Genomes of an entire *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 Oyola[1], Alain Prince Okouga[4], Larson Boundenga[4], Eric Wuillaume[5], Barthélémy Ngoubangoye[4], Nancy Diamella Moukodoum[4], Christophe Paupy[2], Patrick Durand[4], Virginie Rougeron[2,4], Benjamin Ollomo[4], François Renaud[2], Chris Newbold[1,6], Matthew Berriman[1,*] & Franck Prugnolle[2,4,*]

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

[2] Laboratoire MIVEGEC, UMR 5290-224 CNRS-IRD-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*. Although there are further gorilla and chimpanzee-infecting species in the same (*Laverania*) subgenus as *P. falciparum*, none are known to be able to establish repeated infection and transmission in humans. To elucidate underlying mechanisms and the evolutionary history of this subgenus, we have analysed multiple genomes from all known *Laverania* species. Here we estimate the timings of *Laverania* speciation events, placing *P. falciparum* speciation 40,000-60,000 years ago followed by a recent population bottleneck. We show 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 for the first time, features in *P. falciparum* are revealed that made it the only member of the *Laverania* able to infect and spread in humans.

Introduction

The evolutionary history of *Plasmodium falciparum*, the most common and deadliest human malaria parasite, has been the subject of much 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* and 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* only) (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*).

Since the discovery of the *Laverania* a number of studies have provided incremental data on the evolution of this subgenus but the mechanisms underlying host specificity and the reasons why only one productive transfer into humans has taken place are still unclear. Comparisons of the *P. falciparum* genome first to that of *P. reichenowi (6)* and more recently to that of *P. gaboni (7)*, have attempted to identify genes that displayed rapid or specific evolution. However, the lack of whole genome information for the whole subgenus (particularly *P. praefalciparum*, the closest sister species to *P. falciparum*) limited the power of these comparisons. In addition, the studies involving *P. gaboni*, lacked the subtelomeric regions that harbour many gene families involved in host-parasite interactions and antigenic variation (*8*). Only a subset of these multigene families has been studied by PCR-based approaches (*9*) (e.g. the DBLα domain of *var* genes), preventing a complete reconstruction of the paths that led to the evolution of *P. falciparum*.

To investigate the evolutionary history of the entire *Laverania* subgenus and to unravel the genomic evolutionary paths that allowed humans to be colonised, we have sequenced multiple genotypes of all known *Laverania* species.

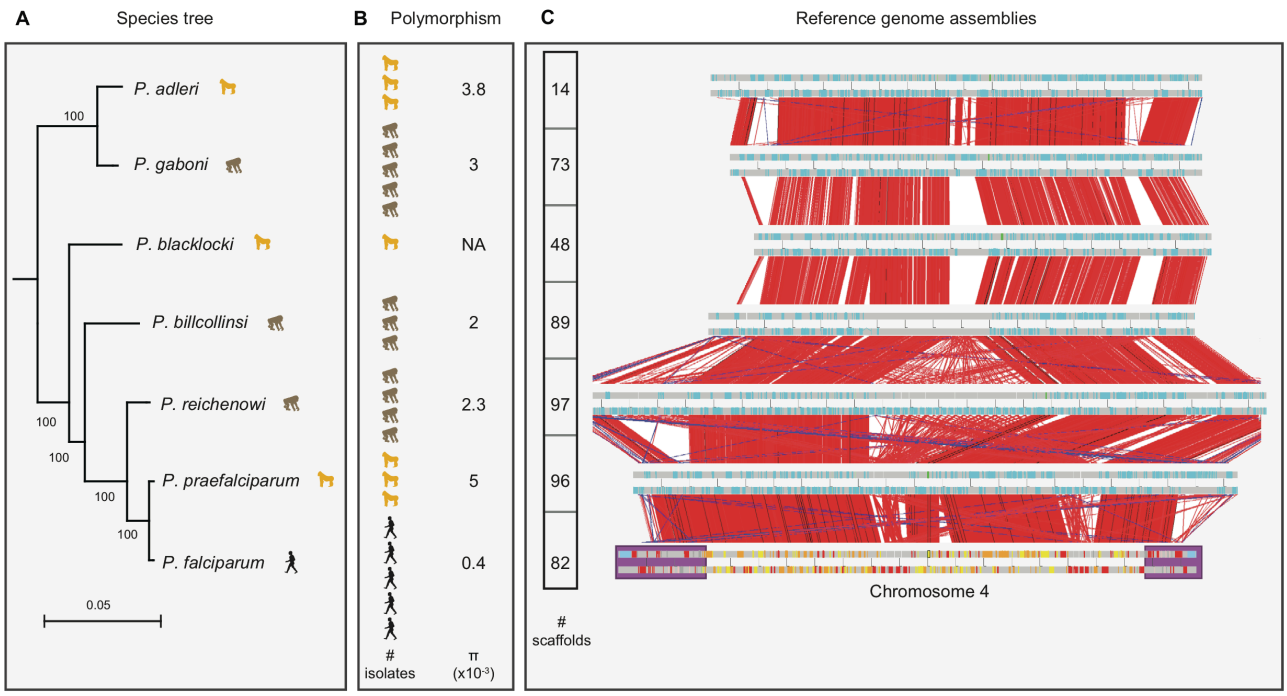


Figure 1: Overview of the *Laverania* genomes analysed. (A) **Maximum likelihood tree of the *Laverania* based on the "Lav12sp" set of orthologues (see Methods). Bootstrap values are provided at nodes.** (B) **The number of isolates per species sequenced is indicated using the symbol of their host species, together with the values of polymorphism within species, as estimated using the nucleotide diversity π ("Lav15st" set of orthologues; see Supplementary Materials).** (C) **The quality of the genome assemblies is indicated through the number of scaffolds and the pairwise BLAST scores within the core chromosome 4 and its subtelomeric regions (shaded boxes).**

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 Supplementary Material and 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 methods enabled sufficient parasite DNA templates to be obtained for short-read (Illumina) and long read (Pacific Bioscience) sequencing (Fig 1B, Table 1). Mixed infections, which were frequent, were resolved by utilising sequence data from single infections, resulting in 19 genotypes (Table S1), and the dominant genotype in each sample was assembled *de novo* into a reference genome (see Supplementary Material and Methods). Each reference genome was assembled into 44-97 scaffolds (Table 1), with large contigs containing the subtelomeric regions or internal *var* gene clusters comprising multigene families (Fig. 1C) that are known in *P. falciparum* and *P. reichenowi* to be involved in virulence and host-pathogen interactions. In total, genomes were produced for six malaria parasite species: *P. praefalciparum*, *P. blacklocki*, *P. adleri, P. billcollinsi*, *P. gaboni* and *P. reichenowi*. 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).

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	Multiple infection, see PadlG03						*P. adleri*
P. praefalciparum II	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**	**Multiple inf. PgabG02**		**23.1**	**306**	**89**	**5,637**	***P. gaboni***
	PbilcG03	Multiple infection, see PgabG04						*P. gaboni*
	PbilcG04	Multiple 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(*6*)[b] Based on percentage of reads mapping to the reference for each Laverania species (see Table S**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 (bold) where assemblies were produced using long-reads from Single-Molecule Real-Time sequencing (Pacific Biosciences), the total assembled size, number of contigs, scaffolds and predicted genes are shown. For those we report the assembly. 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.

Phylogenetic relationships and the emergence of *P. falciparum*

Conservation of gene content and synteny is striking between these complete genomes, enabling us to reconstruct with confidence the relationships between different *Laverania* species, to compare their relative genetic diversity (Supplementary Text, Fig. S1) and to estimate the age of the different speciation events that led to the extant species. Because of uncertainty regarding exact generation time and *in vivo* mutation rates (both of which have a linear influence on time estimates), the values that we quote should be regarded as approximate (Supplementary Materials, Table S2). 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 and B): Clade A that

comprises *P. adleri* and *P. gaboni* and Clade B that includes the remaining species (Fig. 2A). Our range of values is far more recent than previous estimates (*3, 10*). Ancestral state reconstruction did not allow us to solve the nature of the host (gorilla or chimpanzee) in which the ancestor lived. Following the group 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. 2A, Table S2), suggesting that the same 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. 2A), significantly later than the evolution of the first modern humans and their spread throughout Africa (*11*). Our analysis also indicates significant gene flow between these two parasite species after speciation (Table S2).

It has been suggested that *P. falciparum* arose from a single transfer of *P. praefalciparum* into humans (*7*). It has also been proposed (based on the paucity of neutral SNPs within the genome of *P. falciparum*) that *P. falciparum* emerged from a bottleneck of a single parasite around 10,000 thousand years ago, after agriculture was established (Fig. 2B) (*7, 12*). Clearly neither of these hypotheses are correct in light of our results; we estimate that the *P. falciparum* population declined around 8,000-14,000 years ago and reached a minimum about 5,000 years ago (Fig. 2C) with an effective population size (*Ne*) of at least 8,000 (Supplementary Text; generally the census number of parasites is higher than *Ne* (*13*)). Neither are these hypotheses consistent 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 (*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 (Fig. S2A). Most strikingly, a mutually exclusive dimorphism (described as MAD20/K1 (*15*)) in the central 70% of the *msp1* sequence, 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*.

The gene *eba-175* that encodes a parasite surface ligand contains a dimorphism that arose after the emergence of *P. falciparum* (Fig. S2B). The time to the most recent common ancestor 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 recalibrate their estimate of the most recent common ancestor of the *eba-175* alleles to

be 4 thousand years ago, which is in good agreement with our divergence time for *P. falciparum* (Supplementary Text). The recent dimorphism cannot however explain the recent observation of an ancient dimorphism near the human and ape loci for glycophorin (*17*) – an EBA-175 binding protein. Different balancing selection pressures over time may have shaped the formation and maintenance of all of these dimorphic loci. Identifying the interacting host genes (or epistatic interactions with other parasite genes) will shed more light on the underlying mechanisms.

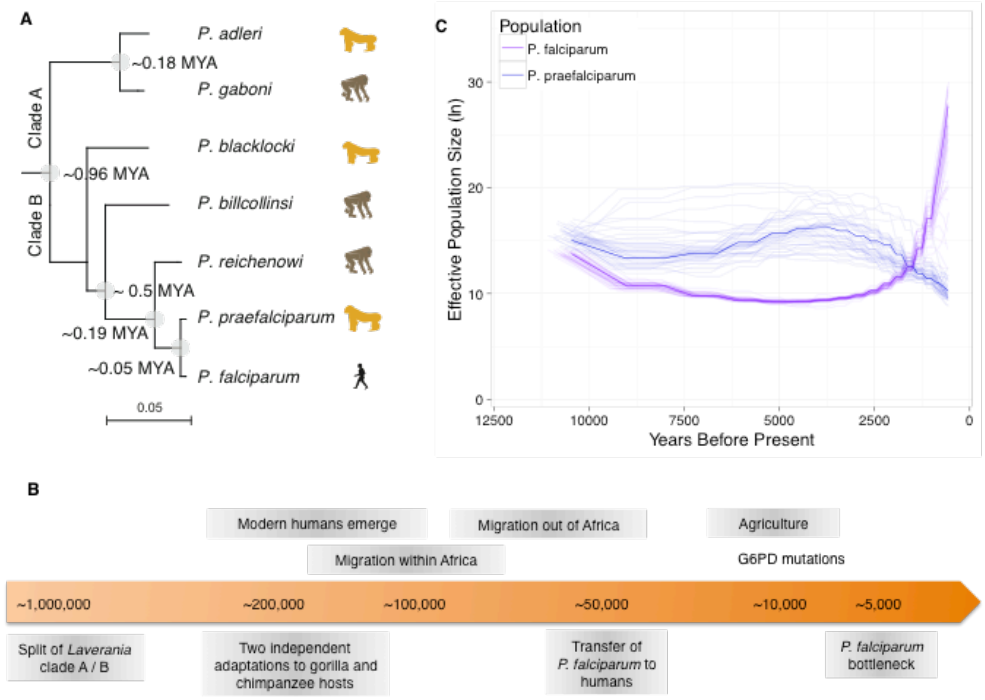


Figure 2 Overview of the dating of the evolution of the *Laverania*. (A) **Coalescence based estimates of the split of the species using intergenic and genic alignments. Dates are displayed on nodes (average of 402–681 mitotic events per year used; MYA - million years ago). The ratios should be robust but the actual dates depend on estimates of generation time and mutation rate (see Table S2 for details).** (B) **Estimated time line of events in the *Laverania* relative to human evolution.** (C) **Multiple sequentially Markovian coalescent estimates of the bottleneck in the *P. falciparum* population. Assuming our estimate of the number of mitotic events per year, the bottleneck occurred 4,000-6,000 years ago. y-axis is the natural logarithm (Ln). Bootstrapping was performed by 50 replicates by randomly resampling from the segregating sites used as input.**

Evolution through introgression, gene transfer and convergence

Frequent mixed infections in apes and mosquitoes (*18*) provide clear opportunities for interspecific gene flow between these parasites. A recent study (*7*) 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 (*rh5* 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. Beyond the region that includes *rh5* (Fig. S3A,B), few signals of interspecific gene flow were obtained (n =14) suggesting that these events were rare or usually strongly deleterious (Supplementary Text, Fig. S4).

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 Text, Fig. S5). 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. 3; Table S3A). 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 (Table S3B) including *rh5* (which has a signal for convergent evolution even when the introgressed tree topology is taken into consideration; Fig. S3C). *Rh5* is the only gene identified in *P. falciparum* that is essential for erythrocyte recognition during invasion, via binding to Basigin. *P. falciparum rh*5 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*) (Fig. S3C).

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 in the other *Laverania* species (Fig. 4) 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 (*22*), possibly through enabling male gametes to bind to erythrocytes prior to exflagellation (*23*). 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.

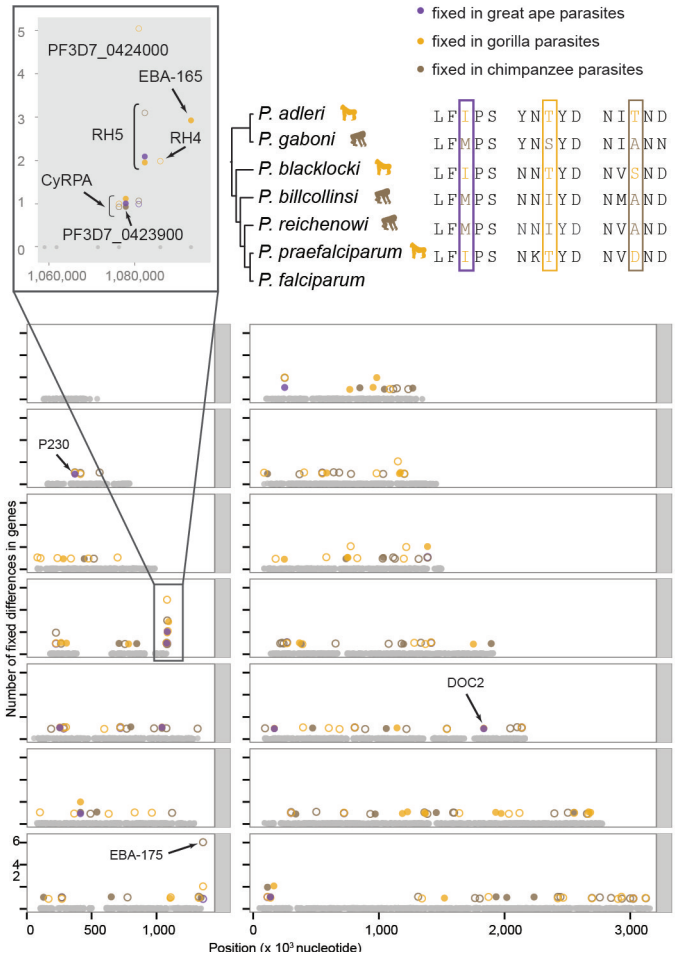


Figure 3: Convergent evolution analyses: fixed differences between great ape-infecting species. Analysis was performed using the "Lav7sp" set of orthologues but filled circles are for the differences that are fixed within all the isolates available ("Lav15st" set) and for which we could reject neutral evolution (For the gene list see Table S3)

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 complete picture of the evolution of these important families (Fig. 4) inside the subgenus.

Gene content and copy number is summarized in Fig. 4 and Table S4A. The first important observation is that 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. 4 and Table S4). For these latter families, two groups of parasites clearly differ in their composition: Clade A on one side (with *P. adleri* and *P. gaboni*) and some species of Clade B on the other side (*P. billcollinsi*, *P. reichenowi* and *P. praefalciparum* and *P. falciparum*). *P. blacklocki* (Clade B) is intermediate in its composition. Such subdivision concerns for instance 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. 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 (*24*) and yet has expanded only in Clade B, after the *P. blacklocki* split. For other gene families, like the group of exported proteins *hyp4, hyp5*, *mc-2tm* and *EPF1*, they seem to have expanded only in *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(*25*), some evolution of function in this organelle may have been an important precursor to human infection.

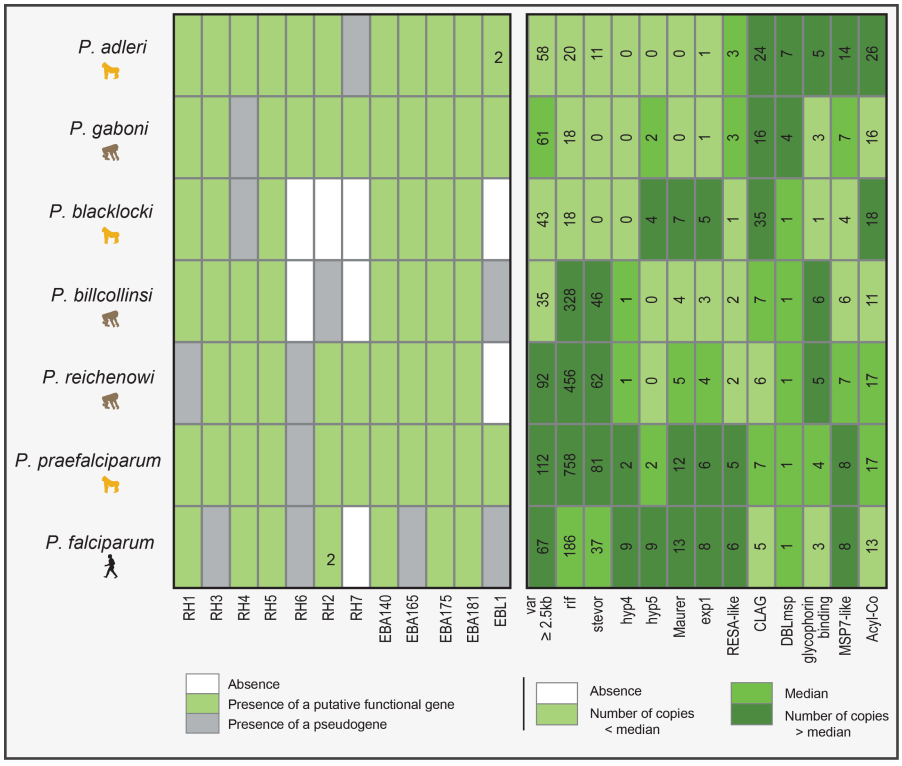
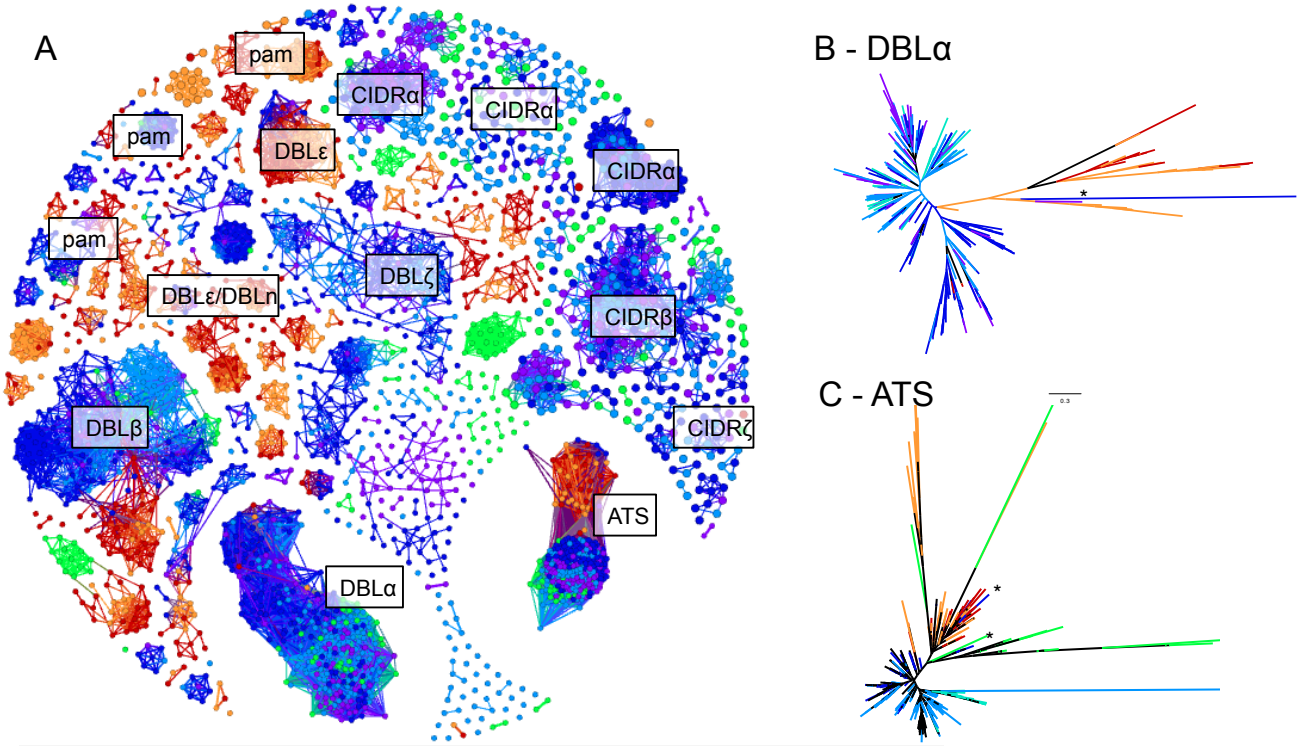


Figure 4 Gene families in the *Laverania.* **Distribution of genes encoding erythrocyte invasion ligands (RH and EBA genes) and other major multigene families.**

Because *Laverania* are known to have varying host tropism and highly specific preferences for their host species, we looked at variations in gene number and relatedness between gorilla and chimpanzee specific parasites. Some *stevor* sequence types are differentially associated with chimpanzee or gorilla infecting parasites (Fig. S6). Interaction with host-factors could be a major role of the *stevor* family; expression on the infected erythrocyte surface and stevor binding to host glycophorin C in *P. falciparum*(*26*) have been reported.

The family of acyl-CoA synthetase genes, reported to be expanded and diversified in *P. falciparum (27)* (Fig. S7A) are in fact expanded across in *Laverania* and have four fewer copies in *P. falciparum*. Three other gene families with significant differences within the *Laverania,* are namely DBLmsp, glycophorin binding protein and CLAG (Fig. S8). Focusing on protein domains within the subtelomeric families, the *Plasmodium* RESA N-terminal domain has clear host-specific differences in copy number variation in gorilla parasites compared to chimpanzee ones (Fig. S7B, Table S4B).



D

Species	# var	NTS	CIDRα	CIDRβ	CIDRδ	CIDRγ	DBLα	DBLβ	DBLδ	DBLε	DBLγ	DBLn	DBLζ	NTS pam	CIDR pam	DBL pam1	DBL pam2	DBL pam3	ATS
Padleri	58	1	1	1	0	0	2	28	0	99	45	58	31	0	2	7	1	7	19
Pgaboni	61	1	1	0	0	0	2	6	0	75	20	45	17	0	16	4	4	19	17
Pblacklocki	43	0	0	0	0	0	0	0	0	0	26	4	6	0	0	8	0	0	0
Pbillcollinsi	35	20	31	25	0	5	27	9	32	4	0	2	1	0	0	0	0	0	21
Preichenowi	92	77	85	60	1	27	87	49	78	18	42	0	8	1	1	2	1	1	62
Ppraefalciparum	112	65	85	47	7	29	92	79	70	96	81	1	29	5	5	6	5	5	94
Pfalciparum	67	56	56	37	2	17	59	17	52	15	15	0	6	1	1	2	1	1	59

Figure 5 species. A BLAST cut-off of 50% global identity was used. More connected domains are more similar. Maximum likelihood trees were generated for DBLα (B), and the conserved C terminal domain termed Acidic Terminal Sequence (ATS), (C). *P. falciparum* domains clustering with Clade A are indicated (*). The heatmap (D) shows the amount of *var* genes longer than 2500bp and the frequency of domains per species

Evolution of *var* genes

The *var* genes, crucial mediators of pathogenesis and virulence through cytoadherence and immune evasion, are the most studied and the best known *P. falciparum* gene family. These genes show striking difference between them in Clades A and B (Fig. 5). This gene family is unique to the *Laverania* and encodes proteins with variable numbers of highly polymorphic Duffy Binding Like (DBL) domains and cysteine rich interdomain regions (CIDR) involved in the interaction with specific host cell surface proteins, and a more conserved intracellular domain (*8*). DBL and CIDR domains have a number of subtypes identified in *P. falciparum* denoted by Greek letters. Our previous analysis of the *P. reichenowi* genome (*6*), and other recent studies on partial *var* genes from a subset of the *Laverania (9)*, established the ancient origin of these sequences, their relationship to the current *var* repertoire of *P. falciparum* and the existence of unusual domains in *P. gaboni*. Here we show that the total domain composition and genomic organisation is conserved in Clade B parasites, with the exception of *P. blacklocki* (Fig. 5a; Fig. S9A).

In contrast, the evolution of these genes in Clade A parasites (*P. adleri* and *P. gaboni*) and in *P. blacklocki* has followed a different course. They almost completely lack CIDR domains, have fewer DBLα domains but contain an excess of other DBL types and untypeable domains named x1 and x2 by Larremore *et al.* (*9*). Analysis of full length sequences however puts these latter domains within the DBLε subtype (Fig. S10). A tree of all of the DBLα domains highlights the fact that those of Clade A parasites are unusual (Fig. 5B). Moreover, clustering within the DBLα of Clade A is a DBLα pseudogene on chromosome 4 of both *P. falciparum* and *P. praefalciparum* that is isolated in these two species but is part of an additional internal *var* gene cluster in the other species. Furthermore, its sequence is highly conserved across all species suggesting that the pseudogene may still be functional.

A further difference on chromosome 4 between Clade A and B parasites relates to the orientation of the internal *var* gene clusters (Fig. S9A). In all of the Clade B parasites they are transcribed from the reverse strand whereas in Clade A they are located on the forward strand. The orientations of these clusters could have implications on their relative capacities to recombine with other family members and create further diversity.

If we compare sequence relatedness between each individual domain across all species then it is apparent that they have diverged widely. This is particularly evident in the CIDR and intracellular domains, suggesting that polymorphism in host receptors, the use of different receptors or changes in signalling may be driving the divergence (Fig. S10).

The number of DBL domains currently associated in *P. falciparum* with *var2csa* (DBLpam) has also increased in Clade A. This gene is involved in cytoadhesion to the placenta in placental malaria but is also a central intermediary in the process of switching between expressed *var* genes during antigenic variation (*28*). The DBLpam domains clearly have a different architecture in Clade A parasites and are absent from *P. billcollinsi* and amplified in *P. praefalciparum* implying that historically they may have had a different function (Fig. 5 and Fig. S10). In addition, though the overall similarity between *P. falciparum var2csa* genes is around 80%, we find in two *P. praefalciparum* parasites, *var2csa* genes that shares a 6kb region of 99.61% identity to a set of current *P. falciparum* field isolates (Fig. S9B) suggesting either introgression or long term functional conservation and strong purifying selection.

***P. falciparum*-specific evolution**

To infer *P. falciparum* specific adaptive changes, we considered the *P. falciparum* / *P. praefalciparum* and *P. reichenowi* genome trios 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 (Table S5). 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 those 171 genes, almost half (n=83) 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* or *emp3* involved during the erythrocytic stages, *P230* involved in gametocytes, *trsp* or *lisp*1 involved in the hepatic stages or *plp4*, *CelTOS* or *Cap380* involved in the mosquito stages) also showed a significant signal of adaptive evolution (Table S5).

Discussion

How did *P. falciparum* arise? We have shown that the successful infection of humans occurred quite recently, around 40,000-60,000 years ago, 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. 2C). Irrespective of the analytical approach used, across the subgenus, we found

entry into host, particularly red cells, to be crucial in determining host specificity. A key historic event in the eventual colonisation of humans is likely to have been the horizontal transfer of a region of chromosome 4 from *P. adleri* to the ancestor of *P. falciparum / P. praefalciparum* that contained both the *rh5* and the *cyrpA* genes, the latter forming an essential complex with *ripr/rh5* in the invasion process. Comparison of *P. falciparum* lines capable of infecting *Aotus* monkeys to those that cannot, suggests that *rh5* is necessary but *not sufficient* for host transfer (*29*). We have identified in this gene sites that are host-specific, species-specific and *praefalciparum/falciparum* specific which taken together suggest that it is crucial to a host switch. To investigate what other changes must have been involved, we looked for genetic signatures shared by parasite species that infect the same primate host and identified some host specific signals within the core genomes. We have identified genes involved in sexual and mosquito stages together with a number of genes of unknown function expressed across the life cycle, the latter of which clearly require further investigation.

Due to the completeness of the genomes, we were able to analyse subtelomeric gene families in detail. Because these families are intimately involved in host-parasite interactions we were surprised to discover that only one of them (*stevor*) showed evidence of host specific evolution. One feature of malaria parasites is their ability to continuously reinfect the same host throughout its life. This ability is due, at least in part, to the size and polymorphism of the *var* repertoire(*30*). It is likely that early in the evolution of this parasite, when host population sizes were relatively small, that this characteristic was essential to continue transmission and overcome herd immunity of a limited host population. We show that the *var* genes must have been present in the common ancestor of the *Laverania* but have evolved differently in the two major branches. However, the differences between the repertoires of Clade A and B parasites are not host specific, despite showing clear signs of species-specific evolution.

The direct comparison of *P. falciparum* with *P. praefalciparum* identified differences in reticulocyte binding proteins (duplication of *rh2* and pseudogenization of *rh3*), in gene dimorphism (*msp1, msp3* and *eba-175*) and in adaptive evolution of genes expressed in the sexual stages. We also observe a reduced number of *pir* genes, a loss of four acyl-coA synthetases as well as a set of genes showing lineage specific signals of adaptive evolution at all stages of the life cycle.

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, 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 200,000 years ago (*31*) and existed as small isolated populations (*11*). 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 implying parasites would need to move between hosts but inefficiently. 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 *Anopheles gambiae*) to feed primarily on humans (*32*). 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 are likely to have occurred to increase human specific reproductive success that both produced a specific parasite genotype expansion (that would also appear as an emergence from a bottleneck) and also resulted in the much lower probability of a direct transfer of *P. falciparum* back to apes. Our estimate of the rapid expansion of *P.falciparum* within the last 5000 years is also consistent with the proposed timing of the population expansion of *Anopheles(33)* and with estimates of the age of haemoglobinopathies known to be protective against malaria(*34*). 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*. Nevertheless, the data and analyses presented here will be invaluable in future studies of host specificity in *Plasmodium*.

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.

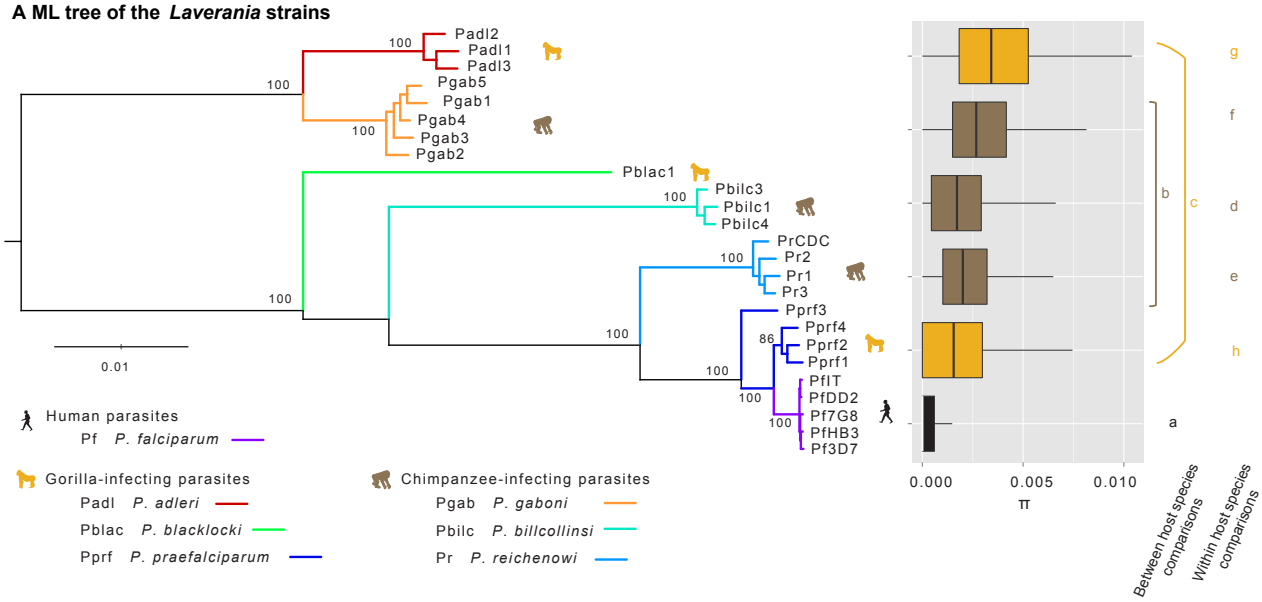
Acknowledgments: This work was funded by ANR ORIGIN JCJC 2012, LMI ZOFAC, CNRS, CIRMF, IRD and the Wellcome Trust (grant WT 098051 to the Sanger Institute, 104792/Z/14/Z CN). TC holds a MRC DTP Studentship. We thank Gavin Rutledge for performing the sWGA and Julian Rayner and Francisco J. Ayala for helpful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

Data Availability

All sequences have been submitted to the European Nucleotide Archive. The accession numbers of the raw reads, and assembly data can be found in Table S6. The genomes are being submitted to EBI, project ID PRJEB13584. As the assemblies are currently private, but are available on request.

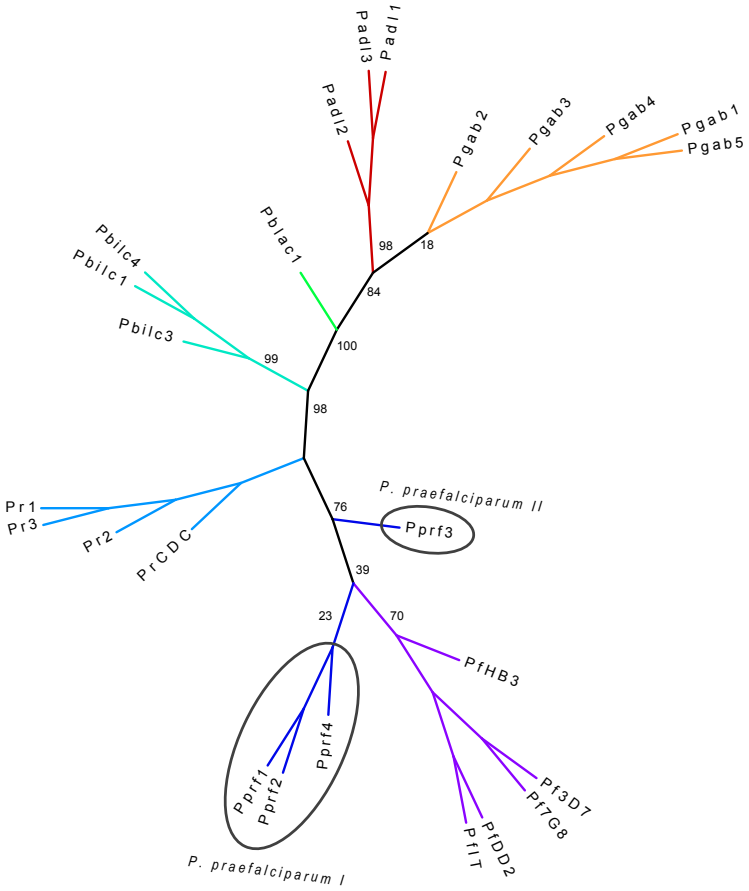
Supplemental Figures & Tables

A ML tree of the *Laverania* strains

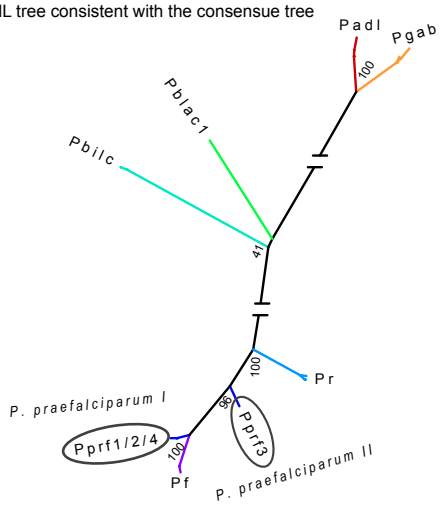


B Identification of the two *P. praefalciparum* lineages

1) Extended majority rule tree



2) ML tree consistent with the consensue tree



3) ML tree showing recombination between the divergent *P. praefalciparum* lineages

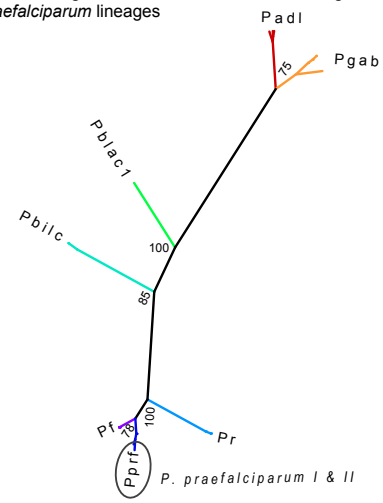


Fig. S1. (A) Maximum Likelihood tree and nucleotide diversity of *Laverania* isolates. Tree was obtained using the sequences of 424 genes ("Lav25st" set of orthologues). The results of the test of comparison of the polymorphism between and within host species are given on the right: polymorphism in *P. falciparum* "a" < polymorphism in the chimpanzee-infecting species "b" < polymorphism in the gorilla-infecting species "c"; within the chimpanzee-infecting species, polymorphism is lower in *P. billcollinsi* "d" and higher in *P. gaboni* "f"; within the gorilla-infecting species, polymorphism is lower in *P. adleri* "g" and higher in *P. praefalciparum* "h" (see Supplementary Text). (**B**) Identification of two *P. praefalciparum* lineages**.** (1) Extended majority rule tree, built using RAxMLv8.1.20(*61*) and the "Lav25st" set of 424 orthologues. The consensus tree, reveals two divergent *praefalciparum* lineages that seems however to recombine. Two examples of the observed single-gene tree topologies are shown. The first (2) is consistent with the consensus tree but the second (3) indicates that recombination has occurred between the two *P. praefalciparum* lineages. For readability we excluded the G0 in the names of the genotypes.

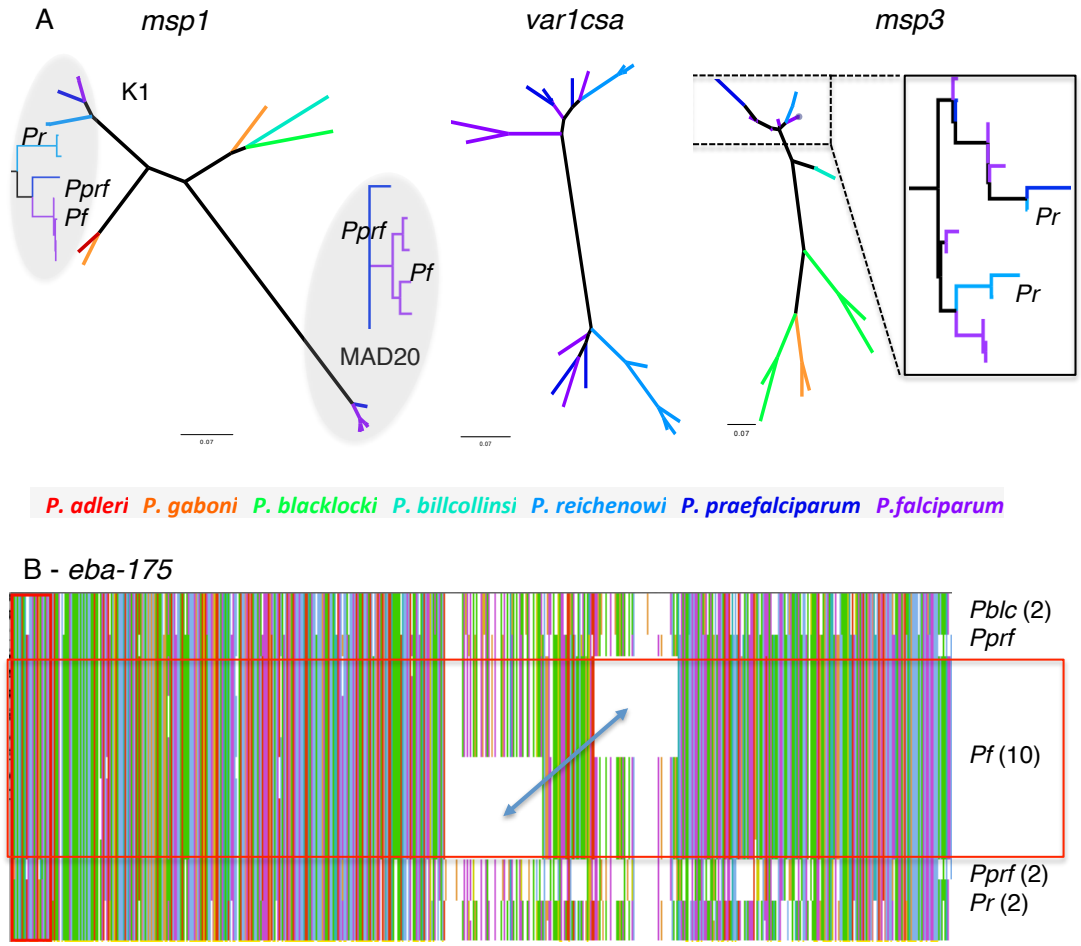


Fig. S2. Dimorphisms in the *Laverania*. (**A**) Examples of ancient dimorphisms based on maximum likelihood phylogenetic trees. Dimorphism in *msp1* arose in the *P. falciparum–P. praefalciparum* ancestor, after the divergence of *P. reichenowi* and dimorphism in *var1csa* evolved in the *P. reichenowi–P. 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 (Supplementary Materials).

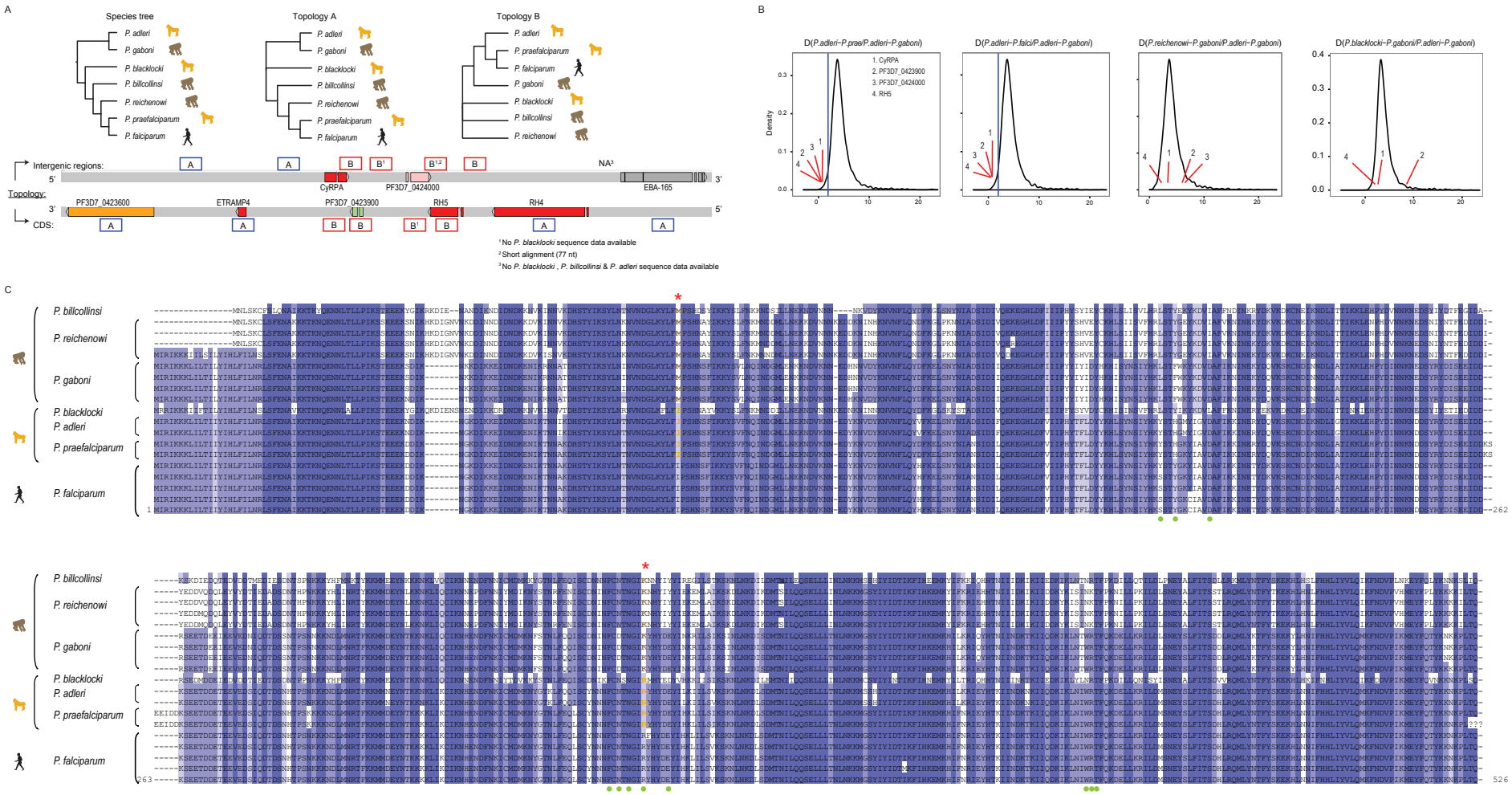


Fig. S3. **Interspecific gene transfer and convergent evolution in the 3' end of the chromosome 4.** (**A**) Support for interspecific gene transfer between the gorilla-infecting species *P. adleri* and the common ancestor of *P. praefalciparum* and *P. falciparum*. The topologies observed in the coding and intergenic regions of the end of chromosome 4 are given. (**B**) The analysis of relative genetic distances in the *rh5* region and in the other regions of the genome also support the interspecies transfer. (**C**) 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.

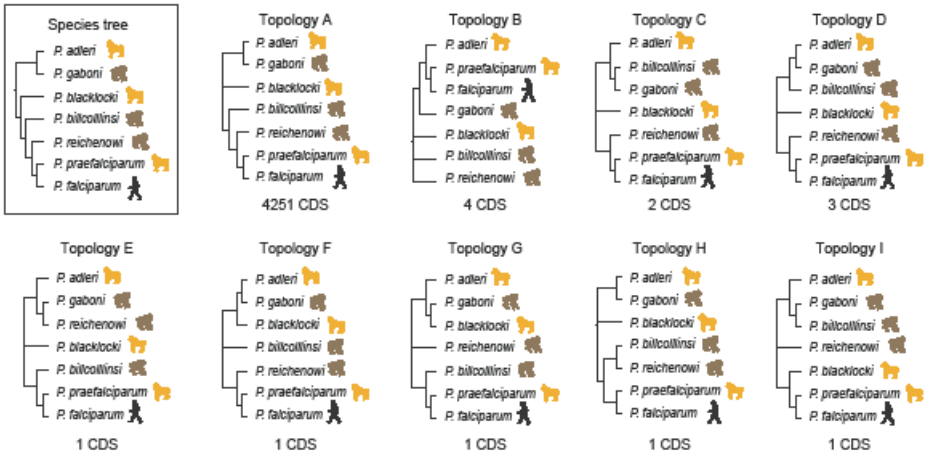


Fig.S4. **Tree topology tests.** The species tree (topology A) and the topologies that differ significantly from the species tree topology – when sufficient phylogenic information was available – are given.

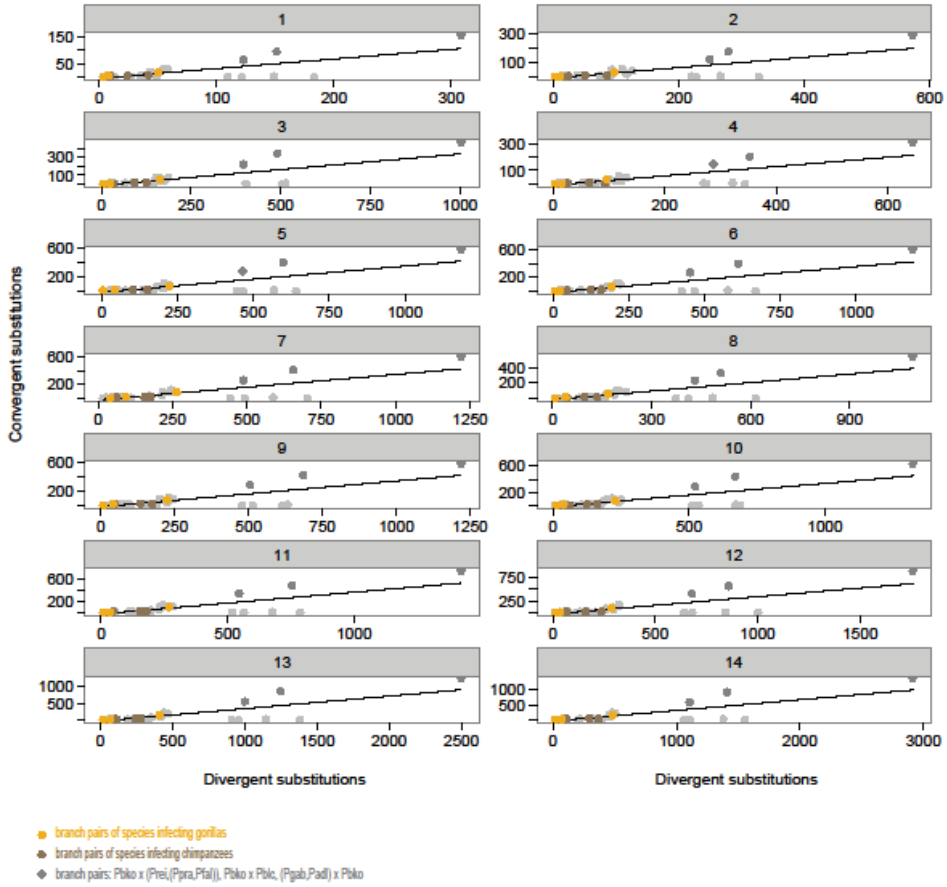


Fig. S5. **Relationships between number of divergent and convergent substitutions for each branch pair of the *Laverania* species tree.** Relationships are provided for each chromosome (1-14). See Material and Methods for details.

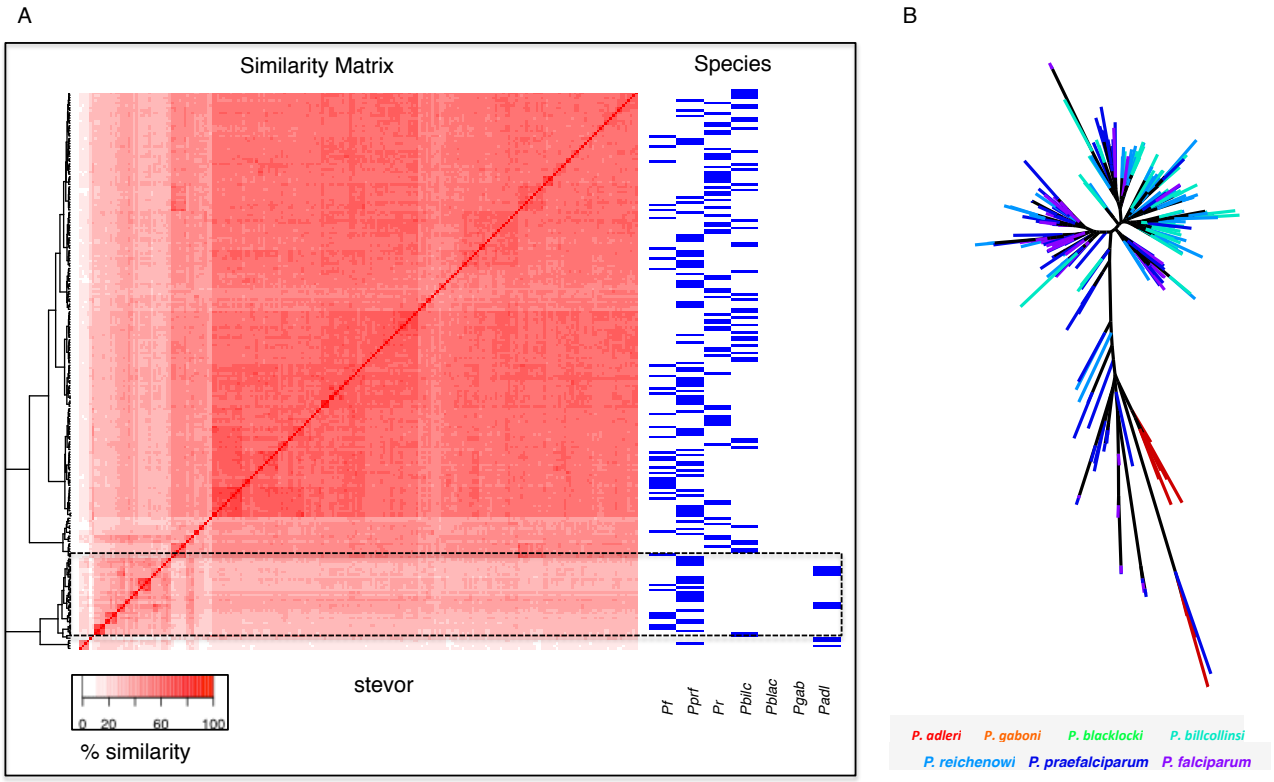


Fig. S6. Similarity between *stevor* genes. (**A**). Similarity matrix of the *stevor* genes. The similarity matrix was generated through a BLASTp (expect-val ≤1e-6) and then clustered through the ward2 algorithm in R. Every row and column represents a gene. (**B**) The dendrogram on the left hand side shows the clustering and similarity between the genes. The binary blue barcode shows in which species each gene copy is present. Different groups can be seen. One cluster has is no present in chimpanzee parasites (black dotted box). *Pf* - *P. falciparum, Pprf P. praefalciparum, Pr* - *P. reichenowi, Pbilc* - *P. billcollinsi, Pblac* - *P. blacklocki, Pgab* - *P. gaboni* and *Padl* - *P. adleri*. b. Maximum likelihood tree of the same data.

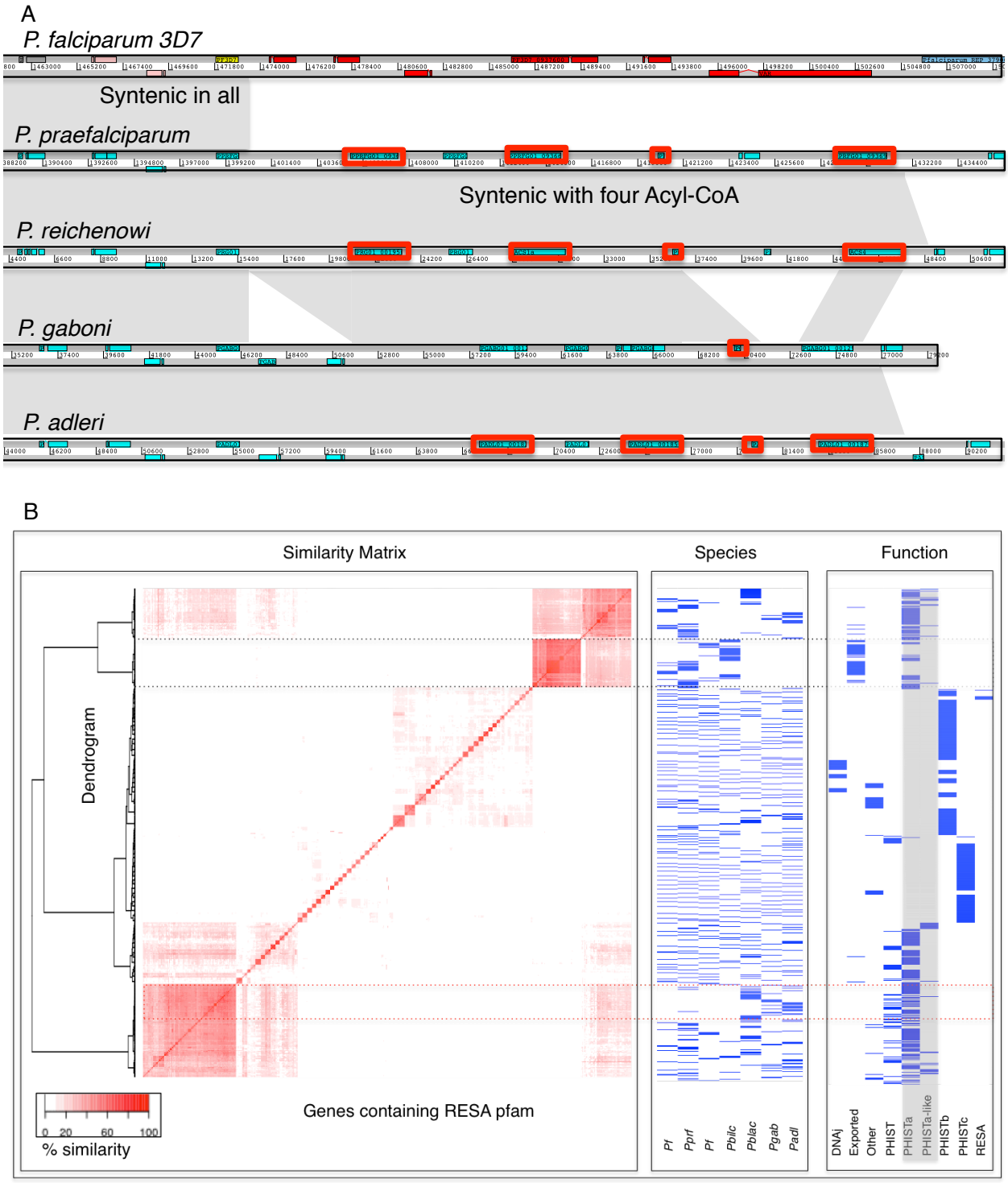


Fig. S7. Acyl-CoA Synthetase expansion on Chromosome 9 and Resa domains. (**A**) ACT view of five genomes, at the right hand side of chromosome 9. The grey areas indicate co-linearity. *P. falciparum* has lost this region with four Acyl-coA synthase genes, as this locus is conserved in the other species. (**B**) Similarity matrix of all genes containing the Resa Pfam domain. The similarity matrix was generated through a BLASTp (expect-val ≤1e-6) and then clustered through the ward2 algorithm in R. Every row and column represents a gene. The dendrogram on the left hand side shows the clustering and similarity between the genes. The

binary blue barcode indicates species and gene annotation. Most of the genes are well distributed between the species, the genes annotated as exported proteins (red dotted box), cluster by species and do not occur in Clade A. Some of the PhistA genes seem also to cluster by species (red dotted box). *Pf* - *P. falciparum, Pprf P. praefalciparum, Pr* - *P. reichenowi, Pbilc* - *P. billcollinsi, Pblac* - *P. blacklocki, Pgab* - *P. gaboni* and *Padl* - *P. adleri*

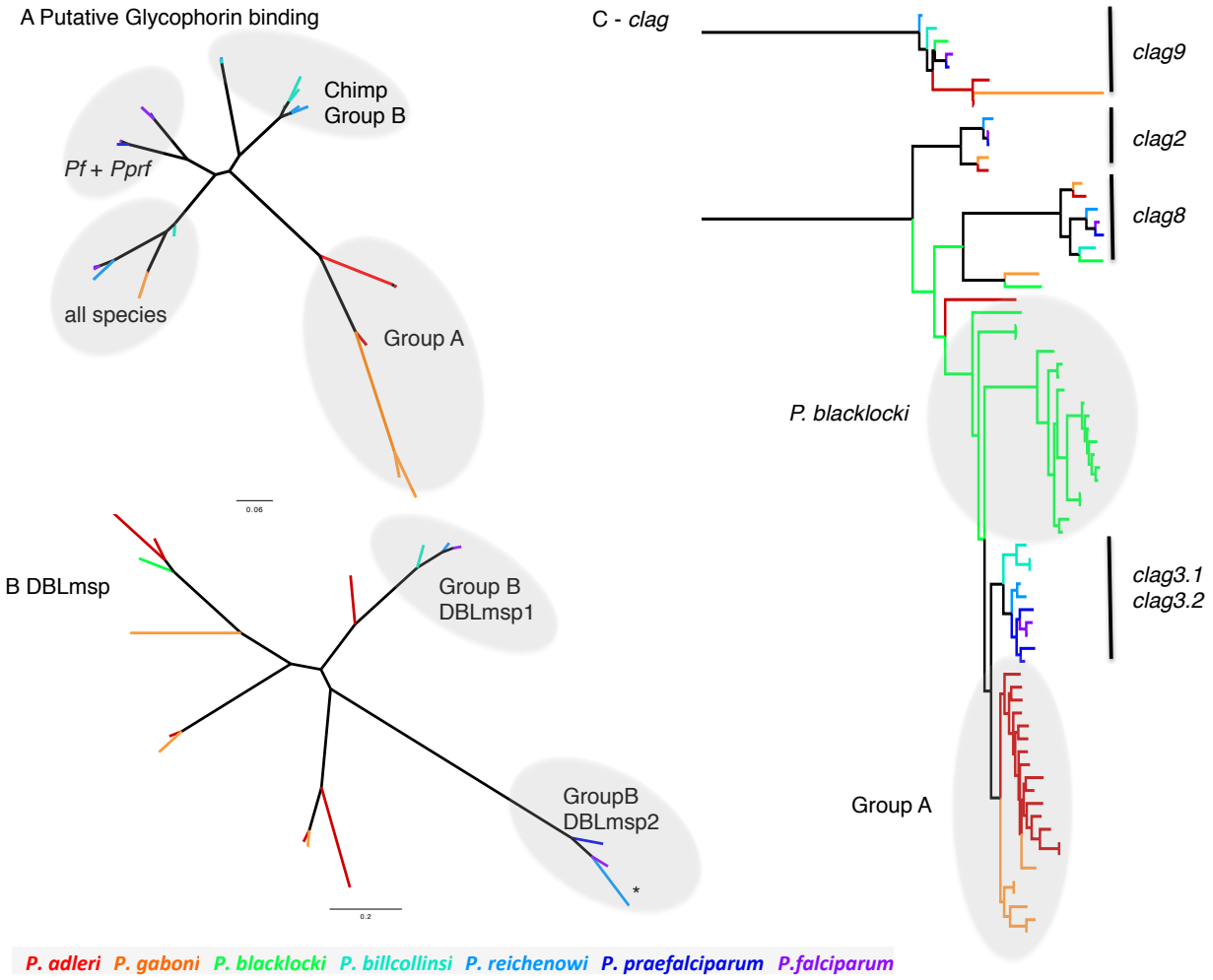
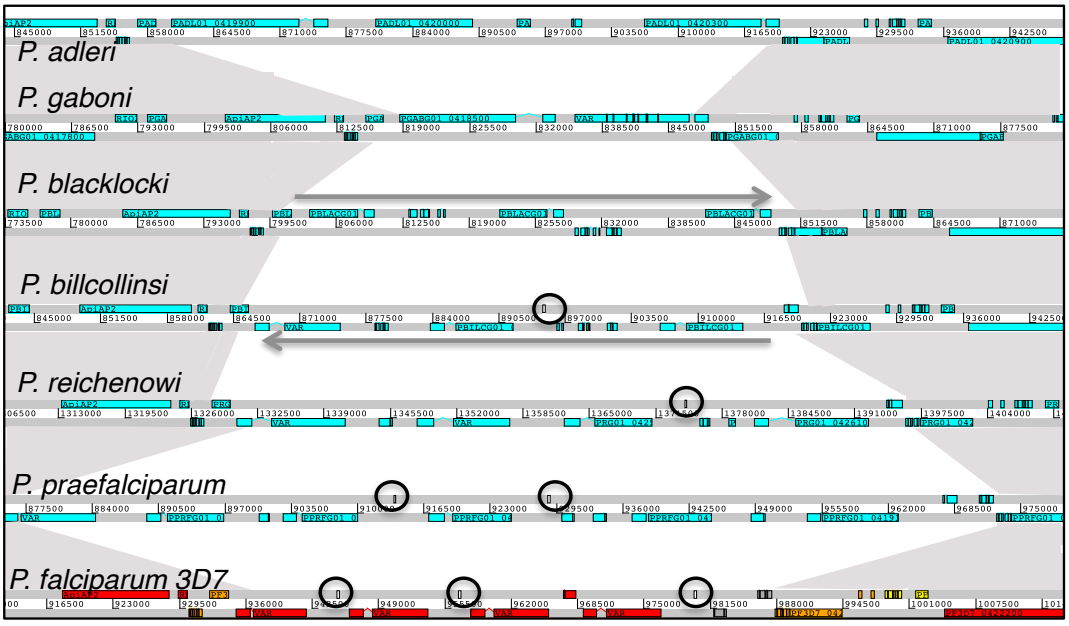


Fig. S8. Phylogenetic analysis of gene families. Example of three families that show differences within the *Laverania*. (**A**) The putative glycophorin binding proteins form four distinct groups. One group contains sequence from all species. The remaining groups are clade, host or species sub-group specific. (**B**) Differences in the DBLmsp that are expanded in Clade A. The DBLmsp2 is a pseudo gene (*) in *Pr*. (c) Expansion of *clag* genes in Clade A.

A



B

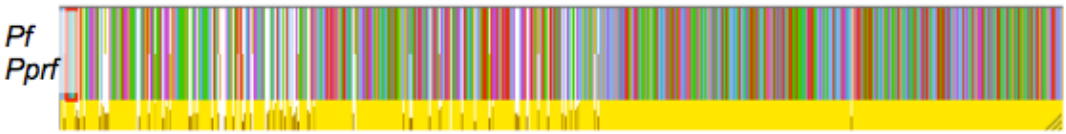
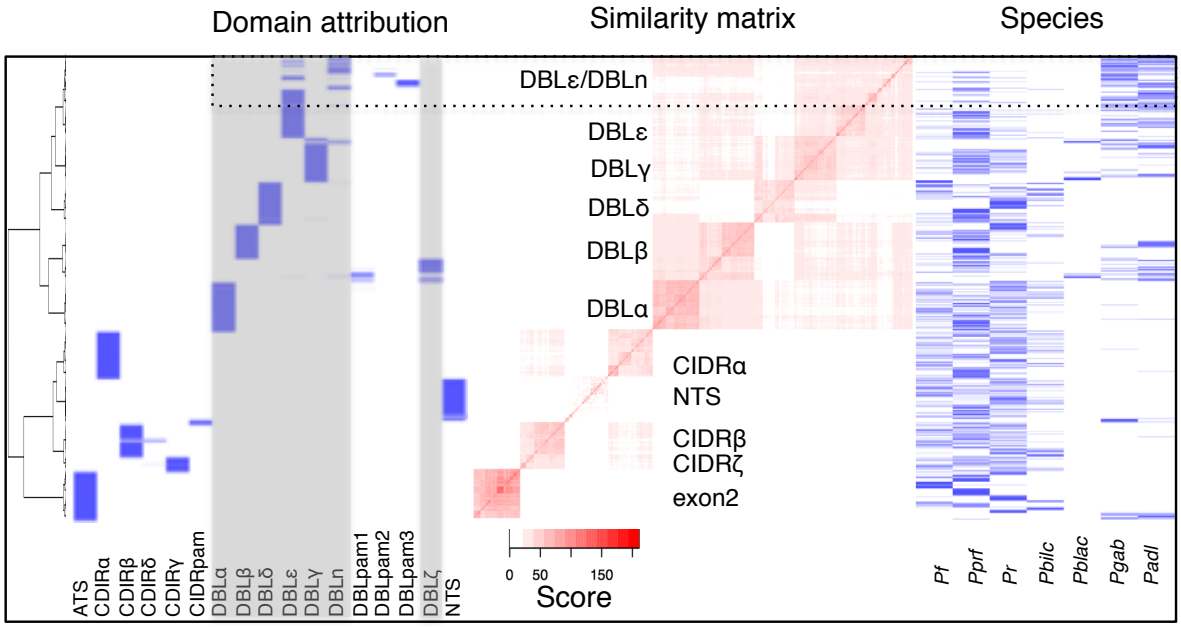


Fig.
S9. Composition, structure and evolution of *var* genes within *Laverania.* (**A**) Schematic of 2nd internal *var* genes cluster of chromosome 4 in the seven *Laverania* species. In the Clade A & *P. blacklocki*, the orientation of the *var* genes is different compared to that of the other species. Also the high GC ruf (RNA of Unknown function) elements, in circles, do not occur in those genomes. Last it can be also seen that the size of the var genes between the species is different. **(B)** shows a conserved *var2csa* between *P. falciparum* and *P. praefalciparum* where 6kb is shared between species with over 99% identity. Total length of the alignment is longer than 12kb.

A


Domain attribution
Similarity matrix
Species
DBLε/DBLn
DBLε
DBLγ
DBLδ
DBLβ
DBLα
CIDRα
NTS
CIDRβ
CIDRζ
exon2
ATS
CDIRα
CDIRβ
CDIRδ
CDIRγ
CIDRpam
DBLα
DBLβ
DBLδ
DBLε
DBLγ
DBLn
DBLpam1
DBLpam2
DBLpam3
DBLζ
NTS
0 50 150
Score
Pf
Pprf
Pr
Pbilc
Pblac
Pgab
Padl

B

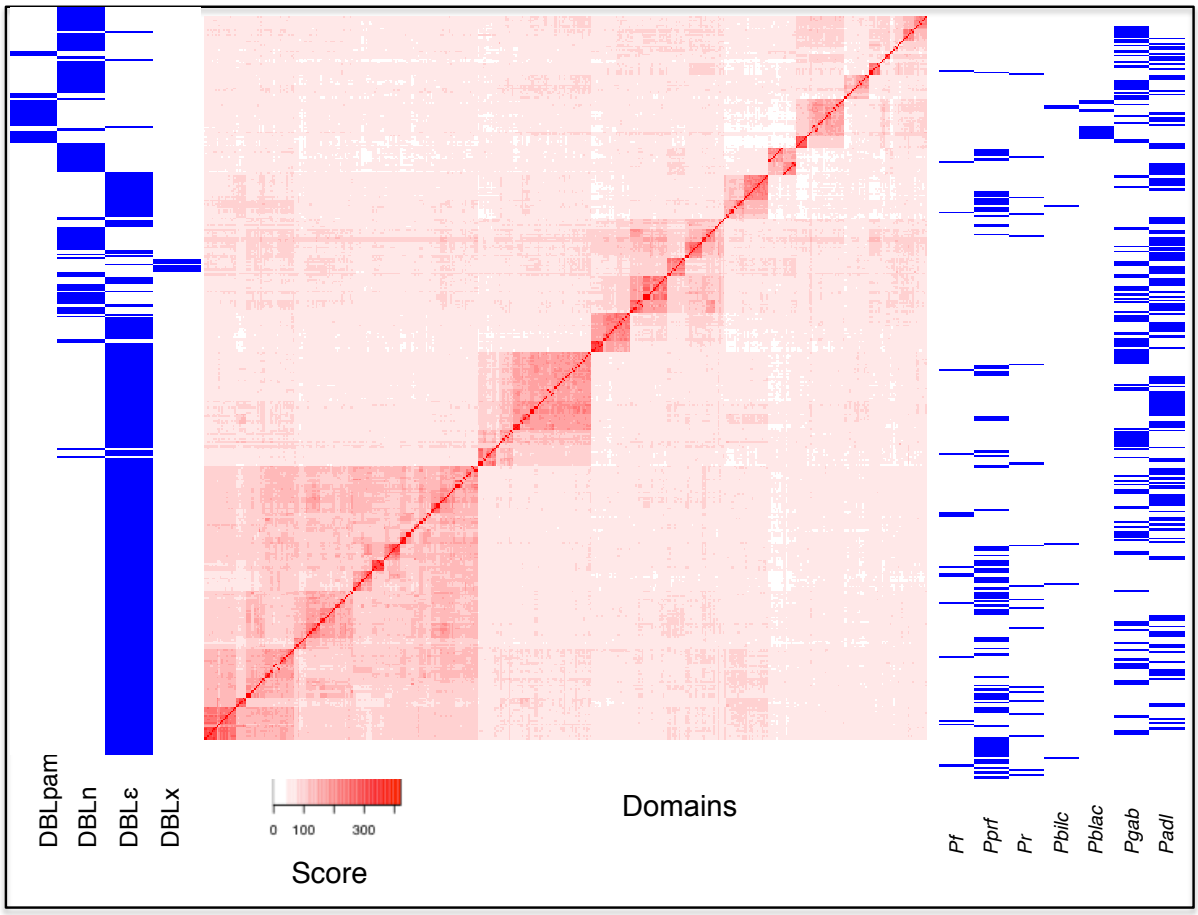

DBLpam
DBLn
DBLε
DBLx
0 100 300
Score
Domains
Pf
Pprf
Pr
Pbilc
Pblac
Pgab
Padl

Fig. S10. Diversity of *var* genes domains. Annotated similarity matrix between all the domains of the *Laverania* of Fig. 5, including their species attribution and their cluster attribution. The similarity matrix shows the score of the BLASTp between the domains. It was clustered with the ward2 algorithm in R. The dendrogram is on the left hand side. Each row (column) presents therefore one domain and shows its score of similarity to the other 2,547 domains and itself. To each row is associated on the left the domain type and on the right the species it comes from. (**A**) Similarity matrix based on the 2548 domains. This figures shows that the ATS (acidic terminal sequence) and NTS (N-terminal sequence) are very distinct. The DBL domains on the other side have more similarity between themselves. Importantly, the genes on the top are similar to each other, but they are annotated as different domains. The DBLn domain is a domain where the duffy binding like Pfam domain had a higher score than the other domains from the vardom server (http://www.cbs.dtu.dk/services/VarDom/). (**B**) To understand the diversity of DBLε and DBLn and to compare it to the newly described DBLx, a similarity matrix of all domains annotated as DBLε, DBLn, DBLpam1 and DBLx labeled sequences, see Supplemental Information C. It can be seen that have domains are similar to each other and that the DBLx labeled sequences as defined by Larremore et al(*9*) 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.*

Sample information					Sequencing				Multiple infection analysis (percentage of mapped reads)										
Species	Sample ID	Collection date	Primate species	Primate name (sex / age in years)	DNA extraction	# SMRT cells (PacBio)	% host contami-nation	Fold coverage Illumina	*P.praefalciparum*	*P. alderi*	*P. blacklocki*	*P.reichenowi*	*P. billcollinsi*	*P. gaboni*	*P.cynomolgi*	*P. knowlesi*	*P. vivax*	*P. berghei*	Other
P. praefalciparum	PprfG01	**Dec-13**	Chimp	**Buenga (F/8)**	CF11 / WGA	29	5.8*	104	42.63	6.89	0.23	0	0.03	0.65	0.01	0	0.01	0.27	49.29
	PprfG02	Sep-12	Gorilla	Djino (M/6)	CF11 / WGA		23.3	1269	20.21	0.26	0.06	1.39	0.08	0.05	0	0	0	0	77.94
	PprfG04		Multiple infection, see PadlG03						9.6	48.14	0.08	1.17	0.09	5.53	0.13	0.11	0.07	0	35.08
P. praefalciparum II	PprfG03	Aug-13	Gorilla	GG06 (F/na)	Cell sorting / noPCR		86.1	668	4.95	0	0.03	3.27	0.04	0.01	0.32	0.06	0	0	91.32
P. reichenowi	PrG01	**Dec-13**	Chimp	**Ebea (F/7)**	CF11 / WGA	15	7.1	106	0.59	0.35	0.67	91.2	0.03	0.27	0.01	0	0.73	0.03	6.12
	PrCDC**	1950s	Chimp	Dennis	standard		45.2	296	5.47	0.15	0.4	63.07	1.78	0.15	0.16	0.31	0.02	0.03	28.45
	PrG02	Dec-13	Chimp	Titus (M/3)	CF11 / WGA		5.7	123	2.58	0.38	0.89	72.36	0.35	0.33	0.01	0.01	0.4	0.02	22.66
	PrG03	Dec-13	Chimp	Cerise (F/7)	CF11 / WGA		7.9	112	2.74	0.41	0.96	73.61	0.31	0.37	0.02	0.01	0.95	0.02	20.61
P. billcollinsi	PbilcG01		Multiple infection, see PgabG02						0.27	4.72	0.25	0.9	26.34	42.13	0.13	0.09	1.03	0	24.14
	PbilcG03		Multiple infection, see PgabG04						0.42	3.98	0.15	8.69	13.59	53.59	0.01	0	3.91	0	15.66
	PbilcG04		Multiple infection, see PgabG05						0.98	5.24	0.26	14.11	12.28	57.83	0.02	0.02	0.82	0	8.46
P. blacklocki	PblacG01	**Jul-11**	Gorilla	**Djino (M/5)**	Frozen Blood sWGA + WGA	10	64.7	221	0.03	0.01	32.16	1.2	0.04	0.01	0.23	0.06	0	0	66.26
P. gaboni	PGAB01	**May-13**	Chimp	**Ebea (F/7)**	CF11 / WGA	16	14.0	367	0.06	1.15	0.03	2.11	0.54	62.55	0.23	0.08	17.98	0	15.28
	PgabG02	**May-13**	Chimp	**Cerise (F/7)**	CF11 / WGA	31	75.3	341	0.27	4.72	0.25	0.9	26.34	42.13	0.13	0.09	1.03	0	24.14
	PgabG03	Sep-12	Chimp	Lolita (F/6)	CF11 / WGA		6.5	669	0.21	4.39	0.06	5.01	1.96	56.63	0.07	0.01	12.86	0	18.8
	PgabG04	Sep-12	Chimp	Toudy (F/6)	CF11 / WGA		8.0	679	0.42	3.98	0.15	8.69	13.59	53.59	0.01	0	3.91	0	15.66
	PgabG05	May-13	Chimp	Nzigou (M/9)	CF11 / WGA		6.9	530	0.98	5.24	0.26	14.11	12.28	57.83	0.02	0.02	0.82	0	8.46
P. adleri	PadlG01	**May-13**	Gorilla	**Djino (M/7)**	CF11 / WGA	31	26.6	372	0.2	69.23	0.02	0.56	0.02	1.54	0.12	0.11	0.03	0	28.18
	PadlG02	Jun-11	Gorilla	Belinga (F/na)	Cell sorting/noPCR		83.7	2262	0.13	15.83	0.03	1.07	0.02	2.17	0.26	0.16	0	0	80.33
	PadlG03	May-13	Gorilla	Rapha (M/7)	CF11 / WGA		27.9	445	9.6	48.14	0.08	1.17	0.09	5.53	0.13	0.11	0.07	0	35.08

* Highest contaminant is *Drosophila melanogaster*

** Data from Otto et al[6]

Table S1. Multiple infections in Laverania samples

Further information on the primates. 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.

Population	Status	Coalescence, years (x 10^5)			Effective Population size, Ne (x 10^5)			Migration target	Probably of Parasite Migration		
		Genic	Intergenic	Intergenic, no UTR	Genic	Intergenic	Intergenic, no UTR		Genic	Intergenic	Intergenic, no UTR
P. falciparum	Present	NA	NA	NA	1.9 - 3.2 (1.8 - 3.4)	0.4 - 0.7 (0 - 2.1)	0.4 - 0.7 (0 - 2.1)	NA	0	0	0
P. praefalciparum	Present	NA	NA	NA	8.8 - 14.8 (8.1 - 16)	12.2 - 20.7 (11.6 - 21.8)	10 - 17 (9.2 - 18.4)	*Pf*	0.02 (0.0 - 0.03)	0.14 (0.09- 0.19)	0.16 (0.08- 0.27
P. reichenowi	Present	NA	NA	NA	11.3 - 19.2 (3.7 - 19.9)	15.9 - 26.9 (15.4 - 27.7)	12.3 - 20.7 (11.7 - 21.8)	NA	0	0	0
P. billcollinsi	Present	NA	NA	NA	9.3 - 15.8 (8.9 - 16.5)	-	-	NA	0	-	-
P. gaboni	Present	NA	NA	NA	18.1 - 30.6 (17.6 - 31.6)	-	-	NA	0	-	-
P. adleri	Present	NA	NA	NA	17.7 - 29.9 (16.9 - 31.2)	-	-	NA	0	-	-
Pf - Pprf	Ancestral	0.2 - 0.3 (0.1 - 0.3)	0.5 - 0.8 (0.4 - 0.8)	0.4 - 0.6 (0.3 - 0.7)	6.5 - 11.1 (6 - 11.9)	16.5 - 27.9 (15.4 - 29.7)	12.4 - 21 (11.1 - 23.2)	NA	0	0	0
Pf - Pprf - Pr	Ancestral	0.9 - 1.5 (0.8 - 1.5)	1.6 - 2.7 (1.5 - 2.7)	1.4 - 2.3 (1.3 - 2.4)	29.7 - 50.3 (27.7 - 53.7)	68.6 - 116.2 (65.4 - 121.6)	42.3 - 71.5 (39.1 - 77.1)	NA	0	0	0
Pf - Pprf - Pr - Pbilc	Ancestral	2.3 - 4 (2.3 - 4)	4.2 - 7.1*	3.7 - 6.2*	65.8 - 111.3 (62 - 118.5)	-	-	*Pgab - Padl*	0.61 (0.45 - 0.77)	-	-
Pgab - Padl	Ancestral	0.9 - 1.4 (0.8 - 1.5)	1.5 - 2.6*	1.3 - 2.2*	28.3 - 47.9 (26.7 - 50.7)	-	-	NA	0	-	-
Pf - Pprf - Pr - Pbilc - Pgab - Padl	Ancestral	4.5 - 7.6 (4.2 - 8)	8.1 - 13.6*	7 - 11.8*	98.9 - 167.4 (93.2 - 178.4)	-	-	NA	0	-	-

Table S2. Estimated dates of speciation (time to coalescence), effective population size (Ne) and migration rates for present and ancestral *Laverania* populations. 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 402–681 mitotic events per year. The number in the brackets is the range based on the prediction (95% confidence interval). Values with (*) are linear obtained estimates, as we were not able

to obtain good alignments 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

References:

1. F. Prugnolle *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 (Jan 26, 2010).
2. W. Liu *et al.*, Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. *Nature* **467**, 420 (Sep 23, 2010).
3. B. Ollomo *et al.*, A New Malaria Agent in African Hominids. *PLoS Pathog* **5**, e1000446 (2009).
4. W. Liu *et al.*, Multigenomic Delineation of Plasmodium Species of the Laverania Subgenus Infecting Wild-Living Chimpanzees and Gorillas. *Genome biology and evolution* **8**, 1929 (2016).
5. L. Boundenga *et al.*, Diversity of malaria parasites in great apes in Gabon. *Malar J* **14**, 111 (2015).
6. T. D. Otto *et al.*, Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nature communications* **5**, 4754 (2014).
7. S. A. Sundararaman *et al.*, Genomes of cryptic chimpanzee Plasmodium species reveal key evolutionary events leading to human malaria. *Nature communications* **7**, 11078 (2016).
8. M. J. Gardner *et al.*, Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498 (2002).
9. D. B. Larremore *et al.*, Ape parasite origins of human malaria virulence genes. *Nature communications* **6**, 8368 (2015).
10. M. A. Pacheco *et al.*, Timing the origin of human malarias: the lemur puzzle. *BMC evolutionary biology* **11**, 299 (2011).
11. D. M. Behar *et al.*, The dawn of human matrilineal diversity. *American journal of human genetics* **82**, 1130 (May, 2008).
12. S. K. Volkman *et al.*, Recent origin of Plasmodium falciparum from a single progenitor. *Science* **293**, 482 (Jul 20, 2001).
13. F. P. Palstra, D. J. Fraser, Effective/census population size ratio estimation: a compendium and appraisal. *Ecology and evolution* **2**, 2357 (Sep, 2012).
14. S. W. Roy, 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 (Dec, 2015).
15. K. Tanabe, M. Mackay, M. Goman, J. G. Scaife, Allelic dimorphism in a surface antigen gene of the malaria parasite Plasmodium falciparum. *Journal of molecular biology* **195**, 273 (May 20, 1987).
16. Y. Yasukochi, I. Naka, J. Patarapotikul, H. Hananantachai, J. Ohashi, Genetic evidence for contribution of human dispersal to the genetic diversity of EBA-175 in Plasmodium falciparum. *Malar J* **14**, 293 (Aug 01, 2015).
17. N. Malaria Genomic Epidemiology, G. Band, K. A. Rockett, C. C. Spencer, D. P. Kwiatkowski, A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature* **526**, 253 (Oct 8, 2015).
18. B. Makanga *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* **113**, 5329 (May 10, 2016).
19. M. Wanaguru, W. Liu, B. H. Hahn, J. C. Rayner, G. J. Wright, RH5-Basigin interaction plays a major role in the host tropism of Plasmodium falciparum. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 20735 (Dec 17, 2013).
20. K. E. Wright *et al.*, Structure of malaria invasion protein RH5 with erythrocyte basigin and blocking antibodies. *Nature* **515**, 427 (Nov 20, 2014).
21. T. Triglia, J. K. Thompson, A. F. Cowman, An EBA175 homologue which is transcribed but not translated in erythrocytic stages of Plasmodium falciparum. *Mol Biochem Parasitol* **116**, 55 (Aug, 2001).
22. R. S. Ramiro *et al.*, Hybridization and pre-zygotic reproductive barriers in Plasmodium. *Proceedings. Biological sciences / The Royal Society* **282**, 20143027 (May 7, 2015).
23. S. Eksi *et al.*, Malaria transmission-blocking antigen, Pfs230, mediates human red blood cell binding to exflagellating male parasites and oocyst production. *Mol Microbiol* **61**, 991 (Aug, 2006).

24. D. Cunningham, J. Lawton, W. Jarra, P. Preiser, J. Langhorne, The pir multigene family of Plasmodium: antigenic variation and beyond. *Mol Biochem Parasitol* **170**, 65 (Apr, 2010).
25. E. Mundwiler-Pachlatko, H. P. Beck, Maurer's clefts, the enigma of Plasmodium falciparum. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 19987 (Dec 10, 2013).
26. M. Niang *et al.*, STEVOR is a Plasmodium falciparum erythrocyte binding protein that mediates merozoite invasion and rosetting. *Cell host & microbe* **16**, 81 (Jul 9, 2014).
27. L. L. Bethke *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 (Nov, 2006).
28. M. Frank, R. Dzikowski, B. Amulic, K. Deitsch, Variable switching rates of malaria virulence genes are associated with chromosomal position. *Mol Microbiol* **64**, 1486 (Jun, 2007).
29. K. Hayton *et al.*, Erythrocyte binding protein PfRH5 polymorphisms determine species-specific pathways of Plasmodium falciparum invasion. *Cell Host Microbe* **4**, 40 (Jul 17, 2008).
30. P. C. Bull *et al.*, Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nat Med* **4**, 358 (Mar, 1998).
31. A. Scally, R. Durbin, Revising the human mutation rate: implications for understanding human evolution. *Nature reviews. Genetics* **13**, 745 (Oct, 2012).
32. R. Carter, K. N. Mendis, Evolutionary and historical aspects of the burden of malaria. *Clinical microbiology reviews* **15**, 564 (Oct, 2002).
33. M. Coluzzi, A. Sabatini, A. della Torre, M. A. Di Deco, V. Petrarca, A polytene chromosome analysis of the Anopheles gambiae species complex. *Science* **298**, 1415 (Nov 15, 2002).
34. J. Flint, R. M. Harding, A. J. Boyce, J. B. Clegg, The population genetics of the haemoglobinopathies. *Bailliere's clinical haematology* **6**, 215 (Mar, 1993).
35. S. Auburn *et al.*, An effective method to purify plasmodium falciparum dna directly from clinical blood samples for whole genome high-throughput sequencing. *PLoS ONE* **6**, (2011).
36. S. O. Oyola *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 (Dec, 2014).
37. S. O. Oyola *et al.*, Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. *bioRxiv*, (2016).
38. A. Boissiere *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).
39. M. A. Quail *et al.*, in *Nat Methods*. (United States, 2012), vol. 9, pp. 10-1.
40. H. Manske, D. Kwiatkowski, SNP-o-matic. *Bioinformatics 25*, 2434 (2009).
41. C. S. Chin *et al.*, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods* **10**, 563 (Jun, 2013).
42. S. Assefa, T. M. Keane, T. D. Otto, C. Newbold, M. Berriman, ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics Vol. 25 No. 15*, 1968 (2009).
43. T. Carver *et al.*, Artemis and ACT: viewing, annotation and comparing sequences stored in relational database. *Bioinformatics* **24**, 2672 (2008).
44. T. D. Otto, M. Sanders, M. Berriman, C. Newbold, Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics 26(14)*, 1704 (2010).
45. A. C. English *et al.*, Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**, e47768 (2012).
46. T. D. Otto, From sequence mapping to genome assemblies. *Methods Mol Biol* **1201**, 19 (2015).
47. T. D. Otto, G. P. Dillon, W. S. Degrave, M. Berriman, RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Research*, 1 (2011).
48. M. Stanke *et al.*, AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435 (Jul 1, 2006).
49. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754 (Jul 15, 2009).
50. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297 (Sep, 2010).
51. P. Danecek *et al.*, The variant call format and VCFtools. *Bioinformatics* **27**, 2156 (Aug 1, 2011).
52. L. Li, C. J. Stoeckert, Jr., D. S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178 (Sep, 2003).

53. G. Jordan, N. Goldman, The Effects of Alignment Error and Alignment Filtering on the Sitewise Detection of Positive Selection. *Molecular biology and evolution* **29**, 1125 (Apr, 2012).
54. A. Loytynoja, N. Goldman, 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 (Jul 26, 2005).
55. A. Loytynoja, N. Goldman, Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632 (Jun 20, 2008).
56. W. Fletcher, Z. Yang, The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection. *Molecular biology and evolution* **27**, 2257 (October 1, 2010, 2010).
57. P. Markova-Raina, D. Petrov, 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 (Jun, 2011).
58. A. Morgulis, E. M. Gertz, A. A. Schäffer, R. Agarwala, A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. *Journal of Computational Biology* **13**, 1028 (2006/06/01, 2006).
59. J. C. Wootton, S. Federhen, Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry* **17**, 149 (1993/06/01, 1993).
60. J. Castresana, Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular biology and evolution* **17**, 540 (April 1, 2000, 2000).
61. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312 (May 1, 2014, 2014).
62. H. Shimodaira, M. Hasegawa, Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Molecular biology and evolution* **16**, 1114 (August 1, 1999, 1999).
63. T. A. Castoe *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 (Jun, 2009).
64. G. W. C. Thomas, M. W. Hahn, Determining the Null Model for Detecting Adaptive Convergence from Genomic Data: A Case Study using Echolocating Mammals. *Molecular biology and evolution* **32**, 1232 (May, 2015).
65. Z. Yang, PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular biology and evolution* **24**, 1586 (August 1, 2007, 2007).
66. J. Zhang, R. Nielsen, Z. Yang, Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Molecular biology and evolution* **22**, 2472 (December 1, 2005, 2005).
67. J. H. McDonald, M. Kreitman, Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652 (1991).
68. A. A. a. J. Rahnenfuhrer, topGO: Enrichment analysis for Gene Ontology. *R package* **version 2.8.0**, (2010).
69. S. Guindon *et al.*, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307 (May, 2010).
70. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792 (2004).
71. M. Gouy, S. Guindon, O. Gascuel, SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution* **27**, 221 (Feb, 2010).
72. T. Carver *et al.*, BamView: visualizing and interpretation of next-generation sequencing read. *Briefings in bioinformatics*, (Jan 24, 2013).
73. T. S. Rask, D. A. Hansen, T. G. Theander, A. Gorm Pedersen, T. Lavstsen, Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes--divide and conquer. *PLoS Comput Biol* **6**, (2010).
74. C. UniProt, UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204 (Jan, 2015).
75. A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, G. J. Barton, Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189 (May 1, 2009).
76. A. Claessens *et al.*, Generation of antigenic diversity in Plasmodium falciparum by structured rearrangement of Var genes during mitosis. *PLoS Genet* **10**, e1004812 (Dec, 2014).
77. I. Gronau, M. J. Hubisz, B. Gulko, C. G. Danko, A. Siepel, Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet* **43**, 1031 (Oct, 2011).
78. S. Schiffels, R. Durbin, Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919 (Aug, 2014).

79. S. C. Nkhoma *et al.*, Population genetic correlates of declining transmission in a human pathogen. *Molecular ecology* **22**, 273 (Jan, 2013).
80. D. A. Joy *et al.*, Early origin and recent expansion of Plasmodium falciparum. *Science* **300**, 318 (Apr 11, 2003).
81. T. Ponnudurai *et al.*, Sporozoite load of mosquitoes infected with Plasmodium falciparum. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **83**, 67 (Jan-Feb, 1989).
82. D. Mazier *et al.*, Complete development of hepatic stages of Plasmodium falciparum in vitro. *Science* **227**, 440 (Jan 25, 1985).
83. N. Gerald, B. Mahajan, S. Kumar, Mitosis in the human malaria parasite Plasmodium falciparum. *Eukaryotic cell* **10**, 474 (Apr, 2011).
84. A. H. Freedman *et al.*, Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet* **10**, e1004016 (Jan, 2014).
85. H. H. Chang *et al.*, Malaria life cycle intensifies both natural selection and random genetic drift. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 20129 (Dec 10, 2013).
86. E. K. Karlsson, D. P. Kwiatkowski, P. C. Sabeti, Natural selection and infectious disease in human populations. *Nature reviews. Genetics* **15**, 379 (Jun, 2014).
87. G. Talavera, J. Castresana, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564 (Aug, 2007).
88. S. Guindon, F. Delsuc, J. F. Dufayard, O. Gascuel, Estimating maximum likelihood phylogenies with PhyML. *Methods Mol Biol* **537**, 113 (2009).
89. A. Rambaut. (2014).
90. R. D. C. Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing (2008).
91. D. H. Haft, J. D. Selengut, O. White, The TIGRFAMs database of protein families. *Nucleic Acids Res* **31**, 371 (Jan 1, 2003).
92. L. S. Johnson, S. R. Eddy, E. Portugaly, Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).