bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Title: Genomes of an entire *Plasmodium* subgenus reveal paths to virulent human malaria

Authors: Thomas D. Otto[1,†,*], Aude Gilabert[2,†], Thomas Crellen[1,3], Ulrike Böhme[1], Céline Arnathau[2], Mandy Sanders[1], Samuel Oyola[1], Alain Prince Okouga[4], Larson Boundenga[4], Eric Willaume[5], Barthélémy Ngoubangoye[4], Nancy Diamella Moukodoum[4], Christophe Paupy[2], Patrick Durand[2], Virginie Rougeron[2,4], Benjamin Ollomo[4], François Renaud[2], Chris Newbold[1,6], Matthew Berriman[1,*] & Franck Prugnolle[2,4,*]

Affiliations:

[1] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, United Kingdom

[2] Laboratoire MIVEGEC, UMR 5290-224 CNRS 5290-IRD 224-UM, Montpellier, France

[3] Department of Infectious Disease Epidemiology, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, United Kingdom

[4] Centre International de Recherches Médicales de Franceville, Franceville, Gabon

[5] Sodepal, Parc of la Lékédi, Bakoumba, Gabon

[6] Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, United Kingdom

[*]Correspondence to: Thomas D. Otto (tdo@sanger.ac.uk), Matthew Berriman (mb4@sanger.ac.uk) or Franck Prugnolle (franck.prugnolle@ird.fr)

[†]These authors contributed equally.

Abstract: *Plasmodium falciparum*, the most virulent agent of human malaria, shares a recent common ancestor with the gorilla parasite *P. praefalciparum*. Little is known about the other gorilla and chimpanzee-infecting species in the same (*Laverania*) subgenus as *P. falciparum* but none of them are capable of establishing repeated infection and transmission in humans. To elucidate underlying mechanisms and the evolutionary history of this subgenus, we have generated multiple genomes from all known *Laverania* species. The completeness of our dataset allows us to conclude that interspecific gene transfers as well as convergent evolution were important in the evolution of these species. Striking copy number and structural variations were observed within gene families and one, *stevor* shows a host specific sequence pattern. The complete genome sequence of the closest ancestor of *P. falciparum* enables us to estimate confidently for the first time the timing of the beginning of speciation to be 40,000-60,000 years ago followed by a population bottleneck around 4,000-6,000 years ago. Our data allow us also to search in detail for the features of *P. falciparum* that made it the only member of the *Laverania* able to infect and spread in humans.

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Main Text:

The evolutionary history of *Plasmodium falciparum*, the most common and deadliest human malaria parasite, has been the subject of uncertainty and debate[1,2]. Recently it has become clear that *P. falciparum* is derived from a group of parasites infecting African Great Apes and known as the *Laverania* subgenus[2]. Until 2009, the only other species known in this subgenus was a parasite of chimpanzees known as *P. reichenowi*, for which only one isolate was available[3]. It is now clear that there are a total of at least seven species in Great Apes that naturally infect chimpanzees (*P. gaboni*, *P. billcollinsi* and *P. reichenowi*), gorillas (*P. praefalciparum*, *P. blacklocki* and *P. adleri*)[4,5], or humans (*P. falciparum*) (Fig. 1a). Within this group, *P. falciparum* is the only parasite that has successfully adapted to humans after a transfer from gorillas and subsequently spread all over the world[2].

There has been much controversy concerning the evolutionary history of *P. falciparum* with the speciation event having been estimated to be anywhere between 10,000 to 5.5 million years ago[6,7]. Others report a bottleneck less than 10,000 years ago[8], but suggest a drop to a single progenitor parasite. The latter seems unlikely due to the presence of allelic dimorphisms that predate speciation events and therefore could not have both been transmitted if a new species were founded by a single individual infection. Also, the dating of the speciation cannot be accurately estimated without the genome sequence of *P. praefalciparum,* the closest living sister species to *P. falciparum*.

The absence of *in vitro* culture or a usable animal model has precluded obtaining sufficient DNA for full genome sequencing and has hindered investigation of the *Laverania.* So far just the full genome of *P. reichenowi* [9] and partial sequence of *P. gaboni*[6] are available. These data together with additional PCR based approaches[10] have provided some insights into the evolution of this subgenus, including the lateral gene transfer of the *rh5* locus and the observation that the common ancestor had also *var* genes. However, the lack of whole genome information for the whole subgenus (particularly *P. praefalciparum*) has severely constrained the scope of subsequent analyses.

To investigate the evolutionary history of the entire *Laverania* subgenus and to address the question of why *P. falciparum* is the only species to adapt successfully to humans, we have sequenced multiple genotypes of all known *Laverania* species.

Genome sequencing from six *Laverania* species

Blood samples were taken during successive routine sanitary controls, from four gorillas and seven chimpanzees living in a sanctuary or quarantine facility prior to release (see Methods). A total of 15 blood samples were positive for ape malaria parasites by PCR. Despite low parasitemia in most animals, a combination of host DNA depletion, parasite cell sorting and amplification

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

methods enabled sufficient parasite DNA templates to be obtained for short-read (Illumina) and long read (Pacific Bioscience) sequencing (Table 1). Mixed-species infections were frequent but resolved by utilising sequence data from single infections, resulting in 19 genotypes (Supplementary Table 1). The dominant genotype in each sample was assembled *de novo* (see Methods) using long read technology into a reference genome for six malaria parasite species: *P. praefalciparum*, *P. blacklocki*, *P. adleri, P. billcollinsi*, *P. gaboni* and *P. reichenowi*. The assemblies comprised 44-97 scaffolds (Table 1), with large contigs containing the subtelomeric regions and internal gene clusters that house multigene families known in *P. falciparum* and *P. reichenowi* to be involved in virulence and host-pathogen interactions. The high quality of the assemblies can be seen in the large number of one-to-one orthologues obtained between the different reference genomes (4,324 among the seven species and 4,818 between *P. falciparum*, *P. praefalciparum* and *P. reichenowi*). Two to four additional genomes were obtained for each species except for *P. blacklocki* (Table 1, Supplementary Table 1).

Speciation history in the *Laverania* sub-genus

Conservation of gene content and synteny is striking between these complete genomes and enabled us to reconstruct with confidence the relationships between different *Laverania* species, to compare their relative genetic diversity (Fig 2a, Supplementary Fig. 1) and to estimate the age of the different speciation events that led to the extant species. The latter has been problematic in the past because of the lack of both complete genome data and accurate estimates of mutation rate and generation time. Here we take the most divergent estimates of generation time and measured mutation rates from the literature and show that the data converge to 0.9-1.5 mutations per year per genome (Supplementary Note 1). We observed a similar mutation rate *in vivo* by examining existing sequence data for 5 geographically diverse isolates, covering a 200-kb region surrounding the PfCRT gene that is relatively conserved due to a selective sweep resulting from chloroquine use (Supplementary Note 1; Supplementary Figure 2). From our Bayesian whole-genome estimates, the ancestor of all current day parasites of this subgenus existed 0.7–1.2 million years ago, a time at which the subgenus divided into two main clades, A (*P. adleri* and *P. gaboni*) and B that includes the remaining species (Fig. 1a). This estimated age coincides with that of the common ancestor of two other human parasites, *P. ovale curtisi* and *P. ovale wallikeri*[11] (Supplementary Note 1). Our range of values is far more recent than previous estimates[3,12]. Following the Clade A/B subdivision, several speciation events occurred leading either to new chimpanzee or gorilla parasites. Interestingly, the divergence between *P. adleri* and *P. gaboni* in one lineage and *P. reichenowi* and the ancestor of *P. praefalciparum/P. falciparum* in the other lineage occurred at approximately the same time (140–230 thousand years ago; Fig. 1a, Supplementary Table 2). This is also the time of

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the split between *P. malariae* (human) from *P. malariae-like* (chimpanzee)[11] suggesting that similar phenomena may have favoured these host switches. Based on our coalescence estimates, *P. falciparum* emerged in humans from *P. praefalciparum* around 40–60 thousand years ago (Fig. 1a), significantly later than the evolution of the first modern humans and their spread throughout Africa[13]. Our analysis also indicates significant gene flow between these two parasite species after speciation (Supplementary Table 2).

P. falciparum has strikingly low diversity (π=0.0004), compared with the other *Laverania* species (0.002–0.0049) (Supplementary Fig. 1). It has been proposed that *P. falciparum* arose from a single transfer of *P. praefalciparum* into humans[6] and based in part on the paucity of neutral SNPs within the genome, that *P. falciparum* emerged from a bottleneck of a single parasite around 10,000 thousand years ago, after agriculture was established[6,8]. Clearly this hypothesis is incorrect in light of our results; we estimate that the *P. falciparum* population declined around 11,000 years ago and reached a minimum about 5,000 years ago (Fig. 1b) with an effective population size (*Ne*) of around 3,000 (Supplementary Note 1; generally the census number of parasites is higher than *Ne*[14]). It is important to note that the methods we used to estimate *Ne* do not rely on the Fisher-Wright model, that could give problems as previously reported[15]. The hypothesis of a single progenitor is also inconsistent with the observation of several ancient gene dimorphisms that have been observed in *P. falciparum*. A previous analysis using *P. reichenowi* and limited *P. gaboni* sequence data, provided some evidence that different dimorphic loci diverged at different points in the tree[16]. Looking at each of these *P. falciparum* loci across the *Laverania*, we found different patterns of evolution at the *msp1*, *var1csa*, and *msp3* loci (Supplementary Fig. 3a). Most strikingly, a mutually exclusive dimorphism (described as MAD20/K1[17]) in the central 70% of the *msp1* sequence, clearly pre-dates the *P. falciparum–P. praefalciparum* common ancestor and dimorphism in *var1csa* (an unusual *var* gene of unknown function that is transcribed late in the asexual cycle) occurred before the split with *P. reichenowi*.

In contrast, the gene *eba-175* that encodes a parasite surface ligand involved in red blood cell invasion contains a dimorphism that arose after the emergence of *P. falciparum* (Supplementary Fig. 3b). The time to the most recent common ancestor of *eba-175* has been estimated as 130–140 thousand years in an analysis[18] that assumed *P. falciparum* and *P. reichenowi* diverged 6 million years ago. However, based on our new estimate for *P. falciparum–P. reichenowi* divergence, we recalibrate their estimate of the most recent common ancestor of the *eba-175* alleles to be around 4,000 years ago, which is in good agreement with our divergence time for *P. falciparum* (Supplementary Note 1). The recent dimorphism cannot however explain the observation of an ancient dimorphism near the human and ape loci for glycophorin[19] – an EBA-175 binding protein.

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The formation and maintenance of all of these dimorphic loci has therefore been shaped by different balancing selection pressures over time.

P. falciparum-specific evolution

During its move away from gorillas, *P. falciparum* had to adapt to new environmental conditions, namely a new vertebrate host (human) and some new vector species (e.g. *Anopheles gambiae*)[20]. To infer *P. falciparum* specific adaptive changes, we considered the *P. falciparum* / *P. praefalciparum* and *P. reichenowi* genome trio and then applied a branch-site test and calculated McDonald Kreitman (MK) ratios to detect events of positive selection that occurred in the *P. falciparum* lineage. The two tests identified 171 genes (out of 4,818) with signatures of positive selection in the human parasite species only (Supplementary Table 3). Of these, 138 genes had a significant d_N/d_S ratio and 35 genes had an MK ratio significantly higher than 1. Two genes (*rop14* and PF3D7_0609900) were significant in both tests. Among the 171 genes, almost half (n=82) encoded proteins of unknown function. Analysis of those with functional annotation indicated that genes involved in pathogenesis and/or the entry into host, in actin movement and organization and in drug response were significantly over-represented. Other genes, expressed in different stages of the *P. falciparum* life cycle (e.g. *sera4* and *emp3*, involved during the erythrocytic stages; *P230*, involved in gametocytes; *trsp* and *lisp*1, involved in the hepatic stages; and *plp4*, *CelTOS* and *Cap380*, involved in the mosquito stages) also showed a significant signal of adaptive evolution (Supplementary Table 3).

Evolution through introgression, gene transfer and convergence

Frequent mixed species infections in apes and mosquitoes[20] provide clear opportunities for interspecific gene flow between these parasites. A recent study[6] reported a gene transfer event between *P. adleri* and the ancestor of *P. falciparum* and *P. praefalciparum* of a region on chromosome 4 including key genes involved in erythrocyte invasion (*rhr5* and *cyrpA*). We systematically examined the evidence for introgression or gene transfer events across the complete subgenus by testing the congruence of each gene tree to the species tree for genes with 1-1 orthologues. Beyond the region that includes *rh5* (Supplementary Fig. 4a), few signals of interspecific gene flow were obtained (n =6) suggesting that these events were rare or usually strongly deleterious (Supplementary Fig. 5).

The *Laverania* subgenus evolved to infect chimpanzees and gorillas several times independently but, on a genome-wide scale, the convergent evolution of host-specific traits has not left a signature (Supplementary Note 2). We therefore examined each CDS independently and were able to identify genes with differences fixed within specific hosts, falling into three categories: 53 in

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

chimpanzee-infective parasites, 49 in gorilla-infective and 12 with fixed traits in both host species (Fig. 2; Supplementary Table 4a). For at least 66 genes, these differences were unlikely to have arisen by chance ($p < 0.05$) and GO term enrichment analysis revealed that several of these genes are involved in erythrocyte invasion (Supplementary Table 4b) including *rh5* (which has a signal for convergent evolution even when the introgressed tree topology is taken into consideration; Supplementary Fig. 4b). *Rh5* is the only gene identified in *P. falciparum* that is essential for erythrocyte recognition during invasion, via binding to Basigin. *P. falciparum rh5* cannot bind to gorilla Basigin and binds poorly to the chimpanzee protein[21]. We notice that one of the convergent sites is known to be a binding site for the host receptor Basigin[22] (Supplementary Fig. 4b). The gene *eba-165* encodes a member of the erythrocyte binding like (EBL) super family of proteins that are involved in erythrocyte invasion. Although *eba-165* is a pseudogene in *P. falciparum*[23], it is not in the other *Laverania* species and may therefore be involved in erythrocyte invasion, like other EBL members. The protein has three convergent sites in gorillas, one falls inside the F2 region, a Duffy Binding-Like (DBL) domain involved in the interactions with erythrocyte receptors. The role of this protein and of these convergent sites in the invasion of gorilla red cells, remains to be determined. Finally, genes involved in gamete fertility (the 6-cysteine protein P230) or previously considered as potential candidate vaccines (doc2) also displayed signals of convergent evolution. Interestingly, among the 12 coding sequences with fixed differences in both great ape parasite species, P230 was the only one found with a position that was different and fixed within the three host species (gorillas, chimpanzees and humans). P230 is involved in gamete development and trans-specific reproductive barriers[24], possibly through enabling male gametes to bind to erythrocytes prior to exflagellation[25]. Host-specific residues observed in P230 might affect the efficiency of the binding to the erythrocyte receptors and result from co-evolution between the parasite molecule and the host receptor.

Subtelomeric gene families

To date, the only in-depth data on the subtelomeres of the *Laverania* have come from *P. reichenowi* and *P. falciparum*. We provide for the first time a comprehensive picture of the evolution of these important families.

Most gene families were likely present in the ancestor of all the *Laverania*, suggesting an ancient origin. In addition, most families displayed the same gene composition throughout the subgenus and only a subset of them displayed species-specific contraction or expansion (Fig. 3 and Supplementary Table 5a). For these latter families, Clade A and most species of Clade B clearly differ in their composition. *P. blacklocki* (Clade B) is intermediate in its composition. Some gene families, like the group of exported proteins *hyp4, hyp5*, *mc-2tm* and *EPF1*, have expanded only in

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

P. praefalciparum and *P. falciparum* (and even more in *P. falciparum* for *hyp4* and *hyp5*). Since the latter three are components of Maurer's clefts, an organelle involved in protein export[26], some evolution of function in this organelle may have been an important precursor to human infection. The family of acyl-CoA synthetase genes, reported to be expanded and diversified in *P. falciparum*[27] are in fact expanded across in *Laverania* and have four fewer copies in *P. falciparum* (Supplementary Fig. 6). Other genes that show clade or group specific expansion include DBLmsp, glycophorin binding protein and CLAG (Supplementary Fig. 7).

One striking inter-clade difference concerns the largest gene family that is likely common to all other malaria species: the *Plasmodium* interspersed repeat family (*pir,* which includes the *rif* and *stevor* families in *P. falciparum*) (Fig. 3, 4). This family has been proposed to be involved in important functions such as antigenic variation, immune evasion, signalling, trafficking, red cell rigidity and adhesion[28] and yet has expanded only in Clade B, after the *P. blacklocki* split (Fig 3). The *rif* genes comprise a small conserved group and a much larger group of more diverse members that contains just 13 genes from Clade A species and at least 180 members per Clade B species (Fig. 4). There is however no evidence for host-specific adaptation in these sequences.

In contrast, a subset of *stevor* genes showed strong host-specific sequence diversification (Supplementary Fig. 8). Based on full-length alignments, there is a deep phylogenetic split between *stevor* genes but when only short conserved motifs are considered, a distinct group clusters that comprises only *stevor* genes from gorilla-infecting species (Supplementary Fig. 8b,c). Since *stevor* genes are known to be involved in host–parasite interactions (such as binding to host glycophorin C in *P. falciparum*[29]), this host specific sequence may reflect sequence differences in host-specific factors in gorillas.

Evolution of *var* genes

The *var* genes, crucial mediators of pathogenesis and the establishment of chronic infection through cytoadherence and immune evasion, are the best studied *P. falciparum* multi-gene family and unique to the *Laverania*[30]. They are two-exon genes and their products have three types of major domain; exon 1 encodes Duffy Binding like (DBL) and Cysteine Rich Interdomain Region (CIDR) and exon 2 encodes Acidic Terminal Sequence (ATS)[31]. Similar to *P. falciparum*, we found that all *Laverania* species have around 60 *var* genes/genome (Fig. 3) that retain a two-exon structure and are organized into subtelomeric or internal *var* gene clusters. There are however three notable features of the evolution of this family within the sub-genus.

First, there is a deep division in how the repertoire is organised between the major clades. The *var* genes of Clade B parasites, with the exception of *P. blacklocki*, resemble those of *P. falciparum* in terms of genomic organisation, domain types and numbers (Fig 5, 6, Supplementary Table 6). In

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

contrast, the repertoires of Clade A parasites and *P. blacklocki* (treated as one group hereafter in this section) differ in their domain composition, contain a novel CIDR-like domain (CIDRn, Fig 5, Supplementary Fig. 9) and have lower sequence diversity per domain but cluster into more sub-groups than Clade B domains (Fig 6a, Supplementary Fig. 10). The paucity of domains similar to those in *P. falciparum* (such as CIDRα) that are involved in cytoadherence to some specific and common host receptors, means that if endothelial cytoadherence was important in Clade A, some alternative receptors must have been utilised.

Second, in total there are 10 internal *var* gene clusters (confirmed by contiguous sequence data) but 8 are oppositely oriented between the two clades (Supplementary Fig. 11, Supplementary Table 7). Clade B parasites also show a much greater number of associated GC-rich RNAs of unknown Function (RUF) elements than Clade A (Supplementary Table 7).

Third, the ATS domains cluster tightly within Clade A. Within Clade B there is clear evidence of species specific diversification, except in *P. praefalciparum* and *P. falciparum* reflecting their recent speciation. There is one intact ATS from *P. falciparum* as well as several pseudogenes that cluster with Clade A (Fig 6b). Moreover, of seven internal *var* arrays (Supplementary Fig. 11) in *P. falciparum*, containing a functional *var* gene, five terminate with one of these pseudogenes (on the opposite DNA strand) suggesting that they may be remnants from ancient rearrangements. The intact *P. falciparum* gene is *var2csa*, a *var*-like gene that is highly conserved between *P. falciparum* isolates[32], involved both in cytoadherence in the placenta in primigravidae, and proposed to be a central intermediate in *var* gene switching during antigenic variation[33]. We therefore propose *var2csa* is a remnant of an ancient multigene family that has been maintained as a single complete gene in *P. falciparum*, for the dual purposes of *var*-switching and placental cytoadherence.

There is other evidence of retention of ancient *var* gene sequence across the subgenus. First, in Clade B we find a nearly full length *var* pseudogene that has highest similarity to *P. adleri* and *P. gaboni var* genes, within an internal *var* cluster on chromosome 4 in *P. falciparum* and *P. praefalciparum* but on the opposite strand to the other *var* genes. It is found in all *P. falciparum* isolates, but not in *P. reichenowi*. Second, in *P. gaboni* and *P. adleri,* three genes have the N-terminal DBLα/CIDRα architecture typical of Clade B genes and their domains cluster within Clade B based on similarity (Fig. 6a, larger nodes). Directly adjacent to two of these *var* genes are two *rif* pseudogenes that also show greatest similarity to those from Clade B. Last, we find a further nine *rif* pseudogenes of Clade A parasites that cluster with Clade B *rif* genes (Fig. 4). If these observations reflect retention of ancient copies, their high sequence conservation suggests that they are under extremely unusual selection pressure. Alternatively, they may represent relics of gene transfer between species that occurred after the Clade A/B split.

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Conclusion

We have produced high quality genomes and used mutation rates and generation times, covering the full range of most recent estimates, to calculate the date of speciation for all members of the *Laverania*, with only a small margin of error. In our analysis, we have shown that the successful infection of humans by *P. falciparum* occurred quite recently and involved numerous parasites rather than a single one as previously proposed. After the establishment in its new host, the parasite population went through a bottleneck around 5,000 years ago during the period of rapid human population expansion due to farming (Fig. 1b). We summarise the major genomic events during the evolution of the *Laverania* in Fig. 7.

As a result of our analyses we propose the following series of events for the emergence of *P. falciparum* as a major human pathogen. First, the crucial lateral transfer event of the *rh5* locus between Clade A and B parasites may also have involved *var* and *rifin* genes in other parts of the genome that, because of their orientation on the opposite strand, were not lost during later recombination. Next, facilitatory mutations are likely to have occurred in *rh5* that in the first instance allowed invasion of both gorilla and human red cells. Modern humans emerged more than 300,000 years ago[34] and existed as small isolated populations[13]. Our evidence suggests that *P. falciparum* and *P. praefalciparum* started to diverge around 40,000-60,000 years ago. In the following 40,000 years with low population densities in humans and gorillas there would have not been high selection pressure to optimise infectivity in either the hosts or vectors, enabling at least some movement of parasites between hosts. We find evidence for gene flow between lineages throughout this period. The expansion of the human population with the advent of farming likely led to strong evolutionary pressure for mosquito species (specifically *An. gambiae*) to feed primarily on humans[35]. Therefore, the existing human infective (*P. falciparum)* genotypes would be selected for human and appropriate vector success and the fittest would rapidly expand. Subsequent rapid accumulation of mutations that favoured growth in humans, and in the anthropophilic vectors such as *An. gambiae*, are likely to have occurred to increase human-specific reproductive success. The resulting specific parasite genotypes that expanded (and appeared as an emergence from a bottleneck), would have had a much lower probability of a direct transfer back to apes. With experiments on gorillas and chimpanzees not possible it will be difficult directly to prove the precise combination of different alleles that allowed the emergence of *P. falciparum*. However, our analysis suggests that the genes involved are expressed throughout the life cycle but that only half have ever been characterised, opening up new opportunities for future studies on host specificity and host adaptation in *Plasmodium*.

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Online Methods

Sample collection

All but two infected blood samples from chimpanzees (*Pan troglodytes troglodytes*) and gorillas (*Gorilla gorilla gorilla*) were obtained from the sanctuary "Parc de La Lékédi", Bakoumba (Haut-Ogooué, Gabon), during routine sanitary controls of the animals. This park holds various primate species, including gorillas, chimpanzees, and monkeys (*Cercopithecinae*), that have been orphaned due to bushmeat-poaching activities and have been confiscated by the Gabonese Government, quarantined at the Centre International de Recherches Médicales de Franceville (CIRMF, Gabon) and finally released into semi-free ranging enclosures in the sanctuary. Every six months, chimpanzees (12 individuals) and gorillas (2 individuals) are anesthetized for medical check up. Blood samples were collected from the animals during sanitary controls (July 2011, September 2012, May 2013 and December 2013). Two additional infected blood samples were obtained from gorilla orphans (GG05, GG06) seized by the Gabonese government in 2011 and 2013 and sent to the CIRMF for a quarantine before being released in a sanctuary. All animal work was conducted according to relevant national and international guidelines. From each animal, 15 ml of whole blood were collected in EDTA tubes. For all samples but three, white blood cell depletion was performed on 10 ml of the freshly collected samples using cellulose columns as described in [36]. Remaining blood was subsequently used for DNA extraction and detection of *Plasmodium* infections as described in Ollomo et al[3]. Overall, 15 blood samples from 7 chimpanzees and 4 gorillas were found to contain the *Laverania* samples used in the present study (Table 1).

Sample preparation

Three methods were used for DNA amplification prior to sequencing (Supplementary Table 1). For all but one sample, whole genome amplification (WGA) was performed with a REPLI-g Mini Kit (Qiagen) following a modified protocol[37] to enrich genomic DNA. The genome of *P. blacklocki* was generated using selective WGA (sWGA) as indicated in[38] using 20 primers, followed by a WGA. Finally, for the PprfG03 (a *P. praefalciparum* isolate) and PadlG02 (a *P. adleri* isolate) samples, we used a cell sorting approach[39].

Sample sequencing

All samples were first sequenced with Illumina. Amplification-free Illumina libraries of 400-600 bp were prepared from the enriched genomic DNA[40] , and run on MiSeq and HiSeq 2000 (v3 chemistry) Illumina machines.

After the Illumina sequencing, six samples with a low number of multiple infections (see below) and low host contamination were chosen for long read sequencing, using Pacific

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Biosciences (PacBio). The DNA of the samples (after WGA) was size-selected to 8 kb and sequenced with the C3/P5 chemistry. The number of SMRT cells (Pacific Bioscience sequencing runs) used varied between samples (Supplementary Table 1).

Genome Assembly, genome QC, split of infection & Annotation

Determination of multiple infections

To initially quantify multiple infections, Illumina reads from each sample were mapped against a concatenation of all available *Cox 3* and *CytB* genes of the *Laverania* from NCBI, using SNP-o-Matic[41] (parameter chop=5) to position reads only where they aligned perfectly. SNP-o-Matic returns all the positions of repetitive mapping reads. This method enabled us to determine the number of different malaria species per sample. Samples were selected for PacBio sequencing from those comprising a low number of species.

WGA bias

The uneven coverage that resulted from WGA bias, host contamination and multiple infections presented a challenge for sequence assembly. To overcome the bias and the host contamination, each DNA sample was therefore sequenced deeper than normally necessary. Lower coverage of the subtelomeres was obtained for the sWGA sample (*P. blacklocki*) meaning that the subtelomeres in that assembly were not as complete as those in the assemblies for other species.

PacBio assembly

Six reference genomes were assembled using HGAP[42], with different settings for the genome size parameter, ranging from 23 Mb (*P. reichenowi*) to 72 Mb (*P. billcollinsi*). This parameter encodes how many long reads are corrected for use in the assembly and depends on the host contamination and the amount of different isolates in the samples. The obtained contigs from HGAP were ordered with ABACAS[43] against a *P. falciparum* 3D7 reference that has no subtelomeric regions. Assembly errors and WGA artefacts were manually corrected using ACT[44] . After this step, three iterations of ICORN2[45] were run, followed by another ABACAS step, allowing overlapping contigs to be merged (parameter: ABA_CHECK_OVERLAP=1). For the PrG01, PgabG01 and PadlG01 assembly, we also ran PBjelly to close some of the sequencing gaps[46].

Host decontamination

To detect and remove sequence data derived from host DNA, contigs were compared with the chimpanzee or gorilla genomes using BLAST. Contigs were considered as host contamination if

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

more than 50% of their BLAST hits had higher than 95% identity to any of the great ape genomes. Unordered contigs with a GC content >32% were searched against the non-redundant nucleotide database, to detect and remove further contaminants.

Resolving multiple infections

The first assembled genome was a single *P. reichenowi* infection, PrG01. We detected low levels of *P. vivax-like* and virus contamination (TT virus, AB038624.1), which were excluded. For quality control, the assembly was compared against the existing PrCDC reference genome. The number of *Plasmodium* interspersed repeats (PIRs) was similar, and there were no breaks in synteny. There were however significantly fewer sequencing gaps and 17 Rep20 regions could be found (a known repeat close to the telomeres in *P. falciparum)*. Thus, the assembly of PacBio data appears to be of higher quality than the existing *P. reichenowi* reference.

The *P. adleri* sample comprised a single infection. Because a large number of cycles of amplification were used, a greater number of SMRT cells were sequenced (Supplementary Table 1) to overcome the problem of uneven coverage resulting in under-represented regions. An estimated genome size of 60 Mb was chosen for the HGAP analysis to ensure that all regions were covered.

PgabG01 was a *P. gaboni* isolate with a *P. vivax-like* co-infection. To detect contigs of *P. vivax*, unordered contigs (those that could not be placed against Pf3D7 using ABACAS) were searched against the protein sequences of *P. falciparum* 3D7 and the *P. vivax* PvP01 reference genome using TBLASTx. For each contig, the relative number of genes hitting against the two genomes was used to assign it to *P. gaboni* or *P. vivax*. In most cases, all genes for a given contig consistently hit only one genome so that the attribution to either species was clear. Overall, 14 Mb of *P. vivax-like* sequences were obtained that will be described elsewhere.

The *P. billcollinsi* genome (PbilcG01) was obtained from a co-infection with a *P. gaboni* genome (PgabG02). Rather than ordering the contigs just against Pf3D7 with ABACAS, contigs were ordered against the combined reference of *P. gaboni* (PgabG01) and the Pf3D7 reference (parameters: overlap 500 bp, identity 90%). The species designation of contigs was confirmed with a TBLASTx searches of annotated genes against a combination of the proteomes of PgabG01 and PrCDC. For subtelomeric gene families, contigs were attributed to species if the hit was significant for one species, not the other. Some of the contigs could not be attributed unambiguously and were discarded. Due to sequencing gaps, some of the core genes are missing from the final assembly.

The sample used to produce the *P. praefalciparum* genome (PprfG01) had a high level of host contamination, a low level of co-infection with *P. adleri* and contained two distinct *P. praefalciparum* genotypes. For the core genome, correcting the reference with short reads iCORN enabled us to distinguish the major from minor alleles and effectively assemble the

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

dominant genotype. In the subtelomeres however, it was not possible to distinguish the two *P. praefalciparum* genotypes. Due to contamination of construction vectors (*E. coli*) and host 29 SMRT cells were sequenced, and the HAGP parameter for the assembly size was set to 60 Mb. Contigs were screened against *P. adleri* and *P. falciparum* to exclude a *P. adleri* co-infection. All of the contigs that had a *P. falciparum* BLAST hit or had no clear hit (such as those containing species-specific gene families) were attributed to the *P. praefalciparum* assembly. Last, all samples (Supplementary Table 1) including five *P. falciparum* genomes were mapped against the Pf3D7, *P. praefalciparum and P. adleri* assemblies. Contigs were excluded where more normalized hits to the three *P. adleri* samples were found than to one of the two other *P. praefalciparum* samples. Similarly, this method was used to confirm that none of the contigs in the *P. praefalciparum* assembly were in fact derived from *P. falciparum* contamination.

The *P. blacklocki* sample was from a single infection. Due to sWGA, the PacBio sequence data covered regions not covered by Illumina but due to the bias of the primers, the subtelomeres were not covered fully. However, the internal *var* gene clusters are all assembled. Some of the core genes from this species are also missing.

Annotation

The genomes were annotated as described in[47]. In short, the annotation of *P. falciparum* (version July 2015) was transferred with RATT[48] and new gene models were called with Augustus[49]. Obvious structural errors in core genes were manually corrected in Artemis[50].

Mapping - generation of further samples

To generate the gene sequence for different samples, Illumina reads were mapped against a set of reference genomes using BWA[51] and default parameters. For the gorilla samples, we mapped against the combined PacBio reference genomes of *P. adleri*, *P. blacklocki* and *P. praefalciparum* and for the chimpanzee samples, the combined references of *P. gaboni* (PgabG01), *P. billcollinsi* and *P. reichenowi* (PrG01). SNPs with Phred score ≥ 100 were called using GATK UnifiedGenotyper[52] v2.0.35 (parameters: -pnrm POOL -ploidy 2 -glm POOLBOTH). From these SNP calls we constructed the new gene set, masking regions in genes with less than 10x coverage of 'properly' (correct distance and orientation) mapped paired reads. To generate the sequences of the other 13 isolates homozygous SNP calls were obtained (consensus program from bcftools-1.2[53]). We quality controlled the SNP calling by regenerating PrCDC and PgabG02 gene set from PrG01 and PgabG01, respectively and confirmed that they were placed with nearly no differences in a phylogenetic tree.

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Orthologous group determination and alignment

Orthologous groups were identified using OrthoMCL v1.4[54] across: (i) the seven core *Laverania* genomes; (ii) the seven core genomes, the *Laverania* isolates PgabG02, PrCDC and *P. falciparum* IT, as well as two outgroup genomes *Plasmodium vivax* Sal1 and *Plasmodium knowlesi* strain H; and (iii) just Pf3D7, PprfG01 and PrG01. *P. praefalciparum* II was excluded due to its partial genome. From these groups, different complete sets of 1:1 orthologues were extracted:

(1) "Lav12sp" set of 3,369 orthologues across the seven core *Laverania* species, the PrCDC and *P. falciparum* IT isolates, *P. vivax* and *P. knowlesi*

(2) "Lav25st" set of 424 1:1 orthologues from across the 25 *Laverania* isolates, including the previously published *P. reichenowi* CDC and five *P. falciparum* isolates (3D7, IT, DD2, HB3 and 7G8[9]).

(3) "Lav7sp" set of 4,269 orthologues from across the seven *Laverania* reference genomes

(4) "Lav15st" set of 3,808 orthologous, with two representative sequences per species, excluding *P. blacklocki* and *P. praefalciparum*.

(5) "Lav3sp" set of 4,818 1:1 orthologues across all the *P. reichenowi*, *P. praefalciparum* and *P. falciparum* isolates

The first two sets were used to reconstruct the species tree, the third one for the comparative genomic analyses (introgression, convergence and gene family evolution), the fourth one for the analyses of within species polymorphism and the fifth one for the analysis of *P. falciparum* adaptive evolution.

To reduce the rate of false positives in the evolutionary analyses due to misalignments (e.g.[55]), codon-based multiple alignments were performed using PRANK[56,57] with the -codon and +F options, as it was shown to outperform other programs in the context of the detection of positive selection[58,59]. Prior to aligning codons, low complexity regions were excluded in the nucleotide sequences using dustmasker[60] and in amino acid sequences using segmasker[61] from NCBI-BLAST. Poorly aligned regions were excluded using Gblocks[62], with default settings.

Interspecific gene flow analyses

Species tree inference

Two ML trees were performed using RAxMLv8.1.20[63] to illustrate the phylogenetic relationships between the *Laverania* species and genotypes studied here using the "Lav12sp" and the "Lav25st" set of orthologues. For each tree, multiple nucleotide alignments of each orthologous group were conducted as described above. Trees were then constructed from the concatenated alignments of the "Lav12sp" set of orthologues for the species tree and the "Lav25st" set for the strain tree using RAxML and the following options "-m GTRGAMMA –f a -# 100". Trees were

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

rooted afterwards using *P. vivax* and *P. knowlesi* for the species tree and the *P. adleri/P. gaboni* clade for the genotype tree.

Tree topology test

Interspecific gene flow was investigated by testing congruence between each gene tree topology and the species tree topology. We performed the Shimodaira-Hasegawa test (SH test[64]) using RAxMLv8.1.20 to test whether the phylogenetic tree for each gene significantly differed from the *Laverania* species tree. Topology tests were based on multiple nucleotide alignments of the 4,269 "Lav7sp" set of orthologues. For each coding sequence, RAxML was called with the options "-m GTRGAMMA –f h".

Convergent evolution analyses

Genome-wide test of convergent evolution

Convergent substitutions can occur by chance and the number of random convergent substitutions between two lineages is correlated with the number of divergent substitutions observed in these two lineages[65,66]. Excess of convergent substitutions in specific branch pairs can thus be identified by analyzing the correlation between the number of convergent and divergent substitutions between all the branch pairs in a phylogeny using orthogonal regression, and looking for outlier branch pairs: branch pairs with a high positive residual show an excess of convergent substitutions relatively to the number of divergent substitutions[65]. We used the software Grand-Convergence (available at https://github.com/dekoning-lab/grand-conv) to estimate for each chromosome the numbers of divergent and convergent substitutions between all branch pairs in the *Laverania* tree and investigate whether branch pairs including *Laverania* species infecting the same host species (gorilla or chimpanzee) presented an excess of convergence. Analyses were performed under different models of amino-acid evolution: LG, WAG, JONES and DAYHOFF.

Gene-based detection of convergent evolution throughout the Laverania

For each orthologue of the "Lav7sp" set, the number and percentage of fixed amino acid differences between parasites infecting the same host were calculated, *i.e.* the number of positions showing the same amino acid within a host species but different amino acid between host species. Alignments of all the available sequences ("Lav15st") from all the sequenced isolates were then used to determine what number of host-specific differences were fixed within each host and each species. To evaluate whether the observed number of host-specific fixed differences in an alignment can be attributed to neutral evolution/purifying selection alone (with no positive selection), we used a simulation-based approach. For each coding sequence, 1,000 sequences of the same size were

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

simulated, evolving along the same tree with the same specified branch lengths, substitution model, codon frequencies and omega (d_N/d_S), using the program Evolver from PAML v4.8a[67]. The program Codeml from PAML v4.8a[67] was first used to estimate the tree, the codon frequencies and the average omega values for each of the coding sequences with fixed amino acid differences. For each simulated dataset, the number of fixed amino-acid differences between the parasites infecting a same host was estimated. The probability of observing *n* fixed differences was then computed as the proportion of simulated dataset showing a higher or equal number of fixed differences as the simulated sequences.

Tests for positive selection

Branch site tests

To search for genes that have been subjected to positive selection in the *P. falciparum* lineage alone after the divergence from *P. praefalciparum*, we used the updated Branch site test [68] implemented in PAML v4.4c [67]. This test detects sites that have undergone positive selection in a specific branch of the phylogenetic tree (foreground branch). The "Lav3sp" set of 4,818 orthologuous groups between *P. reichenowi*, *P. praefalciparum I* and *P. falciparum* was used for the test. Dn/Ds ratio estimates per branch and genes were obtained using Codeml (PAML v4.4c) with a *free-ratio* model of evolution.

McDonald–Kreitman (MK) tests

Selection in *P. falciparum* was also tested using McDonald–Kreitman (MK) tests [69] to compare the polymorphism within species to the divergence between species, using *P. praefalciparum* as the outgroup. Analyses were performed using the 4,818 "Lav3sp" set of orthologues. MK tests were performed as described before[9].

Gene Ontology enrichment analyses

Analysis of Gene Ontology (GO) term-enrichment was performed in R, using TopGO[70] with default parameters. GO annotations from GeneDB were used but with unreviewed automated annotations excluded.

Gene family analyses

To estimate the differential abundance of gene families across species, the Gene products and the Pfam domains were counted and analysed by the variance of the occurrence. Unless otherwise stated, trees were constructed using PhyML[71] (default parameters) or raxML[63] (model estimated) from alignments generated with Muscle[72], and trimming with Gblocks[62] in Seaview[73] with default

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

values. Many of the findings were confirmed manually through ACT and bamview[50]. The analysis of the *var* genes was performed on *var* genes larger than 2.4kb. Domains were called with the HMMer models from varDom[74]. Distance matrices were generated based on BLASTp scores, without filtering low complexity regions. Representation was done in R through the heatmap.2 program from gplot (see also Supplementary Note 3).

Allelic dimorphisms

For the analysis of dimorphism in *msp*, all sequences available for the *Laverania* were downloaded from Uniprot[75]. Data were subsampled to obtain a similar number of sequences for each group. Phylogenetic trees were constructed with PhyML[71], using default parameters and drawn in Figtree. The *eba-175* alignment was visualized with Jalview[76].

Divergence Dating

Alignments of the *Laverania* included intergenic regions where possible. Assuming 402–681 mitotic events per year (Supplementary Note 1) and a mutation rate of 3.78E-10 for 4 mitotic events[77,78](mutation rate from latter paper was taken from Pf3D7 line without drugs), equivalent to around 0.9–1.5 mutations per genome per year. Although we observed similar mutation rates in clinical samples (Supplementary Note 1), these estimates have potential errors and therefore we report ratios of divergence times in the figures that are robust to errors in these parameters, see Supplementary Note 1. For coalescence based estimates of speciation times, G-Phocs[79] was used and multiple sequentially Markovian coalescent (MSMC) on segregating sites[80] was used to estimate the *P. falciparum* bottleneck.

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Author Contributions : TDO**,** BO, FR, CN, MB, FP designed the study. CA, APO, LB, EW, BN, ND, CP, PD, VR, FP collected and assessed samples. CA performed the WGA and cell sorting on one sample. SO performed the WGA on the samples; MS organised the sequencing. TDO did assembly and annotation. UB did manual gene curation; AG, FP performed the evolutionary analyses on core genomes. TDO, CN, MB performed the analyses of gene families and dimorphisms. TC performed the dating analyses. TDO, AG, CN, MB, FP wrote the manuscript. All authors read and approved the paper.

Acknowledgments: This work was funded by ANR ORIGIN JCJC 2012, LMI ZOFAC, CNRS, CIRMF, IRD and the Wellcome Trust (grant WT 098051 to the Sanger Institute, 104792/Z/14/Z to CN). TC holds a MRC DTP Studentship. We thank Gavin Rutledge for performing the sWGA and Julian Rayner and Francisco J. Ayala for helpful discussion.

Accession codes

All sequences have been submitted to the European Nucleotide Archive. The accession numbers of the raw reads, and assembly data can be found in Supplementary Table 8. The genomes are being submitted to EBI, project ID PRJEB13584.

Competing financial interests.

None

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure legends:

Figure 1. Overview of the dating of the evolution of the *Laverania*. (**a**) Maximum likelihood tree of the *Laverania* based on the "Lav12sp" set of orthologues. All bootstrap values are 100. Coalescence based estimates of the timing of speciation events are displayed on nodes (MYA - million years ago), based on intergenic and genic alignments. (**b**) Multiple sequentially Markovian coalescent estimates of the effective population size (*Ne*) in the *P. falciparum* and *P. praefalciparum* population. Assuming our estimate of the number of mitotic events per year, a bottleneck occurred if *P. falciparum* 4,000-6,000 years ago. y-axis is the natural logarithm (Ln) of *Ne*. Bootstrapping was performed by 50 replicates by randomly resampling from the segregating sites used as input.

Figure 2. Overview of the analyses of core genes over all *Laverania* genomes. (a) Summary of evolution of core genes. From outer to inner track: scatterplot of branch site test for each genome, with a summary; d_N/d_S values ($0.5 < d_N/d_S < 2$) and a summary plot; vertical lines under the chromosome track represent orthologues, dots represent *var* genes of *P. falciparum* 3D7 on forward (blue) or reverse track (red) and black if pseudogenized; Polymorphism (π), normalized by the average per species (dataset "Lav25st"); Convergent evolution: host-specific fixed differences analysis (dataset "Lav15st"). For all other analyses we used the dataset "Lav7sp". (b) Magnified view of the *rh5* locus that is enriched with host specific fixed differences (convergent analyses). Analysis was performed using the "Lav7sp" set of orthologues but filled circles are for the differences that are fixed within all the isolates available ("Lav15st" set) and for which we could reject neutral evolution (For the gene list see Supplementary Table 4).

Figure 3. Gene families in the *Laverania.* Distribution of major multigene families. Data from *P. praefalciparum* include the subtelomeric gene families from the two infecting genotypes.

Figure 4: Clustering of *rif* and *stevor* genes. Graphical representation of similarity between *pir* genes (longer than 250aa), coloured by species. A BLAST cut-off of 45% global identity was used. More connected genes are more similar. Black circle highlight Clade A *rif* genes that cluster with Clade B *rif* genes.

Figure 5. Heatmap of frequency of *var* gene domains in each *Laverania* species. Duffy represents regions closest to the Pfam Duffy binding domain. CIDRn is a new domain discovered in this study in Clade A. Only domains from *var* genes longer than 2.5 kb were considered.

Fig. 6. Evolution of *var* genes domains in the *Laverania*. (**a**) Graphical representation of similarity between domains, coloured by species. A BLAST cut-off of 45% global identity was used. More connected domains are more similar. (**b**) Maximum likelihood trees of the Acidic Terminal Sequence (ATS). Domains from *var* genes shorter than 2.5 kb were excluded.

Figure 7. Overview of the genomic evolution of the *Laverania* subgenus. The values of polymorphism (pi) within the species are indicated by triangles of different size at the end of the tree branches, as well the bottleneck in *P. falciparum* (constricted branch width), ~ 5,000 years ago. Also shown are the gene transfers that occurred between certain Clade A and B species and the huge genomic differences that accumulated in Clade B after the divergence with *P. blacklocki*.

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Species	Sample ID	Primate species	Fold coverage	Assembly size Mb	Contigs	Scaffolds	Genes	Primary co-infection [b]
	PprfG01	**Chimp**	**104**	**26**	**142**	**73**	**6,476**	***P. adleri***
P. praefalciparum	PprfG02	Gorilla	1269	-	-	-	-	none
	PprfG04	Co-infection, see PadlG03						*P. adleri*
	PprfG03	Gorilla	668	-	-	-	-	none
P. reichenowi	**PrG01**	**Chimp**	**106**	**24.5**	**66**	**48**	**5,941**	***none***
	PrCDC [a]	Chimp	296	24.1	374	2,465	5,895	none
	PrG02	Chimp	123	-	-	-	-	none
	PrG03	Chimp	112	-	-	-	-	none
P. billcollinsi	**PbilcG01**	**Co-infection PgabG02**		**23.1**	**306**	**89**	**5,637**	***P. gaboni***
	PbilcG03	Co-infection, see PgabG04						*P. gaboni*
	PbilcG04	Co-infection, see PgabG05						*P. gaboni*
P. blacklocki	**PblacG01**	**Gorilla**	**221**	**22**	**311**	**97**	**5,346**	**none**
P. gaboni	**PGAB01**	**Chimp**	**367**	**21.8**	**121**	**96**	**5,421**	***P. vivax***
	PgabG02	**Chimp**	**341**	**20.9**	**76**	**44**	**5,249**	***P. billcollinsi***
	PgabG03	Chimp	669	-	-	-	-	*P. vivax*
	PgabG04	Chimp	679	-	-	-	-	*P. billcollinsi*
	PgabG05	Chimp	530	-	-	-	-	*P. reichenowi*
P. adleri	**PadlG01**	**Gorilla**	**372**	**22.2**	**102**	**82**	**5,515**	**none**
	PadlG02	Gorilla	2262	-	-	-	-	none
	PadlG03	Gorilla	445	-	-	-	-	*P. praefalciparum*

[a] Data from Otto *et al*. [9]

[b] Based on percentage of reads mapping to the reference for each Laverania species (see Supplementary Table 1)

Table 1: Overview of all *Laverania* samples used in study. All samples generated for this study, except PrCDC, were sequenced using Illumina technology with a range of read lengths from 100–250 bp. For isolates in bold, assemblies were produced using long-reads from Single-Molecule Real-Time sequencing (Pacific Biosciences) and the total assembled size, number of contigs, scaffolds and predicted genes are shown. Fold coverage is reported based on mapping reads to the *P. falciparum* 3D7 reference genome (v3.1). *P. billcollinsi* sequences were obtained from *P. gaboni* co-infections, of which some also harboured *P. reichenowi* species.

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1 Prugnolle, F. *et al.* African great apes are natural hosts of multiple related malaria species, including Plasmodium falciparum. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 1458-1463, doi:10.1073/pnas.0914440107 (2010).

2 Liu, W. *et al.* Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. *Nature* **467**, 420-U467, doi:10.1038/nature09442 (2010).

3 Ollomo, B. *et al.* A New Malaria Agent in African Hominids. *PLoS Pathog* **5**, e1000446, doi:10.1371/journal.ppat.1000446 (2009).

4 Liu, W. *et al.* Multigenomic Delineation of Plasmodium Species of the Laverania Subgenus Infecting Wild-Living Chimpanzees and Gorillas. *Genome biology and evolution* **8**, 1929-1939, doi:10.1093/gbe/evw128 (2016).

5 Boundenga, L. *et al.* Diversity of malaria parasites in great apes in Gabon. *Malar J* **14**, 111, doi:10.1186/s12936-015-0622-6 (2015).

6 Sundararaman, S. A. *et al.* Genomes of cryptic chimpanzee Plasmodium species reveal key evolutionary events leading to human malaria. *Nature communications* **7**, 11078, doi:10.1038/ncomms11078 (2016).

7 Silva, J. C., Egan, A., Arze, C., Spouge, J. L. & Harris, D. G. A New Method for Estimating Species Age Supports the Coexistence of Malaria Parasites and Their Mammalian Hosts. *Molecular biology and evolution* **32**, 1354-1364, doi:10.1093/molbev/msv005 (2015).

8 Volkman, S. K. *et al.* Recent origin of Plasmodium falciparum from a single progenitor. *Science* **293**, 482-484, doi:10.1126/science.1059878 (2001).

9 Otto, T. D. *et al.* Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nature communications* **5**, 4754, doi:10.1038/ncomms5754 (2014).

10 Larremore, D. B. *et al.* Ape parasite origins of human malaria virulence genes. *Nature communications* **6**, 8368, doi:10.1038/ncomms9368 (2015).

11 Rutledge, G. G. *et al.* Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. *Nature* **542**, 101-104, doi:10.1038/nature21038 (2017).

12 Pacheco, M. A. *et al.* Timing the origin of human malarias: the lemur puzzle. *BMC evolutionary biology* **11**, 299, doi:10.1186/1471-2148-11-299 (2011).

13 Behar, D. M. *et al.* The dawn of human matrilineal diversity. *American journal of human genetics* **82**, 1130-1140, doi:10.1016/j.ajhg.2008.04.002 (2008).

14 Palstra, F. P. & Fraser, D. J. Effective/census population size ratio estimation: a compendium and appraisal. *Ecology and evolution* **2**, 2357-2365, doi:10.1002/ece3.329 (2012).

15 Chang, H. H. & Hartl, D. L. Recurrent bottlenecks in the malaria life cycle obscure signals of positive selection. *Parasitology* **142 Suppl 1**, S98-S107, doi:10.1017/S0031182014000067 (2015).

16 Roy, S. W. The Plasmodium gaboni genome illuminates allelic dimorphism of immunologically important surface antigens in P. falciparum. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **36**, 441-449, doi:10.1016/j.meegid.2015.08.014 (2015).

17 Tanabe, K., Mackay, M., Goman, M. & Scaife, J. G. Allelic dimorphism in a surface antigen gene of the malaria parasite Plasmodium falciparum. *Journal of molecular biology* **195**, 273-287 (1987).

18 Yasukochi, Y., Naka, I., Patarapotikul, J., Hananantachai, H. & Ohashi, J. Genetic evidence for contribution of human dispersal to the genetic diversity of EBA-175 in Plasmodium falciparum. *Malar J* **14**, 293, doi:10.1186/s12936-015-0820-2 (2015).

19 Malaria Genomic Epidemiology, N., Band, G., Rockett, K. A., Spencer, C. C. & Kwiatkowski, D. P. A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature* **526**, 253-257, doi:10.1038/nature15390 (2015).

20 Makanga, B. *et al.* Ape malaria transmission and potential for ape-to-human transfers in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 1603008113-, doi:10.1073/pnas.1603008113 (2016).

21 Wanaguru, M., Liu, W., Hahn, B. H., Rayner, J. C. & Wright, G. J. RH5–Basigin interaction plays a major role in the host tropism of Plasmodium falciparum. *Proceedings of the National Academy of Sciences* **110**, 20735-20740, doi:10.1073/pnas.1320771110 (2013).

22 Wright, K. E. *et al.* Structure of malaria invasion protein RH5 with erythrocyte basigin and blocking antibodies. *Nature* **515**, 427-+, doi:10.1038/nature13715 (2014).

23 Triglia, T., Thompson, J. K. & Cowman, A. F. An EBA175 homologue which is transcribed but not translated in erythrocytic stages of Plasmodium falciparum. *Mol Biochem Parasitol* **116**, 55-63 (2001).

24 Ramiro, R. S. *et al.* Hybridization and pre-zygotic reproductive barriers in Plasmodium. *Proceedings of the Royal Society B-Biological Sciences* **282**, doi:10.1098/rspb.2014.3027 (2015).

25 Eksi, S. *et al.* Malaria transmission-blocking antigen, Pfs230, mediates human red blood cell binding to exflagellating male parasites and oocyst production. *Molecular Microbiology* **61**, 991-998, doi:10.1111/j.1365-2958.2006.0528.x (2006).

26 Mundwiler-Pachlatko, E. & Beck, H. P. Maurer's clefts, the enigma of Plasmodium falciparum. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 19987-19994, doi:10.1073/pnas.1309247110 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

27 Bethke, L. L. *et al.* 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 (2006).

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

29 Niang, M. *et al.* 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 (2014).

30 Kraemer, S. M. & Smith, J. D. A family affair: var genes, PfEMP1 binding, and malaria disease. *Curr Opin Microbiol* **9**, 374-380, doi:10.1016/j.mib.2006.06.006 (2006).

31 Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511, doi:http://www.nature.com/nature/journal/v419/n6906/suppinfo/nature01097_S1.html (2002).

32 Bordbar, B. *et al.* Genetic diversity of VAR2CSA ID1-DBL2Xb in worldwide Plasmodium falciparum populations: impact on vaccine design for placental malaria. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **25**, 81-92, doi:10.1016/j.meegid.2014.04.010 (2014).

33 Frank, M., Dzikowski, R., Amulic, B. & Deitsch, K. Variable switching rates of malaria virulence genes are associated with chromosomal position. *Mol Microbiol* **64**, 1486-1498, doi:10.1111/j.1365-2958.2007.05736.x (2007).

34 Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nature reviews. Genetics* **13**, 745-753, doi:10.1038/nrg3295 (2012).

35 Carter, R. & Mendis, K. N. Evolutionary and historical aspects of the burden of malaria. *Clinical microbiology reviews* **15**, 564-594 (2002).

36 Auburn, S. *et al.* An effective method to purify plasmodium falciparum dna directly from clinical blood samples for whole genome high-throughput sequencing. *PLoS ONE* **6**, doi:10.1371/journal.pone.0022213 (2011).

37 Oyola, S. O. *et al.* Optimized whole-genome amplification strategy for extremely AT-biased template. *DNA research : an international journal for rapid publication of reports on genes and genomes* **21**, 661-671, doi:10.1093/dnares/dsu028 (2014).

38 Oyola, S. O. *et al.* Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. *bioRxiv*, doi:10.1101/067546 (2016).

39 Boissiere, A. *et al.* Isolation of Plasmodium falciparum by flow-cytometry: implications for single-trophozoite genotyping and parasite DNA purification for whole-genome high-throughput sequencing of archival samples. *Malaria Journal* **11**, 163 (2012).

40 Quail, M. A. *et al.* in *Nat Methods* Vol. 9 10-11 (2012).

41 Manske, H. & Kwiatkowski, D. SNP-o-matic. *Bioinformatics 25*, 2434-2435 (2009).

42 Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods* **10**, 563-569, doi:10.1038/nmeth.2474 (2013).

43 Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics Vol. 25 No. 15*, 1968-1969 (2009).

44 Carver, T. *et al.* Artemis and ACT: viewing, annotation and comparing sequences stored in relational database. *Bioinformatics* **24**, 2672-2676 (2008).

45 Otto, T. D., Sanders, M., Berriman, M. & Newbold, C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics 26(14)*, 1704-1707 (2010).

46 English, A. C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**, e47768, doi:10.1371/journal.pone.0047768 (2012).

47 Otto, T. D. From sequence mapping to genome assemblies. *Methods Mol Biol* **1201**, 19-50, doi:10.1007/978-1-4939-1438-8_2 (2015).

48 Otto, T. D., Dillon, G. P., Degrave, W. S. & Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Research*, 1-7 (2011).

49 Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435-439 (2006).

50 Carver, T. *et al.* BamView: visualizing and interpretation of next-generation sequencing read. *Briefings in bioinformatics*, doi:10.1093/bib/bbr073 (2013).

51 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

52 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

53 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).

54 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189, doi:10.1101/gr.1224503 (2003).

bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

55 Jordan, G. & Goldman, N. The Effects of Alignment Error and Alignment Filtering on the Sitewise Detection of Positive Selection. *Molecular biology and evolution* **29**, 1125-1139, doi:10.1093/molbev/msr272 (2012).

56 Loytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 10557-10562, doi:10.1073/pnas.0409137102 (2005).

57 Loytynoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632-1635, doi:10.1126/science.1158395 (2008).

58 Fletcher, W. & Yang, Z. The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection. *Molecular biology and evolution* **27**, 2257-2267, doi:10.1093/molbev/msq115 (2010).

59 Markova-Raina, P. & Petrov, D. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Research* **21**, 863-874, doi:10.1101/gr.115949.110 (2011).

60 Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. *Journal of Computational Biology* **13**, 1028-1040, doi:10.1089/cmb.2006.13.1028 (2006).

61 Wootton, J. C. & Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry* **17**, 149-163, doi:http://dx.doi.org/10.1016/0097-8485(93)85006-X (1993).

62 Castresana, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular biology and evolution* **17**, 540-552 (2000).

63 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).

64 Shimodaira, H. & Hasegawa, M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Molecular biology and evolution* **16**, 1114 (1999).

65 Castoe, T. A. *et al.* Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 8986-8991, doi:10.1073/pnas.0900233106 (2009).

66 Thomas, G. W. C. & Hahn, M. W. Determining the Null Model for Detecting Adaptive Convergence from Genomic Data: A Case Study using Echolocating Mammals. *Molecular biology and evolution* **32**, 1232-1236, doi:10.1093/molbev/msv013 (2015).

67 Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular biology and evolution* **24**, 1586-1591, doi:10.1093/molbev/msm088 (2007).

68 Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Molecular biology and evolution* **22**, 2472-2479, doi:10.1093/molbev/msi237 (2005).

69 McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652-654 (1991).

70 Rahnenfuhrer, A. A. a. J. topGO: Enrichment analysis for Gene Ontology. *R package* **version 2.8.0** (2010).

71 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321, doi:10.1093/sysbio/syq010 (2010).

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

73 Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution* **27**, 221-224, doi:10.1093/molbev/msp259 (2010).

74 Rask, T. S., Hansen, D. A., Theander, T. G., Gorm Pedersen, A. & Lavstsen, T. Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes--divide and conquer. *PLoS Comput Biol* **6**, doi:10.1371/journal.pcbi.1000933 (2010).

75 UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212, doi:10.1093/nar/gku989 (2015).

76 Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191, doi:10.1093/bioinformatics/btp033 (2009).

77 Claessens, A. *et al.* 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 (2014).

78 Bopp, S. E. *et al.* Mitotic evolution of Plasmodium falciparum shows a stable core genome but recombination in antigen families. *PLoS Genet* **9**, e1003293, doi:10.1371/journal.pgen.1003293 (2013).

79 Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet* **43**, 1031-1034, doi:10.1038/ng.937 (2011).

80 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925, doi:10.1038/ng.3015 (2014).

Fig. 1

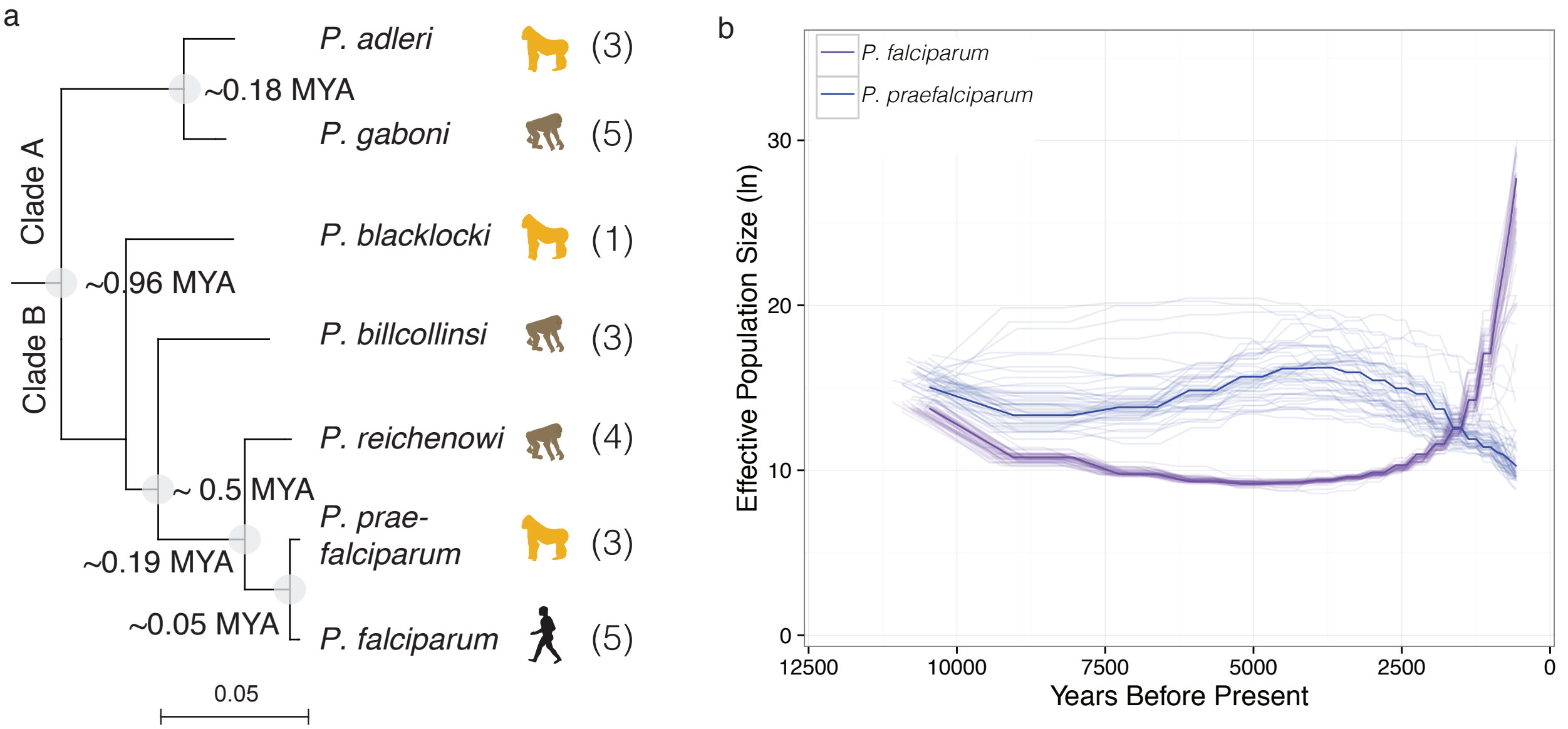
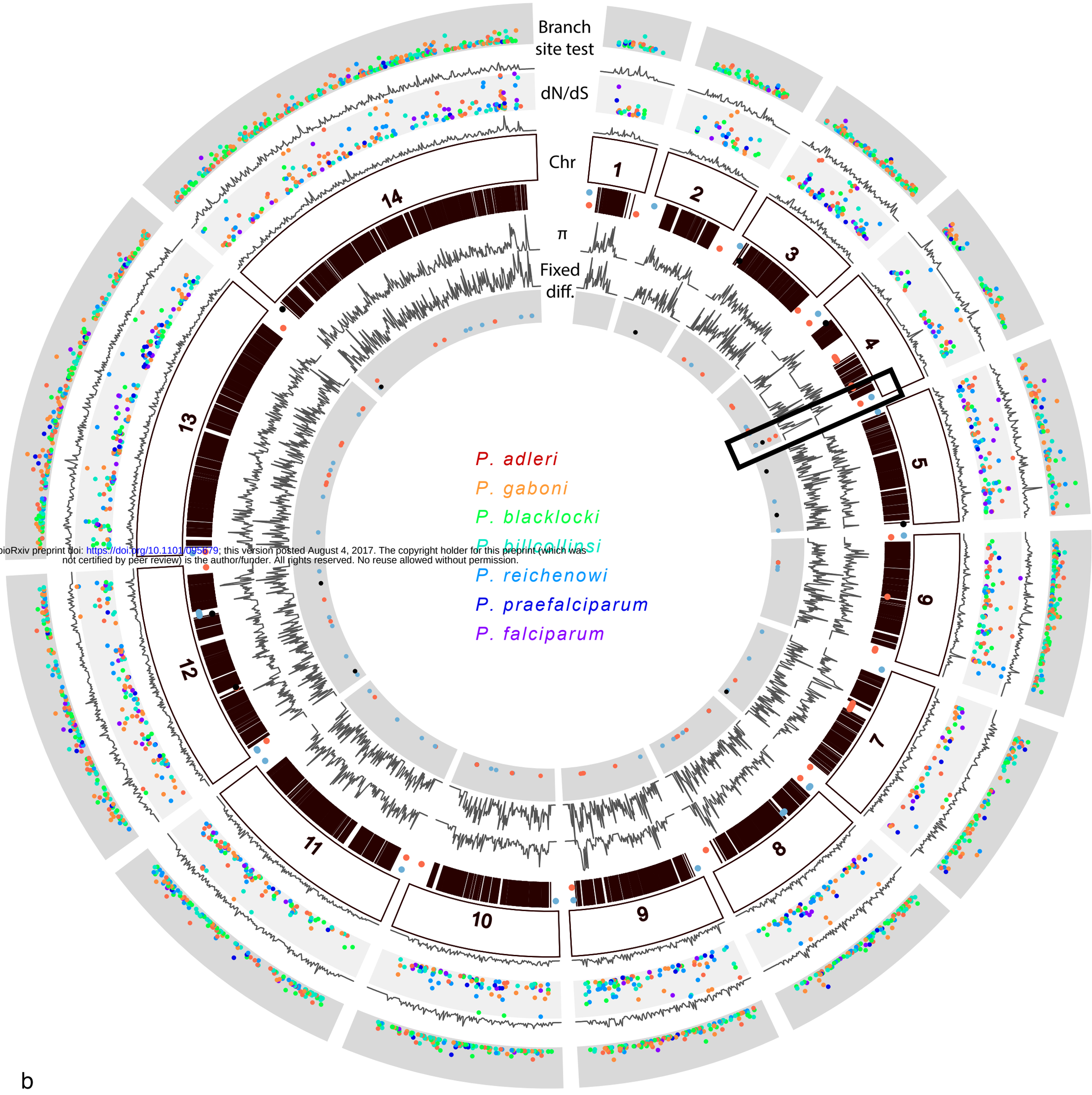

a
P. adleri
(3)
~0.18 MYA
P. gaboni
(5)
Clade A
P. blacklocki
(1)
~0.96 MYA
Clade B
P. billcollinsi
(3)
P. reichenowi
(4)
~ 0.5 MYA
P. prae-
falciparum
(3)
~0.19 MYA
~0.05 MYA
P. falciparum
(5)
0.05
b
P. falciparum
P. praefalciparum
Effective Population Size (ln)
30
20
10
0
12500
10000
7500
5000
2500
0
Years Before Present

Fig. 2

a



bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

b

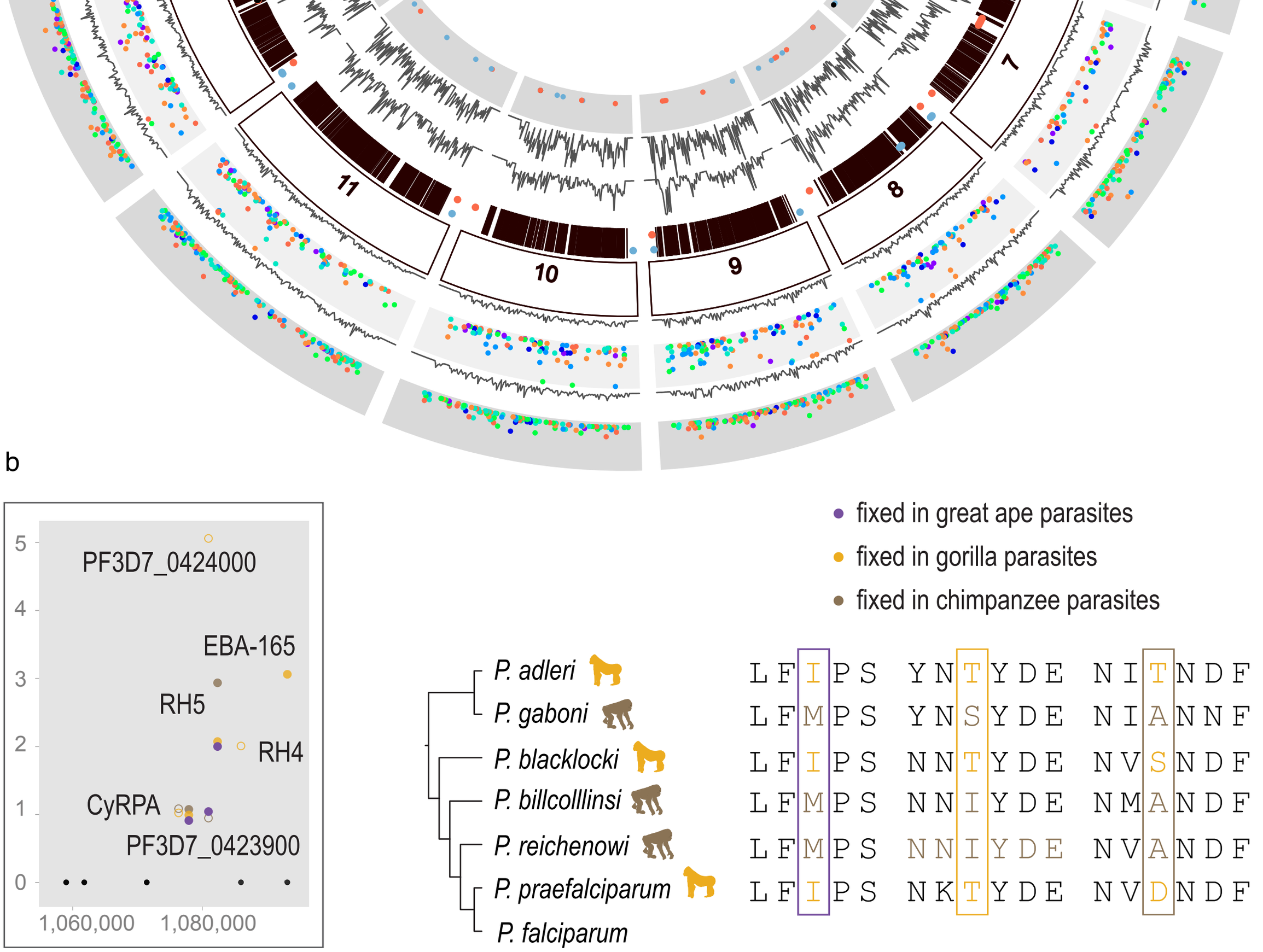
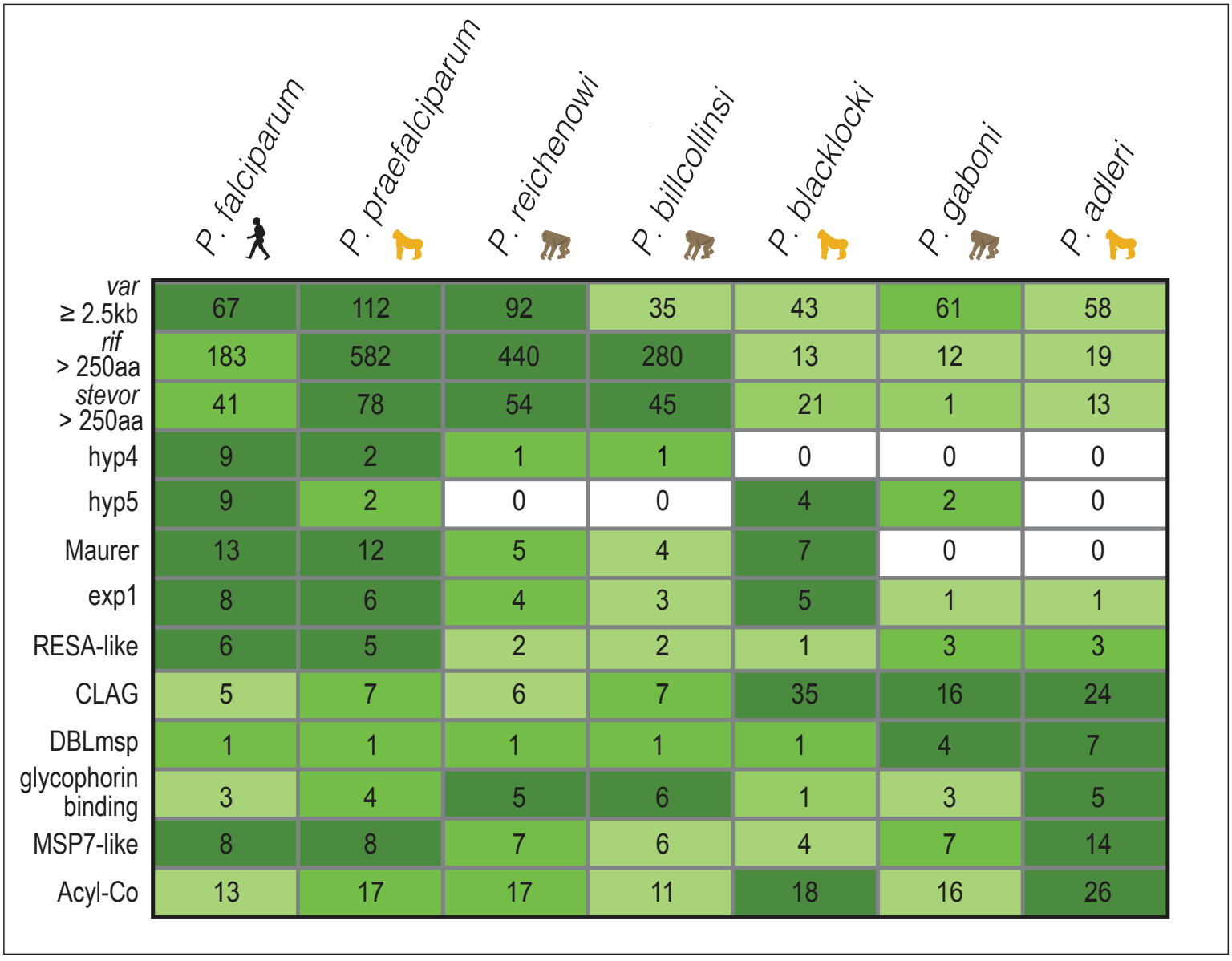


Fig. 3



	P. falciparum	P. praefalciparum	P. reichenowi	P. billcollinsi	P. blacklocki	P. gaboni	P. adleri
var ≥ 2.5kb	67	112	92	35	43	61	58
rif > 250aa	183	582	440	280	13	12	19
stevor > 250aa	41	78	54	45	21	1	13
hyp4	9	2	1	1	0	0	0
hyp5	9	2	0	0	4	2	0
Maurer	13	12	5	4	7	0	0
exp1	8	6	4	3	5	1	1
RESA-like	6	5	2	2	1	3	3
CLAG	5	7	6	7	35	16	24
DBLmsp	1	1	1	1	1	4	7
glycophorin binding	3	4	5	6	1	3	5
MSP7-like	8	8	7	6	4	7	14
Acyl-Co	13	17	17	11	18	16	26

Fig. 4

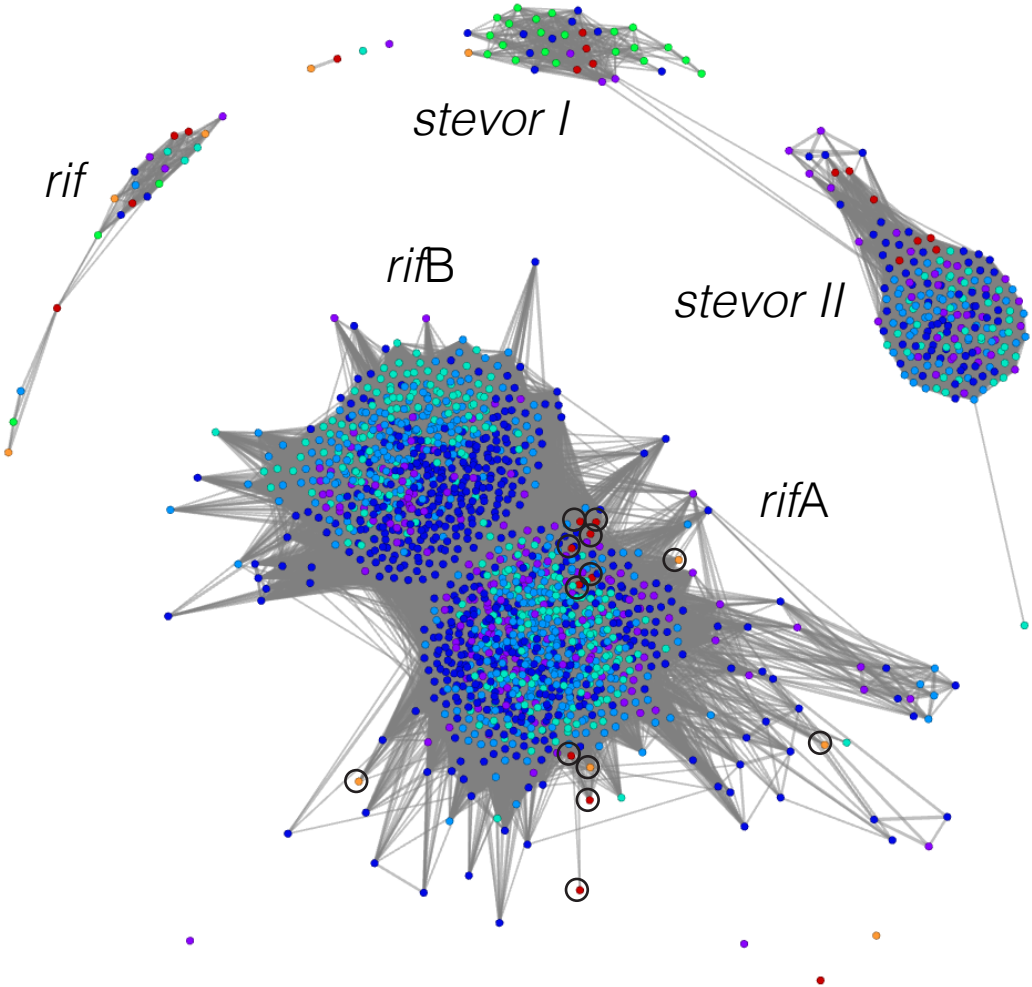

stevor I
rif
rifB
stevor II
rifA
P. adleri
P. blacklocki
P. reichenowi
P. falciparum
P. gaboni
P. billcollinsi
P. praefalciparum

Fig. 5

	CIDRa	CIDRb	CIDRd	CIDRg	CIDRn	CIDRpam	DBLpam1	DBLpam2	DBLpam3	DBLa	DBLb	DBLd	DBLe	DBLg	DBLz	Duffy	ATS
P. adleri	1	1	0	0	14	4	1	15	9	2	63	0	106	67	20	77	39
P. gaboni	1	1	0	0	8	16	3	30	20	2	47	0	84	43	16	48	41
P. blacklocki	0	1	0	0	0	0	0	0	0	1	17	0	16	55	7	0	34
P. billcollinsi	31	28	0	5	0	0	1	0	0	30	9	34	4	2	1	0	28
P. reichenowi	86	61	1	27	0	1	2	1	1	90	50	85	18	43	8	0	85
P. praefalciparum	85	48	5	30	0	5	5	5	5	94	86	72	97	84	34	0	105
P. falciparum	56	37	2	17	0	1	3	1	1	59	18	53	15	16	7	0	65

Fig. 6

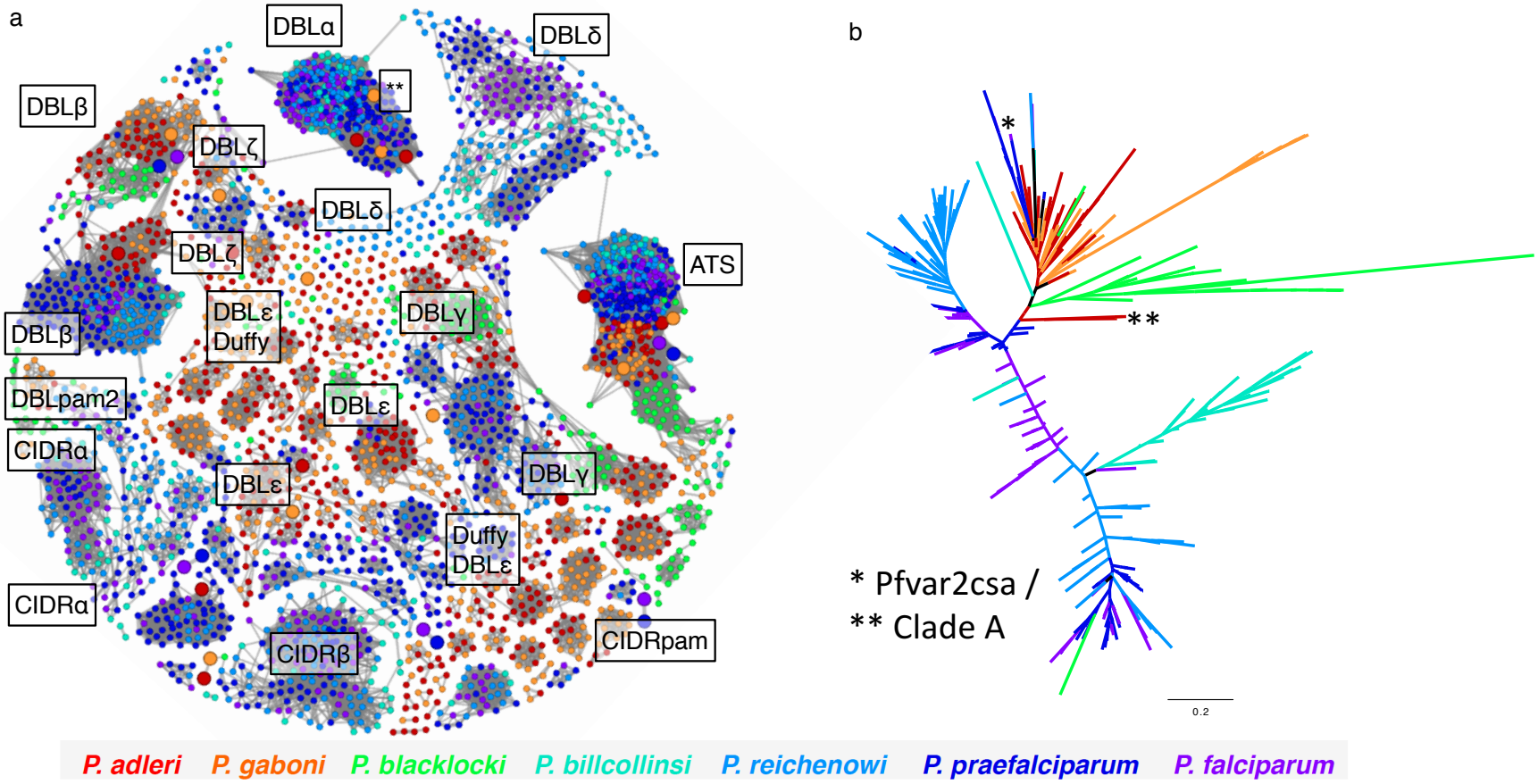
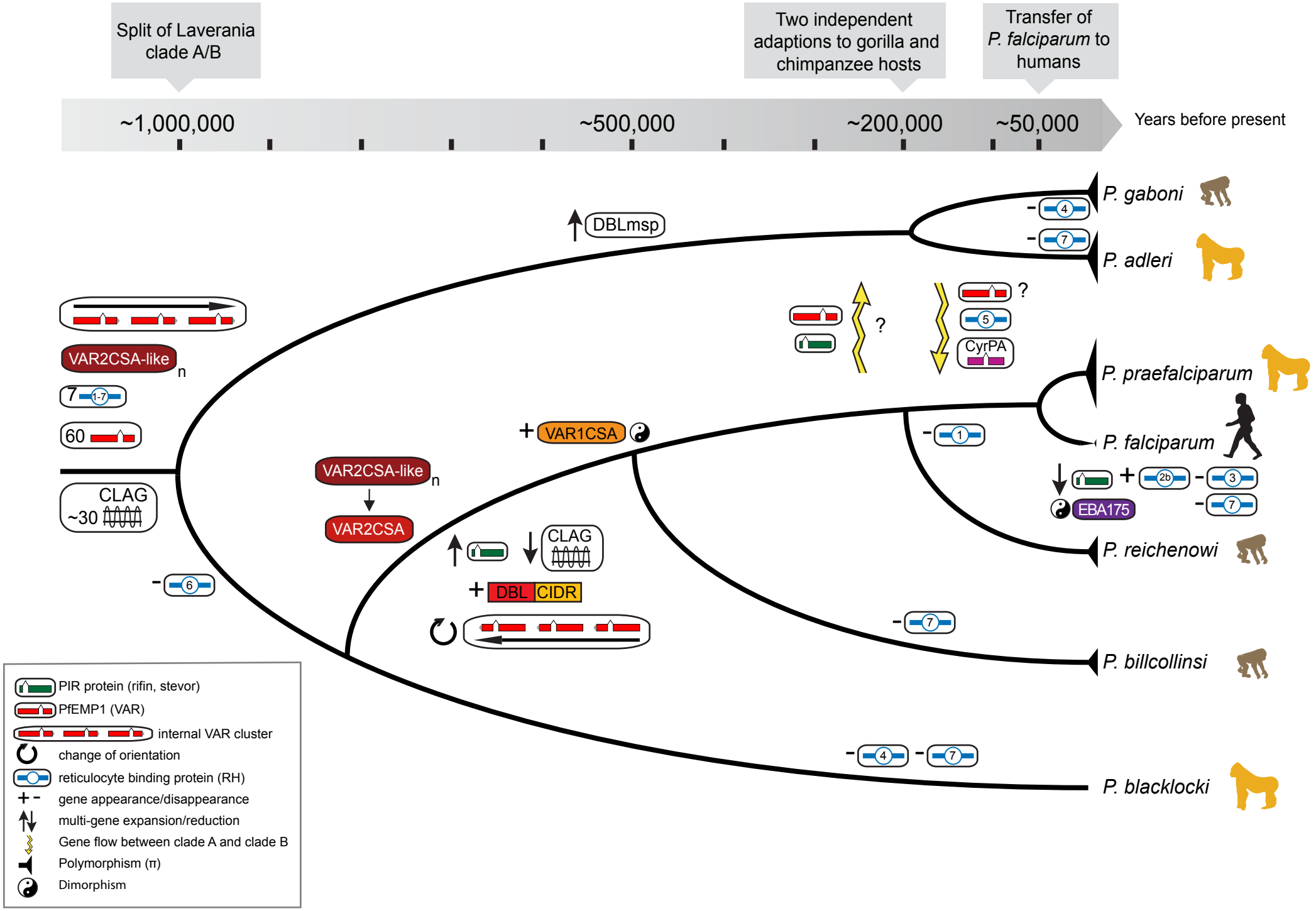

a
DBLα
DBLδ
**
DBLβ
DBLζ
DBLδ
DBLζ
ATS
DBLε
Duffy
DBLγ
DBLβ
DBLpam2
DBLε
CIDRα
DBLε
DBLγ
Duffy
DBLε
CIDRα
CIDRβ
CIDRpam
b
*
**
* Pfvar2csa /
** Clade A
0.2
P. adleri
P. gaboni
P. blacklocki
P. billcollinsi
P. reichenowi
P. praefalciparum
P. falciparum

Fig. 7


Split of Laverania clade A/B
Two independent adaptions to gorilla and chimpanzee hosts
Transfer of P. falciparum to humans
~1,000,000
~500,000
~200,000
~50,000
Years before present
P. gaboni
P. adleri
P. praefalciparum
P. falciparum
P. reichenowi
P. billcollinsi
P. blacklocki
DBLmsp
VAR2CSA-like
n
7 1-7
60
CLAG ~30
6
VAR2CSA
VAR1CSA
CLAG
DBL CIDR
CyrPA
EBA175
PIR protein (rifin, stevor)
PfEMP1 (VAR)
internal VAR cluster
change of orientation
reticulocyte binding protein (RH)
gene appearance/disappearance
multi-gene expansion/reduction
Gene flow between clade A and clade B
Polymorphism (π)
Dimorphism

SUPPLEMENTARY INFORMATION

Title: Genomes of an entire *Plasmodium* subgenus reveal paths to virulent human malaria

Authors: Thomas D. Otto[1,†,*], Aude Gilabert[2,†], Thomas Crellen[1,3], Ulrike Böhme[1], Céline Arnathau[2], Mandy Sanders[1], Samuel Oyola[1], Alain Prince Okouga[4], Larson Boundenga[4], Eric Willaume[5], Barthélémy Ngoubangoye[4], Nancy Diamella Moukodoum[4], Christophe Paupy[2], Patrick Durand[2], Virginie Rougeron[2,4], Benjamin Ollomo[4], François Renaud[2], Chris Newbold[1,6], Matthew Berriman[1,*] & Franck Prugnolle[2,4,*]

Affiliations:

[1] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, United Kingdom

[2] Laboratoire MIVEGEC, CNRS 5290-IRD 224-UM, Montpellier, France

[3] Department of Infectious Disease Epidemiology, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, United Kingdom

[4] Centre International de Recherches Médicales de Franceville, Franceville, Gabon

[5] Sodepal, Parc of la Lékédi, Bakoumba, Gabon

[6] Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, United Kingdom

Contents

Supplementary Note 1: Dating and population size estimates

Estimation of single nucleotide substitution per year for different generation times
in-vivo data
Coalescent models
Adapting dating for *P. ovale* and *P. malariae-like* speciation events
Multiple Sequentially Markovian Coalescent
Estimation of population size
Dating of *eba-175* dimorphism

Supplementary Note 2: Core gene analyses

Within-species polymorphism
Interspecific gene transfer
Genome-wide test of convergent evolution

Supplementary Note 3: Gene family analyses

Differences in gene families

Generation of similarity matrices
The *rif* and *stevor* genes
var gene analysis

Supplementary Fig. 1-12 (Legends, figures are at the end of the document)

Supplementary Table 1-10 (Legends only, Tables are provided in excel files)

Supplementary Note 1: Dating and population size estimates

A major focus of this study has been to understand the population history of the *Laverania* species and in particular the timing of the divergence events as well as the variation of population size. We used two methods: 1) a Bayesian coalescence model, G-PhoCS[1] to estimate the timing of species divergence (Fig. 1a) and 2) the Multiple Sequentially Markovian Coalescent (MSMC)[2] to provide a high resolution estimate of changes of N_e through time, specifically to look for a bottleneck that would explain the low diversity in the *P. falciparum* population (Fig. 1b). To scale the population genetic parameters inferred from these models to 'real time' and N_e, we used a per-base mutation rate of $3.78*10^{-10}$ (for 4 mitotic events in the red blood cycle)[3].

Estimation of single nucleotide substitution per year for different generation times

The total number of mitoses per generation was calculated based on different assumptions about total generation time (time to complete full life cycle), time to complete different stages of the life cycle and number of mitoses per stage.

Data from previous studies:

- Development in mosquito takes 10–22 days and involves 10–12 mitoses[5]
- Development in liver takes 5–7 days and involves 15 mitoses[6]
- Gametocyte development takes 12 days and involves 3 mitoses[4].
- Intra-erythrocytic development involves 2 mitoses per day (undergoes three to four rounds of DNA synthesis, mitosis, and nuclear division to produce a syncytial schizont with 16 to 22 nuclei)[4]

We have assumed that generation times can be within the range 60 –180 days[7,8].

1. Assuming 60-day generation time

	Days	Min. mitoses	Max. mitoses	Days	Min. mitoses	Max. mitoses
Oocyst to salivary gland	10	10	12	22	10	12
Liver	5	15	15	7	15	15
Gametocytes	12	3	3	12	3	3
subtotal	27	28	30	41	28	30
Inferred data, based on 60-day generation time:						
Blood parameters	33	66	66	19	38	38
Total mitoses per gen.		94	96		66	68
Generations per year		6.1	6.1		6.1	6.1
Total mitoses per year		**572**	**584**		**401**	**414**

2. Assuming 180-day generation time

	Days	Min. mitoses	Max. mitoses	Days	Min. mitoses	Max. mitoses
Oocyst to salivary gland	10	10	12	22	10	12
Liver	5	15	15	7	15	15
Gametocytes	12	3	3	12	3	3
subtotal	27	28	30	41	28	30
Inferred data, based on 180-day genome time:						
Blood parameters	153	306	306	139	278	278
Total mitoses per gen.		334	336		306	308
Generations per year		2	2		2	2
Total mitoses per year		**677**	**681**		**621**	**625**

Picking extreme values from **1** and **2** (in red), *total mitoses per year* = 401 to 681

Using data from Claessens *et al*[3]*:*

Average mutation rate = **3.83 x10^{-10}** per base **per 48 hr cycle**

(equivalent to 1.64 mutations per genome per year *in vitro*)

=>Average mutation rate = **9.57 x10^{-11}** per base **per mitosis**

=>Expected **mutations per base per year** = (9.57 x10^{-11} x **401**) to (9.57 x 10^{-11} x **681**)

= **3.84 x 10^{-8} to 6.52 x 10^{-8}**

=>Expected **mutations per genome per year = 0.9 – 1.5**

According to Bopp et al[9], excluding parasites grown in presence of drug the numbers of measured mutations = 5.046, 1.682 and 1.682 per genome per year. The median value from this study is also nearly identical to that described by Claessens *et al*[3]**.**

In-vivo data

In the *Plasmodium falciparum* IT[10] genome, we observed a region of around 225 – 312 kb, covering the PfCRT locus and an internal *var* gene cluster that is highly conserved in a number of field isolates. Since all of these isolates have the chloroquine resistant genotype, the conserved region is likely to have resulted from chloroquine-selective sweep and could be around 50 years old[11]. However, the presence of a *var* gene on the opposite strand differentiates these isolates from others and may have decreased overall recombination rates in this region.

We called SNPs in this region on 5 isolates (from the Pf3k project) including PfIT and detected regions that were nearly SNP free over the PfCRT and the right-hand side *var* gene cluster, Supplementary Fig. 2, table below. We observed between 0-10 substitutions per year, with a median of 1.67 mutations per year.

SNPs were called with mpileup and varfilter from samtools[12], after remapping the reads with bwa[13].

Conserved regions around the PfCRT in five clinical *P. falciparum* isolates.

Assuming that the selection occurred ~50 years ago, we obtain the reported estimate of mutations per year and genome.

Country	Location of valley on PfIT_07	Size of valley	mpileup Called SNPs	SNPs	estimate mutation per year / genome
The Gambia 2	378563..660686	283	6	6	10.38
Thailand	228905..510000	281	4	2	3.48
Senegal	383941..609688	225	0	0	0
The Gambia 1	378352..699841	312	1	1	1.57
Mali	372520..664849	292	1	1	1.68

In conclusion, we assume that *P. falciparum* genomes accumulate on average 0.9-1.5 SNPs per genome per year. We assume that this value is also valid for the other *Laverania* species.

Coalescent models

To estimate key population genetic parameters: effective population size (N_e), dates of divergence (D) and number of migrants per generation from source population to target population (M), we used the Generalised Phylogenetic Coalescent Sampler (G-PhoCS)[1]. As input, we used 1750 alignments from the Lav15sp dataset. We ran two models: (1) split between *P. falciparum* and *P. praefalciparum*, and (2) the entire *Laverania* tree. We incorporated phylogenetic information and modelled bi-directional migration between all extant and ancestral nodes. The MCMC chains were run for a minimum of 10 million iterations, with 20 chains run in parallel. The chains were merged and manually checked for convergence (Tracer version 1.5). We estimated N_e = *theta* / 2.*mu*, D = *g.tau* / *mu* and M = *m_st.tau*, where *theta*, *tau* and *m_st (migration source to target)* are model parameters, *mu* is the mutation rate per base pair per generation (ranges from 6.952×10^{-9} to 1.158×10^{-9} per base-pair per

generation, equivalent to 0.9 - 1.5 mutations per year per genome) and *g* is the generation time (0.18), as described above. The *M* parameter is estimated as the total migration rate, approximately indicating the probability that a given lineage in the source population will migrate into the target population[14]. This migration can be seen in some cases, see Supplementary Table 2, especially from *P. praefalciparum* into *P. falciparum*.

We applied the algorithms to three types of alignments (see Supplementary Table 2): (1) genic regions and (2) intergenic regions with and without assumed 500-bp untranslated regions. These alignments appeared robust for the *P. reichenowi*, *P. praefalciparum* and *P. falciparum* comparison as well as for *P. adleri* and *P. gaboni.* However, alignment of more distantly related species was not possible due to a high number of insertions and deletions and the low GC content. We performed the dating on genic alignments for all possible species for which we had more than 2 samples (thus excluding *P. blacklocki*). Importantly for the estimates of the Fig. 2A and Supplementary Table 2, we approximated some of the estimates of population genetic parameters where we were unable to generate intergenic alignments.

Adapting dating for *P. ovale* and *P. malariae-like speciation events*

Rutledge et al[15] used G-PhoCS to estimate the timing of the speciation of *P. ovale curtisi* and *P. ovale wallikeri* (both human-infecting) and between *P. malariae* (human-infecting) and *P. malariae-like* (chimpanzee-infecting). Rather than estimating the number of mitotic events per year, the number of mitotic events were based on the assumption that *P. reichenowi* and *P. falciparum* split from their common ancestor 3.0–5.5 million years ago. However, the study emphasised the importance of the relative timings of speciation events and, in particular, noted that *P. malariae* and *P. malariae-like* split at the same time as *P. reichenowi* from *P. falciparum*. Further, the study estimated that the *P. ovale* split occurred five times earlier than the *P. reichenowi*-*P. falciparum* split. This is the same factor that relates the *P. reichenowi*-*P. falciparum* split (190,000 years before present) to the split of Clade A and Clade B (around one million years before present), thus we conclude that these two speciation events occurred at approximately the same time.

Multiple Sequentially Markovian Coalescent

To estimate changes in effective population size (N_e) over time in *P. falciparum* (PfGA01 & PfIT, from Pf3K dataset), *P. praefalciparum* (PprfG02 & PprfG01) and the gene-flow between them, we ran the multiple sequentially Markovian coalescent (MSMC) on segregating sites from all

chromosomes[2]. Genome-wide SNPs were generated by firstly mapping raw reads from each sample against the Pf3D7 reference, then piping BAM files through mpileup v. 0.1.9 (parameters -q 20 -Q 20 -C 50) into bcftools call v. 1.1 (see MSMC documentation for more details). Retaining only homozygous SNPs, each *Plasmodium* chromosome was considered a single phased haplotype. MSMC was run for 20 iterations with a fixed recombination rate. Effective population size was calculated as (1/*lambda*)/2.*mu*, *scaled time* was converted into years as (*scaled time* / *mu*) x *g*. The parameters *lambda* and *scaled time* are derived from the model. Values for parameters *mu* and *g* are described above. The error around our estimates was estimated by bootstrapping 50 replicates by randomly resampling from the segregating sites used as input.

Estimation of population size

Effective population size was estimated from 10,000 years before present (BP) until 500 years BP as bootstrapping demonstrated that the model loses resolution for values of N_e more recently than 500 years BP. The effective population size of *P. falciparum* drops from at least 11,000 years BP and steadily declines to reach its lowest value around 6,000-4,000 years BP ($N_e \sim 3000$), before the population size begins to expand thereafter until 500 years BP (Fig. 2b). While others have speculated on the census population size of *P. falciparum* at this time[16] there is no straightforward way to relate N_e to census population size (N) due to complexities in the life-cycle of *P. falciparum* that causes the population to deviate from certain assumptions of the Wright-Fisher model[17]. Though generally the census number of parasites is much higher than *Ne*[18]. The bottleneck is unique to *P. falciparum*; although there is a large degree of error in the bootstrapping around N_e estimates for *P. praefalciparum*, the gorilla-infective species does not appear to go through a bottlenecks during this period. We replicated the analysis with different *P. falciparum* genomes (PfDd2 & PfHB3), which produced near-identical results.

Based on evidence from selection in the human genome, the origin of human malaria has been estimated as ~40,000–60,000 years BP and a population expansion associated with the origins of agriculture is assumed to have taken place ~4,000–6,000 years BP[19]. This scenario is confirmed by our modelling of the speciation event between *P. falciparum* and *P. praefalciparum* with G-PhoCS (40,000–60,000; 30,000–70,000 (95 % CI)) years BP and estimates of N_e of the MSMC through time and a rise in N_e from 4,000-6,000 years BP onwards.

Dating of *eba-175* dimorphism

To adapt the dating of the *eba-175* dimorphism[20], the following calculation was performed. Previous authors used 6 million years BP as the time when *P. reichenowi* split from the ancestor of *P. falciparum* and *P. praefalciparum* and then dated the *eba-175* split to 0.13–0.14 MYA. As those numbers can be scaled linearly, we used time of 0.13 - 0.23 MYA for the *P. reichenowi* split, which puts the data of the *eba-175* split to around 3,000-5,000 thousand years ago. This agrees with our observation that the dimorphism of *eba-175* occurred in *P. falciparum*, not *P. praefalciparum* (Supplementary Fig. 3b), concluding that the dimorphism occurred during the expansion of the *P. falciparum* and its host.

Supplementary Note 2: Evolution of core genes

Within-species polymorphism

The nucleotide diversity per CDS (π), the average number of nucleotide differences per site between two sequences, was calculated for each species and their means compared using *t* tests (Supplementary Fig. 1). Differences in the observed nucleotide diversity may reflect variation in prevalence and different demographic histories of the great ape parasites. The *P. falciparum* nucleotide diversity (computed from 5 worldwide isolates) was significantly lower than the nucleotide diversity observed in any other great ape species (calculated from 2 to 5 genotypes collected in the same localization from Gabon) ($p < 0.0001$). All Wilcoxon rank sum test results comparing nucleotide diversity between the parasites of great apes were significant ($W = 48193000$, $p < 0.0001$; Supplementary Fig. 1). The nucleotide diversity observed in gorilla-infecting species was higher than the diversity observed in the chimpanzee-infecting species ($W_{P.praefal.\text{-}P.\ adleri} = 4963900$, $p < 0.0001$). Among the gorilla-infecting species *P. praefalciparum* presented higher diversity than *P. adleri* ($W_{P.\ adleri.\text{-}P.\ praefal.} = 9540900$, $p < 0.0001$), due to a higher number of genes with relatively high values of nucleotide diversity. When considering only genes with a nucleotide diversity ≤ 0.02, the diversity was higher in *P. adleri* ($W_{P.\ adleri.\text{-}P.\ praefal.} = 9540900$, $p < 0.0001$). Regarding chimpanzee-infecting species, the diversity was significantly higher in *P. gaboni* ($W_{P.\ reichenowi\text{-}P.\ gaboni} = 5719500$, $p < 0.0001$) and lower in *P. billcollinsi* ($W_{P.reichenowi\text{-}P.\ billcollinsi} = 8011900$, $p < 0.0001$). The lowest diversity was observed in the least prevalent species, *P. billcollinsi* that infects chimpanzees.

Interspecific gene transfer

Most of the CDS topologies (4,251 out of 4,269, 99.6%, "Lav7sp" dataset) did not significantly differ from the *Laverania* species tree. For the remaining CDS (n=10, including 4 genes of chromosome 4, Supplementary Fig. 5), we specifically looked at their topology and identified those with possible events of gene transfer between species parasitizing the same host species. We detected a clustering of divergent species infecting the same host for six CDS, but none of them included all the species infecting the same host, *i.e. P. adleri*, *P. blacklocki* and *P. praefalciparum* or *P. gaboni*, *P. billcollinsi* and *P. reichenowi*. Four of them, localized in the same region of the chromosome 4, shared the same topology, with *P. praefalciparum* and *P. falciparum* grouping together with *P. adleri*, and corresponded to the previously reported introgressed genomic island (topology B in Supplementary Fig. 4; see main text). The last two cases involved a clustering of divergent species infecting the

chimpanzees (topology C in Supplementary Fig. 4) and was observed for two CDS, karyopherin beta and a conserved *Plasmodium* protein with unknown function. The intergenic region upstream the karyopherin beta orthologue in *P. billcollinsi* was not available; the topology of the intergenic region downstream corresponded to the species tree topology. Yet the tree topology based on the amino acid alignment did not cluster *P. gaboni* and *P. billcollinsi* together. All these signals remained when considering all sequenced genomes. Beyond these cases, most often, deviations of gene tree topologies from the species tree involved a clustering of *P. billcollinsi* and/or *P. blacklocki* closer to *P. adleri* and/or *P. gaboni* compared to the species phylogeny (Supplementary Fig. 4), or concerned alignments without enough resolution.

Genome-wide test of convergent evolution

We searched for an excess of convergent substitutions in specific branch-pairs by analyzing the correlation between the number of convergent and divergent substitutions between all the branch-pairs in a phylogeny, and looking for outlier branch-pairs that had high positive residuals, indicating an excess of convergent substitutions relative to the number of divergent substitutions[21]. Both for the divergent and convergent substitutions and for all pairwise comparisons, Pearson's correlation coefficients between the number of substitutions estimated under distinct evolutionary models were always higher than 0.99. We therefore only report the results obtained under the LG model of amino-acid substitutions. At a chromosome scale, we did not detect an excess of convergence between parasite species infecting the gorillas or between the parasites infecting the chimpanzees. However, we detected an excess of convergent substitutions relative to divergent substitutions, in three branch-pairs involving *P. blacklocki* but with no association with the host species.

Supplementary Note 3: Gene family analyses

Differences in gene families

The *P. reichenowi, P. gaboni* and *P. adleri* reference genomes are from single *Laverania* infections where a single isolate predominated (see Supplementary Table 1). However, the *P. praefalciparum* sample contained two distinct genotypes of *P. praefalciparum*. For the core region, a single haploid assembly could be resolved into the two genotypes. For more variable regions of the genome, like the subtelomeres, the genotypes could not be completely resolved and the numbers reported for the *rif*, *stevor* and *var* genes therefore contains contributions from both haplotypes. For *P. billcollinsi* and especially for *P. blacklocki*, we could not estimate the extent to which the subtelomeres assembled. Although gene families, like CLAG and the *var* genes from internal clusters, did assemble, the numbers of variable genes families are likely to be underestimated due to amplification biases introduced by the sWGA approach for *P. blacklocki* and the fact that *P. billcollinsi* is obtained from a co-infection with *P. gaboni*.

To estimate the number of genes we used (a) a regular expression to count the genes based on functional annotation and (b) matches to Pfam domains (E-value < 1e-6). To each gene/domain we associated counts and standard deviations (Supplementary Tables 5a, b). Differentially distributed gene families are reported in Fig. 3. For several genes, we performed phylogenetic analyses (Supplementary Fig. 7) to better understand their evolution. This was done by aligning the genes of a specific group with Muscle[22] using default parameters. In Seaview[23], we ran GLOCKS[24] with permissive settings and PhyML[25] (default settings for amino acids) to construct trees. The obtained trees were analysed in Figtree[26].

To perform the alignments of *msp* and *eba-175* dimorphic alleles, the same method was used but the numbers of sequences were reduced by subsampling to visualize dimorphisms.

Generation of similarity matrices

Where sequences were too divergent to perform tree based analyses, we implemented a visualization method based on similarity scores. First, amino acid sequences were compared with a BLASTp (e-value < 1e-6 and low complexity filter set to false). A similarity matrix based on the score or the global identity was built (the alignment length was normalized by the mean sequence length). Using the similarity matrix, the aligned sequences were clustered using the ward.D2 algorithm in the heatmap.2 module of gplots in R[27]. To each gene, we associated their species and in some case their

functional annotation through further heatmaps. We used this approach to analyse domains of *var* genes (see below, Supplementary Fig. 10).

The Rifin and Stevor proteins

To build a BLAST-based network, all Pir proteins were compared with an all-against-all BLASTp (parameter: -e 1e-6 –F F). We clustered the Pir proteins using Gephi[28] and tribeMCL[29] into groups, used in Fig. 3. For the Stevor proteins, we built a phylogenetic tree, using RAxML, with the PROTGAMMAIGT model and 100 bootstraps.

Meme-Motif analysis for Stevor proteins

To predict motifs in this family, we used MEME[30] version 4.9.1. We searched for 96 motifs of 8-15 amino acids using all of the Stevor proteins encoded by the seven reference genomes. Proteins with less than 5 hits were excluded. The output was parsed with a PERL script into a matrix and visualized in R[27], using the heatmap.2 function and the ward2 clustering (Supplementary Fig. 8).

var gene analysis

To analyse full-length *var* genes in the *Laverania* we excluded genes smaller than 2kb and called domains in the genes. The following domains were identified from there conceptual translations: ATS (Acidic Terminal Sequence), NTS (N-Terminal Sequence), DBL (duffy binding like), CIDR (cysteine-rich interdomain region), pam (placenta associated malaria) and the duffy-binding like domain as defined by Pfam (present in invasion related proteins). To call domains, the program hmmscan[31] was used with the HMMer models from the VARdom server using the following parameters: --domT 50 -E 1e-6 to attribute domains to *var* genes. As the domains are similar to each other, we generated a PERL program that ascribed domains based on best scores (at least 80% of the length of the HMMer domains). The regions of *var* genes encoding domains could overlap by up to 20 bp. In some cases, rather than finding one of the known domains (DBL, CIDR, ATS or NTS) the Pfam-defined duffy binding-like domain was found. If this happened, we named that domain Duffy, rather than Duffy Binding-Like. Regions (≥ 300aa) in the *var* genes not covered by known domains were also extracted and first called “Unclassified”. From those “Unclassified” domain a novel domain was found that we termed CIDRn because of the similarity to existing CIDR domains (Supplementary Fig. 9). To better understand the structure of the domains, particularly Duffy, we used a similarity matrix (Supplementary Fig. 10c).

It can be seen that some domains form defined groups, with little similarity to others, like CIDRα or ATS. Other domains share sequence (similarity), like DBLα with DBLβ. Interesting is the classification of the DBLε, DBLpam2/3 and the unclassified Duffy (dotted black lines, top of Supplementary Fig. 10c). Those domains seem to be most common in *P. praefalciparum*, *P. adleri* and *P. gaboni*. They have less similarity to other domains. Rather than representing a new domain (like a DBLx,[32]), we think that those domains might be more ancient.

We also classified the DBLx domain proposed by Larremore *et al*[32]. Their sequences that started with the amino acids specific for the DBLx domain (start NI or DF, end CPQNLDFDRRDQFLR) were compared to our domain dataset. Domains containing those sequences were labelled as DBLx in our set. Next, we generated a similarity matrix with those DBLx, DBLε and Duffy (Supplementary Fig. 12) The DBLx labeled sequences are clustered within the DBLε group. Therefore, we think that the DBLx is not a new domain, but rather part of the diverse DBLε group.

Supplementary Fig. 1. Maximum Likelihood tree and nucleotide diversity of *Laverania* isolates.
The tree was obtained using the sequences of 424 genes ("Lav25st" set of orthologues).

Supplementary Fig. 2. *In-vivo* mutation rate estimation.
The 600 kb conserved region of five clinical *P. falciparum* isolates (Supplementary Table 10), around the *Pf*CRT(*) locus. Large, nearly SNP-free, regions (black boxes) of around 200 kb are found. One *var* gene (**) in the internal cluster is on the opposite strand.

Supplementary Fig. 3. Dimorphisms in the *Laverania*.
(**a**) Examples of ancient dimorphisms based on maximum likelihood phylogenetic trees. Dimorphism in *msp1* arose in the *P. falciparum–P. praefalciparum* ancestor, after the divergence of *P. reichenowi* and dimorphism in *var1csa* evolved in the *P. reichenowi–P. praefalciparum–P. falciparum* ancestor after the divergence of *P. billcollinsi*. There is also evidence of a bi-allelic distribution of msp3 in *P. falciparum, P. praefalciparum* and *P. reichenowi*. (**b**) Dimorphism in *eba-175* is more recent. The alignment shown two mutually exclusive indels (arrow) in the *P. falciparum* sequences, not present in other *Laverania* species. The colours represent different nucleotides. For the *P. falciparum* sequences, we used full sequences from the Pf3K dataset. *Pf*, *P. falciparum; Pprf, P. praefalciparum; Pr, P. reichenowi*; and *Pbilc, P. billcollinsi*

Supplementary Fig. 4 Interspecific gene transfer and convergent evolution in the 3' end of the chromosome 4.
(**a**) Support for interspecific gene transfer between the gorilla-infecting species *P. adleri* and the common ancestor of *P. praefalciparum* and *P. falciparum*. The topologies observed in the coding and intergenic regions of the end of chromosome 4 are given. (**b**) Convergent evolution in the *rh5* gene. Amino acid alignment of the *rh5* region that carries the significant fixed difference between parasites infecting the chimpanzees and those infecting gorillas (red stars). The positions in the alignment of the 19 strains with *rh5* sequence available is given at the top, while the positions at the bottom of the alignments correspond to the position in the *P. falciparum* 3D7 sequence. Green circles indicate positions that are known to be involved in the interaction with the human receptor Basigin.

Supplementary Fig. 5: Tree topology tests.
The topologies confirmed by at least two genes that agree with the species tree (topology A) or that differ – when sufficient phylogenic information was available. The table summarizes all 18 genes that have a different topology from the species topology A.

Supplementary Fig. 6. Acyl-CoA Synthetase expansion on Chromosome 9
ACT view of five genomes, at the right-hand side of chromosome 9. The grey areas indicate co-linearity. *P. falciparum* has lost this region with four Acyl-coA synthetase genes, as this locus is conserved in the other species.

Supplementary Fig. 7. Phylogenetic analysis of multigene families.
Example of three families that show differences within the *Laverania*. (**a**) The putative glycophorin binding proteins form four distinct groups. One group contains sequences from all species. The remaining groups are clade, host or species sub-group specific. (**b**) Differences in the DBLmsp that are expanded in Clade A. The DBLmsp2 is a pseudogene (*) in *P. reichenowi*. (**c**) Expansion of *clag* genes in Clade A.

Supplementary Fig. 8. Analysis of Stevor proteins
(**a**) Length of the Stevor proteins for the seven *Laverania* genomes. (**b**) Occurrence matrix of meme motifs generated for Stevor proteins (Supplementary information). Columns represent the different meme motifs, rows represent all the 301 Stevor proteins. To each gene we associate (binary blue barcode) its length and the species where it is found. The matrix was clustered with the ward2 algorithm. Note that one cluster (*) is not present in chimpanzee parasites. *Pf, P. falciparum; Pprf, P. praefalciparum; Pr, P. reichenowi; Pbilc, P. billcollinsi; Pblac, P. blacklocki; Pgab, P. gaboni;* and *Padl, P. adleri*. (**c**) Maximum likelihood tree of the same data. Bootstrap values of 100 were obtained for all branches.

Supplementary Fig. 9. Phylogenetic Position of the new CIDR domain (CIDRn) specific of Clade A parasites among other domains.
Phylogenetic tree RAxML using the PROTGAMMAIGTR models of the new CIDR domain, together with the other CIDR domains. Bootstrap values of 100 were obtained on all branches.

Supplementary Fig. 10. Diversity of *var* genes domains.
(**a**) Average diversity over all domains apart from ATS between Clade A and Clade B (without *P. blacklocki*). The difference observed between Clade A and Clade B parasites is statistically significant (t-test). (**b**) Diversity observed for each domain in Clade A and B parasites (**c**) annotated similarity matrix between all the domains (> 220aa) of the *Laverania* as defined in Fig. 6, including their species attribution and their cluster attribution. The similarity matrix shows the score of the BLASTp between the domains, clustered with the ward2 algorithm in R. Each row and column represents therefore one domain and shows its score of similarity to the other 2,467 domains and itself. The occurrence of each domain/row across the species is indicated by the blue bars on the right. Although domains above the dotted line are classified differently, they cluster together. *Pf, P. falciparum; Pprf, P. praefalciparum; Pr, P. reichenowi; Pbilc, P. billcollinsi; Pblac, P. blacklocki; Pgab, P. gaboni;* and *Padl, P. adleri*.

Supplementary Fig. 11. Composition, structure and evolution of *var* genes within *Laverania.*
(a) Screenshot from ACT showing the 2nd internal cluster of *var* genes on chromosome 4 in the seven *Laverania* species. In Clade A and *P. blacklocki*, the orientation of the *var* genes is different compared to that of the other species. The GC-rich RUF elements[33] (RNA of Unknown function), highlighted with an R, occur less frequently in Clade A genomes. The size of the *var* genes between the species is different. *var* genes are in red, *pir* genes in green. (b) Bar plot of the number and orientation of *var* genes or pseudogenes, on the forward (blue) or reverse (red) strand,

within internal *var* gene clusters in the *Laverania*. Orientation is relative to the *P. falciparum* 3D7 reference genome.

Supplementary Fig. 12: DBLx is part of a DBLε domain.
To understand the diversity of DBLε and DBLn and to compare them to the newly described DBLx, a similarity matrix of all domains annotated as DBLε, Duffy, DBLpam1 and DBLx labelled sequences, see Supplementary Note 2. It can be seen that all domains are similar to each other and that the DBLx labeled sequences as defined by Larremore *et al*[32] cluster within the DBLε domains. *Pf, P. falciparum; Pprf, P. praefalciparum; Pr, P. reichenowi; Pbilc, P. billcollinsi; Pblac, P. blacklocki; Pgab, P. gaboni;* and *Padl, P. adleri*.

Supplementary Tables 1-8: see associated Excel file.

Supplementary Table 1. Multiple species infections in the *Laverania* samples
Samples were analysed for the presence of multiple-species infections by mapping reads to reference datasets, assembled *de novo* from Pacific Biosciences reads. Host contamination was identified by mapping against the human genome and removed prior to analysis. Information on each primate is shown.

Supplementary Table 2. Estimated dates of speciation (time to coalescence), effective population size (*Ne*) and migration rates for present and ancestral *Laverania* populations. Estimated dates of speciation (time to coalescence), effective population size (Ne) and migration rates for *Laverania* populations, including inferred ancestors. Values inferred using GPhoCS coalescence model. The results are presented from sequence alignments of coding (genic) regions as well as intergenic regions with and without UTR sequences. The coalescence model parameters have been scaled based on an assumption of 401–681 mitotic events per year (Supplementary Note 1). The number in parentheses is the range based on the prediction (95% confidence interval). Values with (*) are linear estimates, as good alignments could not be obtained between those species due to the low GC content. The migration rate refers to the probability of a parasite from the source lineage migrating into the target over the time period. The data from these estimates are presented in Fig. 1.

Supplementary Table 3. *P. falciparum* genome-wide signature of selection.
The branch-site test of positive selection and MacDonald Kreitman (MK) tests calculated for the "Lav3sp" set of orthologues

Supplementary Table 4. Convergent evolution analyses: genes with host-specific fixed differences.
(A) List of genes with fixed amino acid differences between *Laverania* species that infect chimpanzees and gorillas, using all the available *Laverania* strains*. The number of fixed differences observed includes positions with differences fixed in the two great ape hosts. The data "Lav15st" was used.
(B) Gene ontology terms enriched among genes with a significant signal in all strains (genes in bold in Supplementary Table 1a). The Benjamini-Hochberg false discovery rate has been used to correct for multiple tests. **P. blacklocki* is missing an orthologue.

Supplementary Table 5.
(a) Distribution of gene families based on their functional annotation.
Genes were counted based on their annotation term (until the first comma). Annotation terms occurring at least twice in one species are reported. In most cases, functional descriptions were ascribed automatically without manually reviewing the underlying gene structures; count values should therefore be regarded as approximate. For each gene family, the average number of genes and the standard deviation across species are calculated. Alternative estimates are provided in Supplementary Table 5b, where predicted functional domains are counted independent of annotation ascribed

directly to genes. (*) numbers are different from Fig. 3, as they were obtained through domain counting rather than counting annotation terms.

(b) Pfam domains frequency
The proteins of the 7 *Laverania* genomes were searched against Pfam and domains reported using an e-value ≤ 1e-6 as a cut-off. Counts refer to the number of proteins with a given Pfam domain. The cells are coloured by frequency from 0 (white) to 20+ (yellow).

Supplementary Table 6. Domain architecture of *var* genes.
For all *var* genes encoding at least four domains (excluding NTS), the first four domains are counted and grouped by species.

Supplementary Table 7. Summary of internal *var* gene clusters
Cells are coloured by the orientation of the majority of the *var* genes (green forward, orange reverse, white absent or ambiguous, yellow if just pseudogenes occur in cluster). The numbers reported (e.g. 4/2) correspond to 4 functional / 2 pseudogenes. RUF is the ncRNA with high GC content, potentially involved in expression of *var* genes [33]. The field "pseudo *var* end" means if the internal *var* genes cluster is ending with a *var* pseudo gene on the other strand which also clusters with Clade A ATS domains.

Supplementary Table 8. Accession numbers for sequence data
Accession numbers for raw Illumina and Pacific Biosciences (PacBio) data and genome assemblies. For mixed-species infections (analysis results are presented in Supplementary Table 1), data are identified by the major species in the sample. Pending indicates accession numbers that are currently being processed by the European Nucleotide Archive.

References

1 Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet* **43**, 1031-1034, doi:10.1038/ng.937 (2011).

2 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925, doi:10.1038/ng.3015 (2014).

3 Claessens, A. *et al.* 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 (2014).

4 Gerald, N., Mahajan, B. & Kumar, S. Mitosis in the human malaria parasite Plasmodium falciparum. *Eukaryotic cell* **10**, 474-482, doi:10.1128/ec.00314-10 (2011).

5 Ponnudurai, T. *et al.* Sporozoite load of mosquitoes infected with Plasmodium falciparum. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **83**, 67-70 (1989).

6 Mazier, D. *et al.* Complete development of hepatic stages of Plasmodium falciparum in vitro. *Science* **227**, 440-442 (1985).

7 Nkhoma, S. C. *et al.* Population genetic correlates of declining transmission in a human pathogen. *Molecular ecology* **22**, 273-285, doi:10.1111/mec.12099 (2013).

8 Molineux, L. in *Malaria, Principles and Practice of Malariology* Vol. 2 (ed McGregor IA Wernsdorfer WH) 913-998 (London Churchill, Livingston, 1998).

9 Bopp, S. E. *et al.* Mitotic evolution of Plasmodium falciparum shows a stable core genome but recombination in antigen families. *PLoS Genet* **9**, e1003293, doi:10.1371/journal.pgen.1003293 (2013).

10 Udeinya, I. J., Graves, P. M., Carter, R., Aikawa, M. & Miller, L. H. Plasmodium falciparum: effect of time in continuous culture on binding to human endothelial cells and amelanotic melanoma cells. *Experimental parasitology* **56**, 207-214 (1983).

11 Payne, D. Spread of chloroquine resistance in Plasmodium falciparum. *Parasitol Today* **3**, 241-246 (1987).

12 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

13 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:btp698 [pii]
10.1093/bioinformatics/btp698 (2010).

14 Freedman, A. H. *et al.* Genome sequencing highlights the dynamic early history of dogs. P*LoS Genet* 1**0,** e1004016, doi:10.1371/journal.pgen.1004016 (2014).

15 Rutledge, G. G. *et al.* Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. N*ature* 5**42,** 101-104, doi:10.1038/nature21038 (2017).

16 Volkman, S. K. *et al.* Recent origin of Plasmodium falciparum from a single progenitor. S*cience* 2**93,** 482-484, doi:10.1126/science.1059878 (2001).

17 Chang, H. H. *et al.* Malaria life cycle intensifies both natural selection and random genetic drift. P*roceedings of the National Academy of Sciences of the United States of America* 1**10,** 20129-20134, doi:10.1073/pnas.1319857110 (2013).

18 Palstra, F. P. & Fraser, D. J. Effective/census population size ratio estimation: a compendium and appraisal. E*cology and evolution* 2**,** 2357-2365, doi:10.1002/ece3.329 (2012).

19 Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious disease in human populations. N*ature reviews. Genetics* 1**5,** 379-393, doi:10.1038/nrg3734 (2014).

20 Yasukochi, Y., Naka, I., Patarapotikul, J., Hananantachai, H. & Ohashi, J. Genetic evidence for contribution of human dispersal to the genetic diversity of EBA-175 in Plasmodium falciparum. M*alar J* 1**4,** 293, doi:10.1186/s12936-015-0820-2 (2015).

21 Castoe, T. A. *et al.* Evidence for an ancient adaptive episode of convergent molecular evolution. P*roceedings of the National Academy of Sciences of the United States of America* 1**06,** 8986-8991, doi:10.1073/pnas.0900233106 (2009).

22 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. N*ucleic Acids Res* 3**2,** 1792-1797, doi:10.1093/nar/gkh340 (2004).

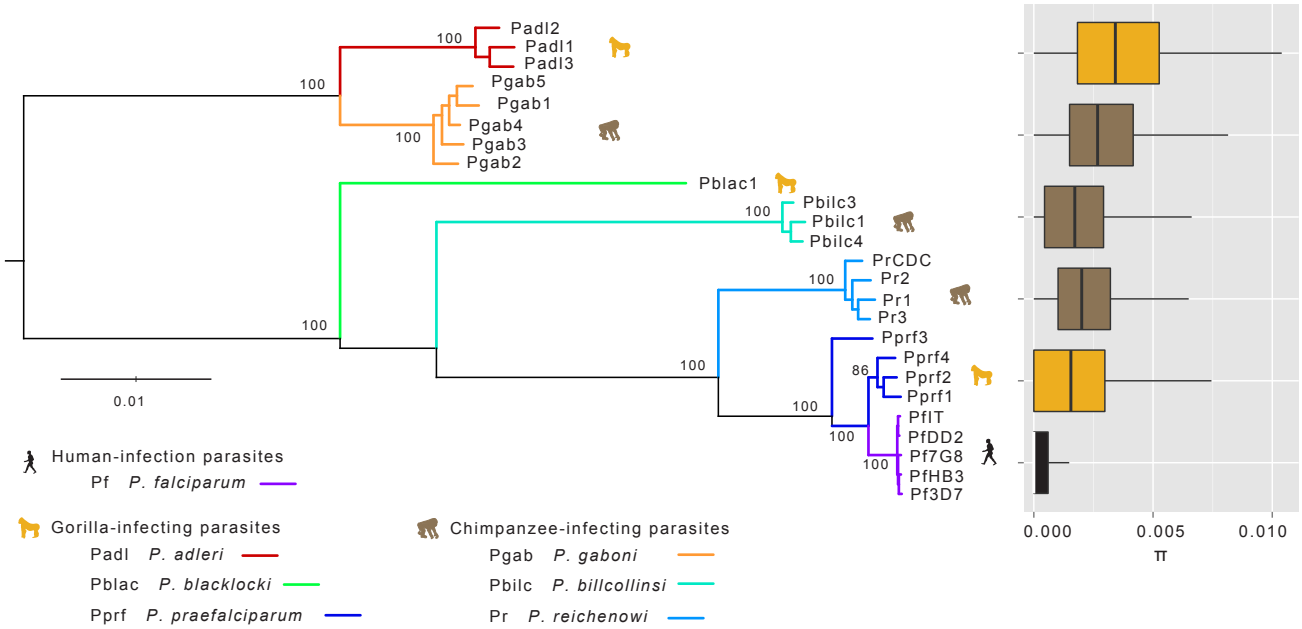
23 Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. M*olecular biology and evolution* 2**7,** 221-224, doi:10.1093/molbev/msp259 (2010).

24 Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. S*yst Biol* 5**6,** 564-577, doi:10.1080/10635150701472164 (2007).

25 Guindon, S., Delsuc, F., Dufayard, J. F. & Gascuel, O. Estimating maximum likelihood phylogenies with PhyML. M*ethods Mol Biol* 5**37,** 113-137, doi:10.1007/978-1-59745-251-9_6 (2009).
26 FigTree v.1.4.2, Available http://tree.bio.ed.ac.uk/software/figtree/ (2014).
27 Team, R. D. C. R*: A Language and Environment for Statistical Computing.* (2008).
28 Bastian, M., Heymann, S. & Jacomy, M. in I*nternational AAAI Conference on Weblogs and Social Media* (2009).
29 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. N*ucleic Acids Res* 3**0,** 1575-1584 (2002).
30 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. N*ucleic Acids Res* 3**7,** W202-208, doi:10.1093/nar/gkp335 (2009).
31 Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. B*MC Bioinformatics* 1**1,** 431, doi:10.1186/1471-2105-11-431 (2010).
32 Larremore, D. B. *et al.* Ape parasite origins of human malaria virulence genes. N*ature communications* 6**,** 8368, doi:10.1038/ncomms9368 (2015).
33 Guizetti, J., Barcons-Simon, A. & Scherf, A. Trans-acting GC-rich non-coding RNA at var expression site modulates gene counting in malaria parasite. N*ucleic Acids Res* 4**4,** 9710-9718, doi:10.1093/nar/gkw664 (2016).

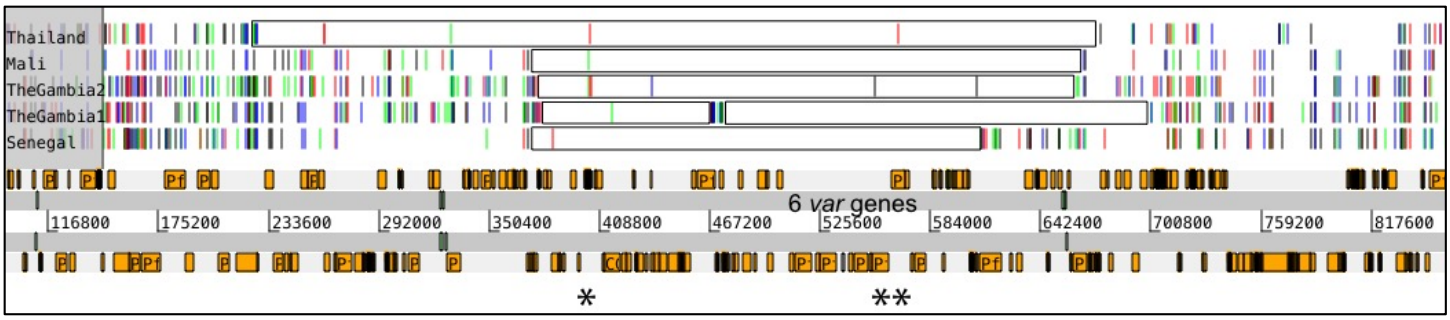
bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figure 1



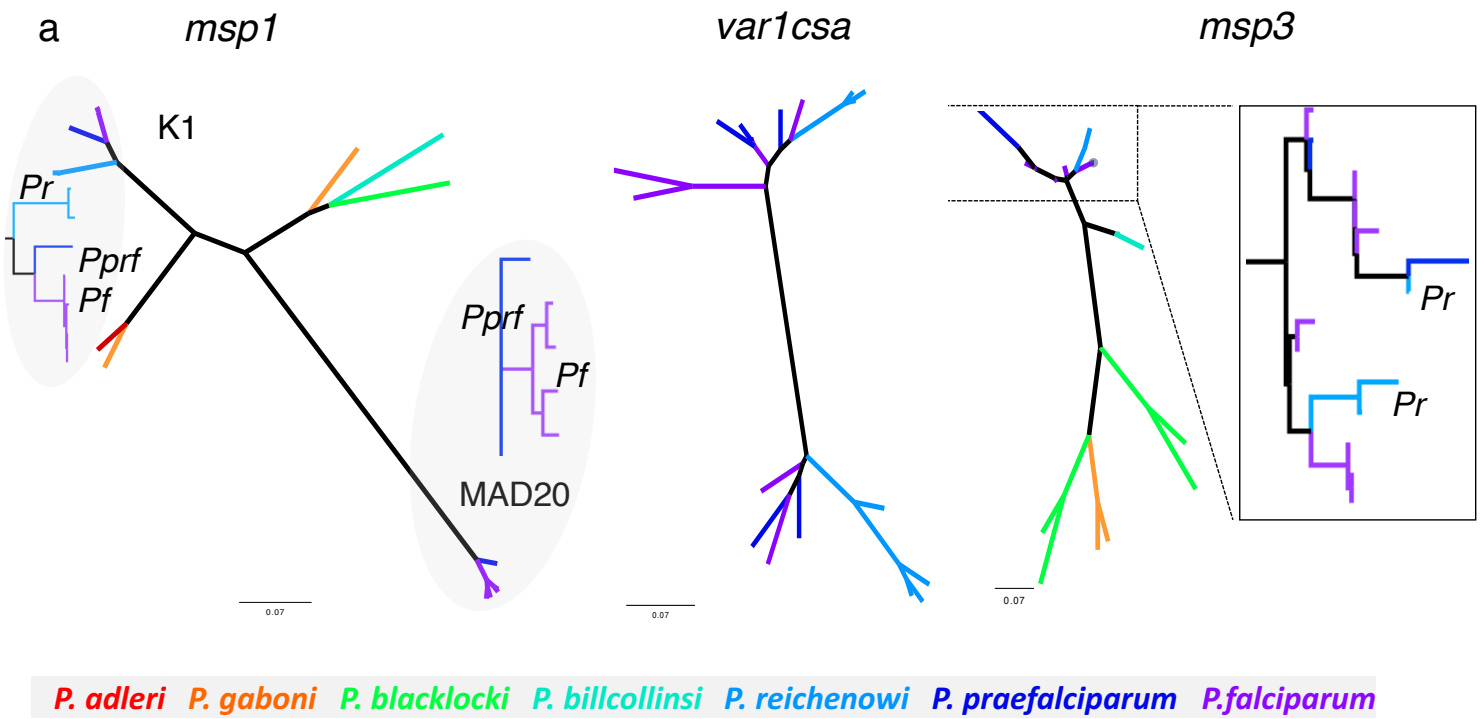
bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figure 2

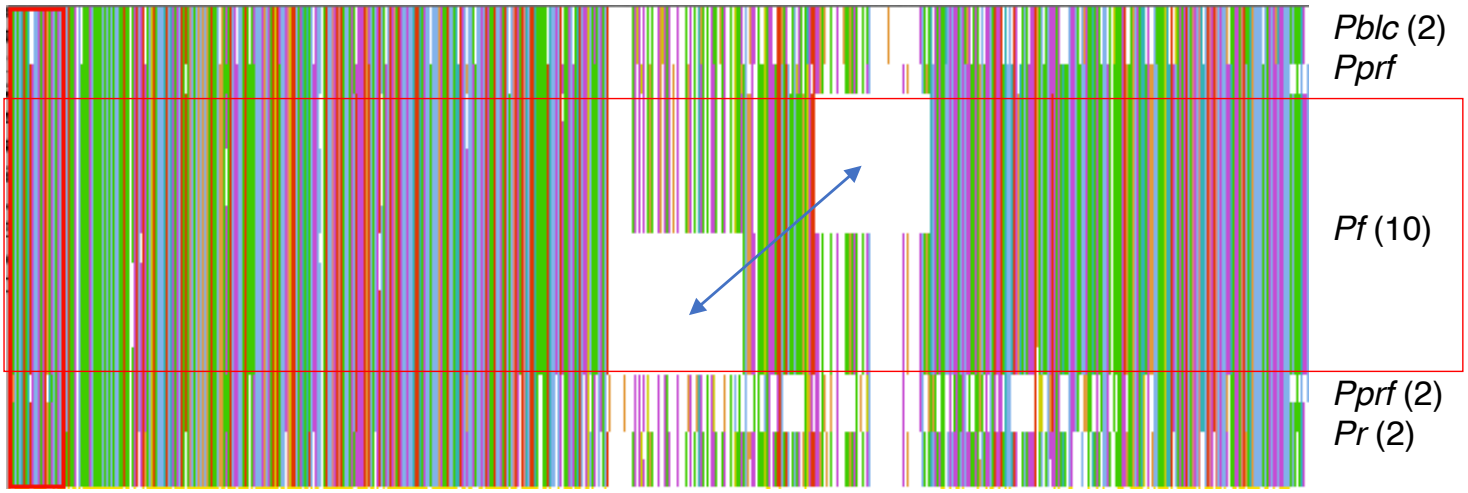


bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figure 3

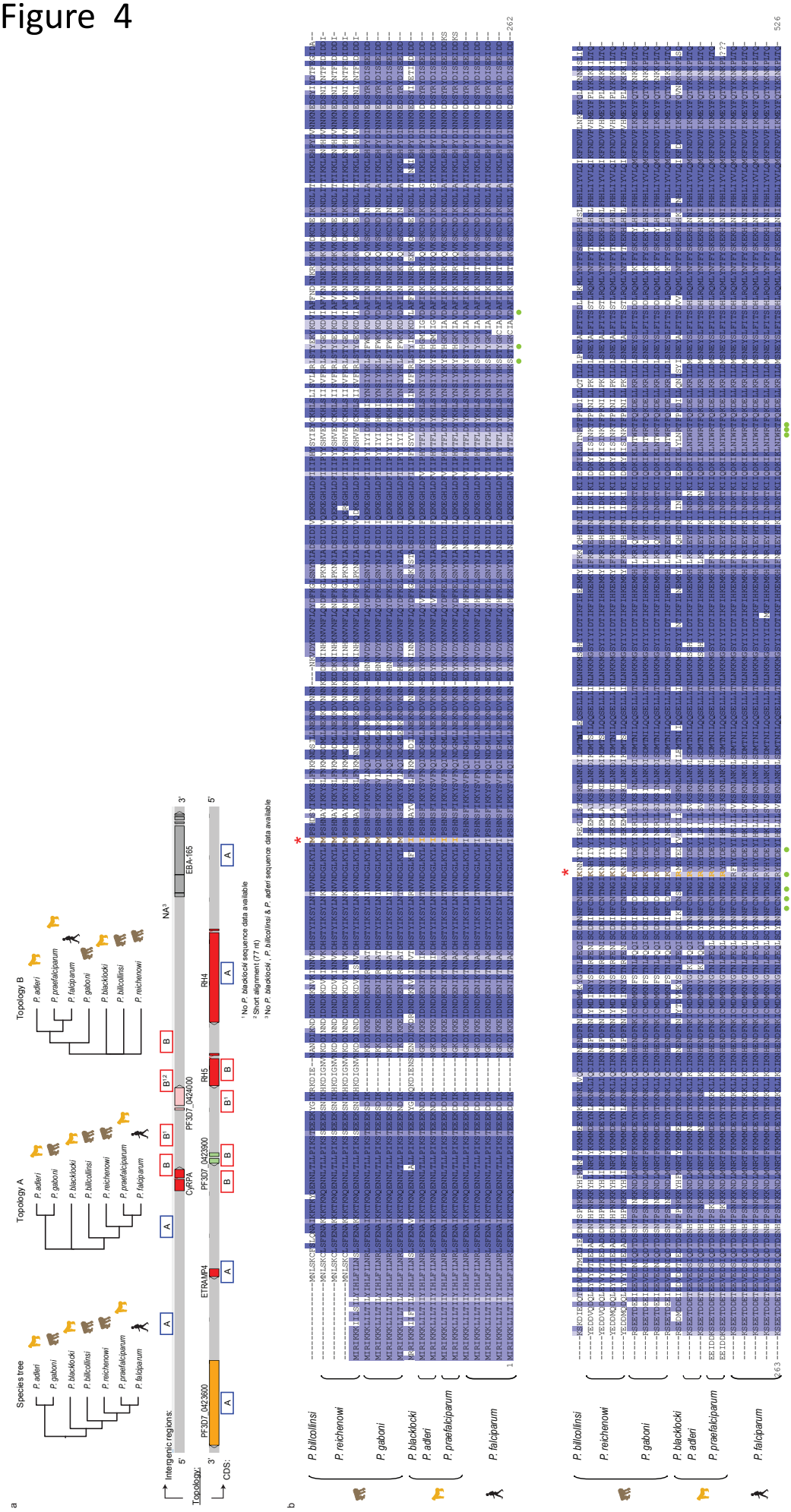


(b) *eba-175*



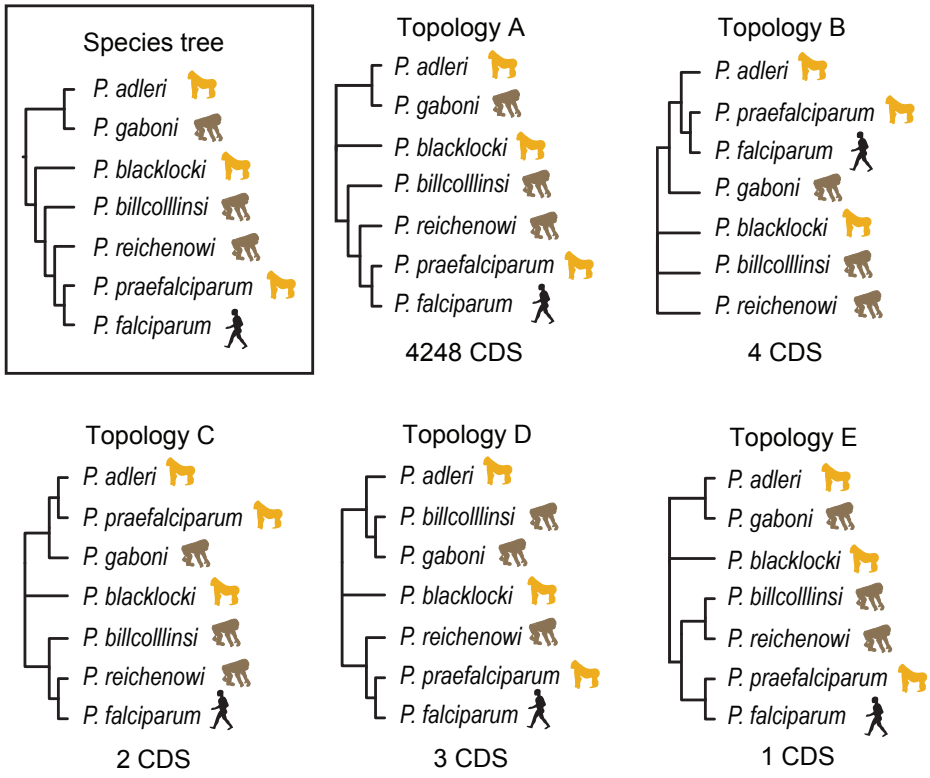
bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figure 4



bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

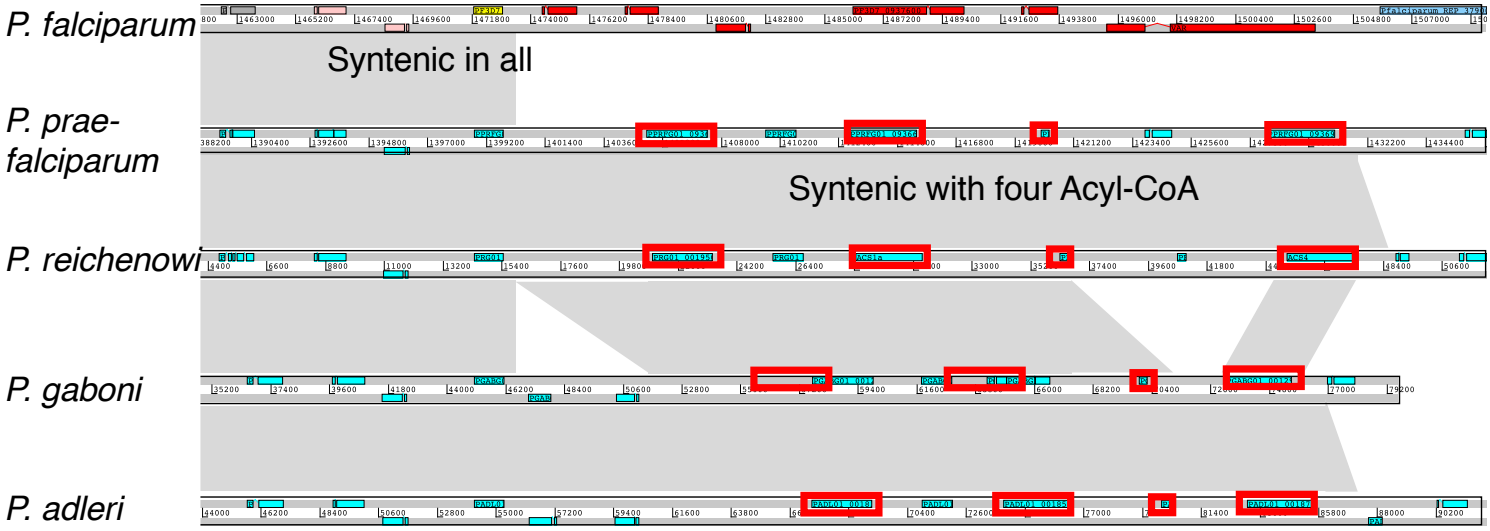
Supplementary Figure 5



Gene ID	Function	Gene Name	Topology see above
PF3D7_0423800	cysteine-rich protective antigen // RH5-Ripr membrane anchoring protein	CyRPA	B
PF3D7_0423900	probable protein, unknown function		B
PF3D7_0424000	Plasmodium exported protein (PHISTc), unknownfunction		B
PF3D7_0424100	reticulocyte binding protein homologue 5	RH5	B
PF3D7_0524000	karyopherin beta	KASbeta	C
PF3D7_0613200	conserved Plasmodium protein, unknown function		C
PF3D7_0902600.1	serine/threonine protein kinase, FIKK family	FIKK9.7	E
PF3D7_1135600	condensin-2 complex subunit D3, putative		D
PF3D7_1327500	conserved Plasmodium protein, unknown function		D
PF3D7_1328000	conserved Plasmodium protein, unknown function		D

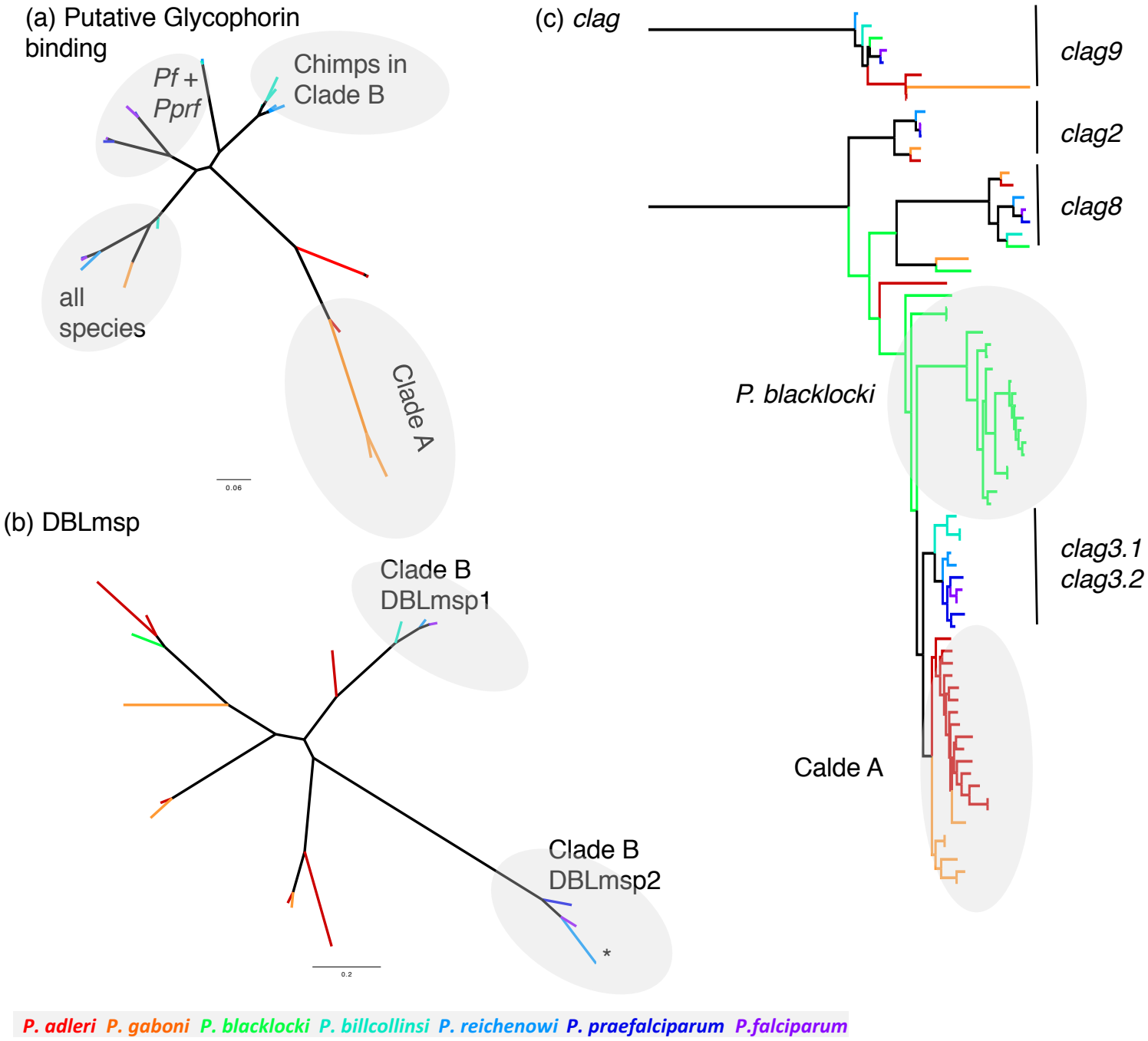
bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figure 6



bioRxiv preprint doi: https://doi.org/10.1101/095679; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

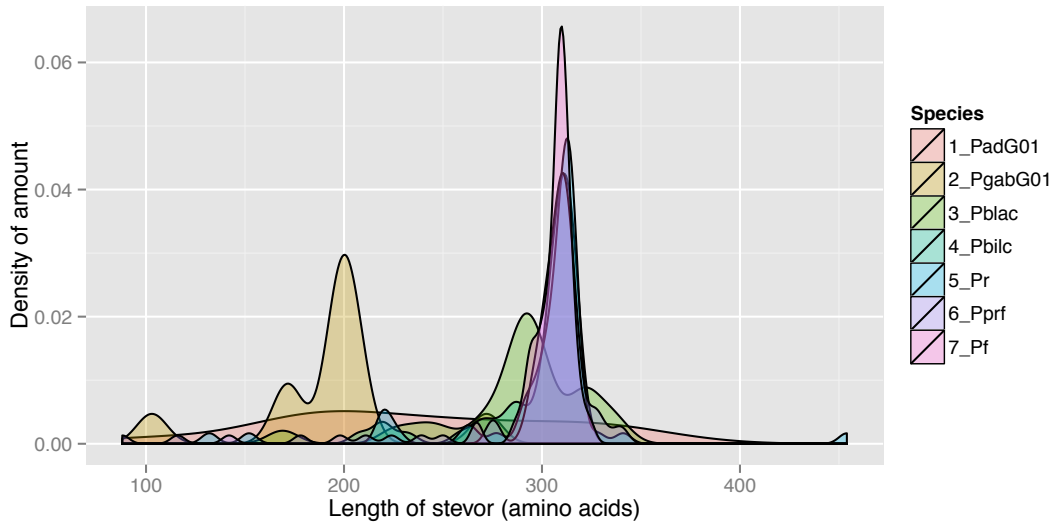
Supplementary Figure 7



P. adleri *P. gaboni* *P. blacklocki* *P. billcollinsi* *P. reichenowi* *P. praefalciparum* *P.falciparum*

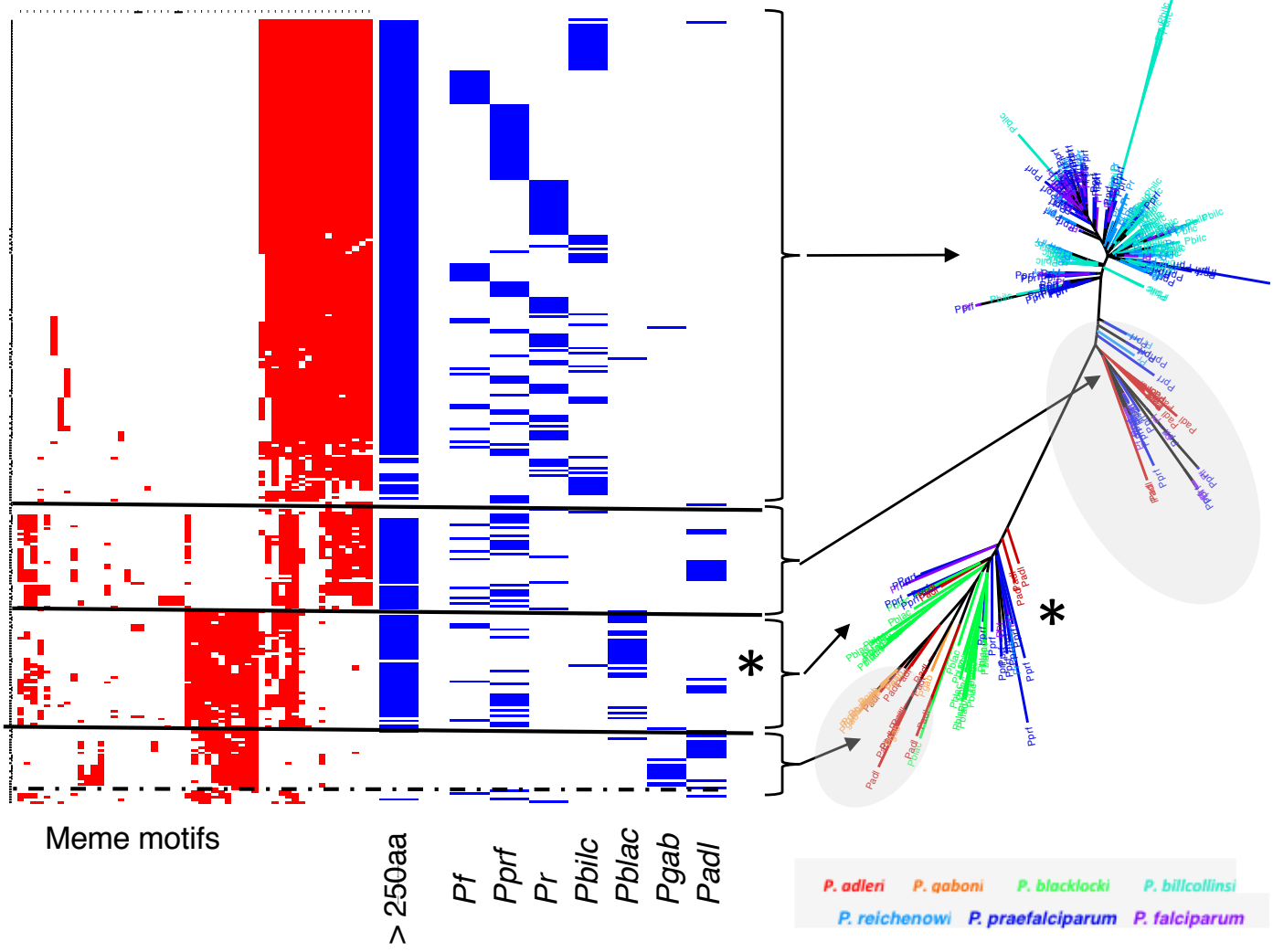
Supplementary Figure 8

a

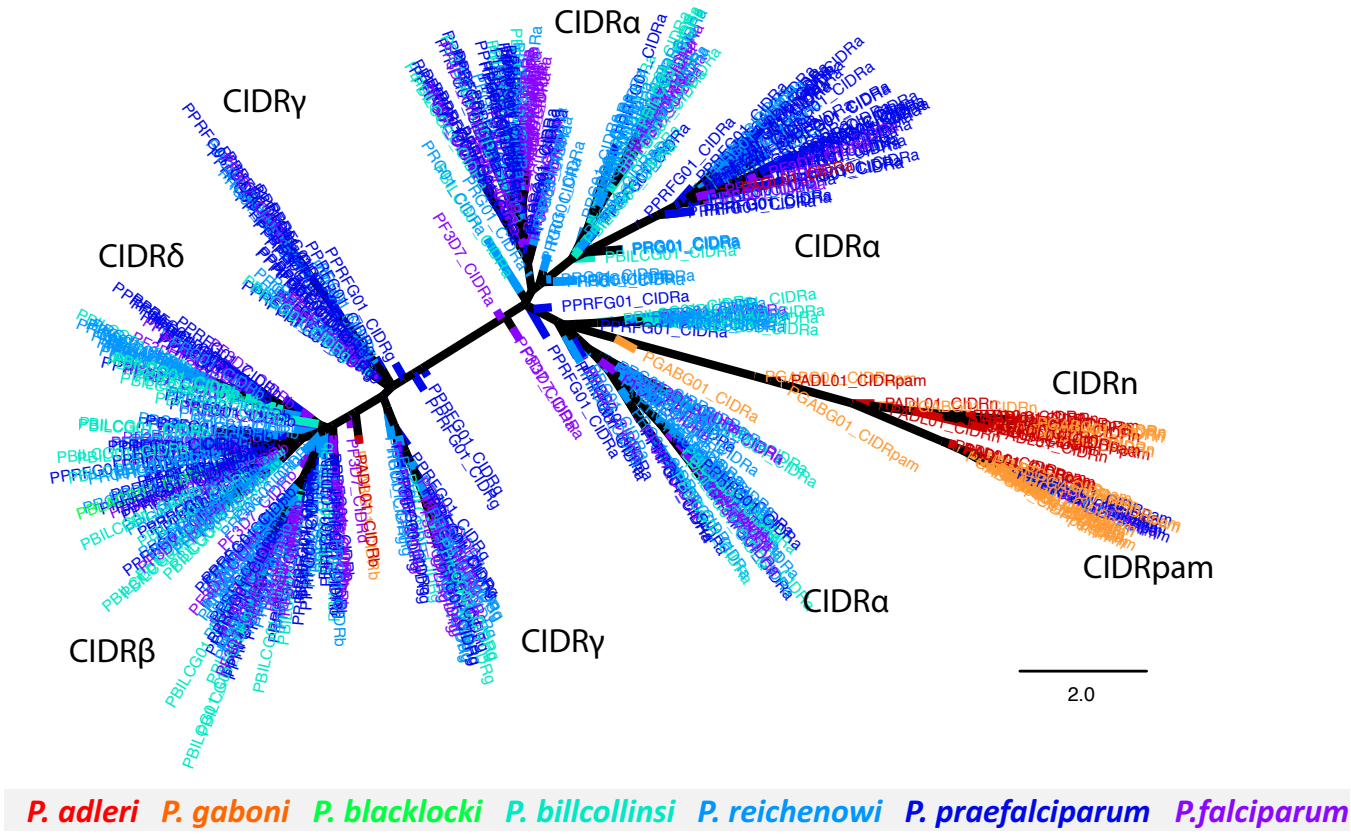


b

c

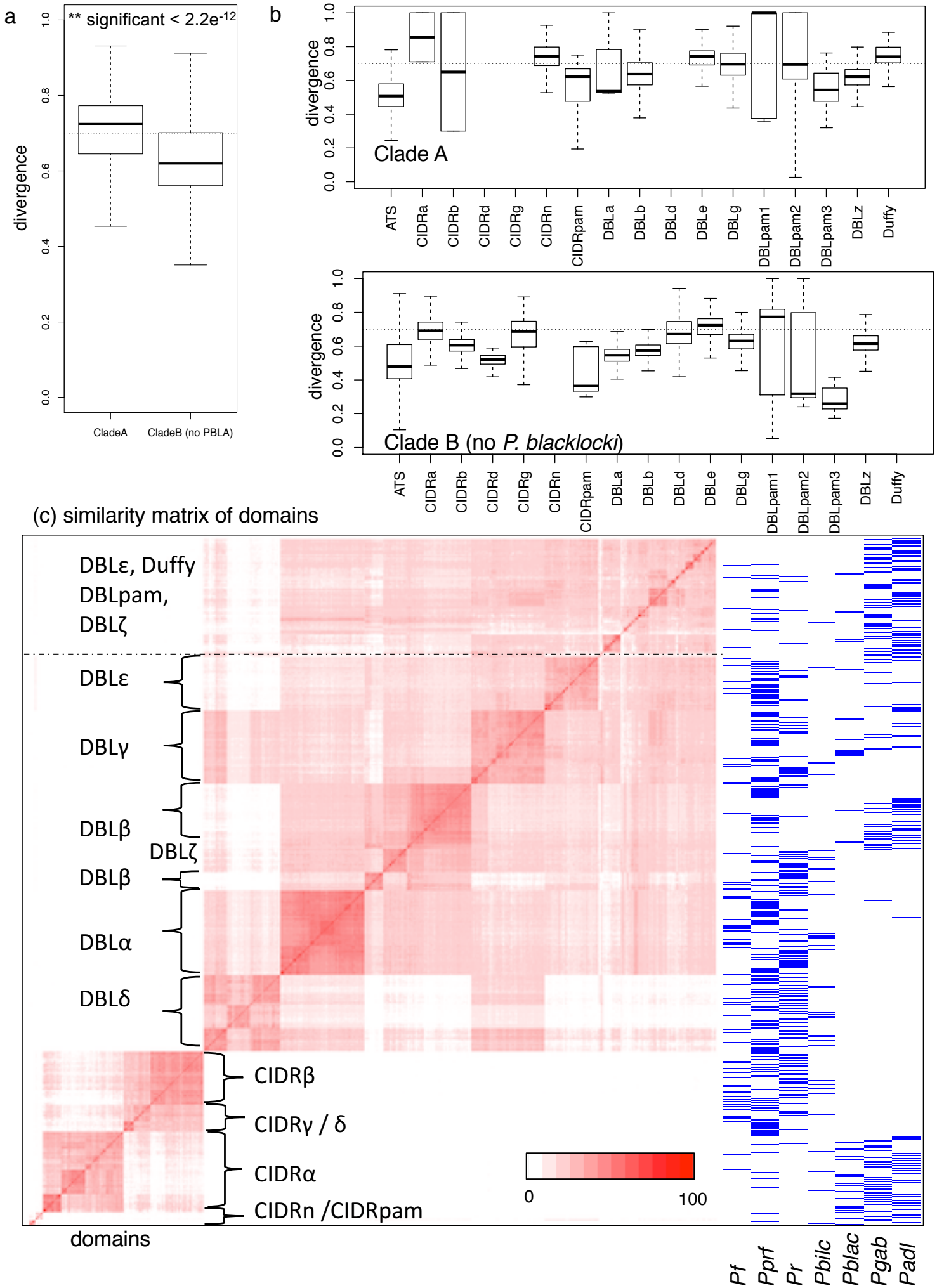


Supplementary Figure 9

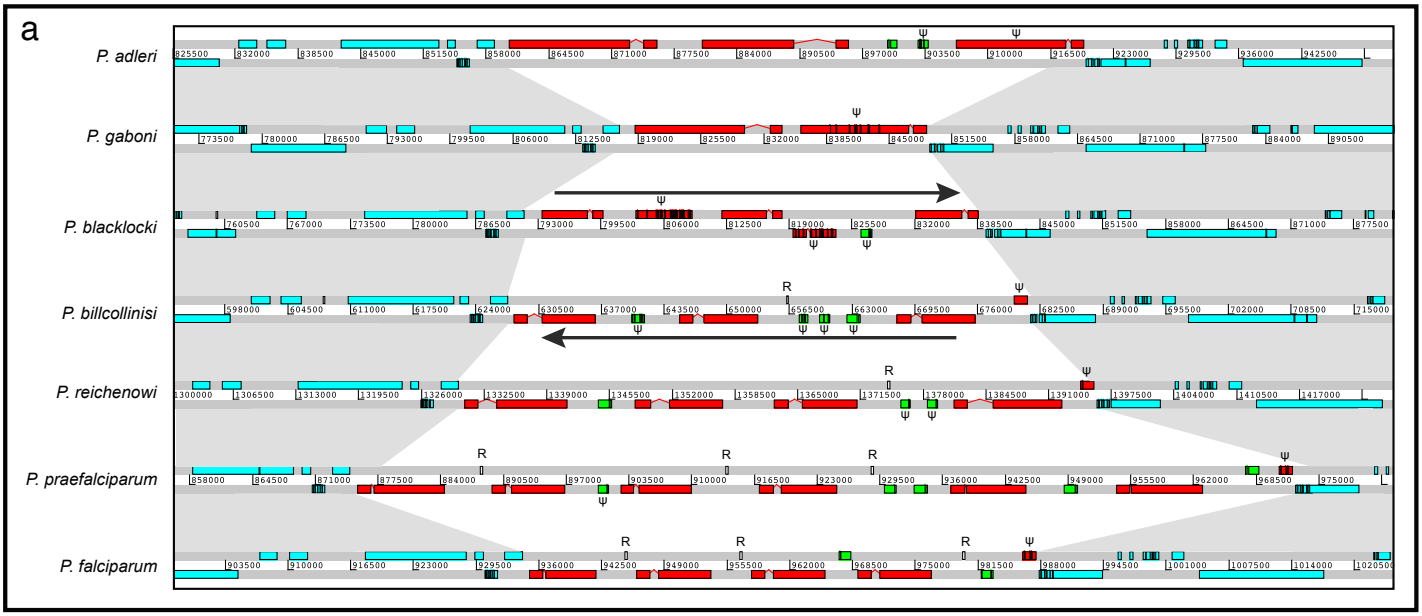


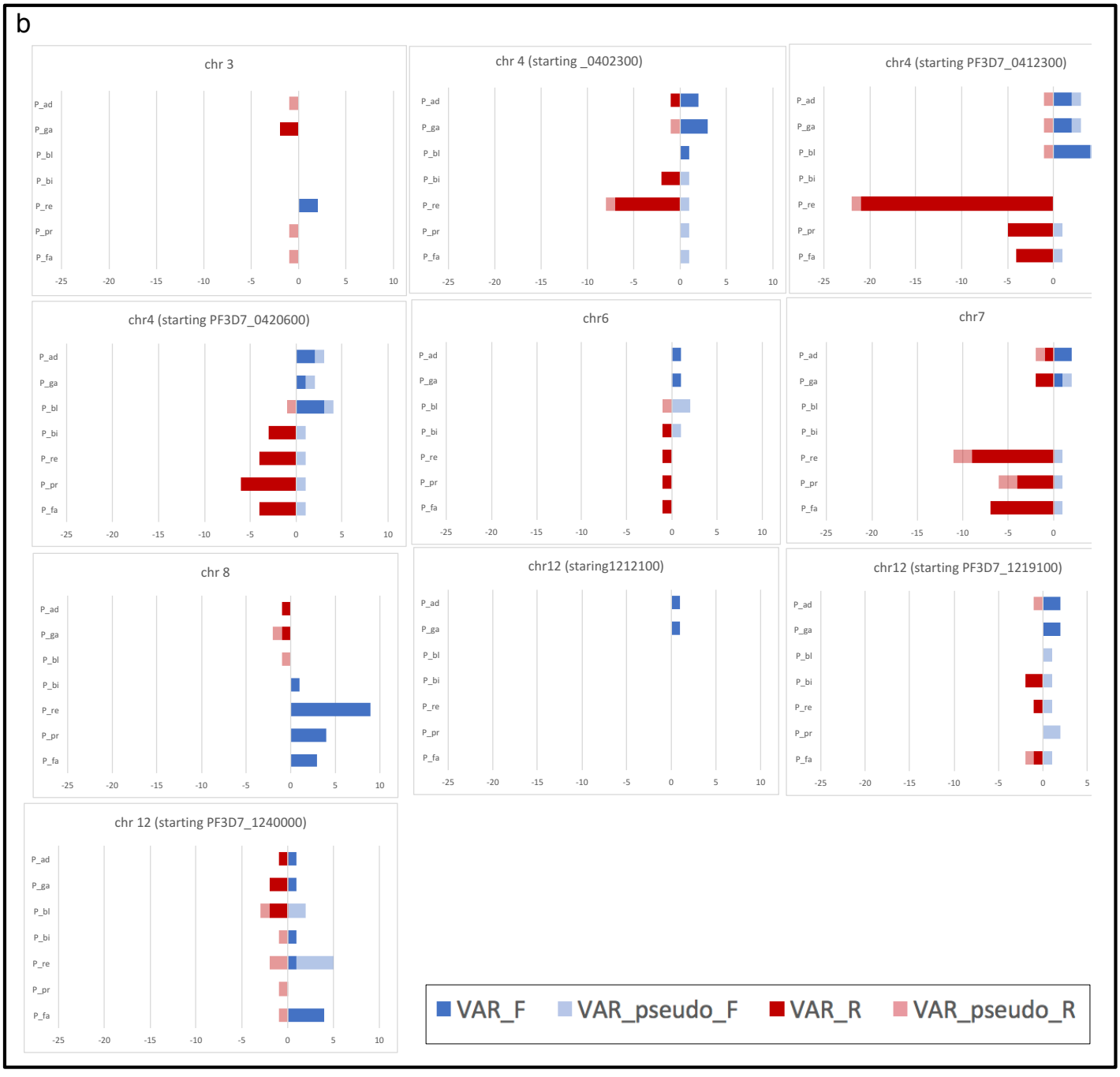
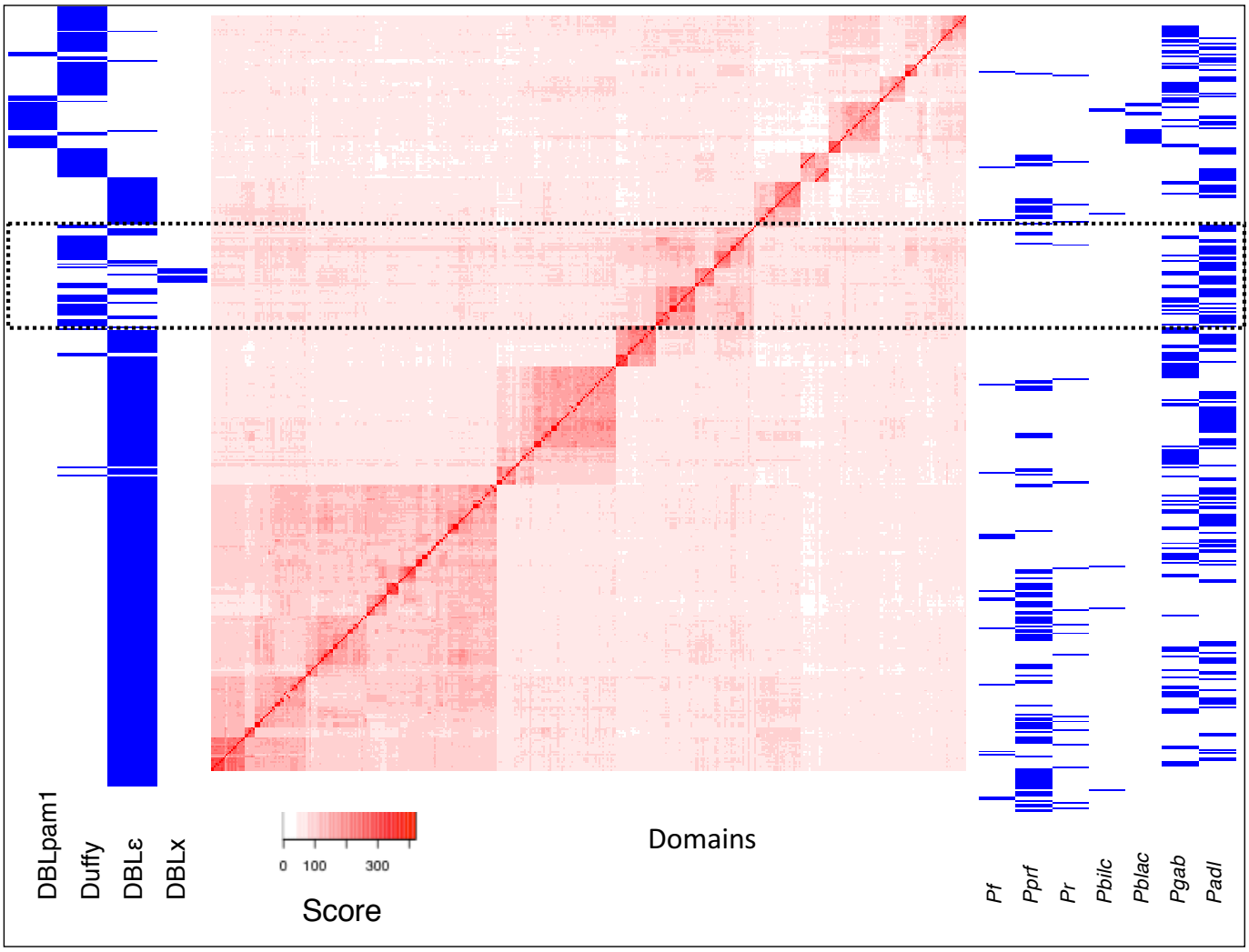
P. adleri *P. gaboni* *P. blacklocki* *P. billcollinsi* *P. reichenowi* *P. praefalciparum* *P.falciparum*

Supplementary Figure 10



Supplementary Figure 11





Supplementary Figure 12