Genomes of all known members of a *Plasmodium* subgenus reveal paths to virulent human malaria

Thomas D. Otto[1,†,#,*], **Aude Gilabert**[2,†], **Thomas Crellen**[1,3], **Ulrike Böhme**[1], **Céline Arnathau**[2], **Mandy Sanders**[1], **Samuel O. Oyola**[1,4], **Alain Prince Okouga**[5], **Larson Boundenga**[5], **Eric Willaume**[6], **Barthélémy Ngoubangoye**[5], **Nancy Diamella Moukodoum**[5], **Christophe Paupy**[2], **Patrick Durand**[2], **Virginie Rougeron**[2,5], **Benjamin Ollomo**[5], **François Renaud**[2], **Chris Newbold**[1,7], **Matthew Berriman**[1,*], and **Franck Prugnolle**[2,5,*]

[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

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

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

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

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

[*]Correspondence to: Thomas D. Otto (thomasdan.otto@glasgow.ac.uk), Matthew Berriman (mb4@sanger.ac.uk) or Franck Prugnolle (franck.prugnolle@ird.fr).
[4]International Livestock Research Institute, Box 30709, Nairobi, Kenya (current address)
[#]Current Address: Centre of Immunobiology, Institute of Infection, Immunity & Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom
[†]These authors contributed equally.

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 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.

The evolutionary history of *Plasmodium falciparum*, the most common and deadly 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, known as the *Laverania* subgenus[2]. Until 2009, chimpanzee-infecting *P. reichenowi* was the only other species known in this subgenus, 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 worldwide[2].

Over time there have been various estimates 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, the latter falsely based on the date of the chimpanzee–human split[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 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 an animal model has precluded obtaining sufficient DNA for full genome sequencing and has hindered investigation of the *Laverania.* So far the full draft genome of *P. reichenowi*[9] and a nearly complete draft sequence of *P. gaboni*[6] are available. These data together with additional PCR based approaches[10] have provided important insights into the evolution of this subgenus, including the lateral gene transfer of the *rh5* locus, the early expansion of the FIKK gene family and the observation that the common ancestor also had *var* genes. Our data confirm and significantly extend these findings. 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 all known members of the *Laverania* subgenus and to address the question of why *P. falciparum* is the only extant species to have adapted successfully to humans, we have sequenced multiple genotypes of all known *Laverania* species.

Genome sequencing from six *Laverania* species

Fifteen blood samples that were positive for ape malaria parasites by PCR 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). Despite low parasitemia, a combination of host DNA depletion, parasite cell sorting and amplification methods enabled sufficient parasite DNA templates to be obtained for short- (Illumina) and long- read (Pacific Bioscience) sequencing (Supplementary 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 (Supplementary 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 compared to those obtained previously is illustrated by the good representation of multi-gene families (Supplementary Table 2) and the large number of one-to-one orthologues obtained between the different reference genomes (4,350 among the seven species and 4,826 between *P. falciparum*, *P. praefalciparum* and *P. reichenowi*). Two to four additional genomes were obtained for each species except for *P. blacklocki* (Supplementary Table 1).

Speciation history in the *Laverania* sub-genus

Conservation of synteny is striking between these complete genomes and enabled us to reconstruct 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 due to the lack of both complete genome data and accurate estimates of mutation rate and generation time. Using the most divergent estimates of generation time and measured mutation rates from the *P. falciparum* literature, we found the data converge to 0.9-1.5 mutations per year per genome (Supplementary Note 1). We observed a similar substitution rate *in vivo* by examining existing sequence data for five 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). The fact that these two figures are similar suggests that the *in vitro* mutation rate may have been underestimated since many mutations will be lost by genetic drift. Since no data is available for the other species we have assumed hereon that these values generalise across the subgenus. From 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). Our range of values is far more recent than previous estimates[3,11]. 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, Table 2). Based on coalescence estimates, *P. falciparum* begun to emerge 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[12]. Our analysis also indicates significant gene flow between these two parasite species after their initial divergence (Supplementary Table 3).

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 years ago, after agriculture was established[6,8]. 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*[13]). 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 *P. gaboni* sequence data, provided some evidence that different dimorphic loci diverged at different points in the tree[14]. 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/K115) 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[16] 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 recalibrated 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[17] – an EBA-175 binding protein. 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 a new vertebrate host (human) and new vector species (e.g. *Anopheles gambiae*)[18]. To infer *P. falciparum* specific adaptive changes, we considered the *P. falciparum* / *P. praefalciparum* / *P. reichenowi* genome trio and then applied two lineage based tests to find positive selection that occurred in the *P. falciparum* branch (see methods). The two tests identified 172 genes (out of 4,826) with signatures of positive selection in the human parasite species only (Supplementary Table 4).

Two genes (*rop14* and PF3D7_0609900) were significant in both tests. Among the 172 genes, almost half (n=82) encoded proteins of unknown function. Analysis of those with functional annotation indicated that genes involved in pathogenesis, entry into host, 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; *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 4).

Evolution through introgression, gene transfer and convergence

Frequent mixed species infections in apes and mosquitoes[18] 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*). Because such events preserve the phylogenetic history of the genes involved, 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 one-to-one orthologues. Beyond the region that includes *rh5* (Fig 2b, Supplementary Fig. 4a), few signals of gene flow between parasites infecting the same host species were obtained (n=11) suggesting that these events were rare or usually strongly deleterious (Supplementary Fig. 5).

The *Laverania* subgenus evolved to infect chimpanzees and gorillas 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 chimpanzee-infective parasites, 49 in gorilla-infective and 12 with fixed traits in both host species (Fig. 2; Supplementary Table 5a). For at least 67 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 host invasion and pathogenesis (Supplementary Table 5b) 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[19]. We notice that one of the convergent sites is known to be a binding site for the host receptor Basigin[20] (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*[21], it is not a pseudogene 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 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 implicated in *Plasmodium* invasion of erythrocytes (doc222) also displayed signals of convergent evolution. Twelve parasite coding sequences had fixed differences at the same

amino-acid position in chimpanzees and gorillas. Of these P230 was the only one found with a position that was different and fixed across all three host species. P230 is involved in gamete development and trans-specific reproductive barriers23, possibly through enabling male gametes to bind to erythrocytes prior to exflagellation24. 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 subtelomeric gene families of the *Laverania* have come from *P. reichenowi* and *P. falciparum*. These important families are well represented in our assemblies (Supplementary Tables 2 and 6A) and we provide a comprehensive picture of their evolution.

Most gene families were likely present in the ancestor of all Laverania. The same general pattern of one-to-one orthology throughout the subgenus indicates that many underwent gene duplication early (e.g. FIKK) or prior (e.g. ETRAMP, PHIST and SURFIN) to the development of a distinct Laverania lineage. Only a subset displayed contractions or expansions between specific Laverania species (Fig. 3 and Supplementary Table 6a, Supplementary Figure 7). 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 *P. praefalciparum* and *P. falciparum* (and even more in *P. falciparum* for *hyp4* and *hyp5*). Since all four are components of Maurer's clefts, an organelle involved in protein export25, 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*26 is in fact expanded across the *Laverania* and has 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 adhesion27 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 (Fig 4 and Supplementary Fig. 8). Based on full-length alignments, there is a deep phylogenetic split between *stevor* genes but when only short conserved protein motifs are considered, a group of Stevor proteins (*stevor II,* Fig 4a) forms a cluster comprising almost entirely of members from gorilla-infecting species. Since *stevor* genes are known to be

involved in host–parasite interactions (such as binding to host glycophorin C in *P. falciparum*28), 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*29. 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 Regions (CIDR) and exon 2 encodes Acidic Terminal Sequence (ATS)30. Similar to *P. falciparum*, our data are consistent with all *Laverania* species having *var* genes (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 *var* evolution 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, Supplementary Table 7). In 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 5a, Supplementary Fig. 9) and have lower sequence diversity per domain but cluster into more sub-groups than Clade B domains (Fig 5b, 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 8). Clade B parasites also show a much greater number of associated GC-rich RNAs of unknown Function (RUF) elements than Clade A (Supplementary Table 8).

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 5c). 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* isolates31, involved both in cytoadherence in the placenta in primigravidae, and proposed to be a central intermediate in *var* gene switching during antigenic variation32. 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. 5b, 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.

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 known 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. 6.

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[33] and existed as small isolated populations[12]. 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[34]. 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, for the genes that we have implicated in this process, existing data (www.genedb.org, plasmodb.org) suggest they are expressed throughout the life cycle but that only half have been characterised. This opens up new opportunities for future studies on host specificity and host adaptation in *Plasmodium*.

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 35. Remaining blood was subsequently used for DNA extraction and detection of *Plasmodium* infections as described in Ollomo et al3. Overall, 15 blood samples from 7 chimpanzees and 4 gorillas were found to contain the *Laverania* samples used in the present study (Supplementary Table 1).

Ethical consideration

The animal well-being was guaranteed by the veterinarians of the "Parc of la Lékédi" and the CIRMF who proceeded to the sanitary controls and the blood sampling. Because these blood samples were collected as part of the standard protocol for the sanitary controls (and not specifically for our experiment), our study did not need the approval of an Institutional Animal Care or Use Committee. Note also that our study did not involve randomisation nor blinding.

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 protocol36 to enrich genomic DNA. The genome of *P. blacklocki* was generated using selective WGA (sWGA) as indicated in37 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 approach38.

Sample sequencing

All samples were first sequenced with Illumina. Amplification-free Illumina libraries of 400-600 bp were prepared from the enriched genomic DNA39 and run on MiSeq and HiSeq 2000 (v3 chemistry) Illumina machines.

After the Illumina sequencing, six samples with a combination of the least number of multiple infections (see below) and the lowest level of host contamination were chosen for long read sequencing, using Pacific 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 and so allow samples to be selected for PacBio sequencing from those comprising a low number of species, 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-Matic40 (parameter chop=5) to position reads only where they aligned perfectly. SNP-o-Matic returns all the positions of repetitive mapping reads. This output allowed us to count the read depth of these two genes across all species and therefore determine the number and relative amount of different malaria species per sample.

Whole genome amplification (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 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.

Long reads (Pacific Bioscience) assemblies—Six reference genomes were assembled using HGAP41, 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 ABACAS42 against a *P. falciparum* 3D7 reference that has no subtelomeric regions. Assembly errors and WGA artefacts were manually corrected using ACT43. After this step, three iterations of ICORN244 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 gaps45.

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 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 PrCDC9 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 (PrG01; Supplementary Table 2) appears to be of higher quality than the existing *P. reichenowi* PrCDC 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 a combined reference comprising *P. gaboni* (PgabG01) and the Pf3D7 (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, we used iCORN to select the dominant genotype at each position. Where it was not possible to phase the genotypes, due to a lack of variation, we assumed that they were identical. In the subtelomeres however, it was possible to distinguish but not phase the two *P. praefalciparum* genotypes resulting in approximately twice the number of *var* genes as seen in *P. falciparum*. 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 eliminate the remote possibility that any of the contigs in the *P. praefalciparum* assembly were in fact derived from *P. falciparum* co-infection.

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 in46. In short, the annotation of *P. falciparum* (version July 2015) was transferred with RATT47 and new gene models were called with Augustus48. Obvious structural errors in core genes were manually corrected in Artemis49.

Mapping - generation of further samples

To generate the gene sequence for different samples, Illumina reads were mapped against a set of reference genomes using BWA50 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 UnifiedGenotyper51 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-1.252). 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.

Orthologous group determination and alignment

Orthologous groups were identified using OrthoMCL v1.453 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 7G89).

(3) "Lav7sp" set of 4,350 orthologues from across the seven *Laverania* reference genomes

(4) "Lav15st" set of 3,808 orthologues, with at least two representative sequences per species, excluding *P. blacklocki* and the most divergent *P. praefalciparum* lineage Pprf3.

(5) "Lav3sp" set of 4,826 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. 54), codon-based multiple alignments were performed using PRANK55,56 with the -codon and +F options, as it was shown to outperform other programs in the context of the detection of positive selection57,58. Prior to aligning codons, low complexity regions were excluded in the nucleotide sequences using dustmasker59 and in amino acid sequences using segmasker60 from NCBI-BLAST. Poorly aligned regions were excluded using Gblocks61, with default settings.

Analysis of interspecific gene flow, introgression or gene transfer

Species-tree inference—Two ML trees were performed using RAxMLv8.1.2062 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 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 test63) 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 4,350 "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[64,65]. 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[64]. 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 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[66]. The program Codeml from PAML v4.8a[66] 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 the simulated dataset of 1000 sequences that showed at least the same number of fixed differences as observed in the real data.

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 67 implemented in PAML v4.4c[66]. This test detects sites that have undergone positive selection in a specific branch of the phylogenetic tree (foreground branch). The "Lav3sp" set of 4,826 orthologuous groups between *P. reichenowi, P. praefalciparum I* and *P. falciparum* was used for the test. d_N/d_S ratio estimates per branch and genes were obtained using Codeml (PAML v4.4c) with a *free-ratio* model of evolution. This identified 139 genes with a significant signal of positive selection in *P. falciparum* only.

A Branch Site test was also applied, for each gene, on each terminal branch of the entire species tree using the "Lav7sp" dataset. d_N/d_S ratio estimates per branch and genes were obtained using Codeml (PAML v4.4c) with a *free-ratio* model of evolution, Figure 2.

McDonald–Kreitman (MK) tests—Selection in *P. falciparum* was also tested using McDonald–Kreitman (MK) tests 68 to compare the polymorphism within species to the divergence between species, using *P. praefalciparum* as the outgroup. Analyses were performed using the 4,826 "Lav3sp" set of orthologues. MK tests were performed as described before9. Thirty-five genes had an MK ratio significantly higher than 1.

Gene Ontology enrichment analyses

Analysis of Gene Ontology (GO) term-enrichment was performed in R, using TopGO69 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 PhyML70 (default parameters) or RAxML62 (model estimated) from alignments generated with Muscle71 and trimmed with Gblocks61 in Seaview72 with default values. Many of the findings were confirmed manually through ACT and bamview49. The analysis of the *var* genes was performed on *var* genes larger than 2.5kb. Domains were called with the HMMer models from varDom73. 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 Uniprot74. Data were subsampled to obtain a similar number of sequences for each group. Phylogenetic trees were constructed with PhyML70, using default parameters and drawn in Figtree. The *eba-175* alignment was visualized with Jalview75.

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 events76,77(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. For coalescence based estimates of speciation times, G-Phocs78 was used and multiple sequentially Markovian coalescent (MSMC) on segregating sites79 was used to estimate the *P. falciparum* bottleneck.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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. 2010; 107:1458–1463. DOI: 10.1073/pnas.0914440107 [PubMed: 20133889]

2. Liu W, et al. Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. Nature. 2010; 467:420–U467. DOI: 10.1038/nature09442 [PubMed: 20864995]

3. Ollomo B, et al. A New Malaria Agent in African Hominids. PLoS Pathog. 2009; 5:e1000446.doi: 10.1371/journal.ppat.1000446 [PubMed: 19478877]

4. Liu W, et al. Multigenomic Delineation of Plasmodium Species of the Laverania Subgenus Infecting Wild-Living Chimpanzees and Gorillas. Genome biology and evolution. 2016; 8:1929–1939. DOI: 10.1093/gbe/evw128 [PubMed: 27289102]

5. Boundenga L, et al. Diversity of malaria parasites in great apes in Gabon. Malaria journal. 2015; 14:111.doi: 10.1186/s12936-015-0622-6 [PubMed: 25889049]

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

7. Silva JC, Egan A, Arze C, Spouge JL, Harris DG. A New Method for Estimating Species Age Supports the Coexistence of Malaria Parasites and Their Mammalian Hosts. Molecular biology and evolution. 2015; 32:1354–1364. DOI: 10.1093/molbev/msv005 [PubMed: 25589738]

8. Volkman SK, et al. Recent origin of Plasmodium falciparum from a single progenitor. Science. 2001; 293:482–484. DOI: 10.1126/science.1059878 [PubMed: 11463913]

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

10. Larremore DB, et al. Ape parasite origins of human malaria virulence genes. Nature communications. 2015; 6:8368.doi: 10.1038/ncomms9368

11. Pacheco MA, et al. Timing the origin of human malarias: the lemur puzzle. BMC evolutionary biology. 2011; 11:299.doi: 10.1186/1471-2148-11-299 [PubMed: 21992100]

12. Behar DM, et al. The dawn of human matrilineal diversity. American journal of human genetics. 2008; 82:1130–1140. DOI: 10.1016/j.ajhg.2008.04.002 [PubMed: 18439549]

13. Palstra FP, Fraser DJ. Effective/census population size ratio estimation: a compendium and appraisal. Ecology and evolution. 2012; 2:2357–2365. DOI: 10.1002/ece3.329 [PubMed: 23139893]

14. Roy SW. The Plasmodium gaboni genome illuminates allelic dimorphism of immunologically important surface antigens in P. falciparum. Infection, Genetics and Evolution. 2015; 36:441–449. DOI: 10.1016/j.meegid.2015.08.014

15. Tanabe K, Mackay M, Goman M, Scaife JG. Allelic dimorphism in a surface antigen gene of the malaria parasite Plasmodium falciparum. Journal of molecular biology. 1987; 195:273–287. DOI: 10.1016/0022-2836(87)90649-8 [PubMed: 3079521]

16. 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. 2015; 14:293.doi: 10.1186/s12936-015-0820-2 [PubMed: 26231699]

17. Malaria Genomic Epidemiology, N. Band G, Rockett KA, Spencer CC, Kwiatkowski DP. A novel locus of resistance to severe malaria in a region of ancient balancing selection. Nature. 2015; 526:253–257. DOI: 10.1038/nature15390 [PubMed: 26416757]

18. 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. 2016; 1603008113. doi: 10.1073/pnas.1603008113

19. Wanaguru M, Liu W, Hahn BH, Rayner JC, Wright GJ. RH5–Basigin interaction plays a major role in the host tropism of Plasmodium falciparum. Proceedings of the National Academy of Sciences. 2013; 110:20735–20740. DOI: 10.1073/pnas.1320771110

20. Wright KE, et al. Structure of malaria invasion protein RH5 with erythrocyte basigin and blocking antibodies. Nature. 2014; 515:427.doi: 10.1038/nature13715 [PubMed: 25132548]

21. Triglia T, Thompson JK, Cowman AF. An EBA175 homologue which is transcribed but not translated in erythrocytic stages of Plasmodium falciparum. Mol Biochem Parasitol. 2001; 116:55–63. DOI: 10.1016/S0166-6851(01)00303-6 [PubMed: 11463466]

22. Farrell A, et al. A DOC2 Protein Identified by Mutational Profiling Is Essential for Apicomplexan Parasite Exocytosis. Science. 2012; 335:218–221. DOI: 10.1126/science.1210829 [PubMed: 22246776]

23. Ramiro RS, et al. Hybridization and pre-zygotic reproductive barriers in Plasmodium. Proceedings of the Royal Society B-Biological Sciences. 2015; 282doi: 10.1098/rspb.2014.3027

24. Eksi S, et al. Malaria transmission-blocking antigen, Pfs230, mediates human red blood cell binding to exflagellating male parasites and oocyst production. Molecular Microbiology. 2006; 61:991–998. DOI: 10.1111/j.1365-2958.2006.0528.x [PubMed: 16879650]

25. Mundwiler-Pachlatko E, Beck HP. Maurer's clefts, the enigma of Plasmodium falciparum. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110:19987–19994. DOI: 10.1073/pnas.1309247110 [PubMed: 24284172]

26. Bethke LL, et al. Duplication, gene conversion, and genetic diversity in the species-specific acyl-CoA synthetase gene family of Plasmodium falciparum. Mol Biochem Parasitol. 2006; 150:10–24. DOI: 10.1016/j.molbiopara.2006.06.004 [PubMed: 16860410]

27. Cunningham D, Lawton J, Jarra W, Preiser P, Langhorne J. The pir multigene family of Plasmodium: antigenic variation and beyond. Mol Biochem Parasitol. 2010; 170:65–73. DOI: 10.1016/j.molbiopara.2009.12.010 [PubMed: 20045030]

28. Niang M, et al. STEVOR is a Plasmodium falciparum erythrocyte binding protein that mediates merozoite invasion and rosetting. Cell Host Microbe. 2014; 16:81–93. DOI: 10.1016/j.chom.2014.06.004 [PubMed: 25011110]

29. Kraemer SM, Smith JD. A family affair: var genes, PfEMP1 binding, and malaria disease. Curr Opin Microbiol. 2006; 9:374–380. DOI: 10.1016/j.mib.2006.06.006 [PubMed: 16814594]

30. Gardner MJ, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature. 2002; 419:498–511. DOI: 10.1038/nature01097 [PubMed: 12368864]

31. 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. 2014; 25:81–92. DOI: 10.1016/j.meegid.2014.04.010

32. Frank M, Dzikowski R, Amulic B, Deitsch K. Variable switching rates of malaria virulence genes are associated with chromosomal position. Mol Microbiol. 2007; 64:1486–1498. DOI: 10.1111/j.1365-2958.2007.05736.x [PubMed: 17555435]

33. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nature reviews. Genetics. 2012; 13:745–753. DOI: 10.1038/nrg3295

34. Carter R, Mendis KN. Evolutionary and historical aspects of the burden of malaria. Clinical microbiology reviews. 2002; 15:564–594. DOI: 10.1128/CMR.15.4.564-594.2002 [PubMed: 12364370]

35. 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. 2011; 6doi: 10.1371/journal.pone.0022213

36. Oyola SO, 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. 2014; 21:661–671. DOI: 10.1093/dnares/dsu028 [PubMed: 25240466]

37. Oyola SO, et al. Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. bioRxiv. 2016; doi: 10.1101/067546

38. 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. 2012; 11:163.doi: 10.1186/1475-2875-11-163 [PubMed: 22583664]

39. Quail MA, et al. Optimal enzymes for amplifying sequencing libraries. Nat Methods. 2012; 9:10–11. DOI: 10.1038/nmeth.1814

40. Manske H, Kwiatkowski D. SNP-o-matic. Bioinformatics. 2009; 25:2434–2435. DOI: 10.1093/bioinformatics/btp403 [PubMed: 19574284]

41. Chin CS, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013; 10:563–569. DOI: 10.1038/nmeth.2474 [PubMed: 23644548]

42. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics. 2009; 25(15):1968–1969. [PubMed: 19497936]

43. Carver T, et al. Artemis and ACT: viewing, annotation and comparing sequences stored in relational database. Bioinformatics. 2008; 24:2672–2676. DOI: 10.1093/bioinformatics/btn529 [PubMed: 18845581]

44. Otto TD, Sanders M, Berriman M, Newbold C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. Bioinformatics. 2010; 26:1704–1707. DOI: 10.1093/bioinformatics/btq269 [PubMed: 20562415]

45. English AC, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One. 2012; 7:e47768.doi: 10.1371/journal.pone.0047768 [PubMed: 23185243]

46. Otto TD. From sequence mapping to genome assemblies. Methods Mol Biol. 2015; 1201:19–50. DOI: 10.1007/978-1-4939-1438-8_2 [PubMed: 25388106]

47. Otto TD, Dillon GP, Degrave WS, Berriman M. RATT: Rapid Annotation Transfer Tool. Nucleic Acids Research. 2011; 39:e57.doi: 10.1093/nar/gkq1268 [PubMed: 21306991]

48. Stanke M, et al. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Research. 2006; 34:W435–439. DOI: 10.1093/nar/gkl200 [PubMed: 16845043]

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

50. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. DOI: 10.1093/bioinformatics/btp324 [PubMed: 19451168]

51. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–1303. DOI: 10.1101/gr.107524.110 [PubMed: 20644199]

52. Danecek P, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27:2156–2158. DOI: 10.1093/bioinformatics/btr330 [PubMed: 21653522]

53. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003; 13:2178–2189. DOI: 10.1101/gr.1224503 [PubMed: 12952885]

54. Jordan G, Goldman N. The Effects of Alignment Error and Alignment Filtering on the Sitewise Detection of Positive Selection. Molecular biology and evolution. 2012; 29:1125–1139. DOI: 10.1093/molbev/msr272 [PubMed: 22049066]

55. 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. 2005; 102:10557–10562. DOI: 10.1073/pnas.0409137102 [PubMed: 16000407]

56. Loytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 2008; 320:1632–1635. DOI: 10.1126/science.1158395 [PubMed: 18566285]

57. Fletcher W, Yang Z. The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection. Molecular biology and evolution. 2010; 27:2257–2267. DOI: 10.1093/molbev/msq115 [PubMed: 20447933]

58. 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. 2011; 21:863–874. DOI: 10.1101/gr.115949.110 [PubMed: 21393387]

59. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. Journal of Computational Biology. 2006; 13:1028–1040. DOI: 10.1089/cmb.2006.13.1028 [PubMed: 16796549]

60. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. Computers & Chemistry. 1993; 17:149–163. DOI: 10.1016/0097-8485(93)85006-X

61. Castresana J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Molecular biology and evolution. 2000; 17:540–552. DOI: 10.1093/oxfordjournals.molbev.a026334 [PubMed: 10742046]

62. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30:1312–1313. DOI: 10.1093/bioinformatics/btu033 [PubMed: 24451623]

63. Shimodaira H, Hasegawa M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Molecular biology and evolution. 1999; 16:1114.doi: 10.1093/oxfordjournals.molbev.a026201

64. Castoe TA, 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. 2009; 106:8986–8991. DOI: 10.1073/pnas.0900233106 [PubMed: 19416880]

65. Thomas GWC, Hahn MW. Determining the Null Model for Detecting Adaptive Convergence from Genomic Data: A Case Study using Echolocating Mammals. Molecular biology and evolution. 2015; 32:1232–1236. DOI: 10.1093/molbev/msv013 [PubMed: 25631926]

66. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular biology and evolution. 2007; 24:1586–1591. DOI: 10.1093/molbev/msm088 [PubMed: 17483113]

67. 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. 2005; 22:2472–2479. DOI: 10.1093/molbev/msi237 [PubMed: 16107592]

68. McDonald JH, Kreitman M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature. 1991; 351:652–654. DOI: 10.1038/351652a0 [PubMed: 1904993]

69. Rahnenfuhrer, AAaJ. topGO: Enrichment analysis for Gene Ontology. R package. 2010; doi: 10.18129/B9.bioc.topGO

70. Guindon S, et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010; 59:307–321. DOI: 10.1093/sysbio/syq010 [PubMed: 20525638]

71. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. DOI: 10.1093/nar/gkh340 [PubMed: 15034147]

72. 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. 2010; 27:221–224. DOI: 10.1093/molbev/msp259 [PubMed: 19854763]

73. Rask TS, Hansen DA, Theander TG, Gorm Pedersen A, Lavstsen T. Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes--divide and conquer. PLoS Comput Biol. 2010; 6doi: 10.1371/journal.pcbi.1000933

74. UniProt, C. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43:D204–212. DOI: 10.1093/nar/gku989 [PubMed: 25348405]

75. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009; 25:1189–1191. DOI: 10.1093/bioinformatics/btp033 [PubMed: 19151095]

76. Claessens A, et al. Generation of antigenic diversity in Plasmodium falciparum by structured rearrangement of Var genes during mitosis. PLoS Genet. 2014; 10:e1004812.doi: 10.1371/journal.pgen.1004812 [PubMed: 25521112]

77. Bopp SE, et al. Mitotic evolution of Plasmodium falciparum shows a stable core genome but recombination in antigen families. PLoS Genet. 2013; 9:e1003293.doi: 10.1371/journal.pgen. 1003293 [PubMed: 23408914]

78. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human demography from individual genome sequences. Nat Genet. 2011; 43:1031–1034. DOI: 10.1038/ng.937 [PubMed: 21926973]

79. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014; 46:919–925. DOI: 10.1038/ng.3015 [PubMed: 24952747]
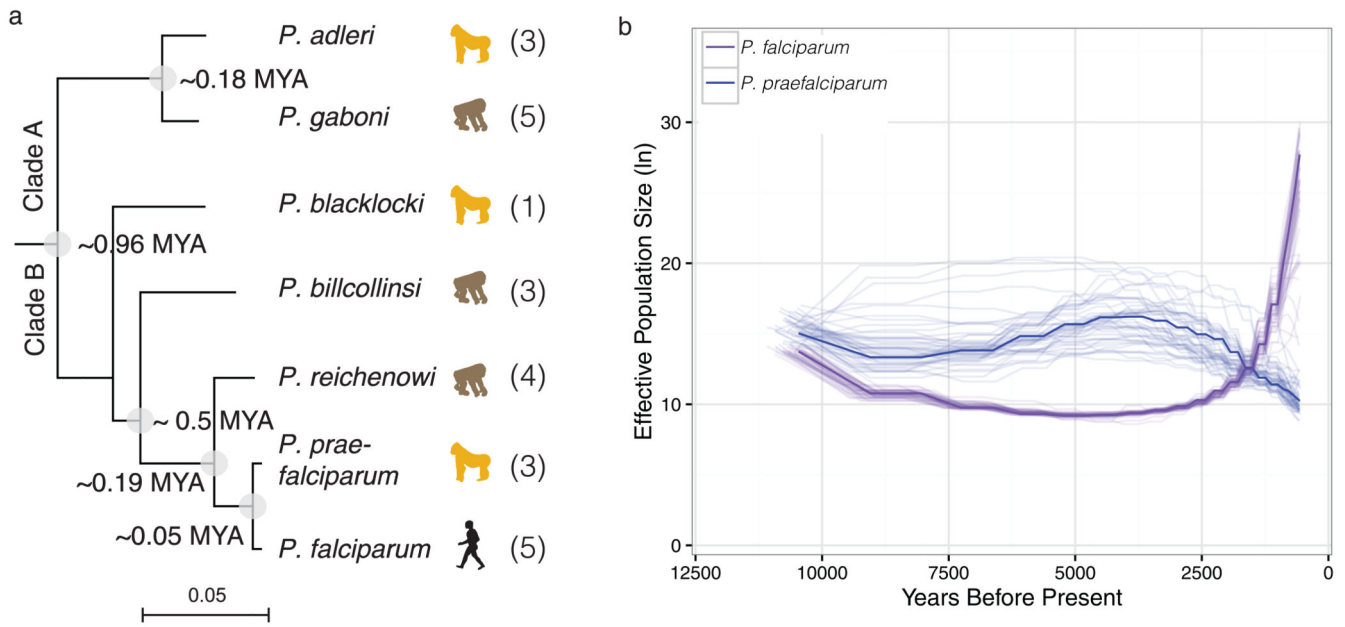
Figure 1. Overview of the dating of the evolution of the *Laverania*.
(**a**) Maximum likelihood tree of the *Laverania* based on the "Lav12sp" set of orthologues.
All bootstrap values are 100. Coalescence based estimates of the timing of speciation events
are displayed on nodes (MYA - million years 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 in *P. falciparum*
4,000-6,000 years ago. The y-axis shows the natural logarithm (Ln) of *Ne*. Bootstrapping
(pale lines) was performed by randomly resampling segregating sites from the input 50
times.
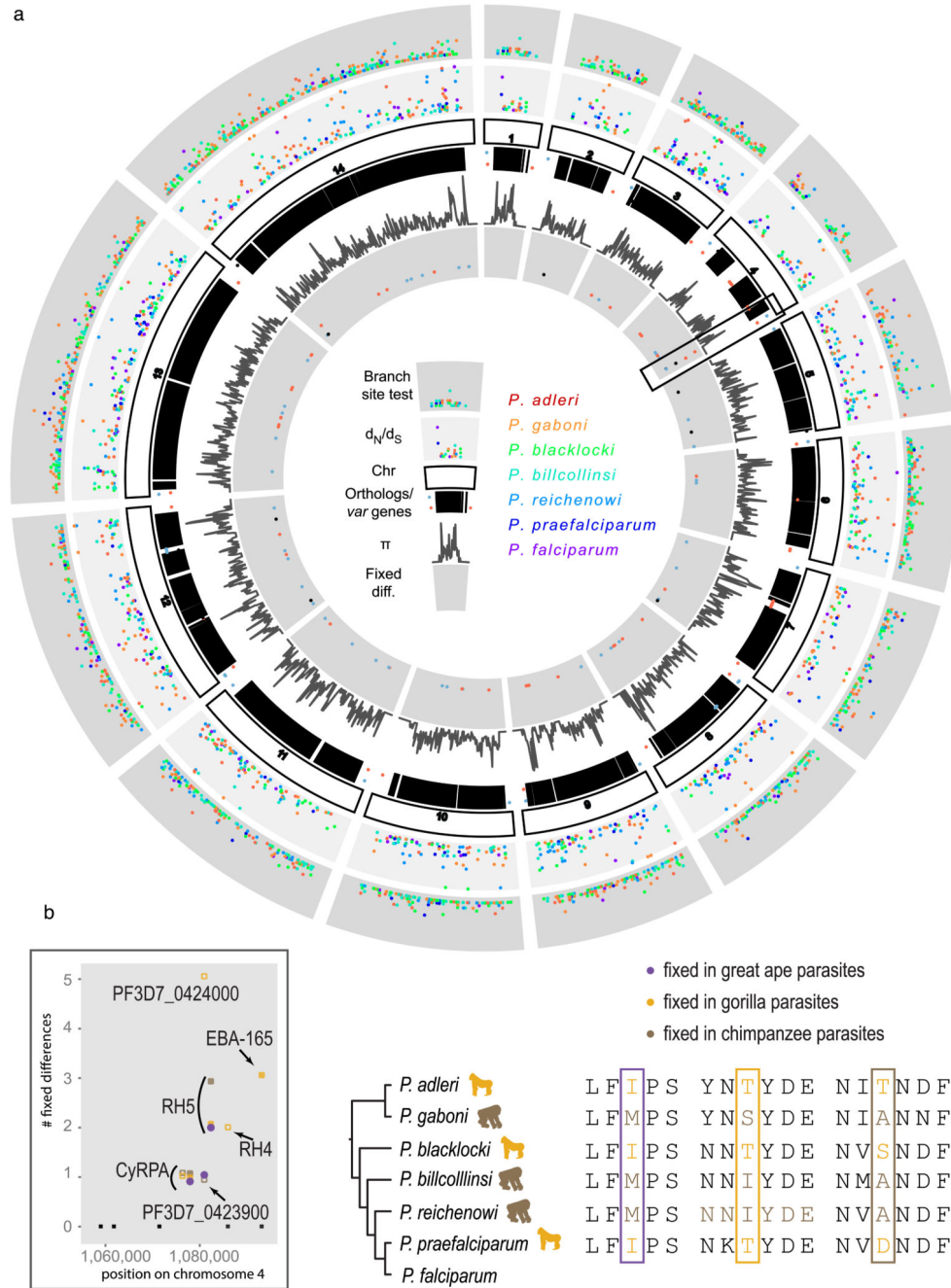
Figure 2. Overview of the analyses of core genes over all *Laverania* genomes.
(a) Summary of evolution of core genes. From outer to inner track: scatterplot of branch site test for each genome (see Supplementary Table 4 for *P. falciparum* data); per-species d_N/d_S values ($0.5 < d_N/d_S < 2$); orthologues represented by vertical black lines under the chromosome track represent, with dots representing *P. falciparum* 3D7 *var* genes on the forward (blue) or reverse strands (red), or *var* pseudogenes (black); average of the relative polymorphism (π) across species, with the underlying π for each species calculated from multiple strains ("Lav15st" dataset) and normalized by the average for that species;

signatures of convergent evolution based on host-specific fixed differences analysis with the chromosome 4 region that includes the *Rh5* locus highlighted (black box). (b) Magnified view of the *Rh5* region that is enriched with host specific fixed differences. Convergent evolution analysis was performed using orthologues conserved across the *Laverania*. Filled circles represent the subset of differences that were fixed within all the isolates available ("Lav15st" set) and for which we could reject neutral evolution (for the gene list see Supplementary Table 5).

	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	2	1	5	3	0
Maurer	13	12	5	4	7	0	0
EPF1	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-CoA	13	17	17	11	18	16	26

Figure 3. Gene families in the *Laverania*.
Distribution of major multigene families including *var* and those that show significant copy number variation among lineages. Data from *P. praefalciparum* include the subtelomeric gene families from the two infecting genotypes. Assembly of *P. billcollinsi* is incomplete in the subtelomeres.
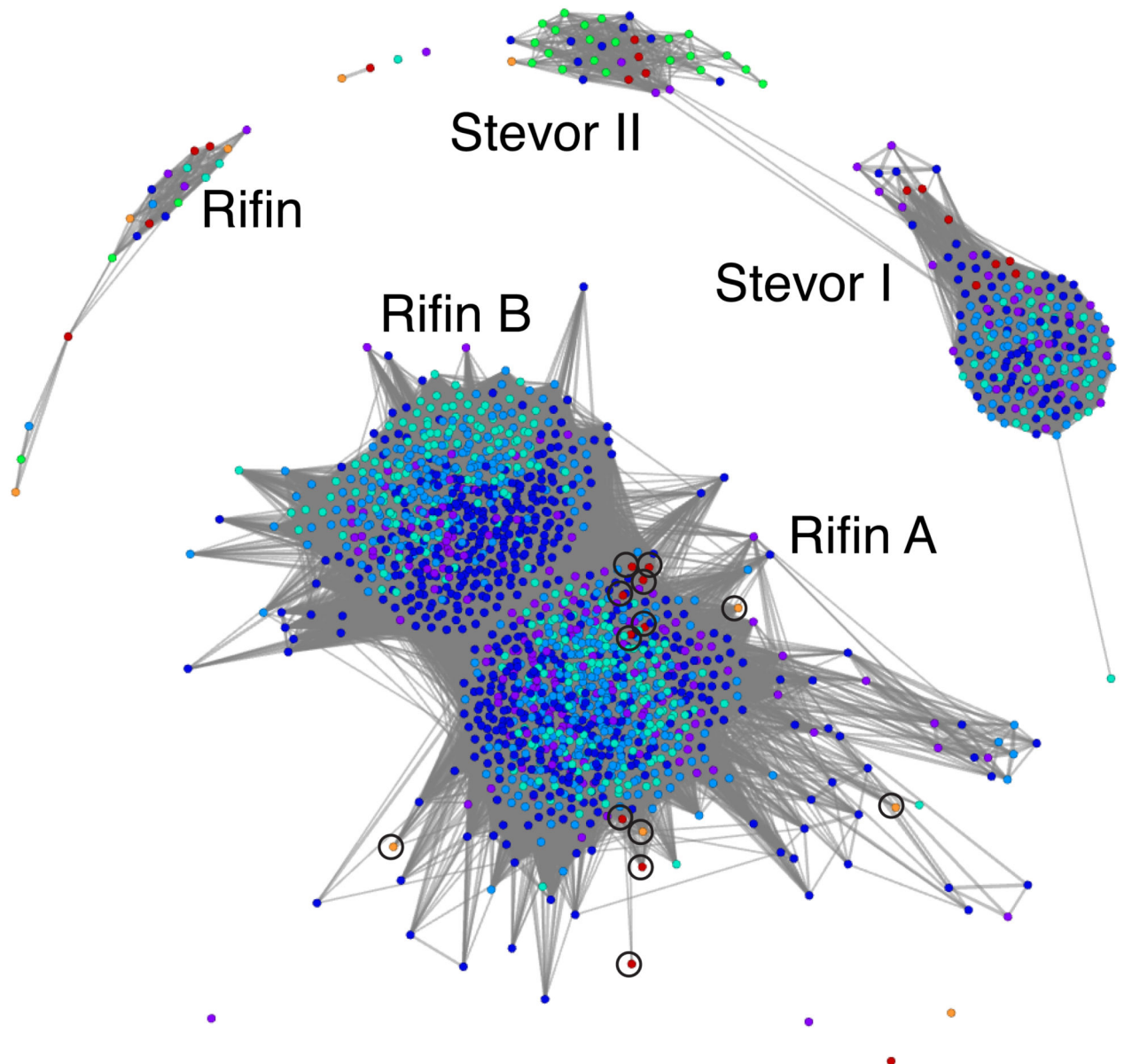
Figure 4. Clustering of Pir (Rifin and Stevor) proteins families.
Graphical representation of similarity between all *pir* proteins > 250aa, coloured by species.
A BLAST cut-off of 45% global identity was used (see methods). More connected genes are
more similar. Black circles highlight Clade A rifin proteins that cluster with Clade B rifin
proteins.

a

	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

P. adleri   *P. gaboni*   *P. blacklocki*   *P. billcollinsi*   *P. reichenowi*   *P. praefalciparum*   *P. falciparum*
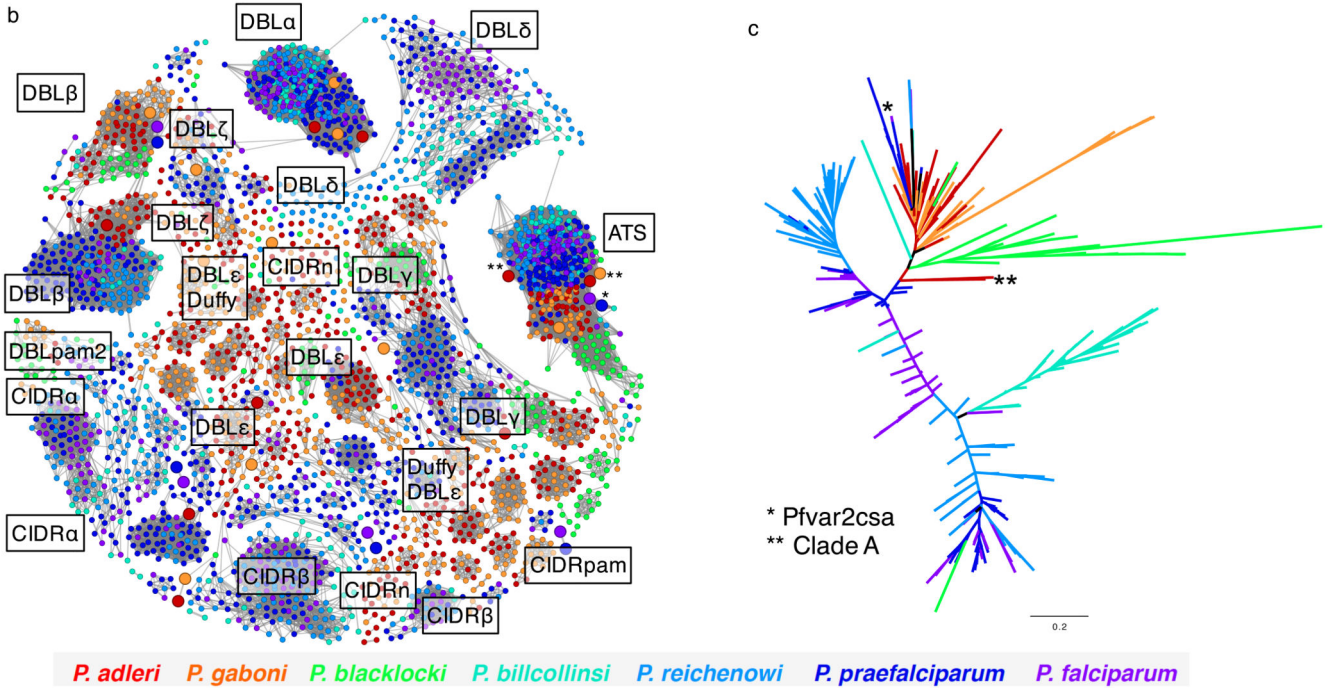
Figure 5. Evolution of *var* gene domains in the *Laverania*

(**a**) Heatmap of numbers 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. Heat map colours blue-yellow-white indicate decreasing copy numbers. (**b**) Graphical representation of similarity between domains, using domains from *var* genes longer than 2.5kb. Domains are coloured by species and clustered by a minimum BLAST cut-off of 45% global identity. Larger circles denote *var* genes in the opposite orientation. (**c**) Maximum likelihood trees of the Acidic Terminal Sequence (ATS). Apparent ATS sequences from clade A that cluster with clade B are indicated (**).
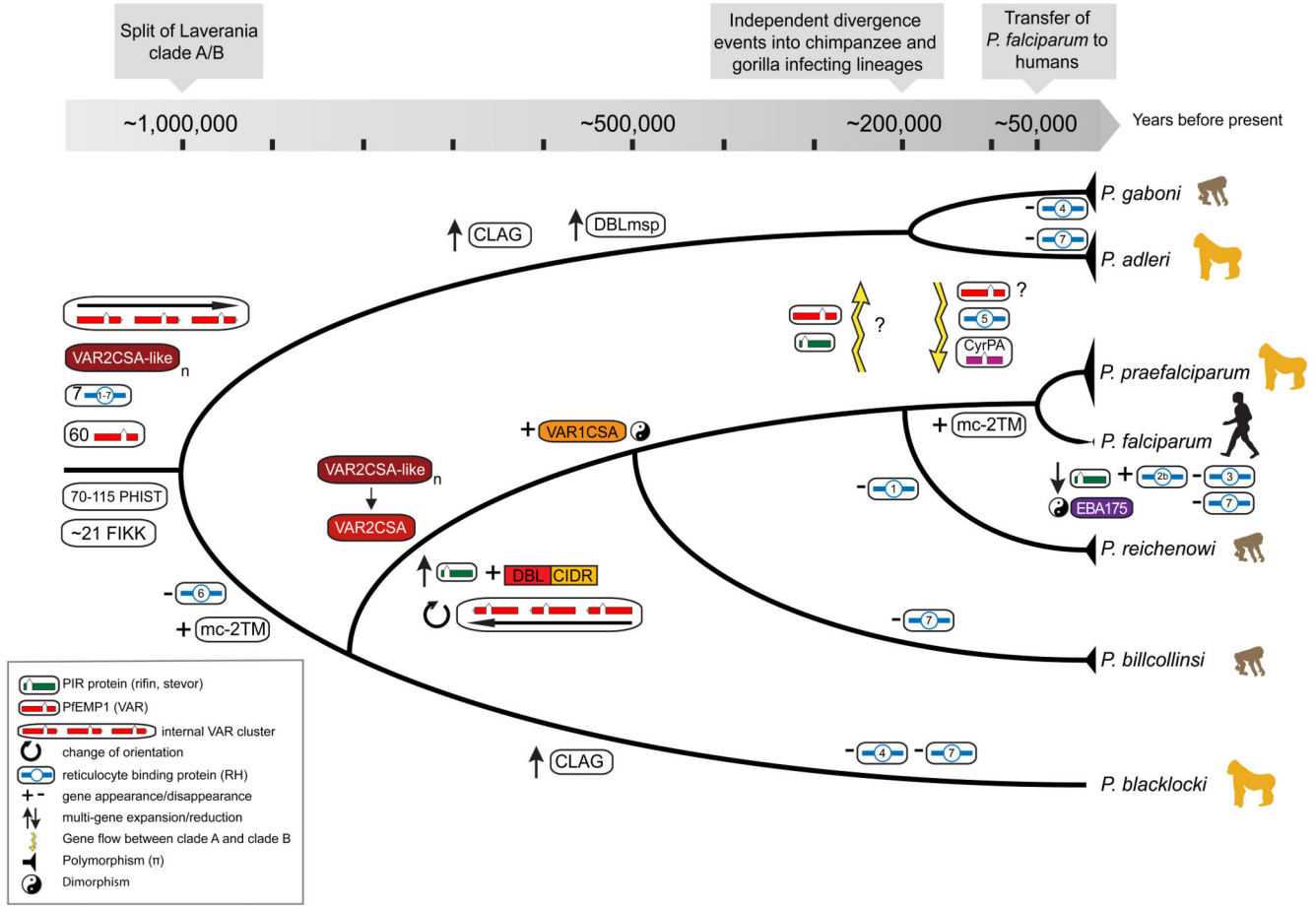
Figure 6. Overview of the genomic evolution of the *Laverania* subgenus.
The values of polymorphism (π) within the species are indicated by triangles of different size at the end of the tree branches, as well the bottleneck in *P. falciparum* (constricted branch width), ~ 5,000 years ago. Also shown 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*.