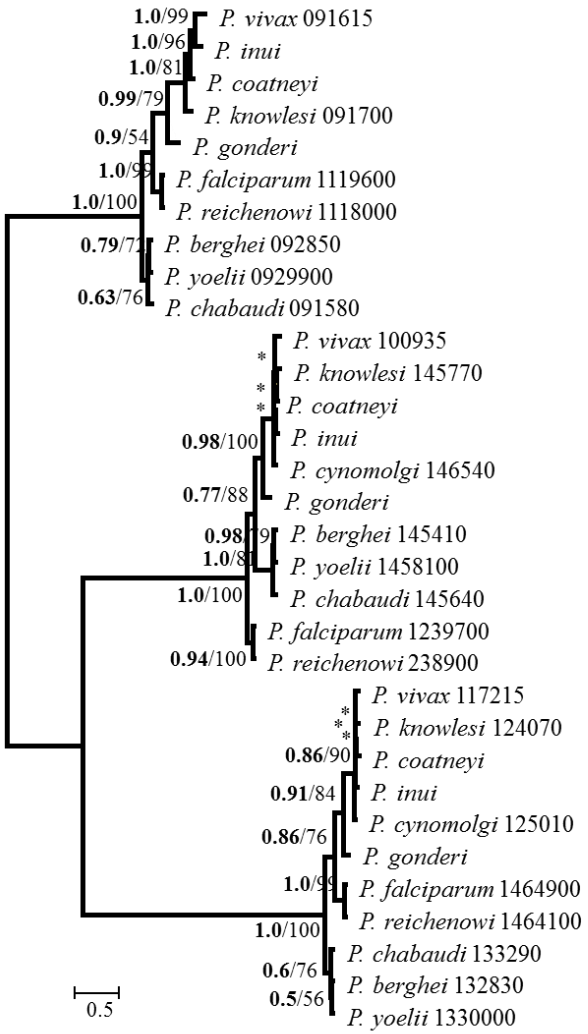
Calcium transporting ATPase (SERCA)



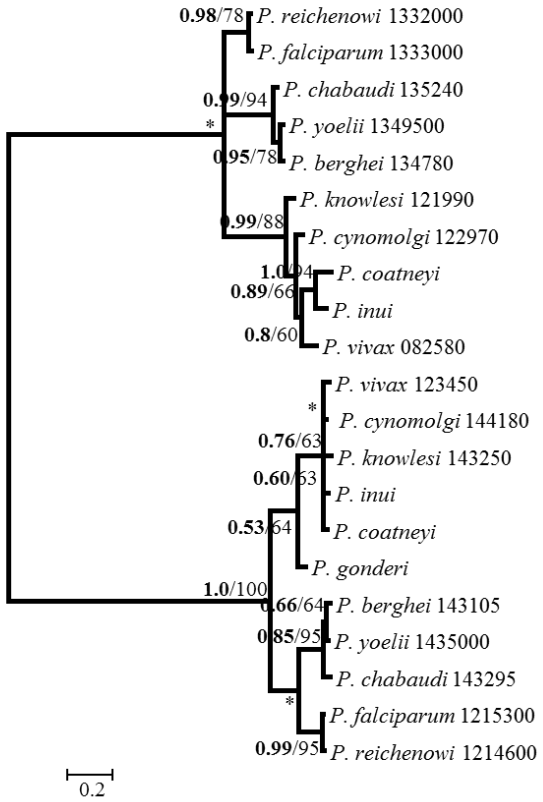
Casein kinase II beta chain



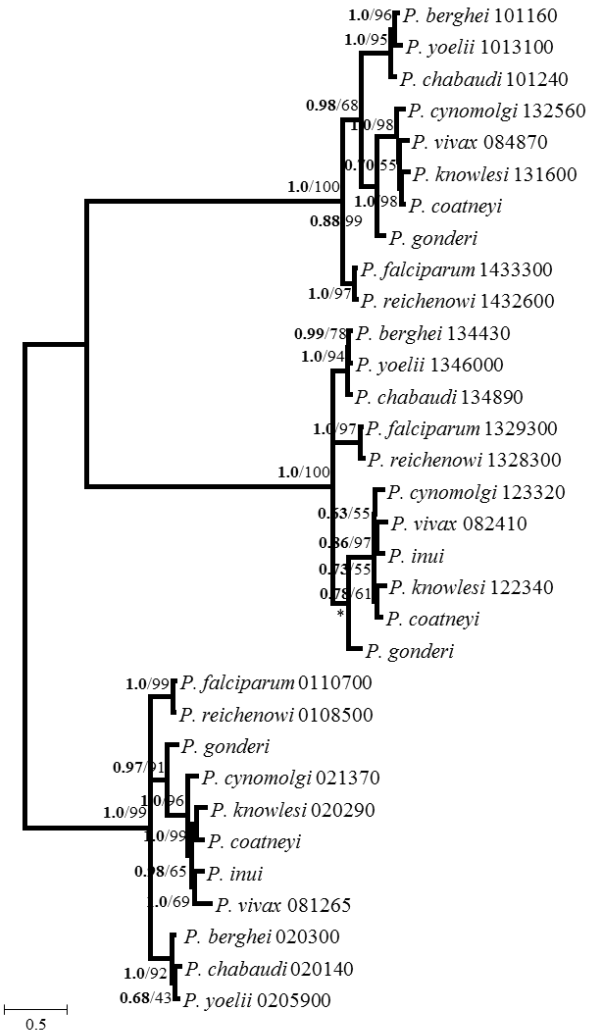
Calcium Dependent Protein Kinase (CDPK)



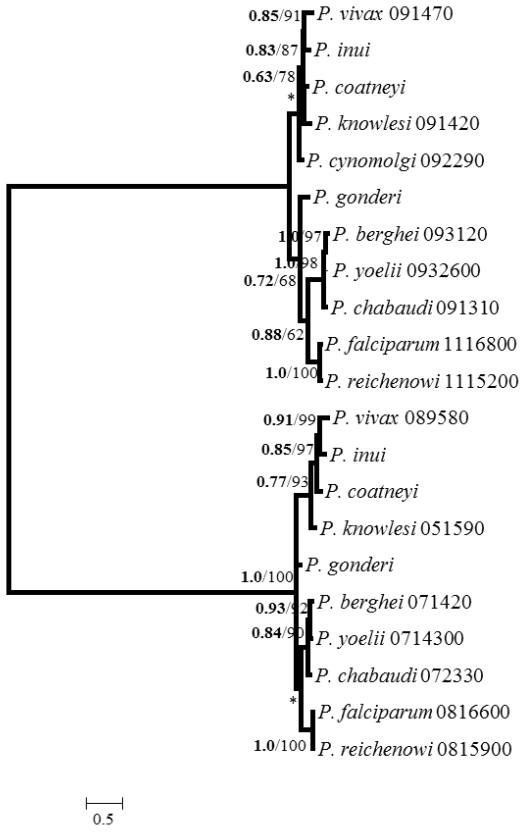
Cell division protein



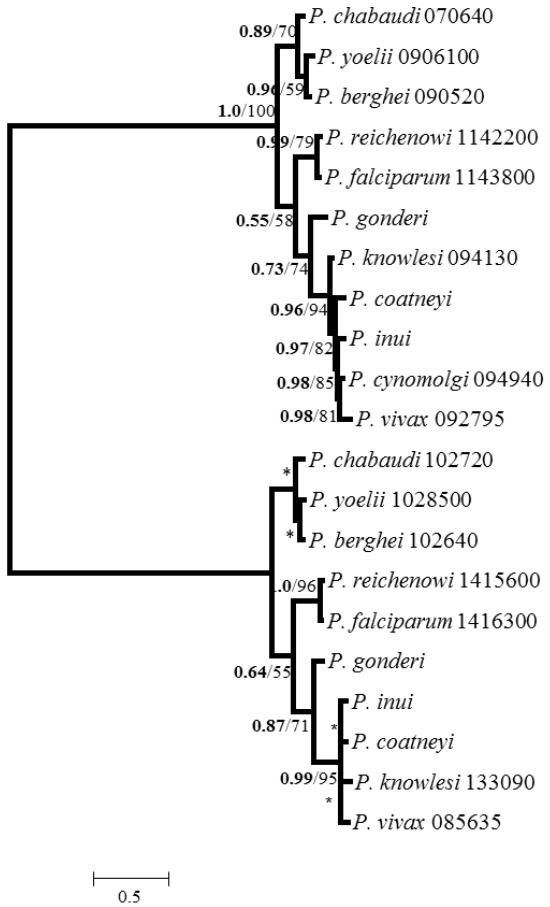
Chaperonin putative CPN



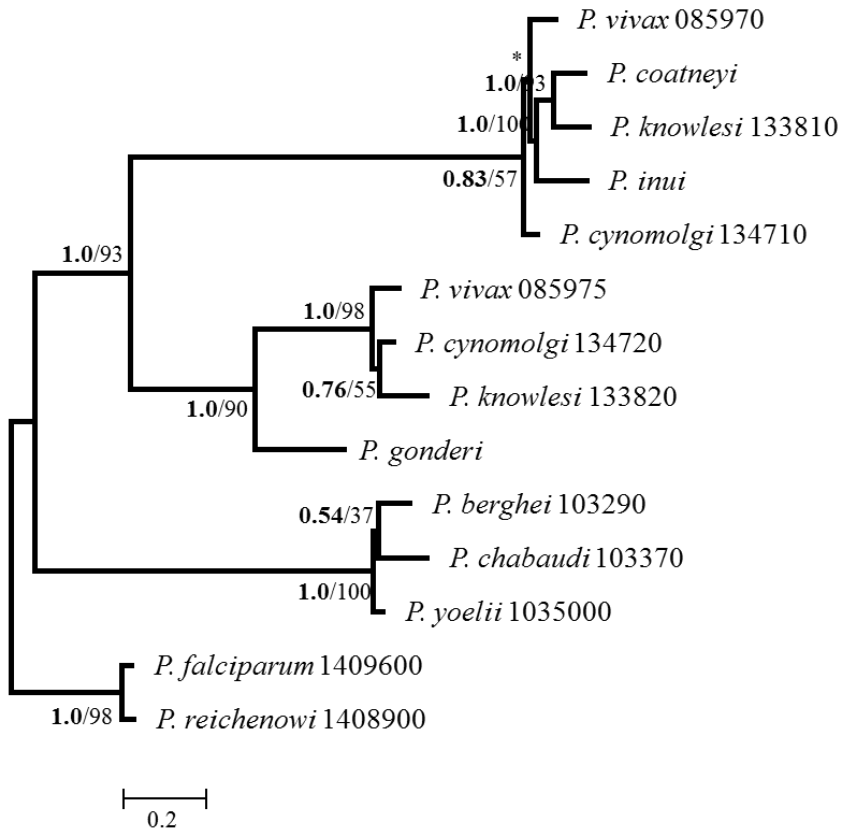
Chromatin assembly factor 1subunit



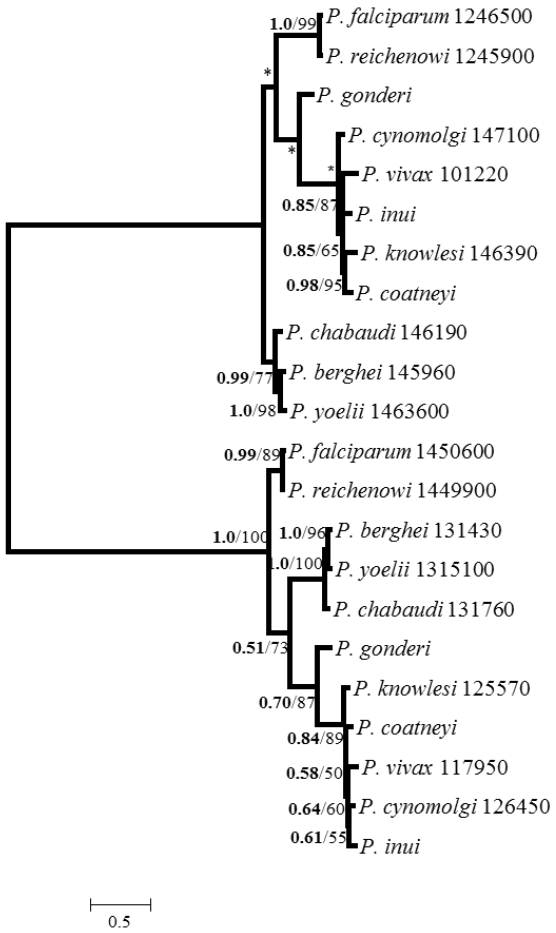
ClpB protein



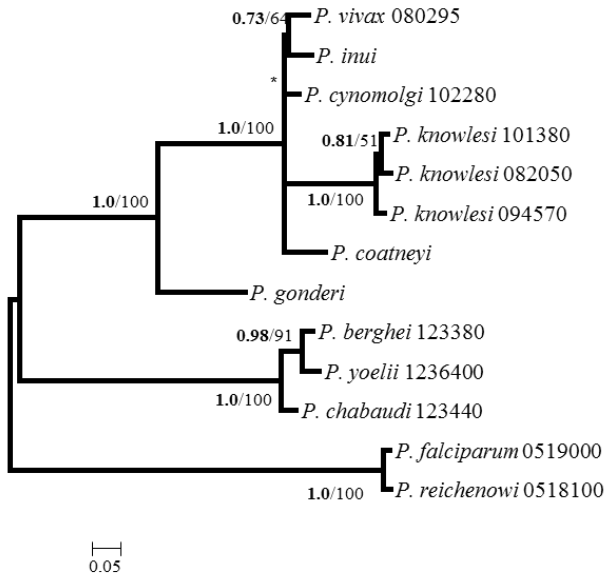
Conserved and hypothetical *Plasmodium* protein



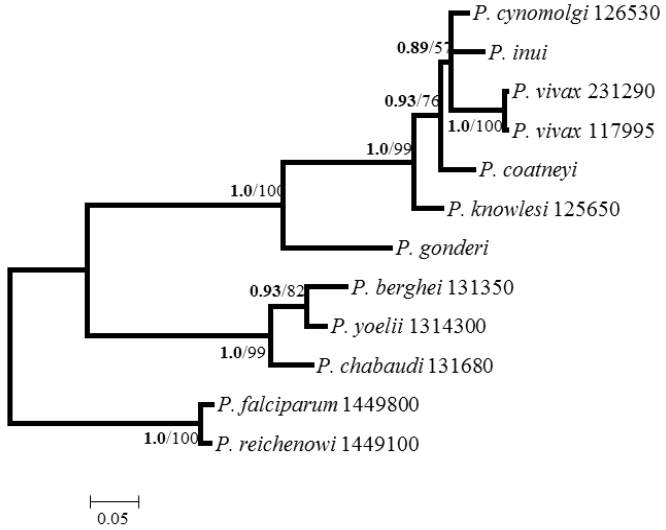
Conserved *Plasmodium* protein



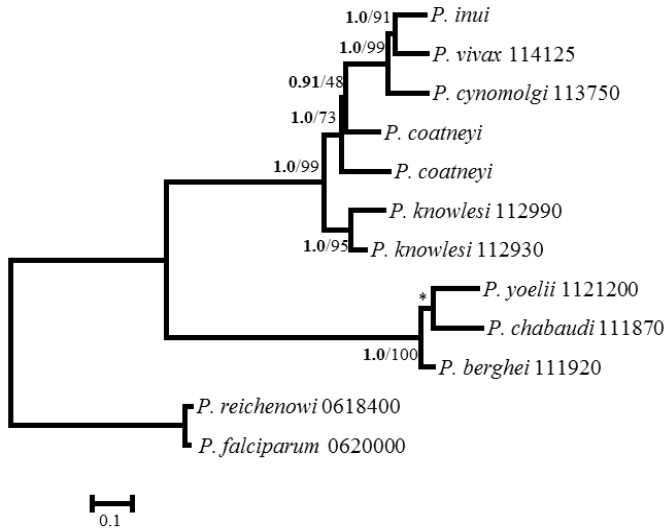
Conserved *Plasmodium* protein unknown function 2



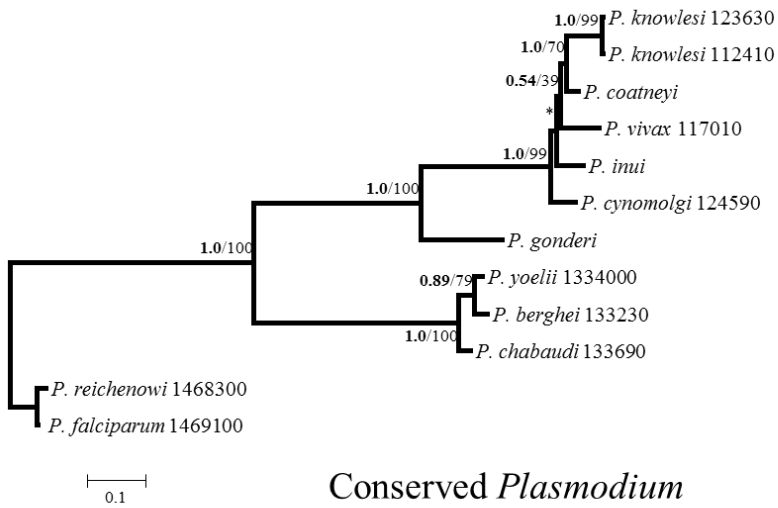
Conserved *Plasmodium* protein unknown function 3



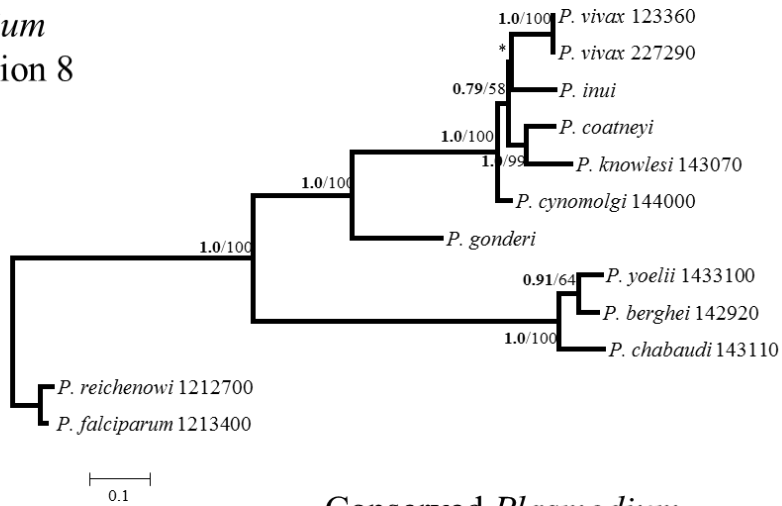
Conserved *Plasmodium* protein unknown function 4



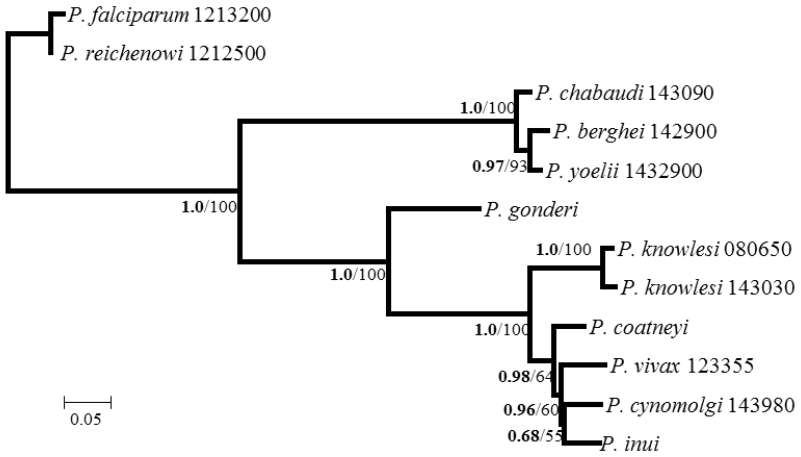
Conserved *Plasmodium* protein unknown function 6



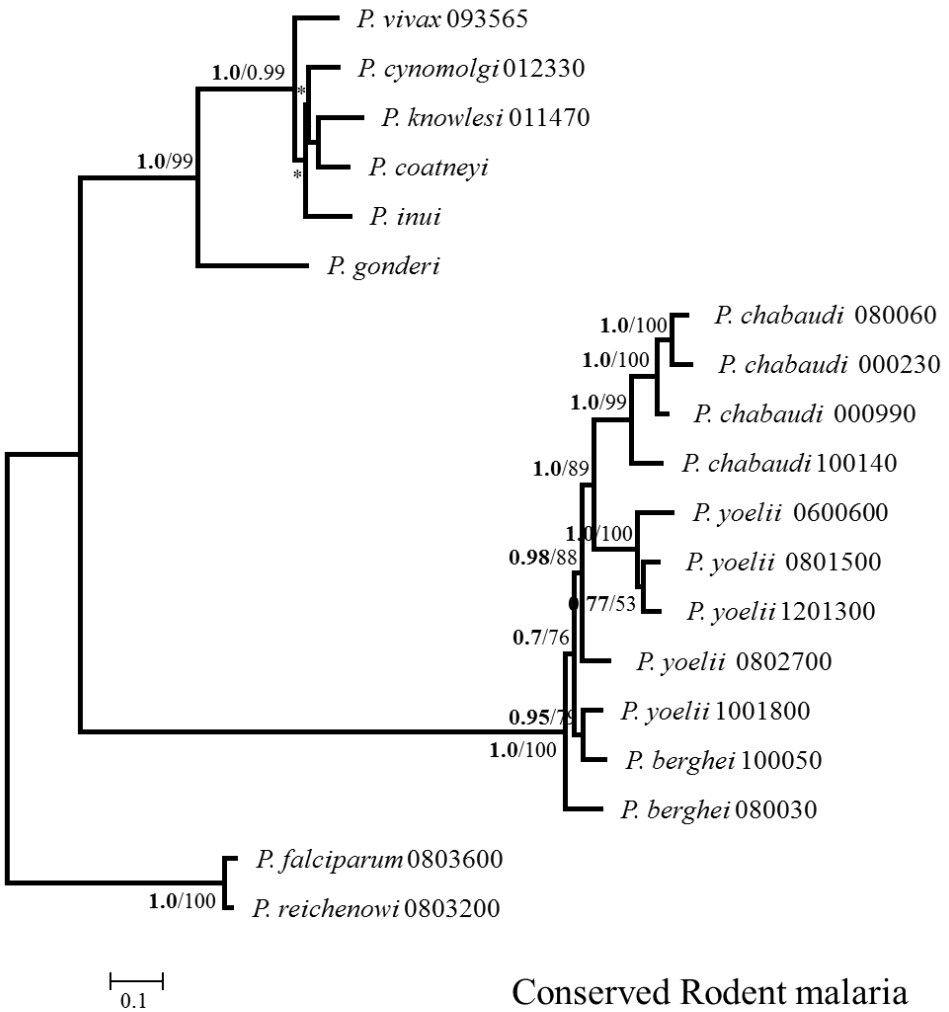
Conserved *Plasmodium*
protein unknown function 8



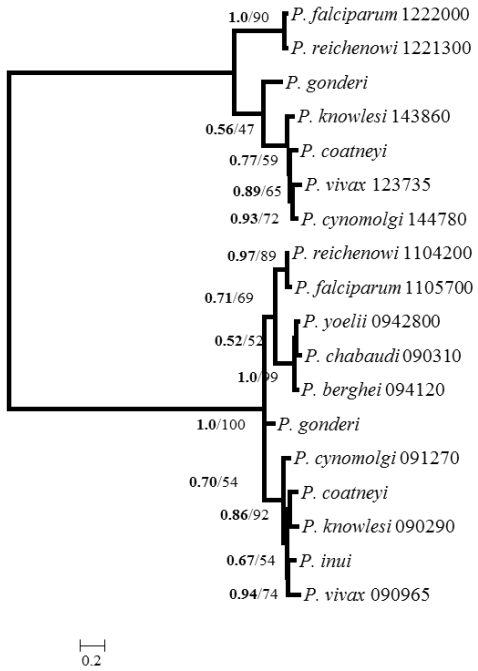
Conserved *Plasmodium*
protein unknown function 11



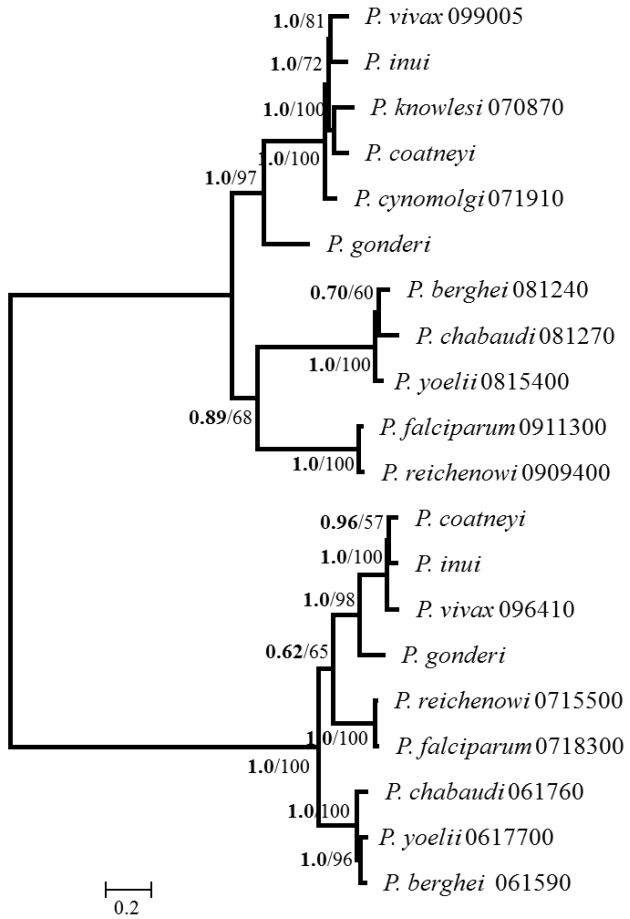
Conserved *Plasmodium*
protein unknown function 12



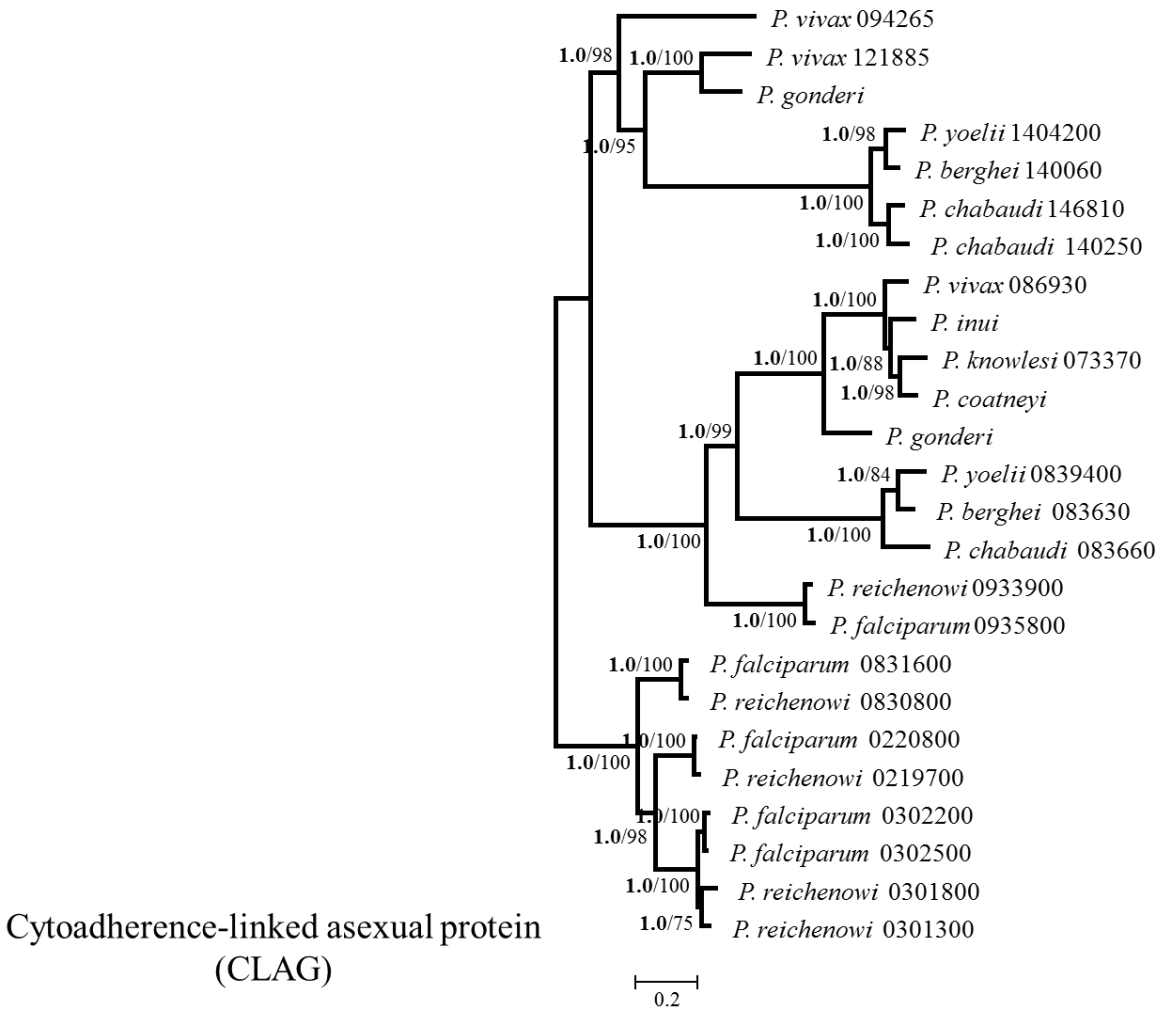
Conserved Rodent malaria protein unknown function



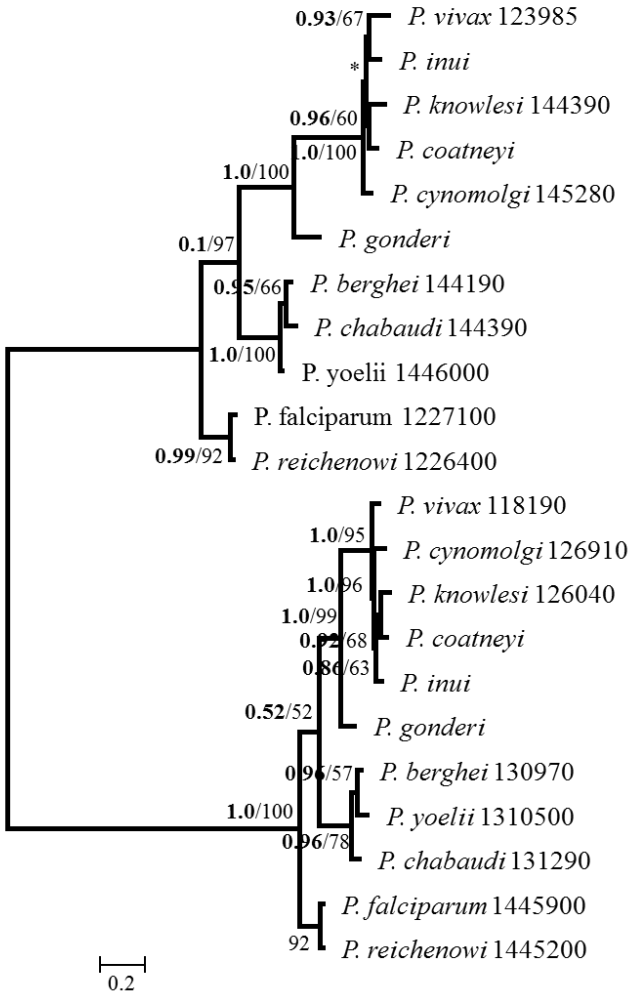
Conserved protein unknown function



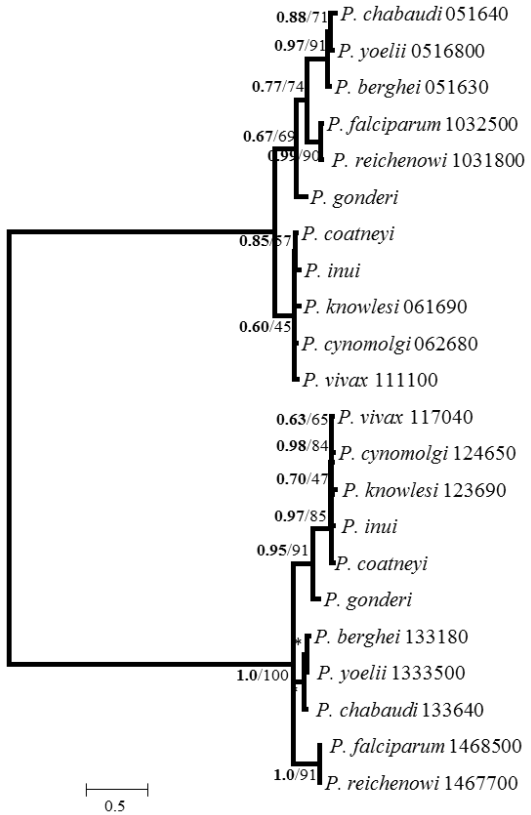
Cysteine repeat modular protein (CRMP)



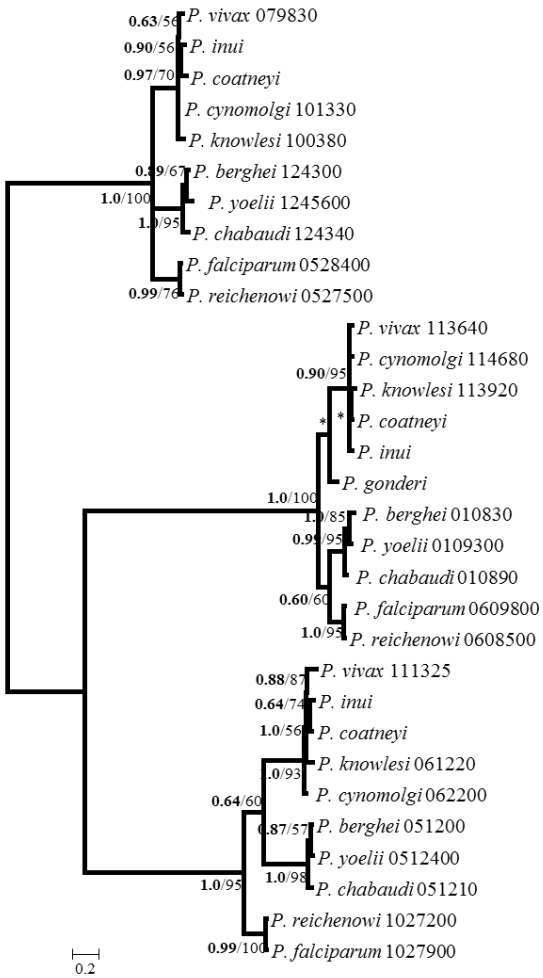
Cytoadherence-linked asexual protein (CLAG)



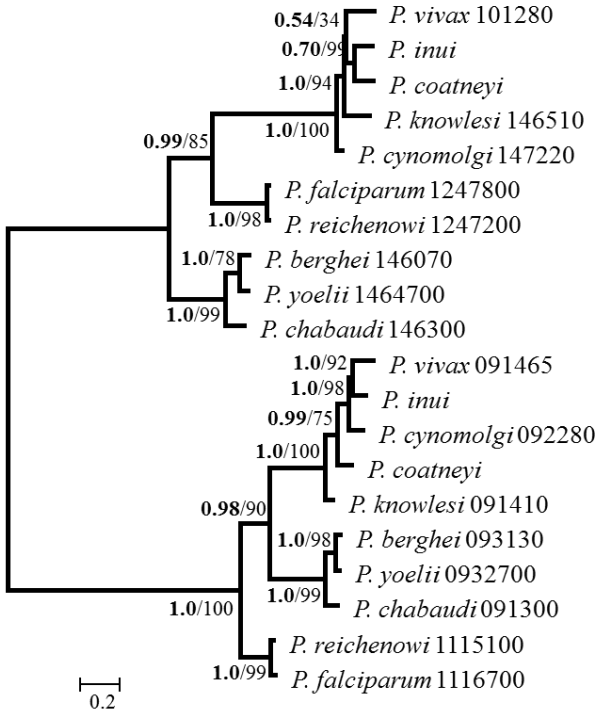
DEAD DEAH box ATP dependent RNA helicase



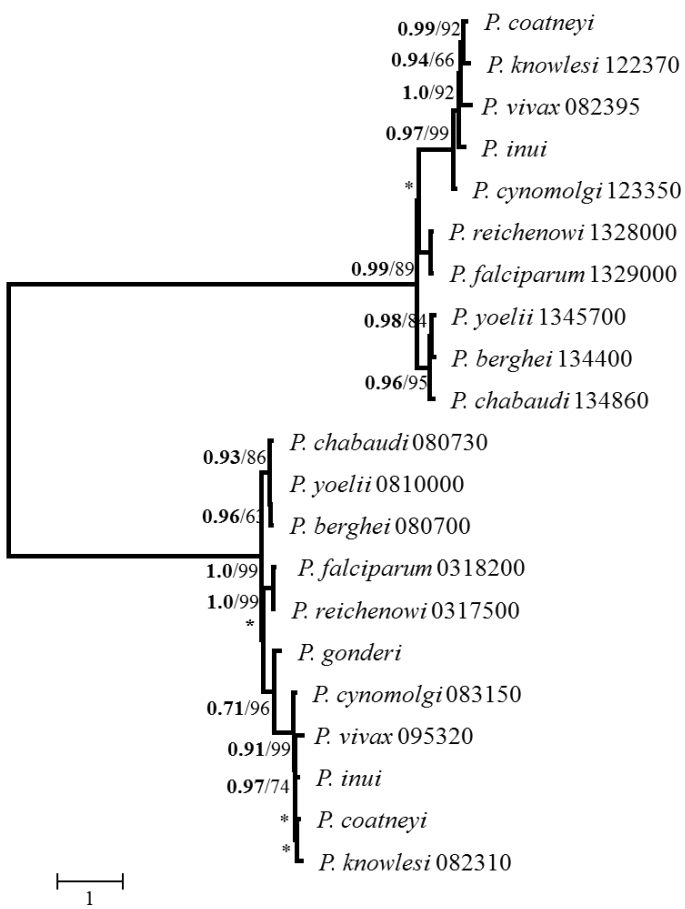
DER1 like protein



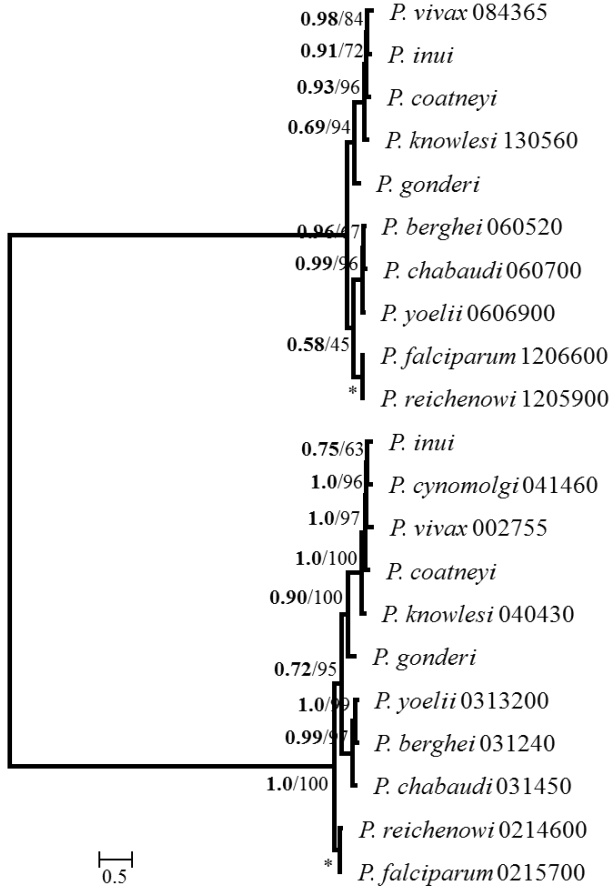
DHHC type zinc finger protein



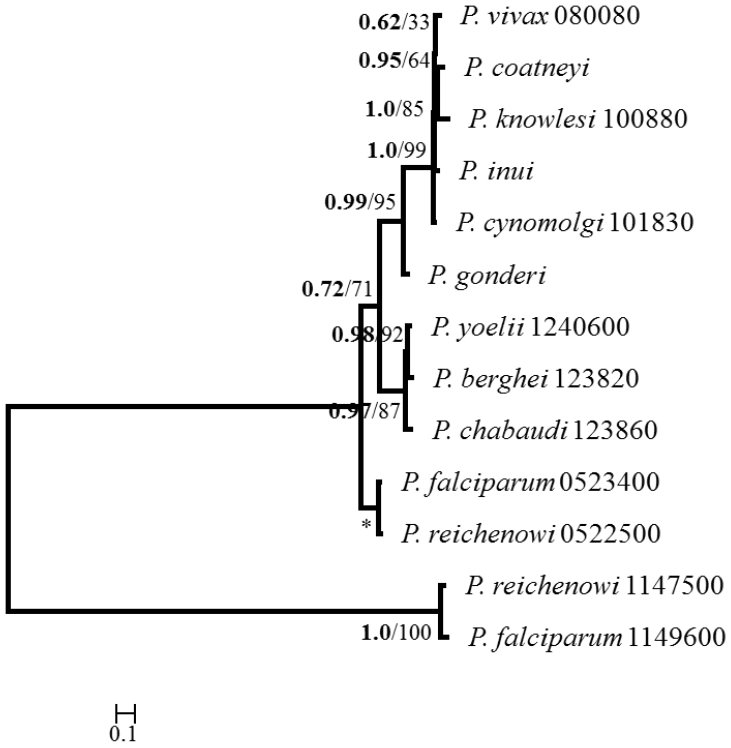
Dipeptidyl amino peptidase DPAP



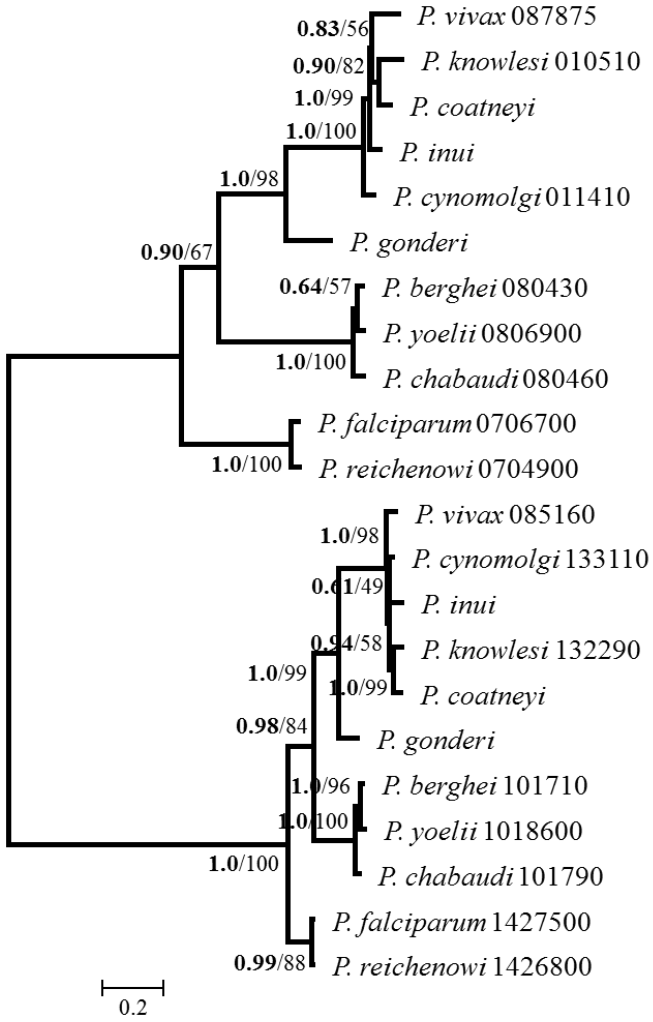
DNA directed RNA polymerase II



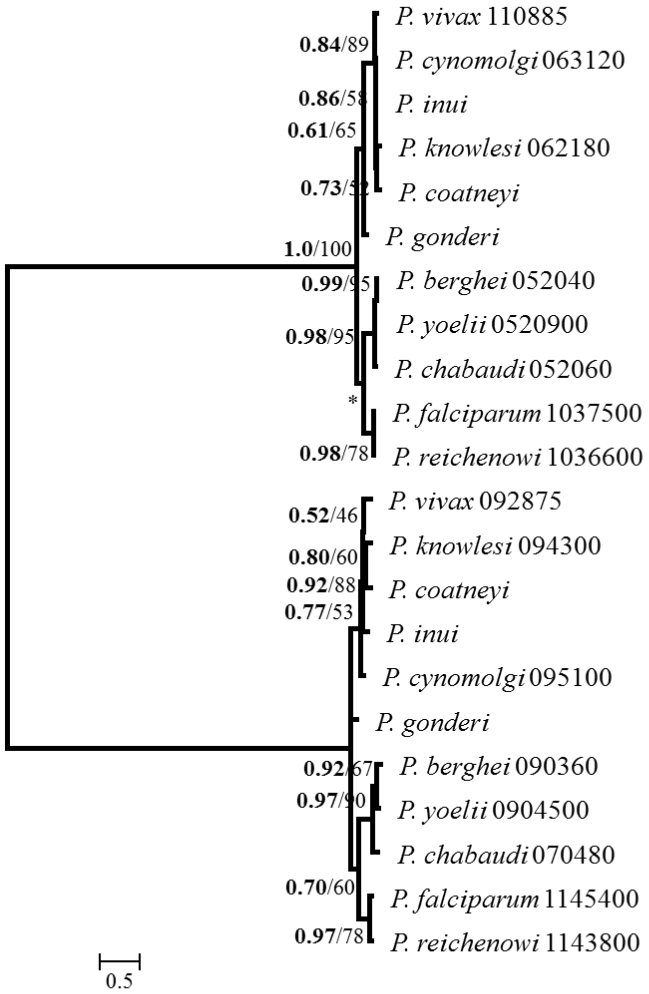
DNA directed RNA polymerase second largest subunit



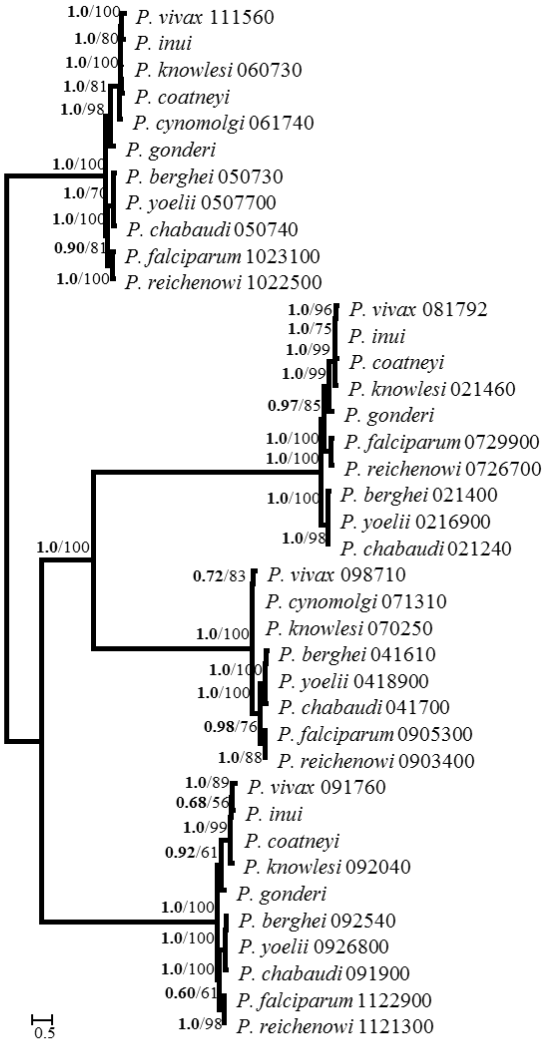
DnaJ protein 2



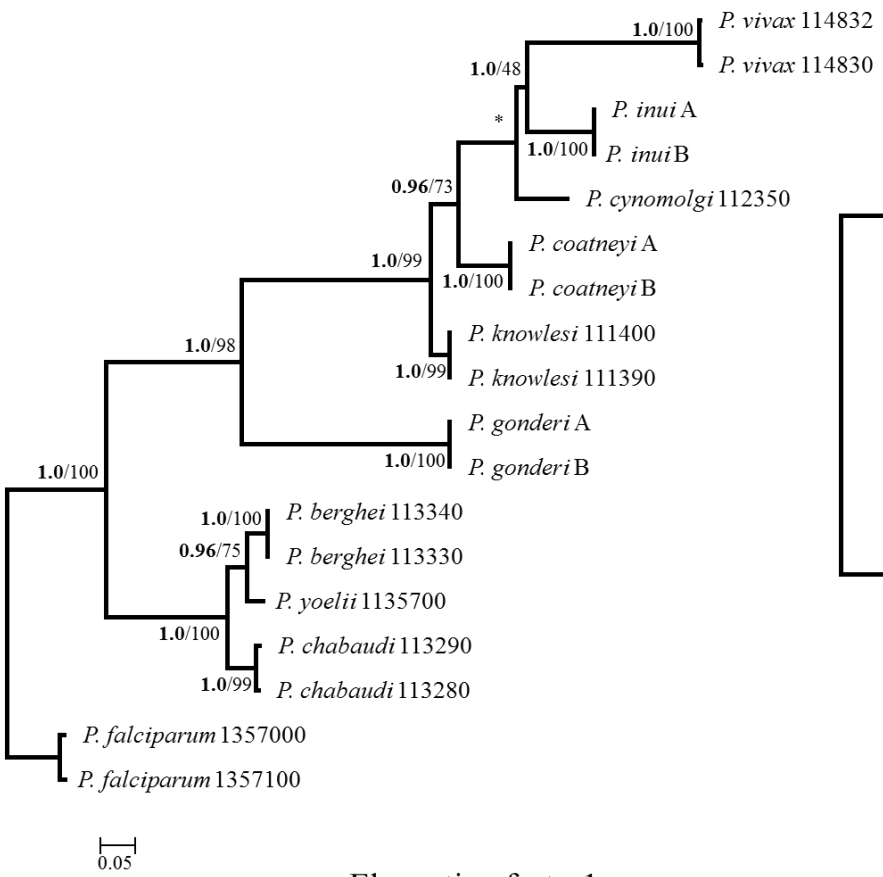
DNA mismatch repair protein



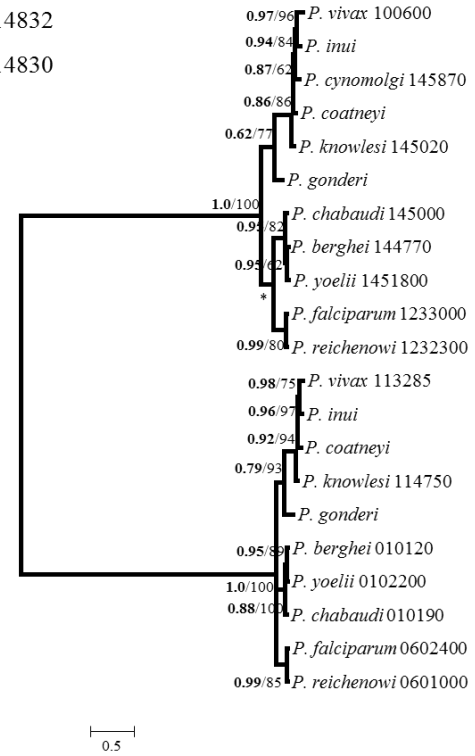
Dynamin like protein



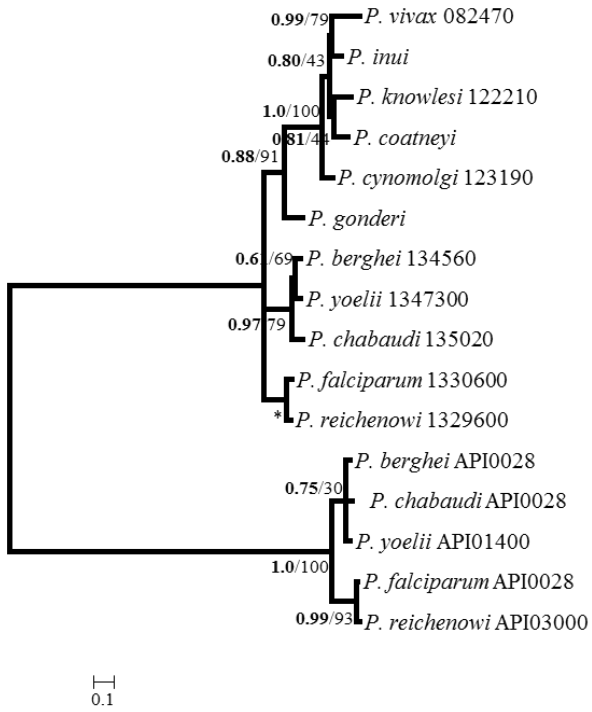
Dynein heavy chain



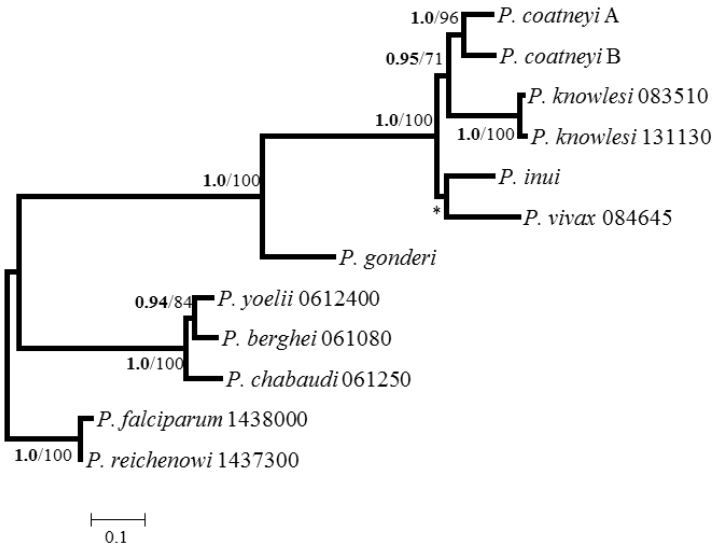
Elongation factor1



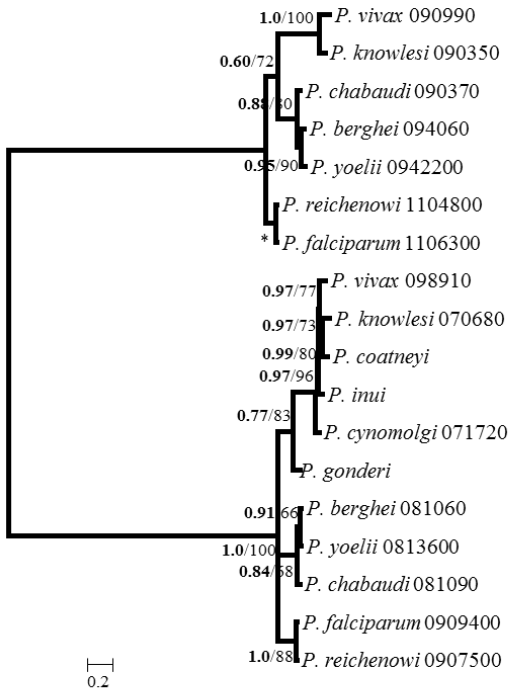
Elongation factor G



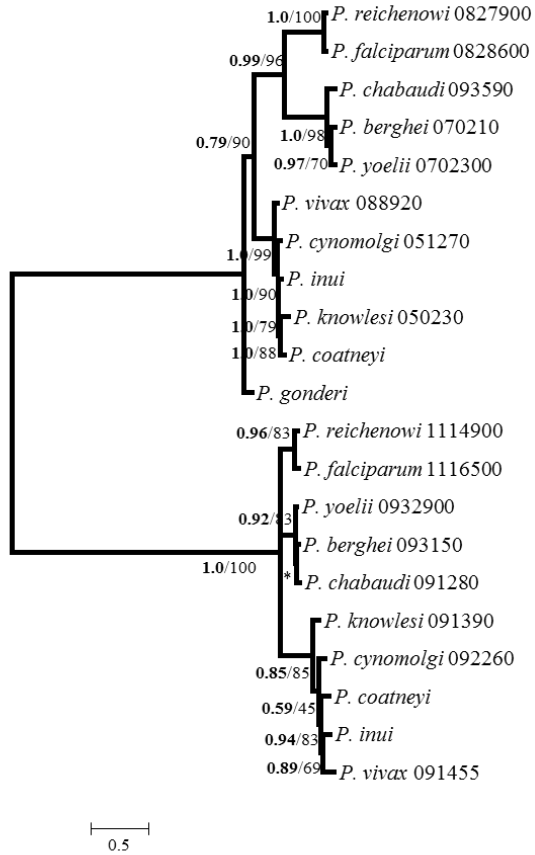
Elongation factor tufA



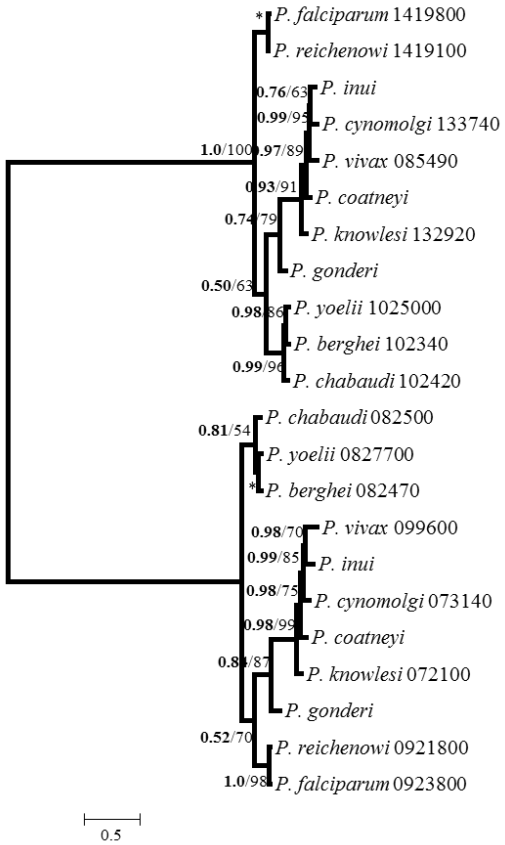
Eukaryotic initiation factor 2a



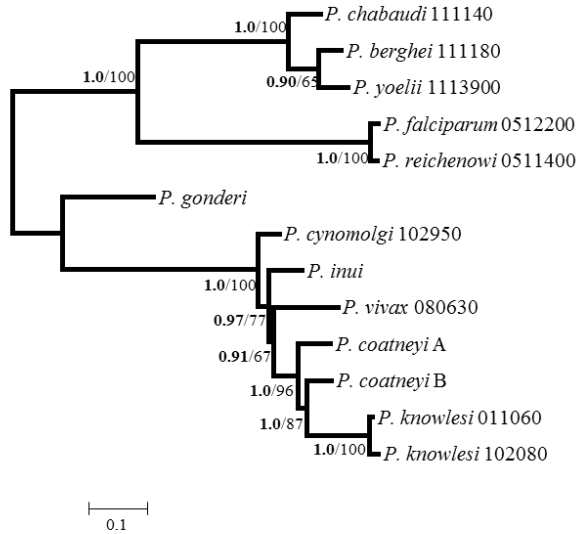
Exonuclease



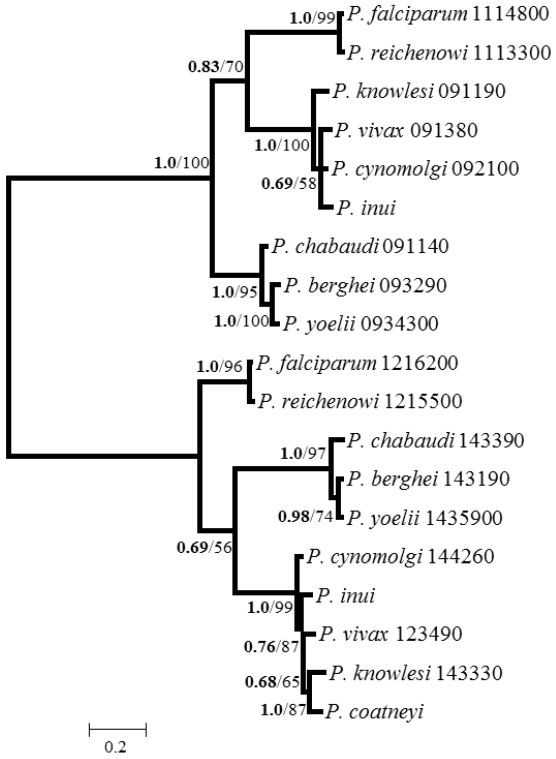
Folate transporter FT



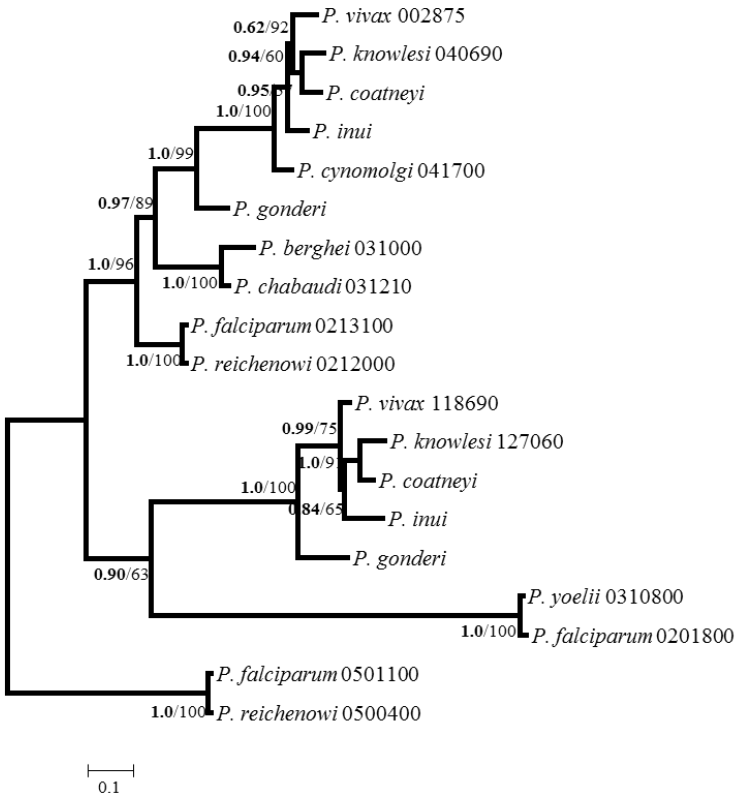
Glutathione reductase GR



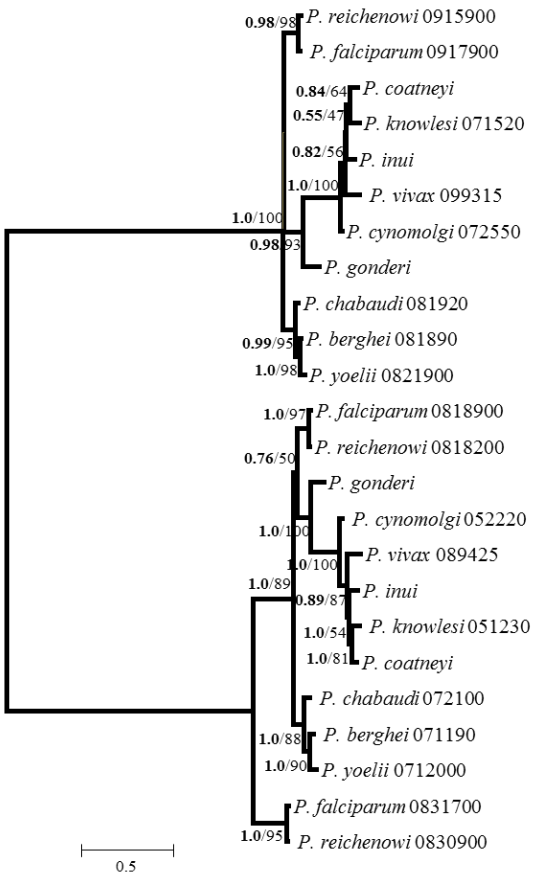
Glutathione synthetase



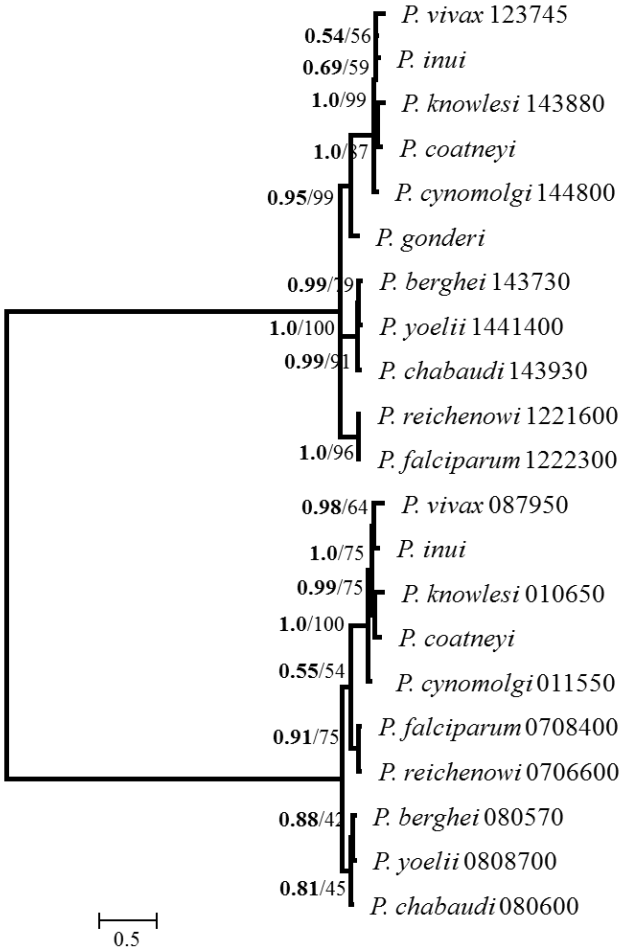
Glycerol 3 phosphate dehydrogenase



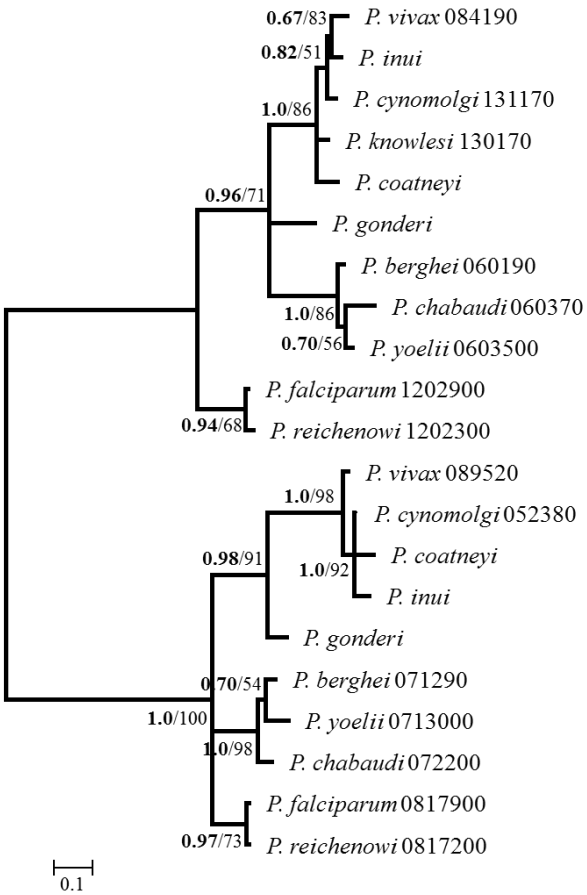
Heat shock protein 40



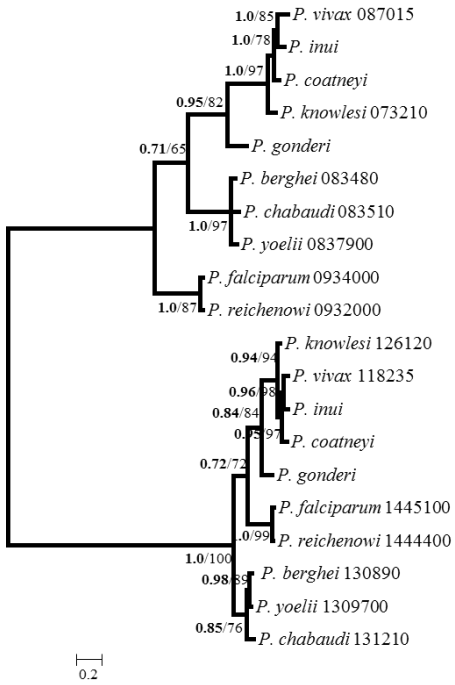
Heat Shock protein 70



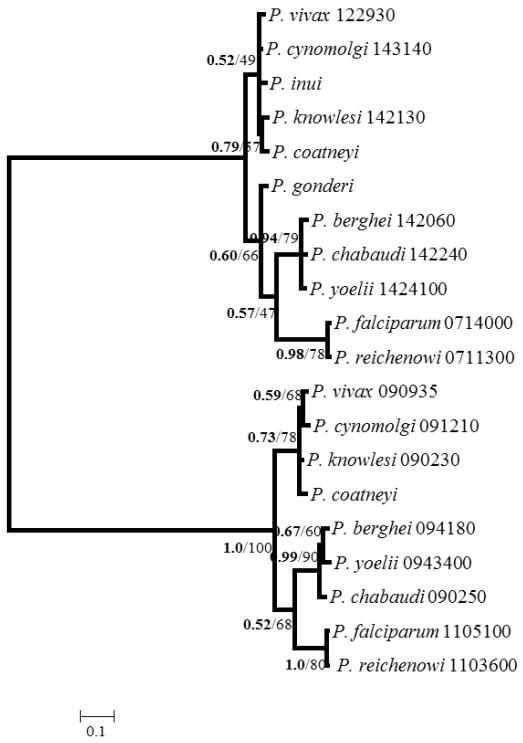
Heat shock protein 90



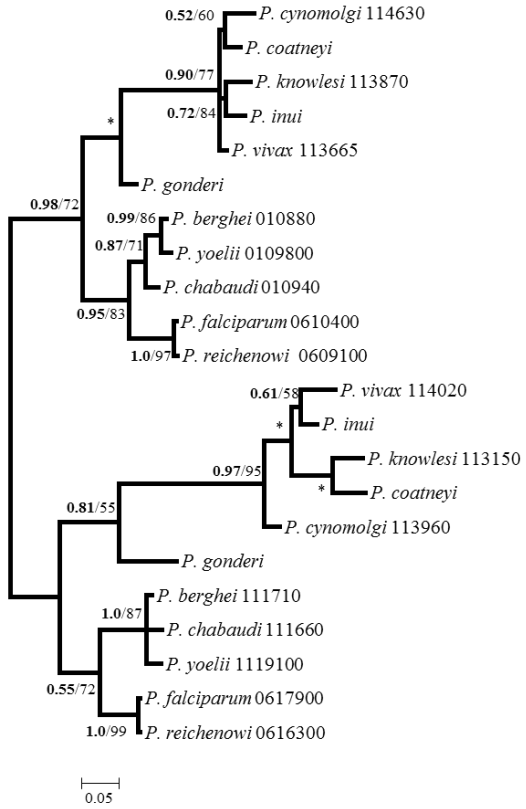
High mobility group protein B HMGB



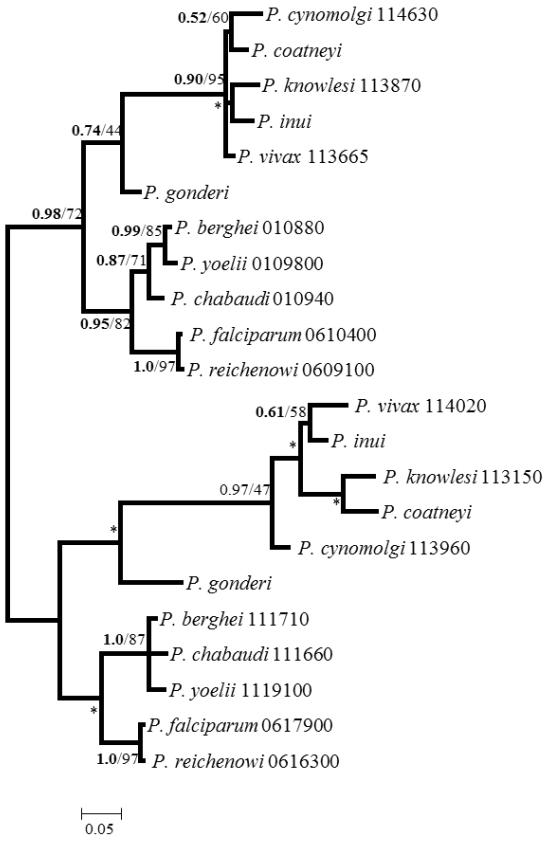
Histidine tRNA ligase



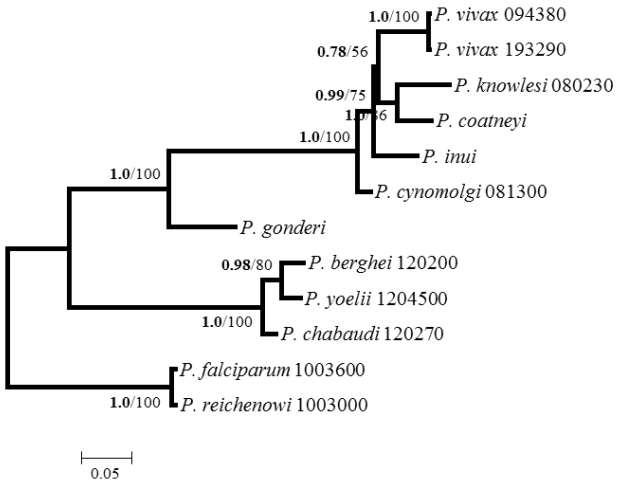
Histone H2B



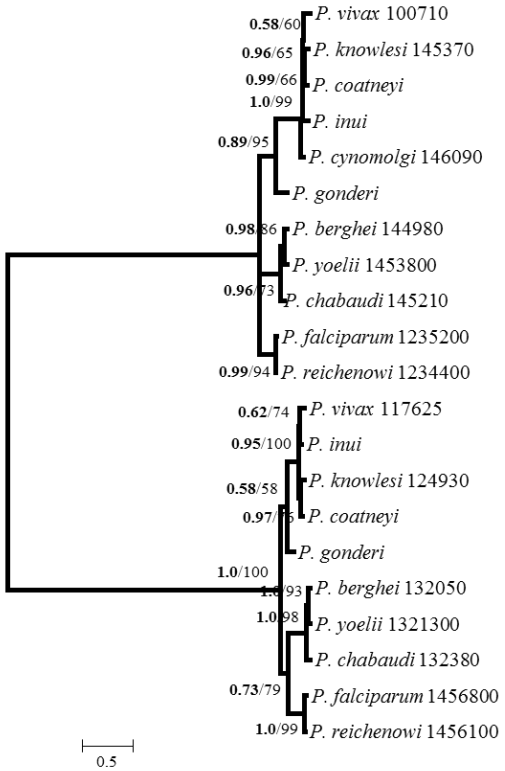
Histone H3



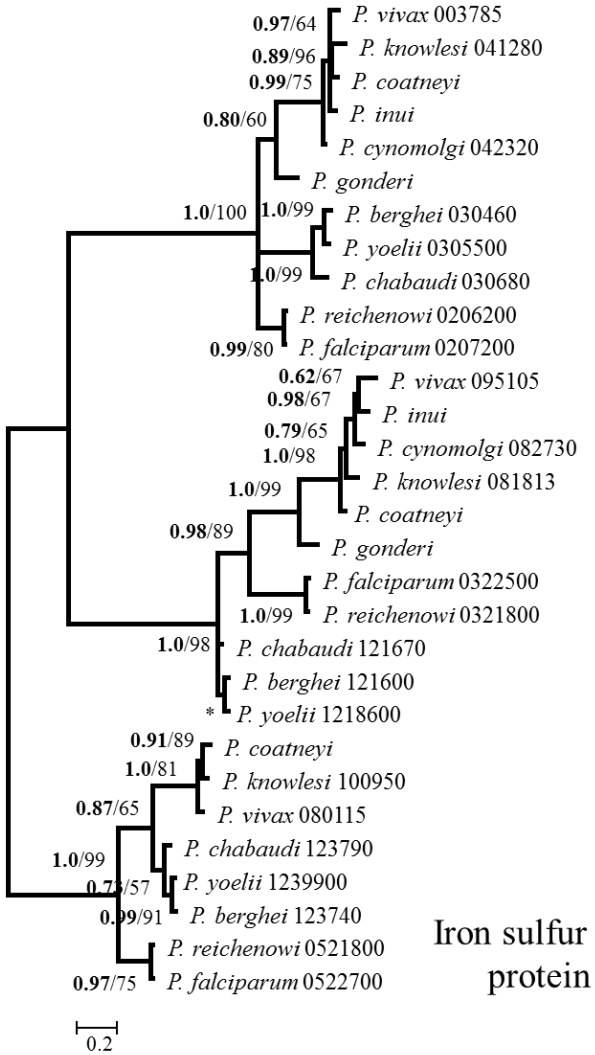
Hypothetical protein



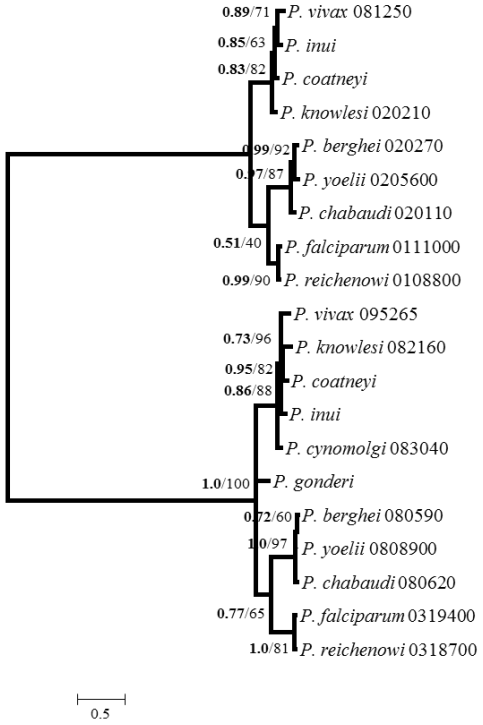
Inner membrane complex protein 1c (IMC1c)



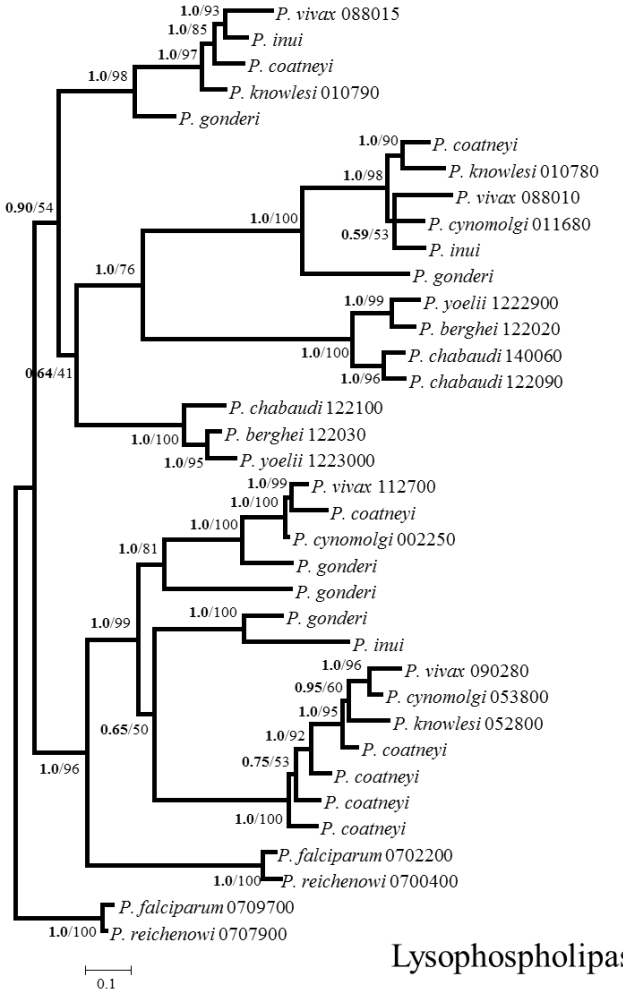
Inorganic pyrophosphatase VP



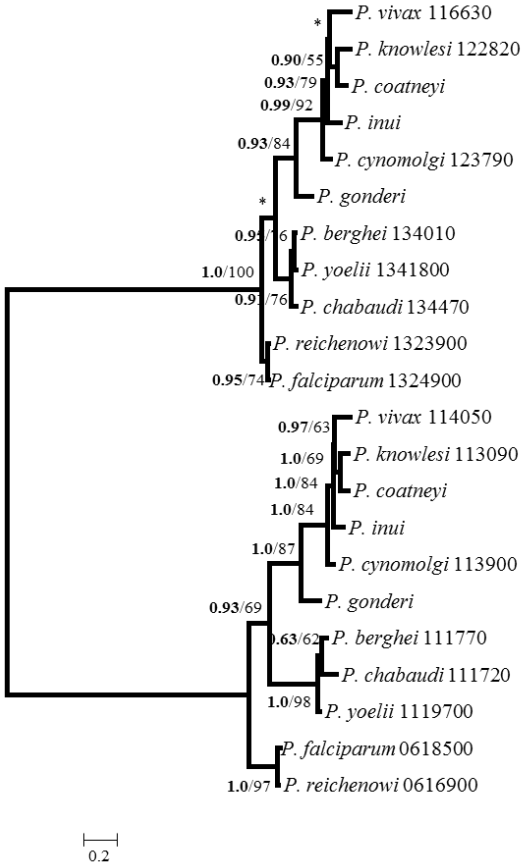
Iron sulfur assembly protein SufA



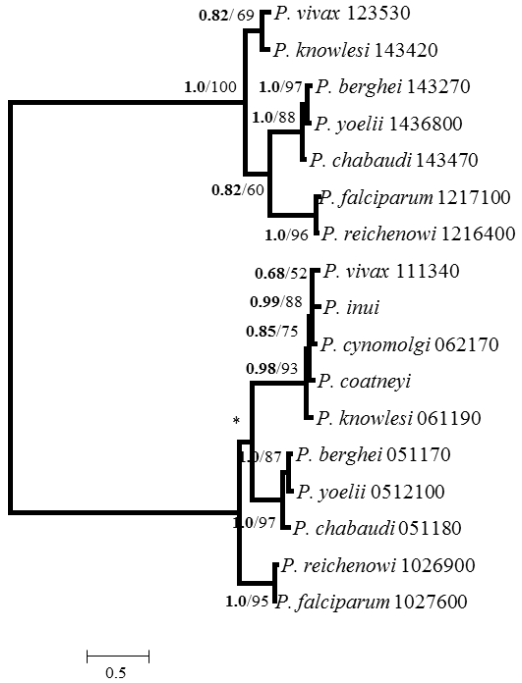
Kinesin 8



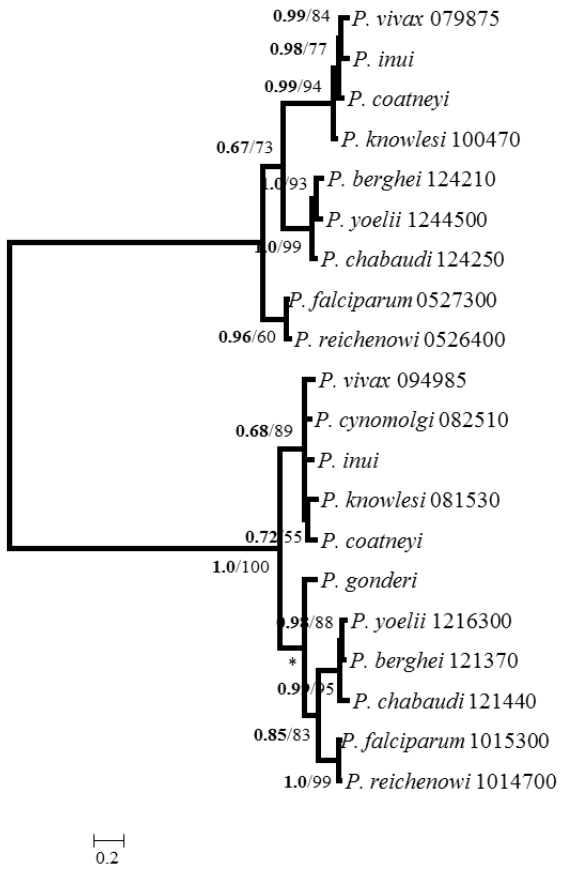
Lysophospholipase



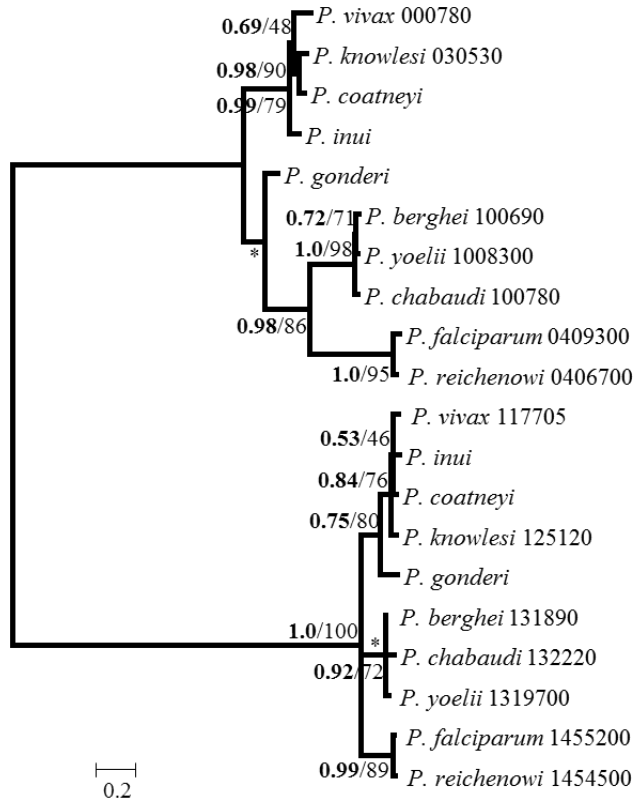
Malate dehydrogenase MDH



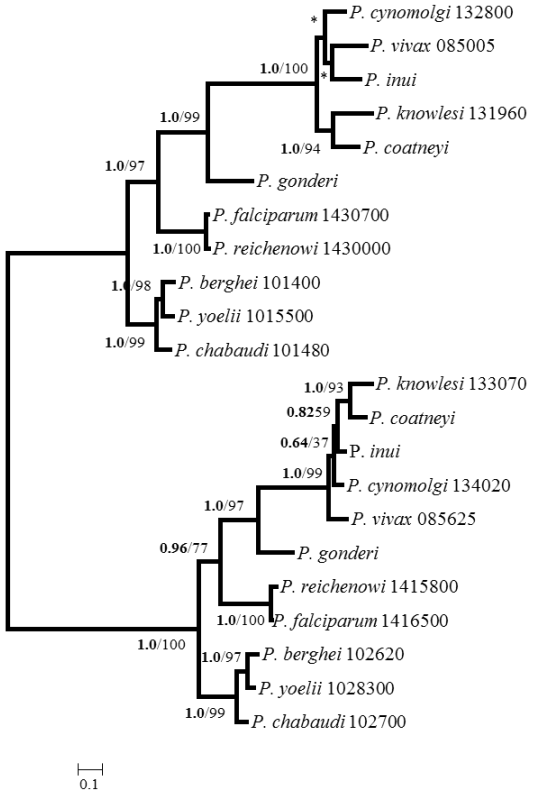
Meiotic recombination protein SPO11



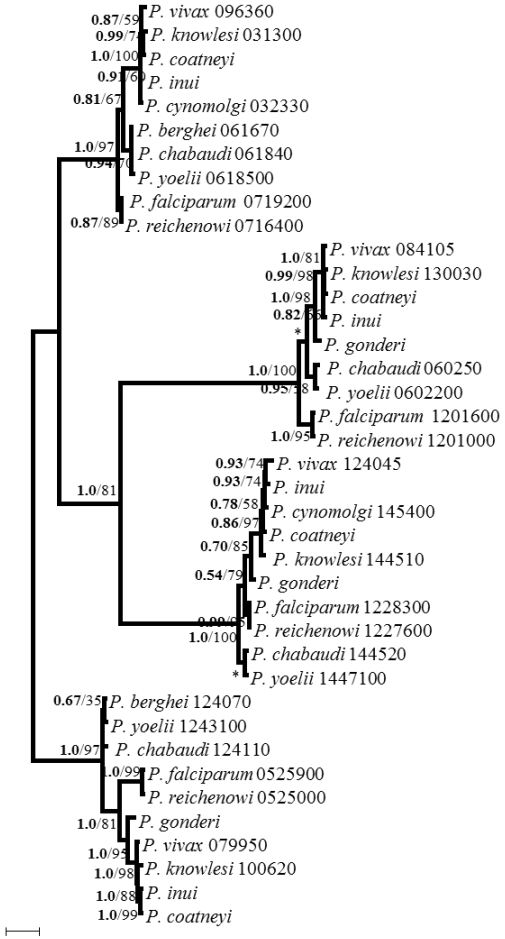
Methionine amino peptidase



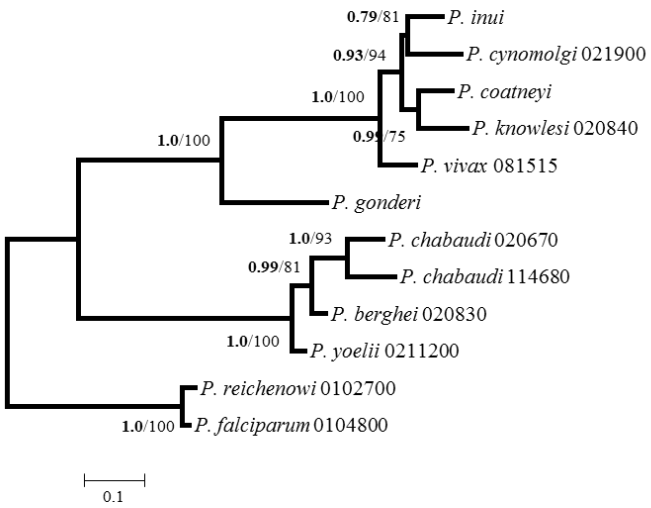
Methyltransferase



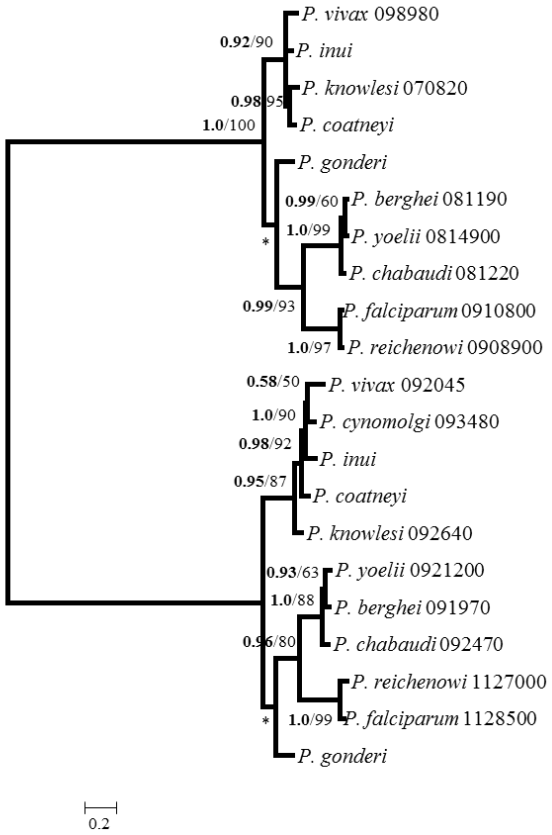
NADP specific glutamate dehydrogenase GDH



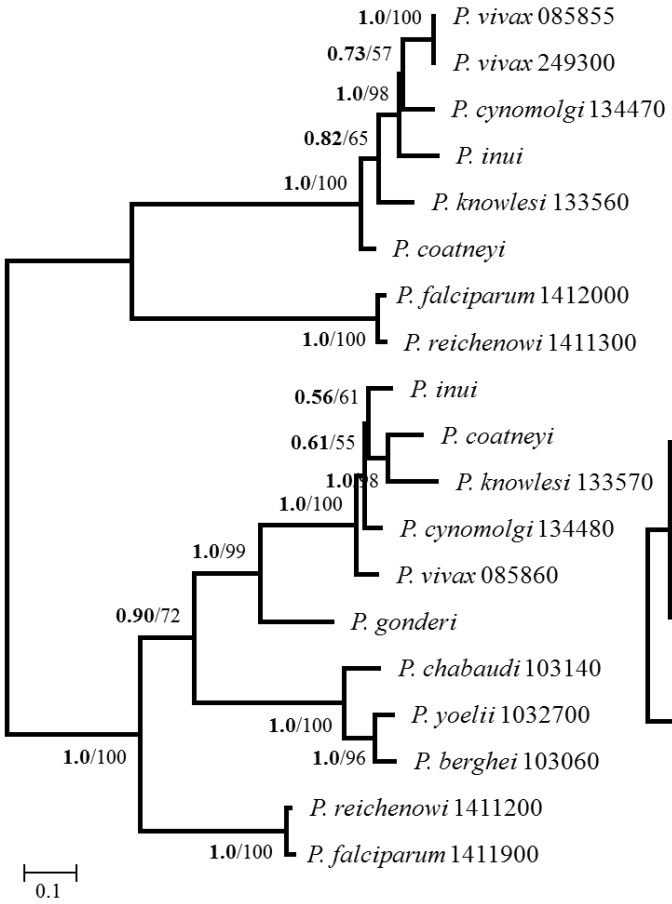
NIMA related kinase (NEK)



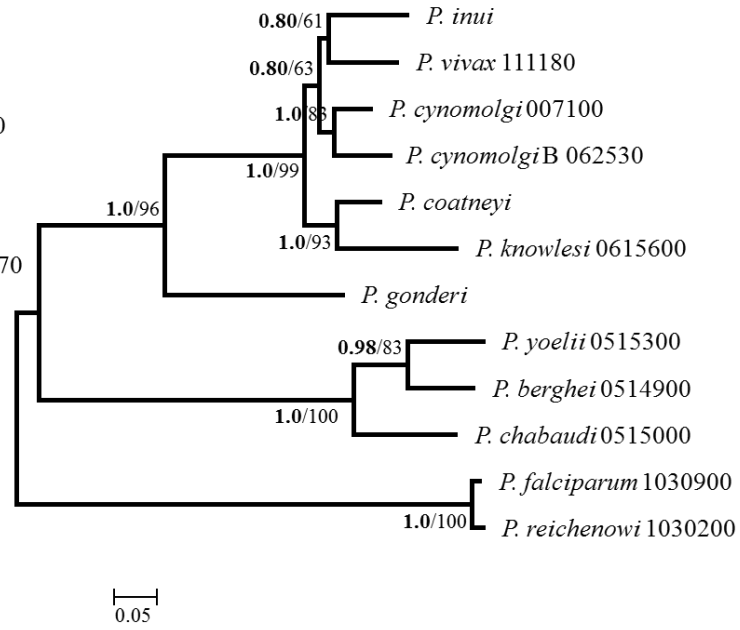
Novel putative transporter 1 NPT1



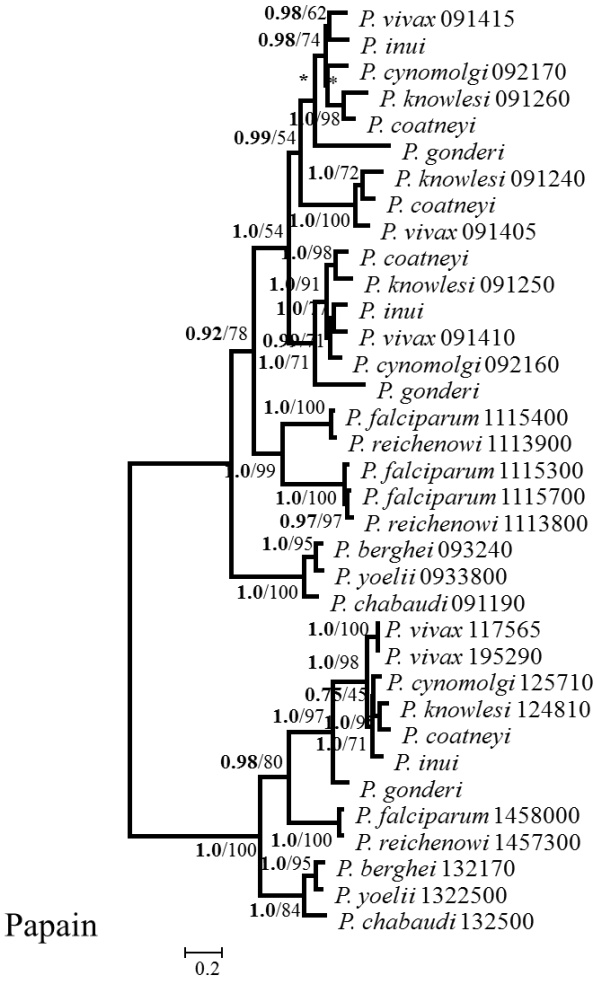
Nucleotide binding protein

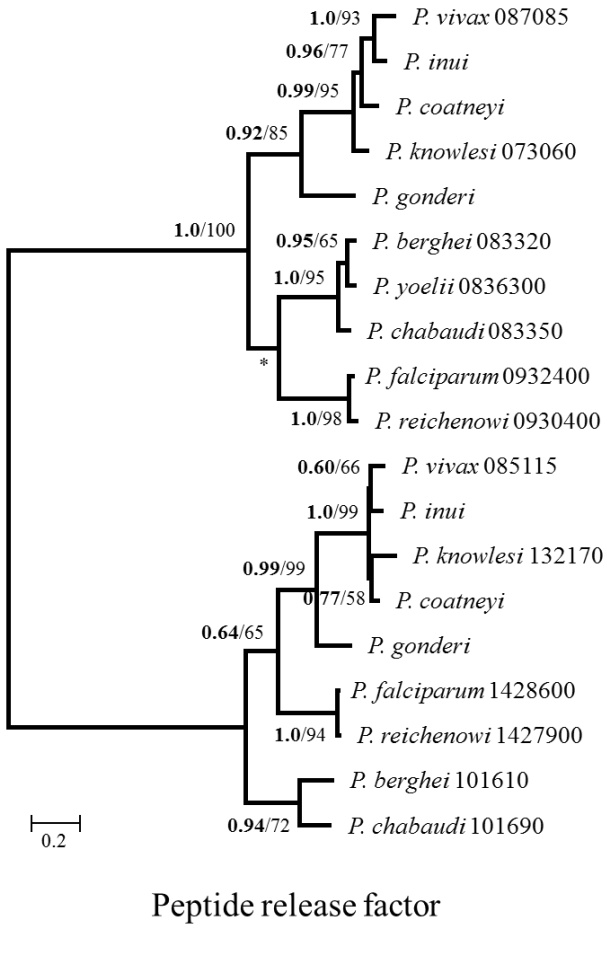


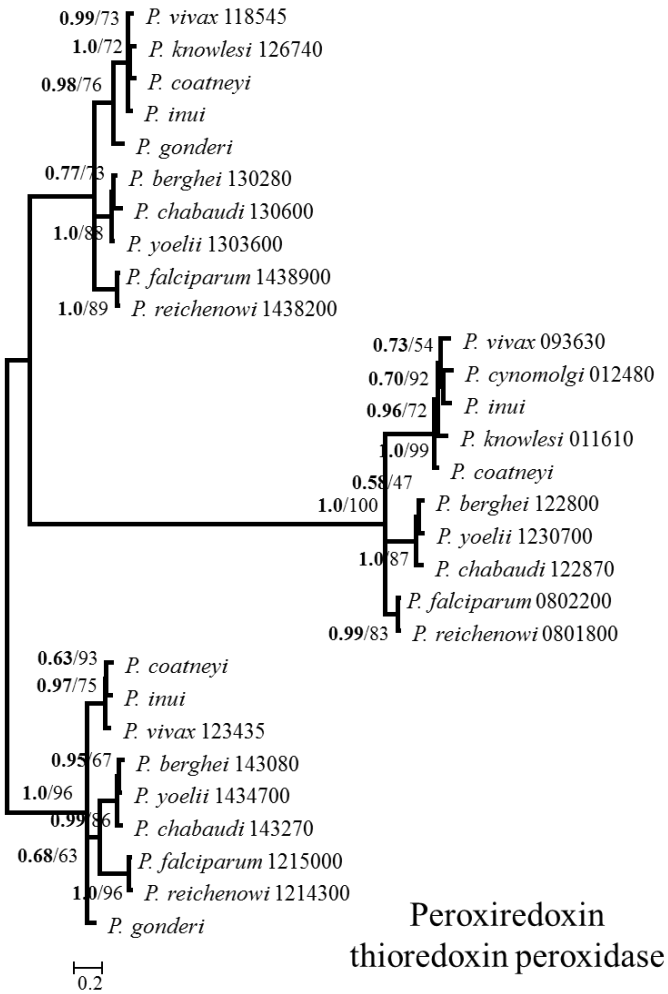
P1s1 nuclease



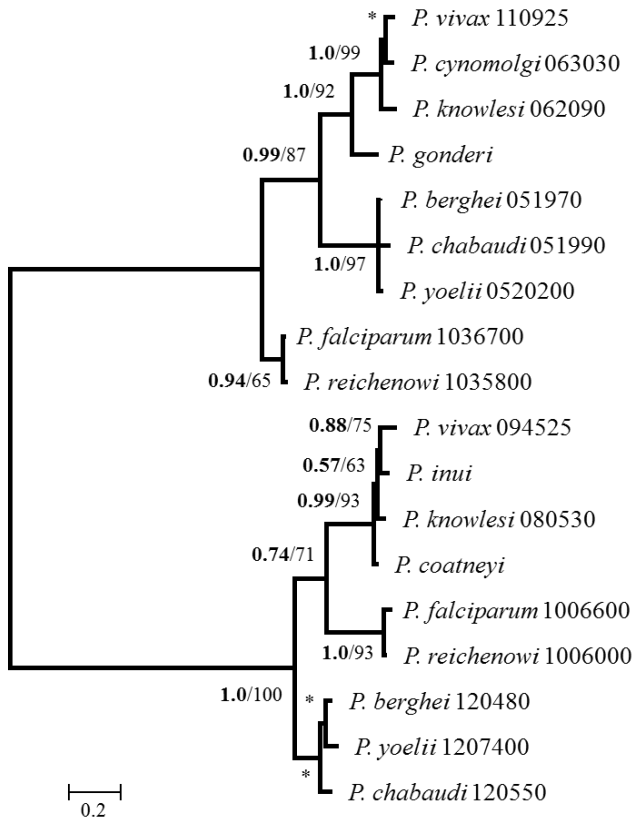
P28


0.98/62 P. vivax 091415
0.98/74 P. inui
P. cynomolgi 092170
P. knowlesi 091260
1.0/98 P. coatneyi
0.99/54 P. gonderi
1.0/72 P. knowlesi 091240
P. coatneyi
1.0/100 P. vivax 091405
1.0/54 P. coatneyi
1.0/91 P. knowlesi 091250
P. inui
0.92/78 P. vivax 091410
1.0/71 P. cynomolgi 092160
P. gonderi
1.0/100 P. falciparum 1115400
P. reichenowi 1113900
1.0/99 P. falciparum 1115300
1.0/100 P. falciparum 1115700
0.97/97 P. reichenowi 1113800
1.0/95 P. berghei 093240
P. yoelii 0933800
1.0/100 P. chabaudi 091190
1.0/100 P. vivax 117565
1.0/98 P. vivax 195290
P. cynomolgi 125710
P. knowlesi 124810
1.0/97 P. coatneyi
1.0/71 P. inui
0.98/80 P. gonderi
P. falciparum 1458000
1.0/100 P. reichenowi 1457300
1.0/100 1.0/95 P. berghei 132170
P. yoelii 1322500
1.0/84 P. chabaudi 132500
0.2
Papain

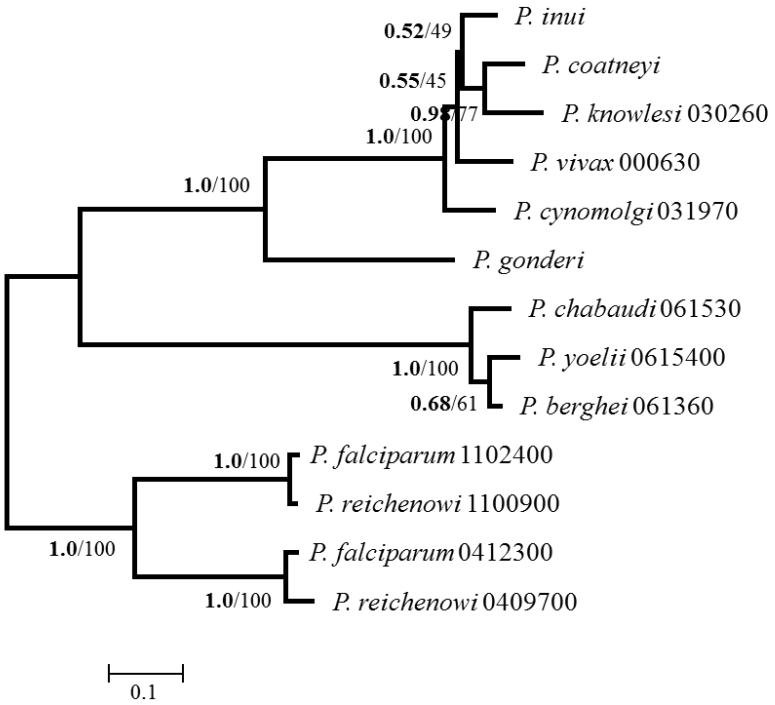

1.0/93 P. vivax 087085
0.96/77 P. inui
0.99/95 P. coatneyi
0.92/85 P. knowlesi 073060
P. gonderi
1.0/100 0.95/65 P. berghei 083320
1.0/95 P. yoelii 0836300
P. chabaudi 083350
P. falciparum 0932400
1.0/98 P. reichenowi 0930400
0.60/66 P. vivax 085115
1.0/99 P. inui
P. knowlesi 132170
0.99/99 0.77/58 P. coatneyi
0.64/65 P. gonderi
P. falciparum 1428600
1.0/94 P. reichenowi 1427900
P. berghei 101610
0.94/72 P. chabaudi 101690
0.2
Peptide release factor



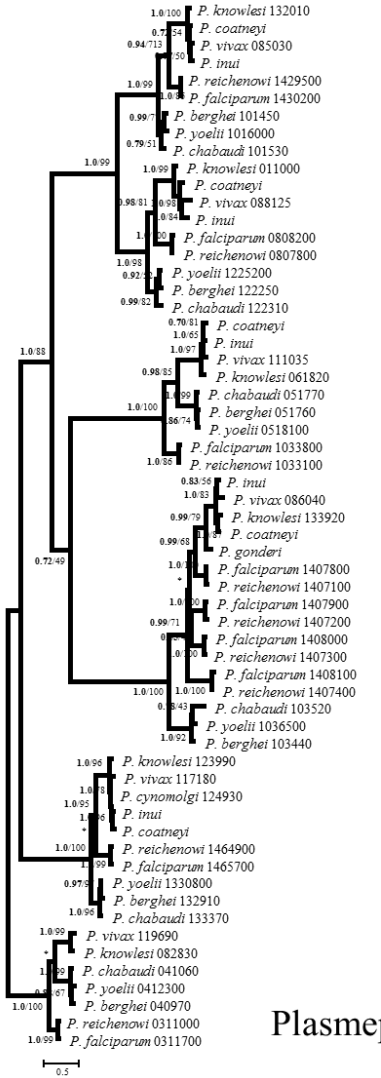
Peroxiredoxin
thioredoxin peroxidase



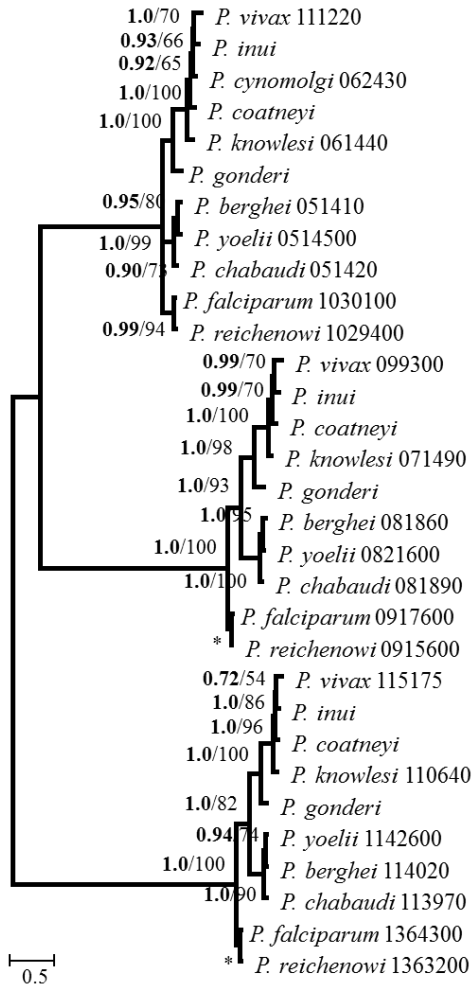
Phosducin like protein PhLP



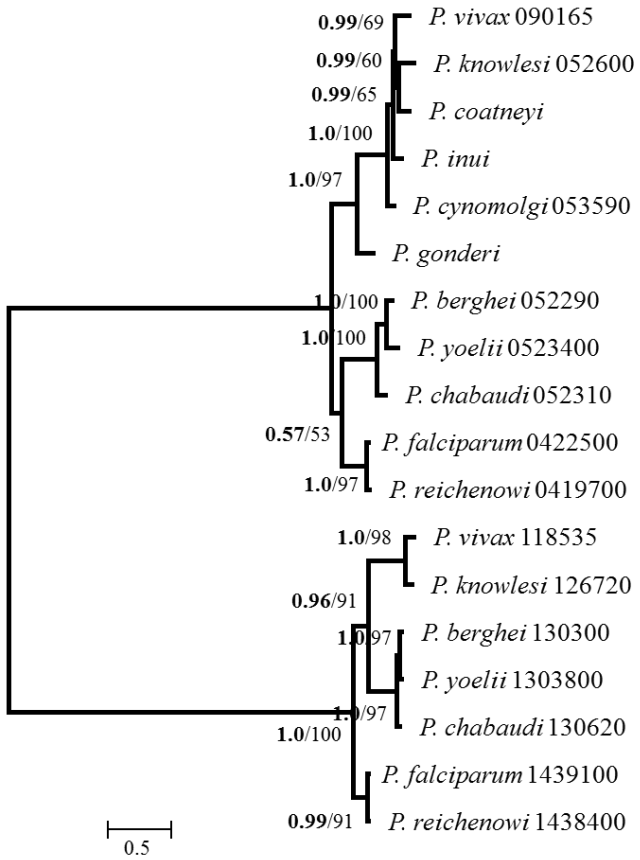
Phosphopantothenol cysteine synthetase



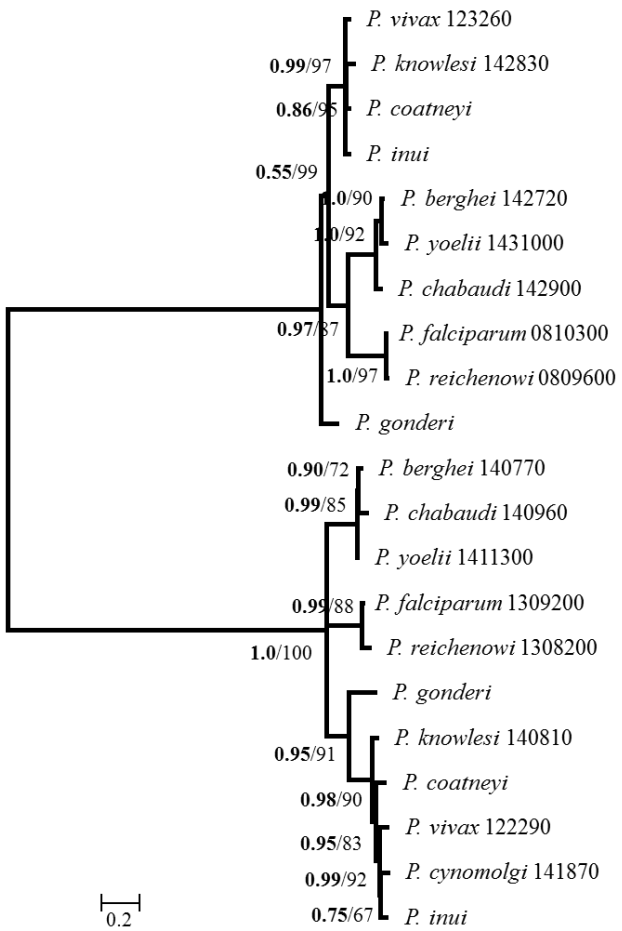
Plasmepsin



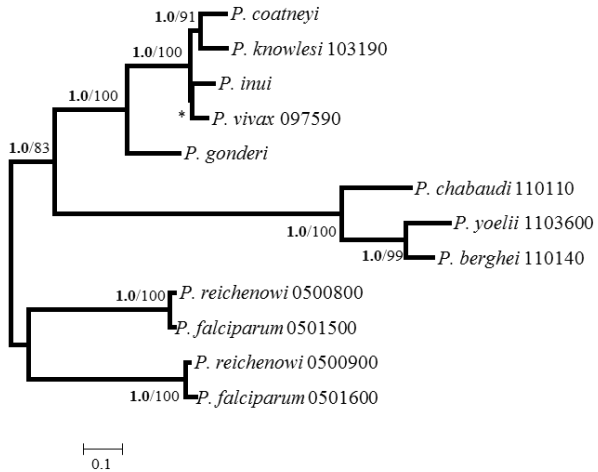
Pre mRNA splicing factor ATP dependent RNA helicase



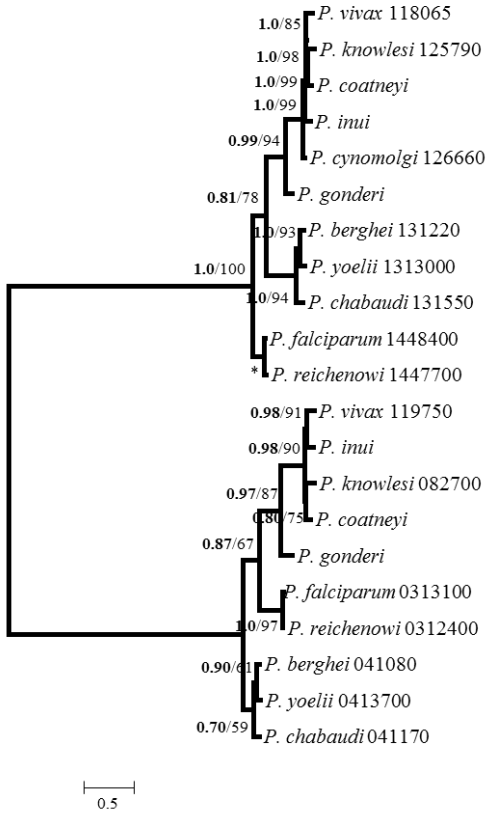
Pre mRNA splicing helicase



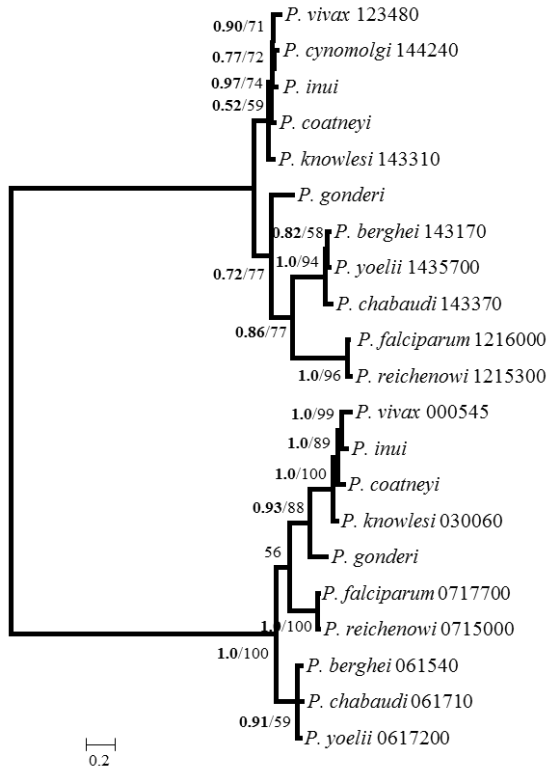
Protein phosphatase 2C



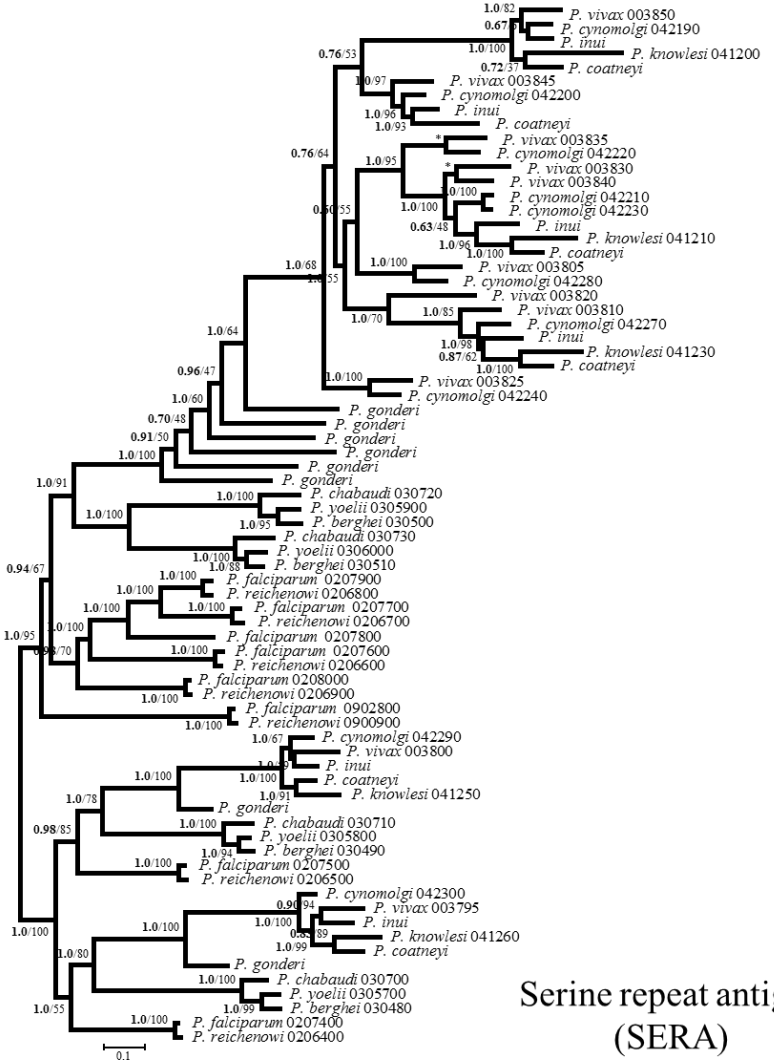
Rhoptry associated protein 23



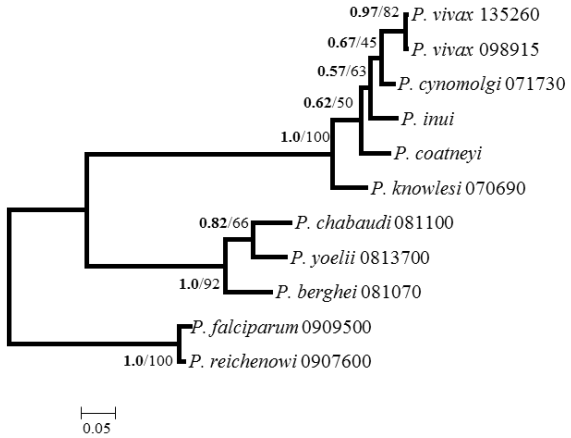
SEL1 protein



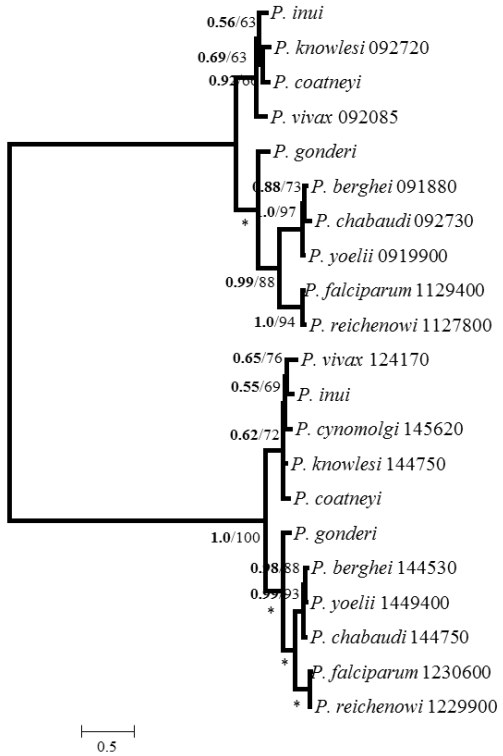
Serine tRNA ligase



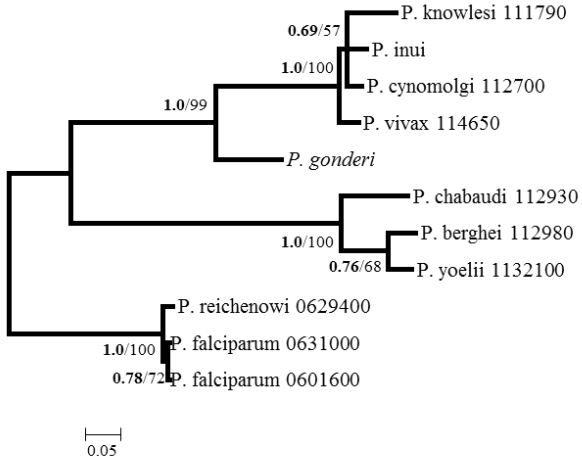
Serine repeat antigen
(SERA)



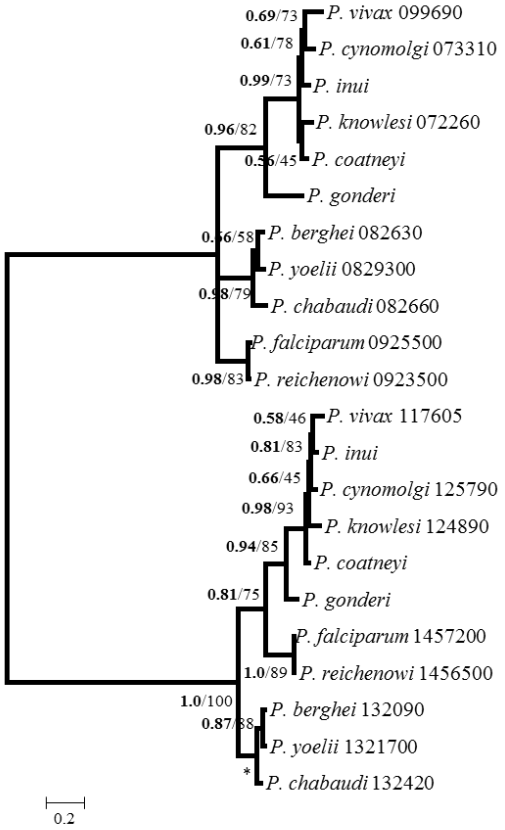
Subpellicular microtubule protein 1



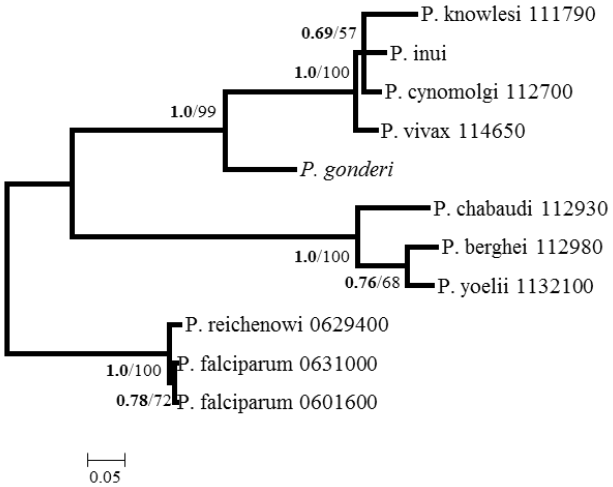
Sun family protein



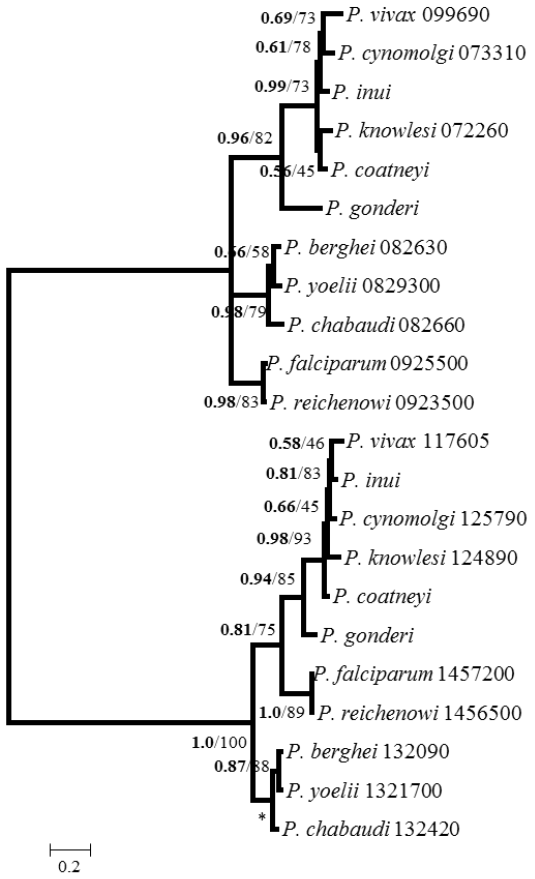
Tetratricopeptide repeat protein



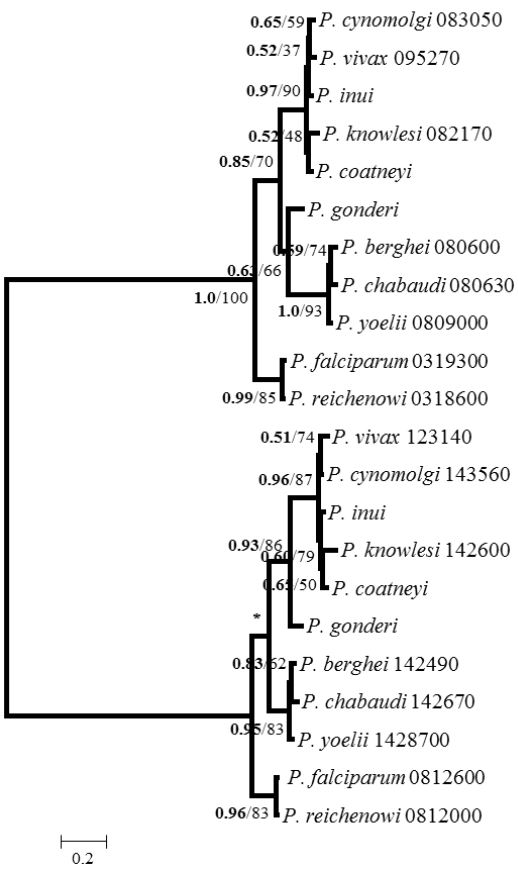
Thioredoxin



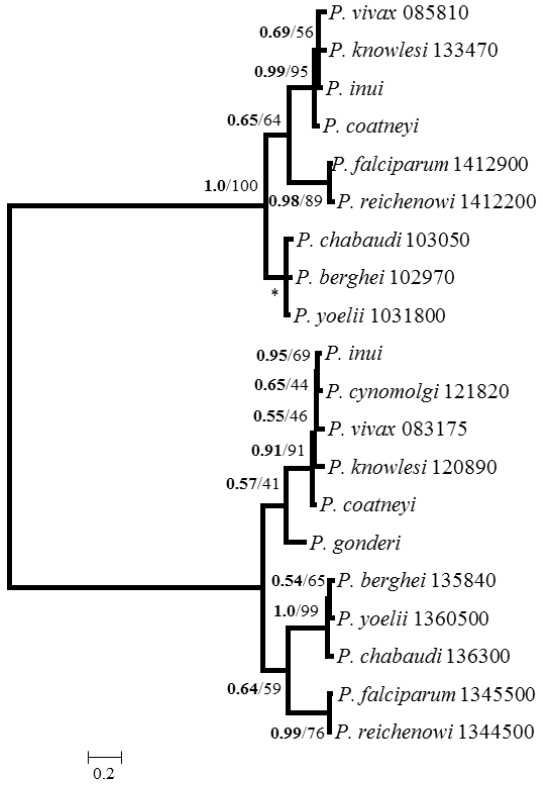
Tetratricopeptide repeat protein



Thioredoxin

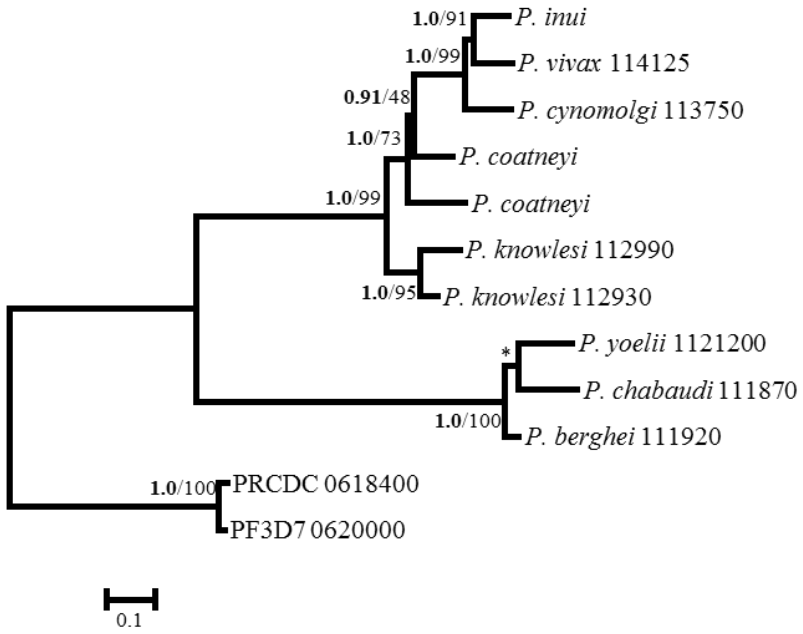


Ubiquitin conjugating enzyme

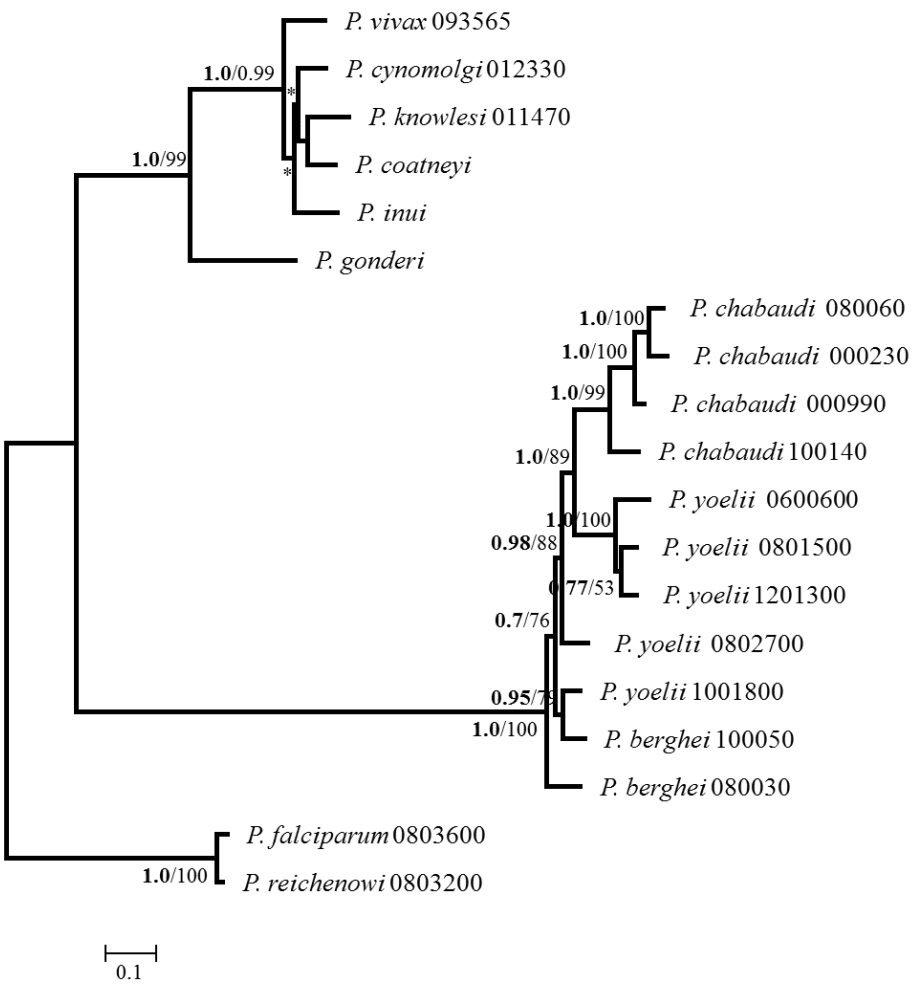


Ubiquitin conjugating enzyme 2

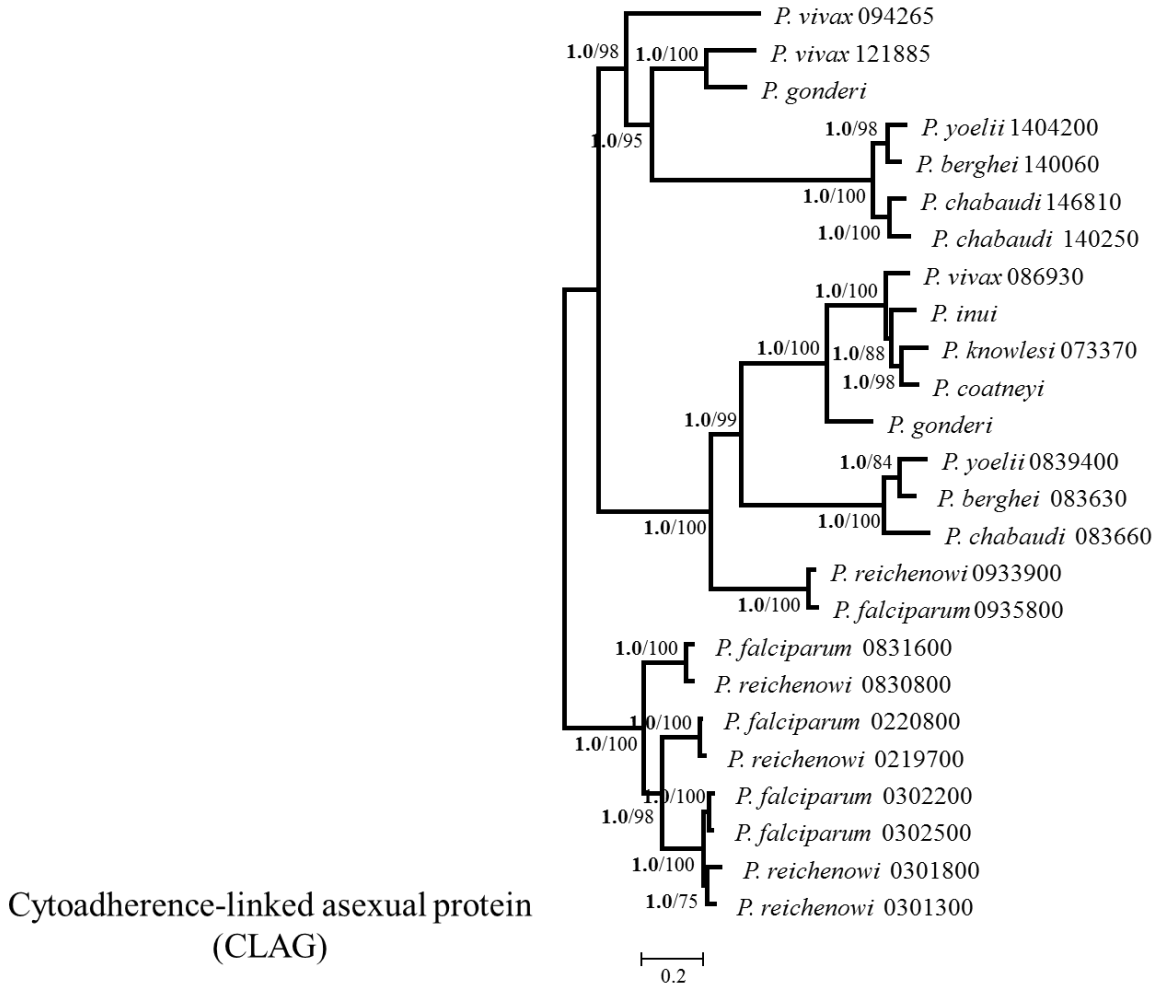
Figure S4-1. Bayesian Inference (BI) and Maximum Likelihood (ML) trees showed almost identical topologies, so only BI topology is shown. Asterisks (*) indicate conflicting branching patterns. Posterior probabilities (PP) and bootstrap values (BV) are shown next to the phylogenetic tree nodes (PP/BV). Paralogs identities are indicated by a combination of the species name and PlasmoDB identification numbers. The name of each multigene family is indicated at the bottom of each tree. The number of sequences and nucleotide positions varied among aligned multigene families. The most informative nucleotide substitution model was estimated for each multigene family alignment.



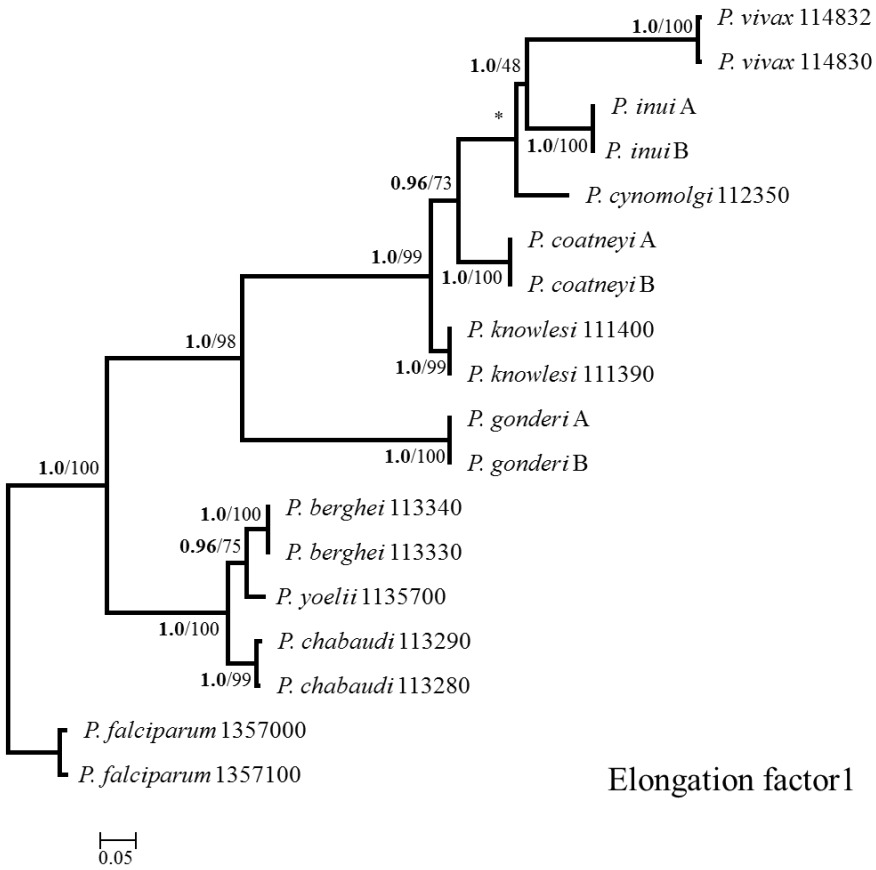
Conserved *Plasmodium* protein unknown function 6

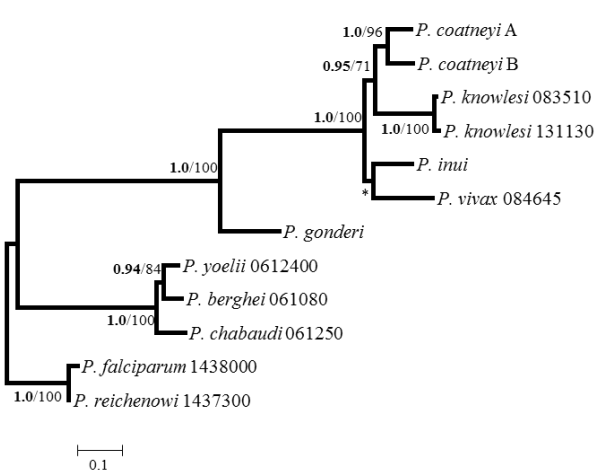


Conserved Rodent malaria protein unknown function

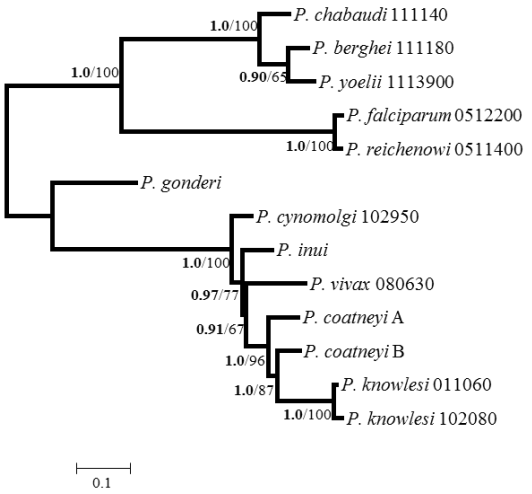


Cytoadherence-linked asexual protein (CLAG)


1.0/100
P. vivax 114832
P. vivax 114830
1.0/48
*
P. inui A
1.0/100
P. inui B
0.96/73
P. cynomolgi 112350
1.0/99
P. coatneyi A
1.0/100
P. coatneyi B
1.0/98
P. knowlesi 111400
1.0/99
P. knowlesi 111390
P. gonderi A
1.0/100
P. gonderi B
1.0/100
1.0/100
P. berghei 113340
0.96/75
P. berghei 113330
P. yoelii 1135700
1.0/100
P. chabaudi 113290
1.0/99
P. chabaudi 113280
P. falciparum 1357000
P. falciparum 1357100
0.05
Elongation factor1



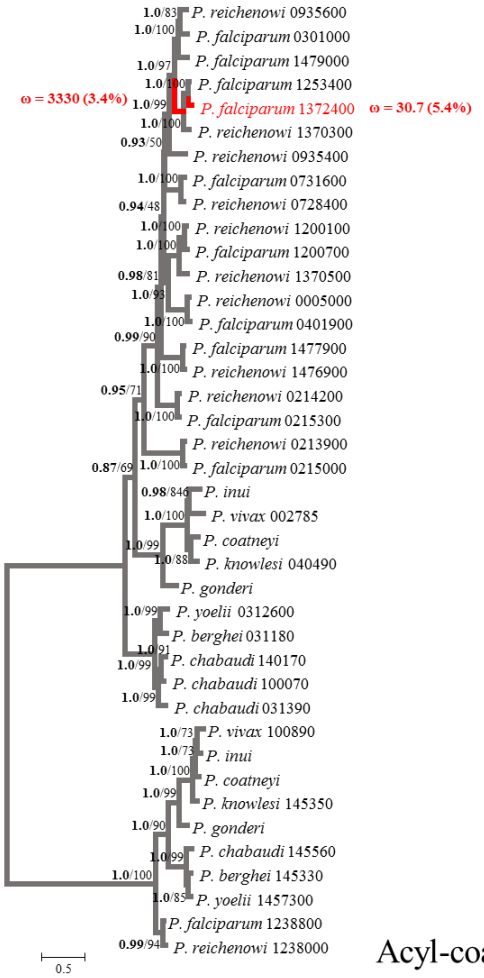
Eukaryotic initiation factor 2a



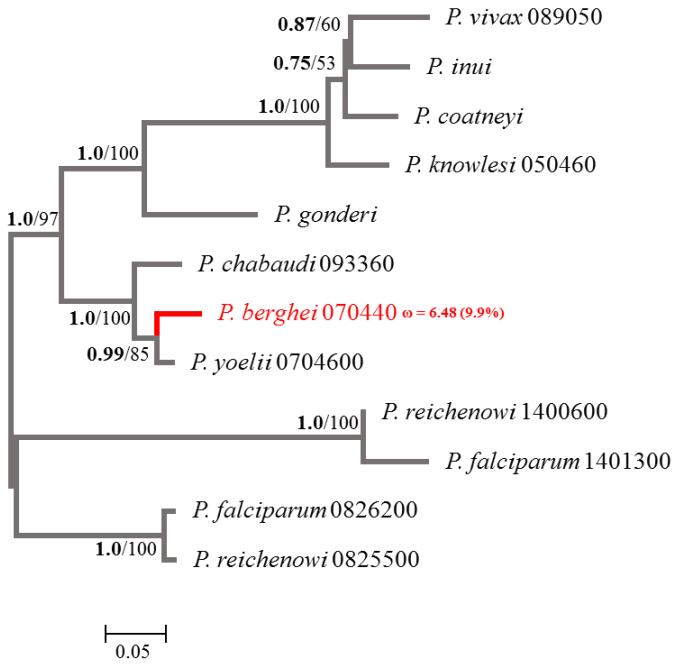
Glutathione synthetase

Figure S4-2. Bayesian Inference (BI) and Maximum Likelihood (ML) trees showed almost identical topologies, so only BI topology is shown. Asterisks (*) indicate conflicting branching patterns. Posterior probabilities (PP) and

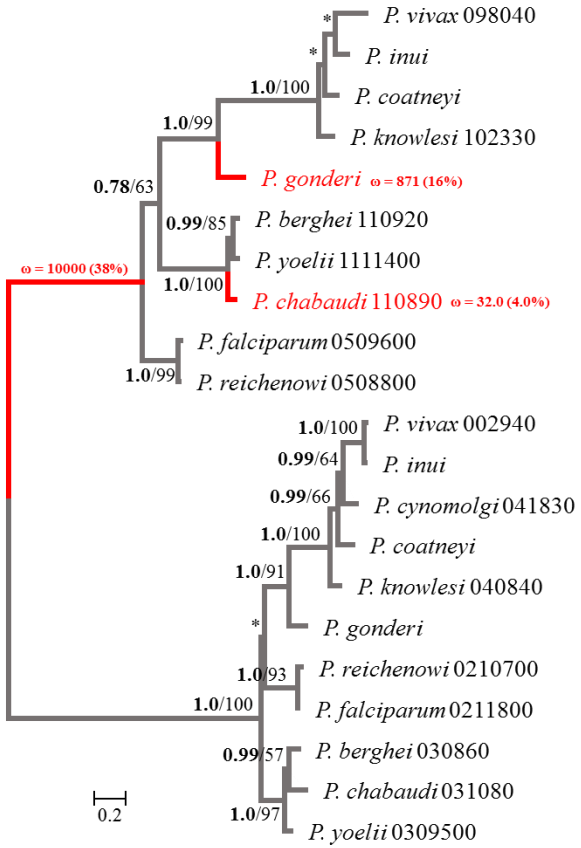
bootstrap values (BV) are shown next to the phylogenetic tree nodes (PP/BV). Paralogs identities are indicated by a combination of the species name and PlasmoDB identification numbers. The name of each multigene family is indicated at the bottom of each tree. The number of sequences and nucleotide positions varied among aligned multigene families. The most informative nucleotide substitution model was estimated for each multigene family alignment.



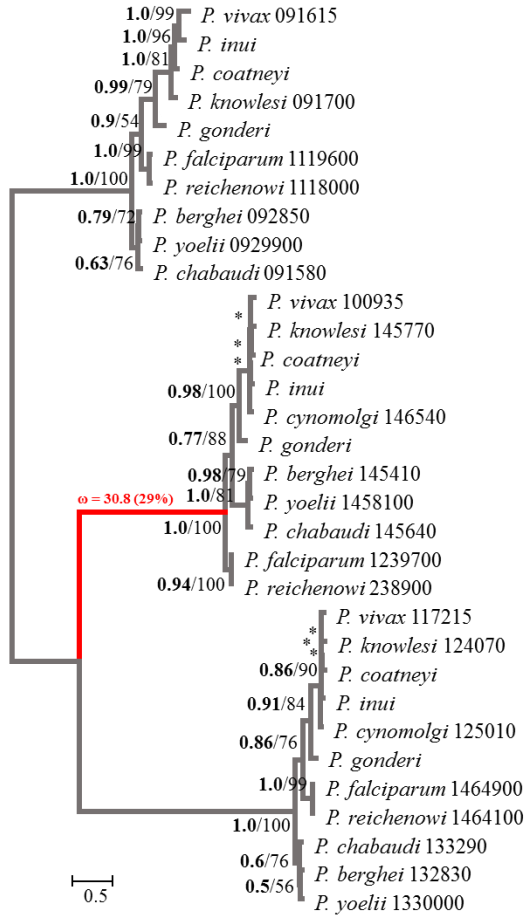
Acyl-coa synthase



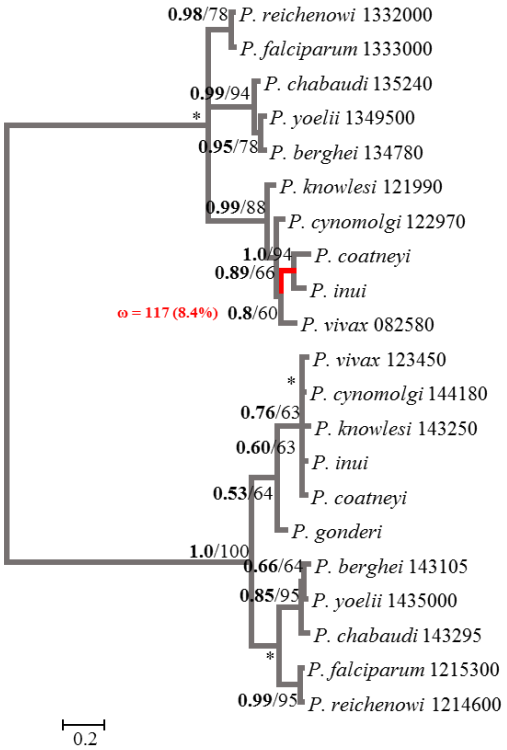
Alpha beta hydrolase



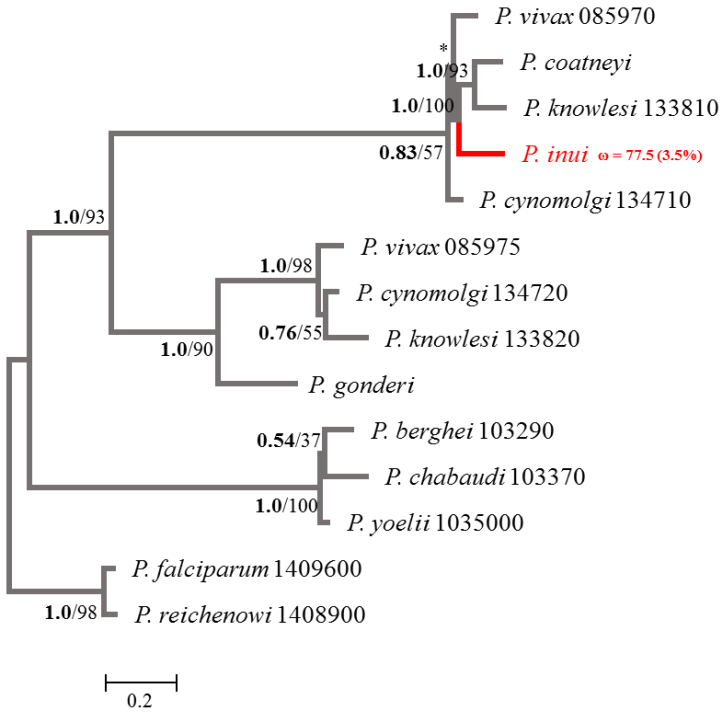
Asparagine tRNA ligase



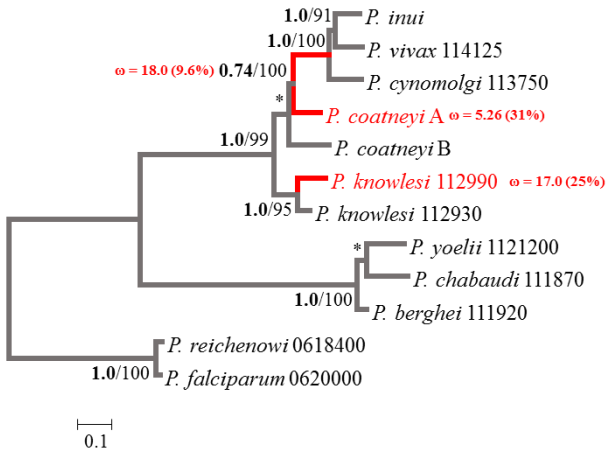
Cell division protein



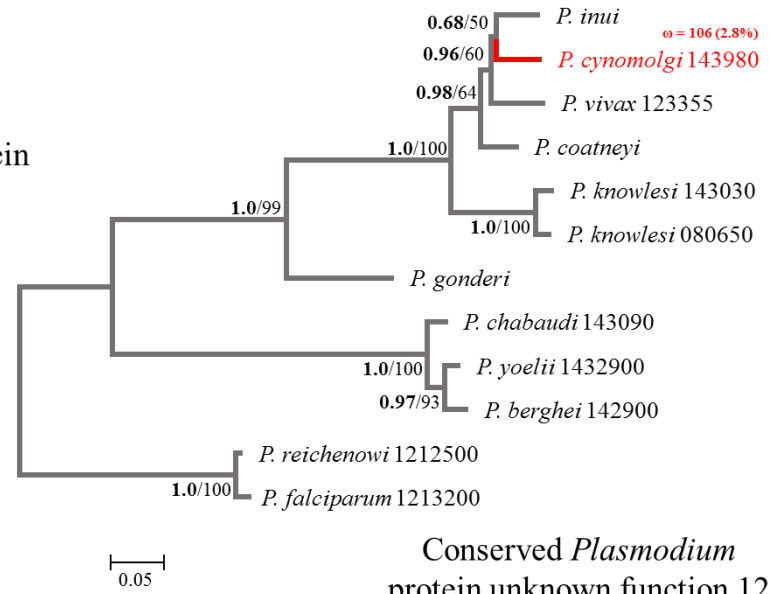
Chaperonin putative CPN



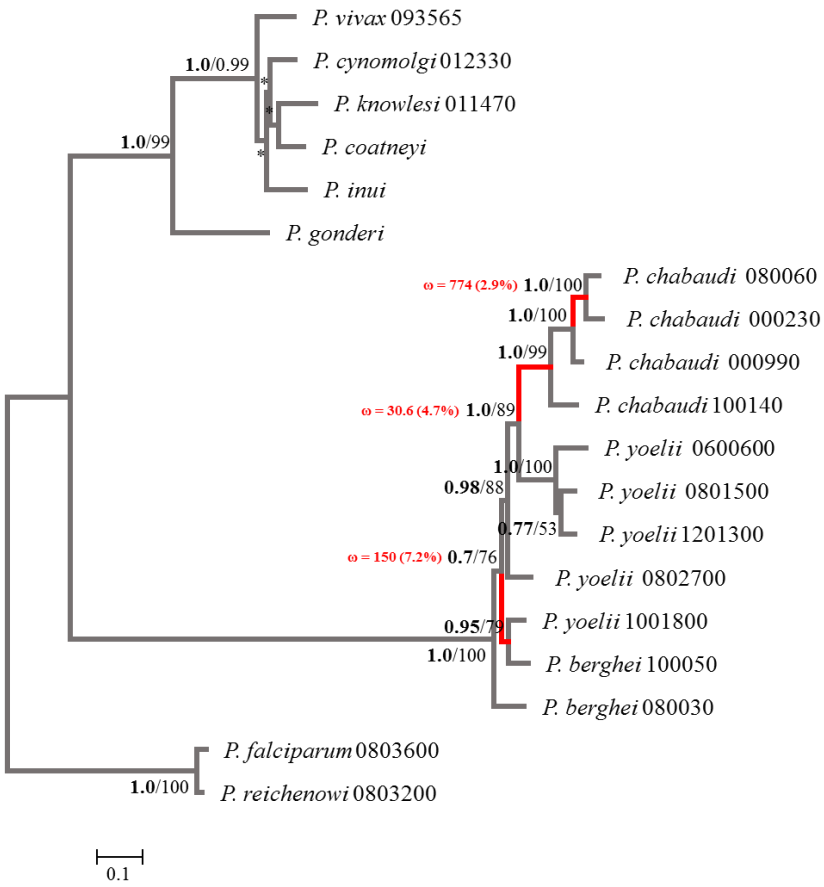
Conserved *Plasmodium* protein



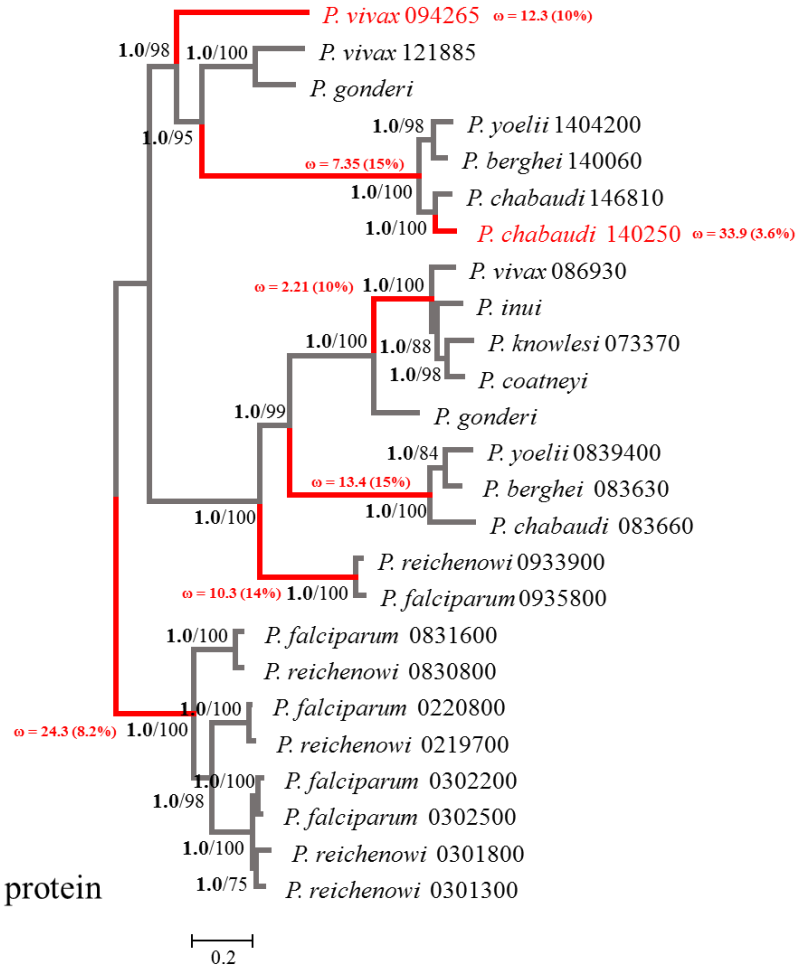
Conserved *Plasmodium* protein unknown function 6



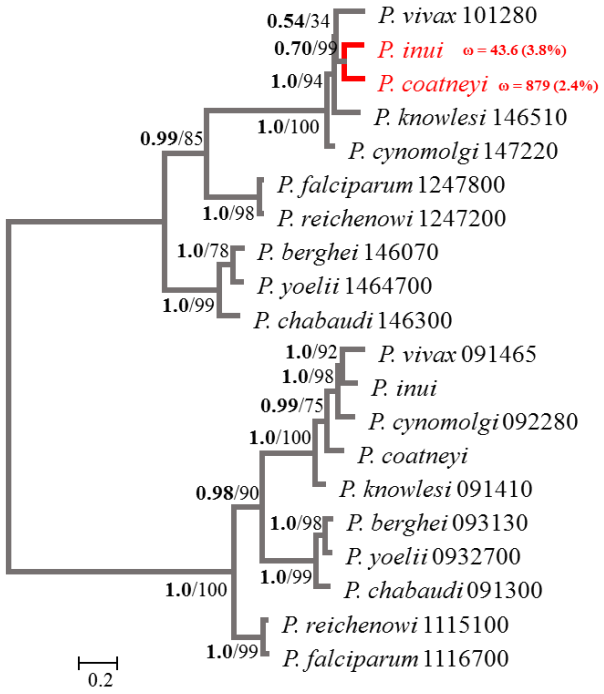
Conserved *Plasmodium* protein unknown function 12



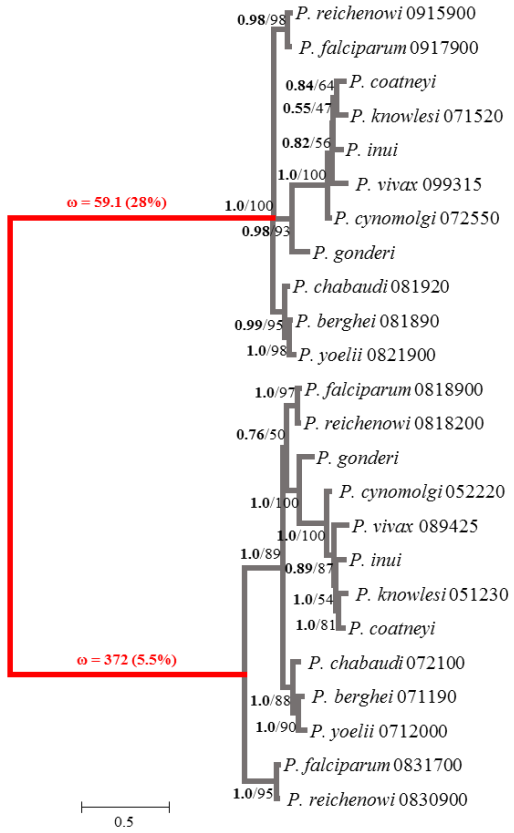
Conserved Rodent malaria protein unknown function



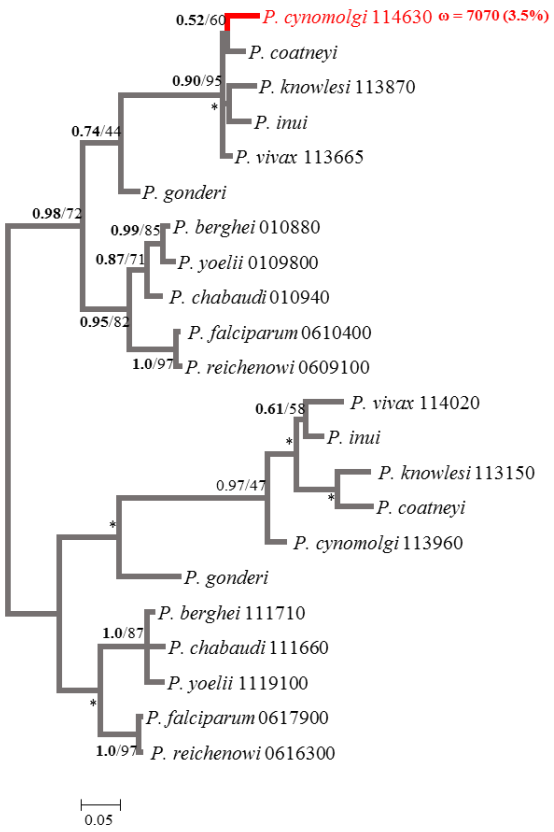
Cytoadherence-linked asexual protein (CLAG)



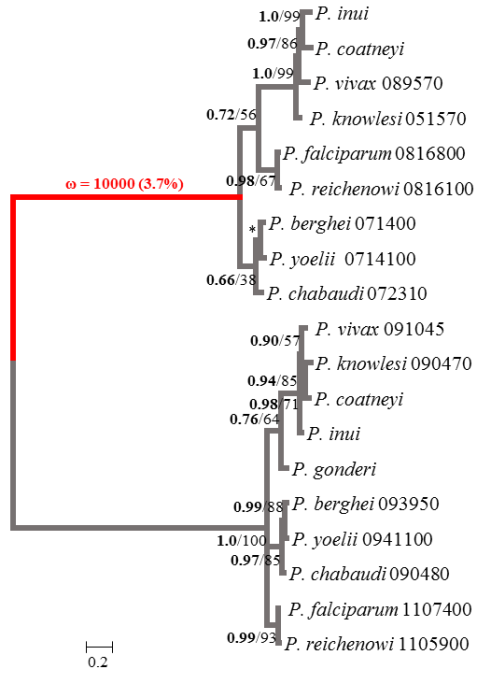
Dipeptidyl amino peptidase DPAP



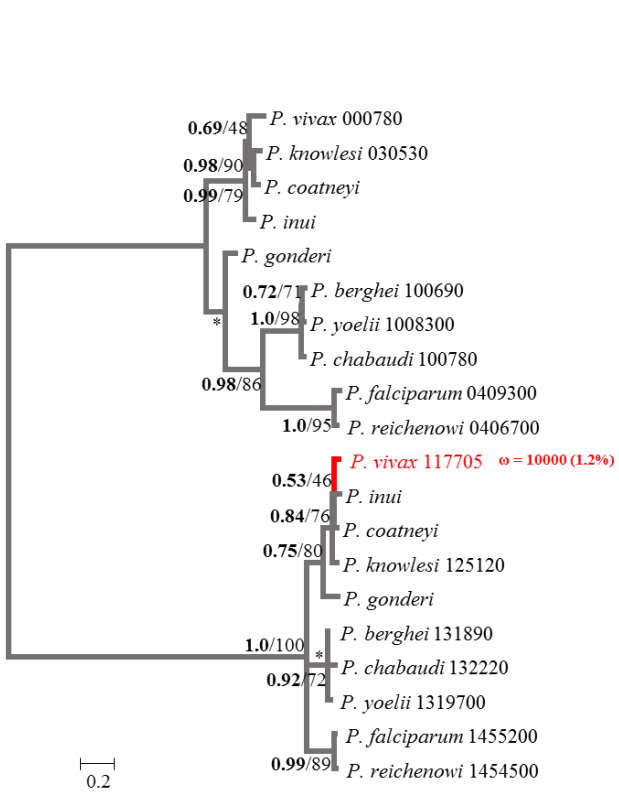
Heat Shock protein 70



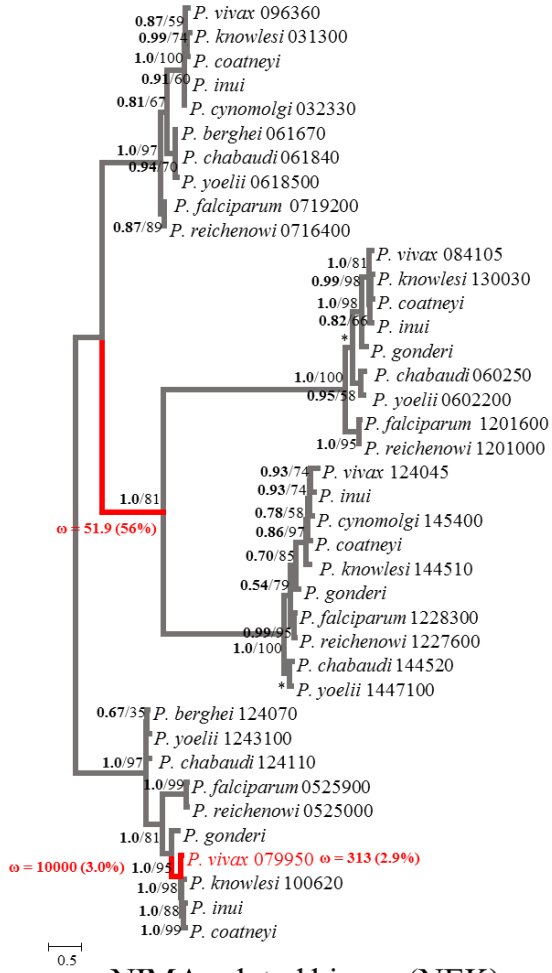
Hypothetical protein



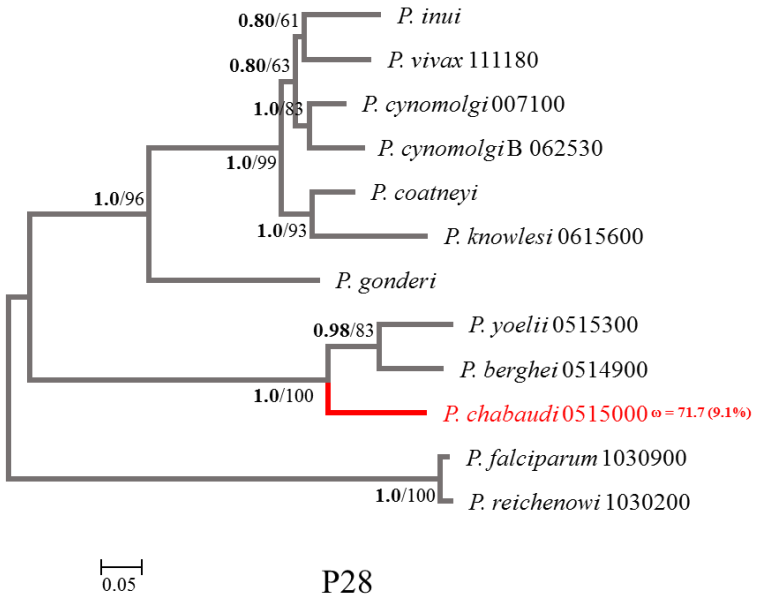
Meiotic recombination protein DMC



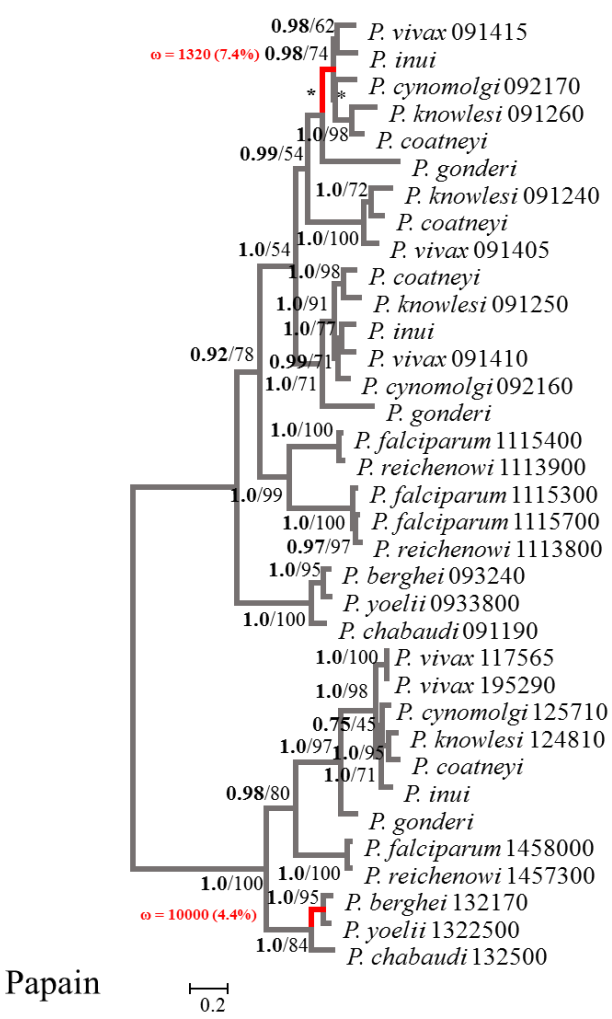
Methyltransferase



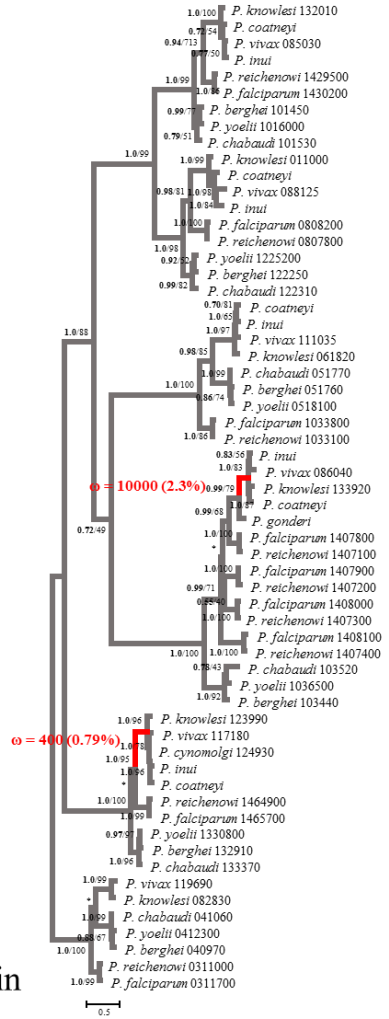
NIMA related kinase (NEK)


0.80/61
P. inui
0.80/63
P. vivax 111180
1.0/83
P. cynomolgi 007100
P. cynomolgi B 062530
1.0/99
P. coatneyi
1.0/96
1.0/93
P. knowlesi 0615600
P. gonderi
0.98/83
P. yoelii 0515300
P. berghei 0514900
1.0/100
P. chabaudi 0515000 ω = 71.7 (9.1%)
P. falciparum 1030900
1.0/100
P. reichenowi 1030200
0.05

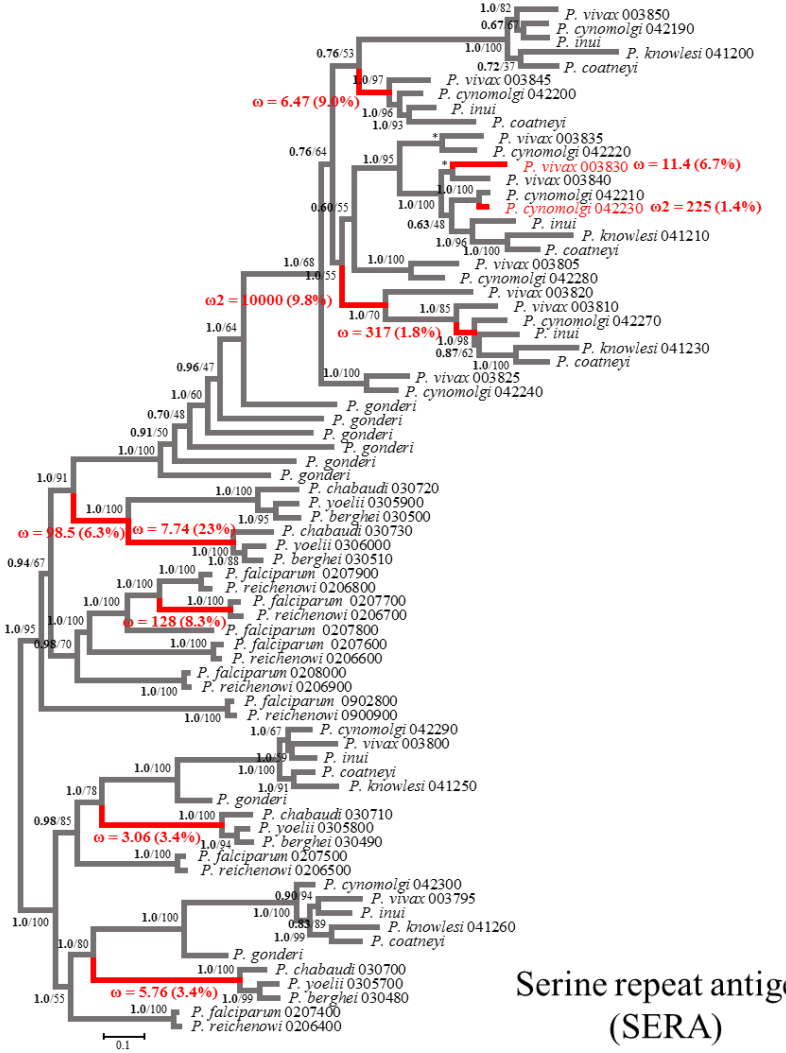
P28



Papain



Plasmepsin



Serine repeat antigen (SERA)

Figure S4-3. Bayesian Inference (BI) and Maximum Likelihood (ML) trees showed almost identical topologies, so only BI topology is shown. Asterisks (*) indicate conflicting branching patterns. Posterior probabilities (PP) and bootstrap values (BV) are shown next to the phylogenetic tree nodes (PP/BV). Paralogs identities are indicated by a combination of the species name and PlasmoDB identification numbers. The name of each multigene family is indicated at the bottom of each tree. The number of sequences and nucleotide positions varied among aligned multigene families. The most informative nucleotide model was estimated for each multigene family alignment. Branches under significant episodic selection are marked in red. Paralogs names are indicated in red fonts when terminal branches showed significant signs of episodic selection. The strength of the selective signal (ω) and the percentage of positively selected sites are shown alongside branches under episodic selection.