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

Authors: Thomas D. Otto[1,†,#,*], Aude Gilabert[2,†], Thomas Crellen[1,3], Ulrike Böhme[1], Céline Arnathau[2], Mandy Sanders[1], Samuel Oyola[1,4], 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,*] & Franck Prugnolle[2,5,*]

Affiliations:

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

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

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

[4] International Livestock Research Institute, Box 30709, Nairobi, Kenya (current address)

[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

Current Address: Centre of Immunobiology, Institute of Infection, Immunity & Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom

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

[†]These authors contributed equally.

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

Main Text:

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

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 therefore could not have both been transmitted if a new species were founded by a single individual infection. Also, the dating of the speciation cannot be accurately estimated without the genome sequence of *P. praefalciparum,* the closest living sister species to *P. falciparum*.

The absence of *in vitro* culture or a usable animal model has precluded obtaining sufficient DNA for full genome sequencing and has hindered investigation of the *Laverania.* So far just the full 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.

73 **Genome sequencing from six *Laverania* species**

74 Blood samples were taken during successive routine sanitary controls, from four gorillas and
75 seven chimpanzees living in a sanctuary or quarantine facility prior to release (see Methods). A
76 total of 15 blood samples were positive for ape malaria parasites by PCR. Despite low parasitemia
77 in most animals, a combination of host DNA depletion, parasite cell sorting and amplification
78 methods enabled sufficient parasite DNA templates to be obtained for short-read (Illumina) and
79 long read (Pacific Bioscience) sequencing (Table 1). Mixed-species infections were frequent but
80 resolved by utilising sequence data from single infections, resulting in 19 genotypes
81 (Supplementary Table 1). The dominant genotype in each sample was assembled *de novo* (see
82 Methods) using long read technology into a reference genome for six malaria parasite species: *P.*
83 *praefalciparum*, *P. blacklocki*, *P. adleri, P. billcollinsi*, *P. gaboni* and *P. reichenowi*. The
84 assemblies comprised 44-97 scaffolds (Table 1), with large contigs containing the subtelomeric
85 regions and internal gene clusters that house multigene families known in *P. falciparum* and *P.*
86 *reichenowi* to be involved in virulence and host-pathogen interactions. The high quality of the
87 assemblies compared to those obtained previously is illustrated by the good representation of multi-
88 gene family members (Supplementary Table 2) and the large number of one-to-one orthologues
89 obtained between the different reference genomes (4,350 among the seven species and 4,826
90 between *P. falciparum*, *P. praefalciparum* and *P. reichenowi*). Two to four additional genomes
91 were obtained for each species except for *P. blacklocki* (Table 1, Supplementary Table 1).
92

93 **Speciation history in the *Laverania* sub-genus**

94 Conservation of gene content and synteny is striking between these complete genomes and
95 enabled us to reconstruct with confidence the relationships between different *Laverania* species, to
96 compare their relative genetic diversity (Fig 2a, Supplementary Fig. 1) and to estimate the age of
97 the different speciation events that led to the extant species. The latter has been problematic in the
98 past due to the lack of both complete genome data and accurate estimates of mutation rate and
99 generation time. Using the most divergent estimates of generation time and measured mutation
100 rates from the *P. falciparum* literature, we found the data converge to 0.9-1.5 mutations per year per
101 genome (Supplementary Note 1). We observed a similar substitution rate *in vivo* by examining
102 existing sequence data for five geographically diverse isolates, covering a 200-kb region
103 surrounding the PfCRT gene that is relatively conserved due to a selective sweep resulting from
104 chloroquine use (Supplementary Note 1; Supplementary Figure 2). The fact that these two figures
105 are similar suggests that the *in vitro* mutation rate may have been underestimated since many

106 mutations will be lost by genetic drift. Since no data is available for the other species we have
107 assumed hereon that these values generalise across the subgenus. From our Bayesian whole-genome
108 estimates, the ancestor of all current day parasites of this subgenus existed 0.7–1.2 million years
109 ago, a time at which the subgenus divided into two main clades, A (*P. adleri* and *P. gaboni*) and B
110 that includes the remaining species (Fig. 1a). Our range of values is far more recent than previous
111 estimates[3,11]. Following the Clade A/B subdivision, several speciation events occurred leading
112 either to new chimpanzee or gorilla parasites. Interestingly, the divergence between *P. adleri* and *P.*
113 *gaboni* in one lineage and *P. reichenowi* and the ancestor of *P. praefalciparum/P. falciparum* in the
114 other lineage occurred at approximately the same time (140–230 thousand years ago; Fig. 1a, Table
115 2). Based on our coalescence estimates, *P. falciparum* begun to emerge in humans from *P.*
116 *praefalciparum* around 40–60 thousand years ago (Fig. 1a), significantly later than the evolution of
117 the first modern humans and their spread throughout Africa[12]. Our analysis also indicates
118 significant gene flow between these two parasite species after their initial divergence (Table 2).
119 *P. falciparum* has strikingly low diversity (π =0.0004), compared with the other *Laverania*
120 species (0.002–0.0049) (Supplementary Fig. 1). It has been proposed that *P. falciparum* arose from
121 a single transfer of *P. praefalciparum* into humans[6] and based in part on the paucity of neutral SNPs
122 within the genome, that *P. falciparum* emerged from a bottleneck of a single parasite around 10,000
123 years ago, after agriculture was established[6,8]. In light of our results, we estimate that the
124 *P. falciparum* population declined around 11,000 years ago and reached a minimum about 5,000
125 years ago (Fig. 1b) with an effective population size (*Ne*) of around 3,000 (Supplementary Note 1;
126 generally the census number of parasites is higher than *Ne*[13]). It is important to note that the
127 methods we used to estimate *Ne* do not rely on the Fisher-Wright model, that could give problems
128 as previously reported[14]. The hypothesis of a single progenitor is also inconsistent with the
129 observation of several ancient gene dimorphisms that have been observed in *P. falciparum*. A
130 previous analysis using *P. reichenowi* and *P. gaboni* sequence data, provided some evidence that
131 different dimorphic loci diverged at different points in the tree[15]. Looking at each of these
132 *P. falciparum* loci across the *Laverania*, we found different patterns of evolution at the *msp1*,
133 *var1csa*, and *msp3* loci (Supplementary Fig. 3a). Most strikingly, a mutually exclusive dimorphism
134 (described as MAD20/K1[16]) in the central 70% of the *msp1* sequence, clearly pre-dates the
135 *P. falciparum–P. praefalciparum* common ancestor and dimorphism in *var1csa* (an unusual *var*
136 gene of unknown function that is transcribed late in the asexual cycle) occurred before the split with
137 *P. reichenowi*.

138 In contrast, the gene *eba-175* that encodes a parasite surface ligand involved in red blood cell
139 invasion contains a dimorphism that arose after the emergence of *P. falciparum* (Supplementary
140 Fig. 3b). The time to the most recent common ancestor of *eba-175* has been estimated as 130–140
141 thousand years in an analysis[17] that assumed *P. falciparum* and *P. reichenowi* diverged 6 million
142 years ago. However, based on our new estimate for *P. falciparum–P. reichenowi* divergence, we
143 recalibrated their estimate of the most recent common ancestor of the *eba-175* alleles to be around
144 4,000 years ago, which is in good agreement with our divergence time for *P. falciparum*
145 (Supplementary Note 1). The recent dimorphism cannot however explain the observation of an
146 ancient dimorphism near the human and ape loci for glycophorin[18] – an EBA-175 binding protein.
147 The formation and maintenance of all of these dimorphic loci has therefore been shaped by different
148 balancing selection pressures over time.
149
150 ***P. falciparum*-specific evolution**
151
152 During its move away from gorillas, *P. falciparum* had to adapt to new environmental
153 conditions, namely a new vertebrate host (human) and new vector species (e.g. *Anopheles*
154 *gambiae*)[19]. To infer *P. falciparum* specific adaptive changes, we considered the *P. falciparum* /
155 *P. praefalciparum* /*P. reichenowi* genome trio and then applied two lineage based tests to find
156 positive selection that occurred in the *P. falciparum* branch (see methods). The two tests identified
157 172 genes (out of 4,826) with signatures of positive selection in the human parasite species only
158 (Supplementary Table 3). Two genes (*rop14* and PF3D7_0609900) were significant in both tests.
159 Among the 172 genes, almost half (n=82) encoded proteins of unknown function. Analysis of those
160 with functional annotation indicated that genes involved in pathogenesis, entry into host, actin
161 movement and organization and in drug response were significantly over-represented. Other genes,
162 expressed in different stages of the *P. falciparum* life cycle (e.g. *sera4* and *emp3*, involved during
163 the erythrocytic stages; *trsp* and *lisp*1, involved in the hepatic stages; and *plp4*, *CelTOS* and
164 *Cap380*, involved in the mosquito stages) also showed a significant signal of adaptive evolution
165 (Supplementary Table 3).
166
167 **Evolution through introgression, gene transfer and convergence**
168 Frequent mixed species infections in apes and mosquitoes[19] provide clear opportunities for
169 interspecific gene flow between these parasites. A recent study[6] reported a gene transfer event
170 between *P. adleri* and the ancestor of *P. falciparum* and *P. praefalciparum* of a region on
171 chromosome 4 including key genes involved in erythrocyte invasion (*rhr5* and *cyrpA*). Because

172 such events preserve the phylogenetic history of the genes involved, we systematically examined
173 the evidence for introgression or gene transfer events across the complete subgenus by testing the
174 congruence of each gene tree to the species tree for genes with one-to-one orthologues. Beyond the
175 region that includes *rh5* (Fig 2b, Supplementary Fig. 4a), few signals of gene flow between
176 parasites infecting the same host species were obtained (n=11) suggesting that these events were
177 rare or usually strongly deleterious (Supplementary Fig. 5).
178 The *Laverania* subgenus evolved to infect chimpanzees and gorillas but, on a genome-wide
179 scale, the convergent evolution of host-specific traits has not left a signature (Supplementary Note
180 2). We therefore examined each CDS independently and were able to identify genes with
181 differences fixed within specific hosts, falling into three categories: 53 in chimpanzee-infective
182 parasites, 49 in gorilla-infective and 12 with fixed traits in both host species (Fig. 2; Supplementary
183 Table 4a). For at least 67 genes, these differences were unlikely to have arisen by chance ($p < 0.05$)
184 and GO term enrichment analysis revealed that several of these genes are involved in host invasion
185 and pathogenesis (Supplementary Table 4b) including *rh5* (which has a signal for convergent
186 evolution even when the introgressed tree topology is taken into consideration; Supplementary Fig.
187 4b). *Rh5* is the only gene identified in *P. falciparum* that is essential for erythrocyte recognition
188 during invasion, via binding to Basigin. *P. falciparum rh5* cannot bind to gorilla Basigin and binds
189 poorly to the chimpanzee protein[20]. We notice that one of the convergent sites is known to be a
190 binding site for the host receptor Basigin[21] (Supplementary Fig. 4b). The gene *eba-165* encodes a
191 member of the erythrocyte binding like (EBL) super family of proteins that are involved in
192 erythrocyte invasion. Although *eba-165* is a pseudogene in *P. falciparum*[22], it is not a pseudogene
193 in the other *Laverania* species and may therefore be involved in erythrocyte invasion, like other
194 EBL members. The protein has three convergent sites in gorillas. One falls inside the F2 region, a
195 domain involved in the interactions with erythrocyte receptors. The role of this protein and of these
196 convergent sites in the invasion of gorilla red cells remains to be determined. Finally, genes
197 involved in gamete fertility (the 6-cysteine protein P230) or implicated in *Plasmodium* invasion of
198 erythrocytes (doc2[23]) also displayed signals of convergent evolution. Twelve parasite coding
199 sequences had fixed differences at the same amino-acid position in chimpanzees and gorillas. Of
200 these P230 was the only one found with a position that was different and fixed across all three host
201 species. P230 is involved in gamete development and trans-specific reproductive barriers[24], possibly
202 through enabling male gametes to bind to erythrocytes prior to exflagellation[25]. Host-specific
203 residues observed in P230 might affect the efficiency of the binding to the erythrocyte receptors and
204 result from co-evolution between the parasite molecule and the host receptor.

205

206 **Subtelomeric gene families**

207 To date, the only in-depth data on the subtelomeric gene families of the *Laverania* have come
208 from *P. reichenowi* and *P. falciparum*. As a result of using long read sequencing, these families are
209 well represented in our assemblies (Supplementary tables 2 and 5A). We provide for the first time a
210 comprehensive picture of the evolution of these important families.

211 Most gene families were likely present in the ancestor of all Laverania. The same general
212 pattern of one-to-one orthology throughout the subgenus indicates that many underwent gene
213 duplication early (e.g. FIKK) or prior (e.g. ETRAMP, PHIST and SURFIN) to the development of a
214 distinct Laverania lineage. Only a subset displayed contractions or expansions between specific
215 Laverania species (Fig. 3 and Supplementary Table 5a, Supplementary Figure 7). For these latter
216 families, Clade A and most species of Clade B clearly differ in their composition. *P. blacklocki*
217 (Clade B) is intermediate in its composition. Some gene families, like the group of exported
218 proteins *hyp4, hyp5*, *mc-2tm* and *EPF1*, have expanded only in *P. praefalciparum* and *P.*
219 *falciparum* (and even more in *P. falciparum* for *hyp4* and *hyp5*). Since all four are components of
220 Maurer's clefts, an organelle involved in protein export[26], some evolution of function in this
221 organelle may have been an important precursor to human infection. The family of acyl-CoA
222 synthetase genes, reported to be expanded and diversified in *P. falciparum*[27] is in fact expanded
223 across the *Laverania* and has four fewer copies in *P. falciparum* (Supplementary Fig. 6). Other
224 genes that show clade or group specific expansion include DBLmsp, glycophorin binding protein
225 and CLAG (Supplementary Fig. 7).

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

234 In contrast, a subset of *stevor* genes showed strong host-specific sequence diversification (Fig
235 4 and Supplementary Fig. 8). Based on full-length alignments, there is a deep phylogenetic split
236 between *stevor* genes but when only short conserved protein motifs are considered, a group of
237 Stevor proteins (*stevor II,* Fig 4a) forms a cluster comprising almost entirely of members from

238 gorilla-infecting species. Since *stevor* genes are known to be involved in host–parasite interactions
239 (such as binding to host glycophorin C in *P. falciparum*[29]), this host specific sequence may reflect
240 sequence differences in host-specific factors in gorillas.
241
242 **Evolution of *var* genes**
243 The *var* genes, crucial mediators of pathogenesis and the establishment of chronic infection through
244 cytoadherence and immune evasion, are the best studied *P. falciparum* multi-gene family and
245 unique to the *Laverania*[30]. They are two-exon genes and their products have three types of major
246 domain; exon 1 encodes Duffy Binding like (DBL) and Cysteine Rich Interdomain Regions (CIDR)
247 and exon 2 encodes Acidic Terminal Sequence (ATS)[31]. Similar to *P. falciparum*, our data are
248 consistent with all *Laverania* species having *var* genes (Fig. 3) that retain a two-exon structure and
249 are organized into subtelomeric or internal *var* gene clusters. There are however three notable
250 features of the evolution of this family within the sub-genus.
251 First, there is a deep division in how the repertoire is organised between the major clades. The
252 *var* genes of Clade B parasites, with the exception of *P. blacklocki*, resemble those of *P. falciparum*
253 in terms of genomic organisation, domain types and numbers (Fig 5, Supplementary Table 6). In
254 contrast, the repertoires of Clade A parasites and *P. blacklocki* (treated as one group hereafter in
255 this section) differ in their domain composition, contain a novel CIDR-like domain (CIDRn, Fig 5a,
256 Supplementary Fig. 9) and have lower sequence diversity per domain but cluster into more sub-
257 groups than Clade B domains (Fig 5b, Supplementary Fig. 10). The paucity of domains similar to
258 those in *P. falciparum* (such as CIDRα) that are involved in cytoadherence to some specific and
259 common host receptors, means that if endothelial cytoadherence was important in Clade A, some
260 alternative receptors must have been utilised.
261 Second, in total there are 10 internal *var* gene clusters (confirmed by contiguous sequence data)
262 but 8 are oppositely oriented between the two clades (Supplementary Fig. 11, Supplementary Table
263 7). Clade B parasites also show a much greater number of associated GC-rich RNAs of unknown
264 Function (RUF) elements than Clade A (Supplementary Table 7).
265 Third, the ATS domains cluster tightly within Clade A. Within Clade B there is clear evidence
266 of species-specific diversification, except in *P. praefalciparum* and *P. falciparum* reflecting their
267 recent speciation. There is one intact ATS from *P. falciparum* as well as several pseudogenes that
268 cluster with Clade A (Fig 5c). Moreover, of seven internal *var* arrays (Supplementary Fig. 11) in
269 *P. falciparum*, containing a functional *var* gene, five terminate with one of these pseudogenes (on
270 the opposite DNA strand) suggesting that they may be remnants from ancient rearrangements. The

271 intact *P. falciparum* gene is *var2csa*, a *var*-like gene that is highly conserved between *P. falciparum*
272 isolates[32], involved both in cytoadherence in the placenta in primigravidae, and proposed to be a
273 central intermediate in *var* gene switching during antigenic variation[33]. We therefore propose
274 *var2csa* is a remnant of an ancient multigene family that has been maintained as a single complete
275 gene in *P. falciparum*, for the dual purposes of *var*-switching and placental cytoadherence.
276 There is other evidence of retention of ancient *var* gene sequence across the subgenus. First,
277 in Clade B we find a nearly full length *var* pseudogene that has highest similarity to *P. adleri* and *P.*
278 *gaboni var* genes, within an internal *var* cluster on chromosome 4 in *P. falciparum* and *P.*
279 *praefalciparum* but on the opposite strand to the other *var* genes. It is found in all *P. falciparum*
280 isolates, but not in *P. reichenowi*. Second, in *P. gaboni* and *P. adleri,* three genes have the N-
281 terminal DBLα/CIDRα architecture typical of Clade B genes and their domains cluster within Clade
282 B based on similarity (Fig. 5b, larger nodes). Directly adjacent to two of these *var* genes are two *rif*
283 pseudogenes that also show greatest similarity to those from Clade B. Last, we find a further nine
284 *rif* pseudogenes of Clade A parasites that cluster with Clade B *rif* genes (Fig. 4). If these
285 observations reflect retention of ancient copies, their high sequence conservation suggests that they
286 are under extremely unusual selection pressure. Alternatively, they may represent relics of gene
287 transfer between species that occurred after the Clade A/B split.
288
289 **Conclusion**
290 We have produced high quality genomes and used mutation rates and generation times, covering the
291 full range of most recent estimates, to calculate the date of speciation for all known members of the
292 *Laverania*, with only a small margin of error. In our analysis, we have shown that the successful
293 infection of humans by *P. falciparum* occurred quite recently and involved numerous parasites
294 rather than a single one as previously proposed. After the establishment in its new host, the parasite
295 population went through a bottleneck around 5,000 years ago during the period of rapid human
296 population expansion due to farming (Fig. 1b). We summarise the major genomic events during the
297 evolution of the *Laverania* in Fig. 6.
298 As a result of our analyses we propose the following series of events for the emergence of *P.*
299 *falciparum* as a major human pathogen. First, the crucial lateral transfer event of the *rh5* locus
300 between Clade A and B parasites may also have involved *var* and *rifin* genes in other parts of the
301 genome that, because of their orientation on the opposite strand, were not lost during later
302 recombination. Next, facilitatory mutations are likely to have occurred in *rh5* that in the first
303 instance allowed invasion of both gorilla and human red cells. Modern humans emerged more than

300,000 years ago[34] and existed as small isolated populations[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[35]. Therefore, the existing human infective (*P. falciparum)* genotypes would be selected for human and appropriate vector success and the fittest would rapidly expand. Subsequent rapid accumulation of mutations that favoured growth in humans, and in the anthropophilic vectors such as *An. gambiae*, are likely to have occurred to increase human-specific reproductive success. The resulting specific parasite genotypes that expanded (and appeared as an emergence from a bottleneck), would have had a much lower probability of a direct transfer back to apes. With experiments on gorillas and chimpanzees not possible it will be difficult directly to prove the precise combination of different alleles that allowed the emergence of *P. falciparum*. However, existing data suggest that the genes that we have implicated in this process are expressed throughout the life cycle but that half remain uncharacterised (www.genedb.org, www.plasmodb.org), opening up new opportunities for future studies on host specificity and host adaptation in *Plasmodium*.

323 **Online Methods**

324 **Sample collection**

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

343

344 **Sample preparation**

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

351

352 **Sample sequencing**

353 All samples were first sequenced with Illumina. Amplification-free Illumina libraries of 400-
354 600 bp were prepared from the enriched genomic DNA[40] and run on MiSeq and HiSeq 2000 (v3
355 chemistry) Illumina machines.

356 After the Illumina sequencing, six samples with a combination of the least number of multiple
357 infections (see below) and the lowest level of host contamination were chosen for long read
358 sequencing, using Pacific Biosciences (PacBio). The DNA of the samples (after WGA) was size-
359 selected to 8 kb and sequenced with the C3/P5 chemistry. The number of SMRT cells (Pacific
360 Bioscience sequencing runs) used varied between samples (Supplementary Table 1).
361
362 **Genome assembly, genome QC, split of infection & annotation**
363 *Determination of multiple infections*
364 To initially quantify multiple infections and so allow samples to be selected for PacBio
365 sequencing from those comprising a low number of species, Illumina reads from each sample were
366 mapped against a concatenation of all available *Cox 3* and *CytB* genes of the *Laverania* from NCBI,
367 using SNP-o-Matic[41] (parameter chop=5) to position reads only where they aligned perfectly. SNP-
368 o-Matic returns all the positions of repetitive mapping reads. This output allowed us to count the
369 read depth of these two genes across all species and therefore determine the number and relative
370 amount of different malaria species per sample.
371
372 *Whole genome amplification (WGA) bias*
373 The uneven coverage that resulted from WGA bias, host contamination and multiple infections
374 presented a challenge for sequence assembly. To overcome the bias and the host contamination,
375 each DNA sample was sequenced deeper than normally necessary. Lower coverage of the
376 subtelomeres was obtained for the sWGA sample (*P. blacklocki*) meaning that the subtelomeres in
377 that assembly were not as complete as those in the assemblies for other species.
378
379 *Long reads (Pacific Bioscience) assemblies*
380 Six reference genomes were assembled using HGAP[42], with different settings for the genome
381 size parameter, ranging from 23 Mb (*P. reichenowi*) to 72 Mb (*P. billcollinsi*). This parameter
382 encodes how many long reads are corrected for use in the assembly and depends on the host
383 contamination and the amount of different isolates in the samples. The obtained contigs from
384 HGAP were ordered with ABACAS[43] against a *P. falciparum* 3D7 reference that has no
385 subtelomeric regions. Assembly errors and WGA artefacts were manually corrected using ACT[44].
386 After this step, three iterations of ICORN2[45] were run, followed by another ABACAS step,
387 allowing overlapping contigs to be merged (parameter: ABA_CHECK_OVERLAP=1). For the

388 PrG01, PgabG01 and PadlG01 assembly, we also ran PBjelly to close some of the sequencing
389 gaps[46].
390
391 *Host decontamination*
392 To detect and remove sequence data derived from host DNA, contigs were compared with the
393 chimpanzee or gorilla genomes using BLAST. Contigs were considered as host contamination if
394 more than 50% of their BLAST hits had higher than 95% identity to any of the great ape genomes.
395 Unordered contigs with a GC content >32% were searched against the non-redundant nucleotide
396 database, to detect and remove further contaminants.
397
398 *Resolving multiple infections*
399 The first assembled genome was a single *P. reichenowi* infection, PrG01. We detected low
400 levels of *P. vivax-like* and virus contamination (TT virus, AB038624.1), which were excluded. For
401 quality control, the assembly was compared against the existing PrCDC[9] reference genome. The
402 number of *Plasmodium* interspersed repeats (PIRs) was similar, and there were no breaks in
403 synteny. There were however significantly fewer sequencing gaps and 17 Rep20 regions could be
404 found (a known repeat close to the telomeres in *P. falciparum)*. Thus, the assembly of PacBio data
405 (PrG01; Supplementary Table 2) appears to be of higher quality than the existing *P. reichenowi*
406 PrCDC reference.
407 The *P. adleri* sample comprised a single infection. Because a large number of cycles of
408 amplification were used, a greater number of SMRT cells were sequenced (Supplementary Table 1)
409 to overcome the problem of uneven coverage resulting in under-represented regions. An estimated
410 genome size of 60 Mb was chosen for the HGAP analysis to ensure that all regions were covered.
411 PgabG01 was a *P. gaboni* isolate with a *P. vivax-like* co-infection. To detect contigs of *P.*
412 *vivax*, unordered contigs (those that could not be placed against Pf3D7 using ABACAS) were
413 searched against the protein sequences of *P. falciparum* 3D7 and the *P. vivax* PvP01 reference
414 genome using TBLASTx. For each contig, the relative number of genes hitting against the two
415 genomes was used to assign it to *P. gaboni* or *P. vivax*. In most cases, all genes for a given contig
416 consistently hit only one genome so that the attribution to either species was clear. Overall, 14 Mb
417 of *P. vivax-like* sequences were obtained that will be described elsewhere.
418 The *P. billcollinsi* genome (PbilcG01) was obtained from a co-infection with a *P. gaboni*
419 genome (PgabG02). Rather than ordering the contigs just against Pf3D7 with ABACAS, contigs
420 were ordered against a combined reference comprising *P. gaboni* (PgabG01) and the Pf3D7

421 (parameters: overlap 500 bp, identity 90%). The species designation of contigs was confirmed with
422 a TBLASTx searches of annotated genes against a combination of the proteomes of PgabG01 and
423 PrCDC. For subtelomeric gene families, contigs were attributed to species if the hit was significant
424 for one species, not the other. Some of the contigs could not be attributed unambiguously and were
425 discarded. Due to sequencing gaps, some of the core genes are missing from the final assembly.
426 The sample used to produce the *P. praefalciparum* genome (PprfG01) had a high level of host
427 contamination, a low level of co-infection with *P. adleri* and contained two distinct
428 *P. praefalciparum* genotypes. For the core genome, we used iCORN to identify and utilise the
429 dominant genotype at each position. Where it was not possible to phase the genotypes, due to a lack
430 of variation, we assumed that they were identical. In the subtelomeres however, it was possible to
431 distinguish but not phase the two *P. praefalciparum* genotypes resulting in approximately twice the
432 number of *var* genes as seen in *P. falciparum*. Due to contamination of construction vectors (*E.*
433 *coli*) and host, 29 SMRT cells were sequenced and the HAGP parameter for the assembly size was
434 set to 60 Mb. Contigs were screened against *P. adleri* and *P. falciparum* to exclude a *P. adleri* co-
435 infection. All of the contigs that had a *P. falciparum* BLAST hit or had no clear hit (such as those
436 containing species-specific gene families) were attributed to the *P. praefalciparum* assembly. Last,
437 all samples (Supplementary Table 1) including five *P. falciparum* genomes were mapped against
438 the Pf3D7, *P. praefalciparum and P. adleri* assemblies. Contigs were excluded where more
439 normalized hits to the three *P. adleri* samples were found than to one of the two other *P.*
440 *praefalciparum* samples. Similarly, this method was used to eliminate the remote possibility that
441 any of the contigs in the *P. praefalciparum* assembly were in fact derived from *P. falciparum* co-
442 infection.
443 The *P. blacklocki* sample was from a single infection. Due to sWGA, the PacBio sequence data
444 covered regions not covered by Illumina but due to the bias of the primers, the subtelomeres were
445 not covered fully. However, the internal *var* gene clusters are all assembled. Some of the core genes
446 from this species are also missing.
447
448 *Annotation*
449 The genomes were annotated as described in[47]. In short, the annotation of *P. falciparum*
450 (version July 2015) was transferred with RATT[48] and new gene models were called with
451 Augustus[49]. Obvious structural errors in core genes were manually corrected in Artemis[50].
452
453 **Mapping - generation of further samples**

454 To generate the gene sequence for different samples, Illumina reads were mapped against a set
455 of reference genomes using BWA[51] and default parameters. For the gorilla samples, we mapped
456 against the combined PacBio reference genomes of *P. adleri*, *P. blacklocki* and *P. praefalciparum*
457 and for the chimpanzee samples, the combined references of *P. gaboni* (PgabG01), *P. billcollinsi*
458 and *P. reichenowi* (PrG01). SNPs with Phred score ≥ 100 were called using GATK
459 UnifiedGenotyper[52] v2.0.35 (parameters: -pnrm POOL -ploidy 2 -glm POOLBOTH). From these
460 SNP calls we constructed the new gene set, masking regions in genes with less than 10x coverage
461 of 'properly' (correct distance and orientation) mapped paired reads. To generate the sequences of
462 the other 13 isolates, homozygous SNP calls were obtained (consensus program from bcftools-
463 1.2[53]). We quality controlled the SNP calling by regenerating PrCDC and PgabG02 gene set from
464 PrG01 and PgabG01, respectively and confirmed that they were placed with nearly no differences
465 in a phylogenetic tree.
466
467 **Orthologous group determination and alignment**
468 Orthologous groups were identified using OrthoMCL v1.4[54] across: (i) the seven core
469 *Laverania* genomes; (ii) the seven core genomes, the *Laverania* isolates PgabG02, PrCDC and *P.*
470 *falciparum* IT, as well as two outgroup genomes *Plasmodium vivax* Sal1 and *Plasmodium knowlesi*
471 strain H; and (iii) just Pf3D7, PprfG01 and PrG01. *P. praefalciparum* II was excluded due to its
472 partial genome. From these groups, different complete sets of 1:1 orthologues were extracted:
473 (1) "Lav12sp" set of 3,369 orthologues across the seven core *Laverania* species, the PrCDC
474 and *P. falciparum* IT isolates, *P. vivax* and *P. knowlesi*
475 (2) "Lav25st" set of 424 1:1 orthologues from across the 25 *Laverania* isolates, including the
476 previously published *P. reichenowi* CDC and five *P. falciparum* isolates (3D7, IT, DD2,
477 HB3 and 7G8[9]).
478 (3) "Lav7sp" set of 4,350 orthologues from across the seven *Laverania* reference genomes
479 (4) "Lav15st" set of 3,808 orthologues, with two representative sequences per species,
480 excluding *P. blacklocki* and *P. praefalciparum*.
481 (5) "Lav3sp" set of 4,826 1:1 orthologues across all the *P. reichenowi*, *P. praefalciparum* and
482 *P. falciparum* isolates
483 The first two sets were used to reconstruct the species tree, the third one for the comparative
484 genomic analyses (introgression, convergence and gene family evolution), the fourth one for the
485 analyses of within species polymorphism and the fifth one for the analysis of *P. falciparum*
486 adaptive evolution.

487 To reduce the rate of false positives in the evolutionary analyses due to misalignments (e.g.[55]),
488 codon-based multiple alignments were performed using PRANK[56,57] with the -codon and +F
489 options, as it was shown to outperform other programs in the context of the detection of positive
490 selection[58,59]. Prior to aligning codons, low complexity regions were excluded in the nucleotide
491 sequences using dustmasker[60] and in amino acid sequences using segmasker[61] from NCBI-BLAST.
492 Poorly aligned regions were excluded using Gblocks[62], with default settings.
493
494 **Analysis of interspecific gene flow, introgression or gene transfer**
495 *Species-tree inference*
496 Two ML trees were performed using RAxMLv8.1.20[63] to illustrate the phylogenetic
497 relationships between the *Laverania* species and genotypes studied here using the "Lav12sp" and
498 the "Lav25st" set of orthologues. For each tree, multiple nucleotide alignments of each orthologous
499 group were conducted as described above. Trees were then constructed from the concatenated
500 alignments of the "Lav12sp" set of orthologues for the species tree and the "Lav25st" set for the
501 strain tree using RAxML and the following options "-m GTRGAMMA –f a -# 100". Trees were
502 rooted afterwards using *P. vivax* and *P. knowlesi* for the species tree and the *P. adleri/P. gaboni*
503 clade for the genotype tree.
504
505 *Tree topology test*
506 Interspecific gene flow was investigated by testing congruence between each gene tree
507 topology and the species tree topology. We performed the Shimodaira-Hasegawa test (SH test[64])
508 using RAxMLv8.1.20 to test whether the phylogenetic tree for each gene significantly differed from
509 the *Laverania* species tree. Topology tests were based on multiple nucleotide alignments of the
510 4,350 "Lav7sp" set of orthologues. For each coding sequence, RAxML was called with the options
511 "-m GTRGAMMA –f h".
512
513 **Convergent evolution analyses**
514 *Genome-wide test of convergent evolution*
515 Convergent substitutions can occur by chance and the number of random convergent
516 substitutions between two lineages is correlated with the number of divergent substitutions
517 observed in these two lineages[65,66]. Excess of convergent substitutions in specific branch pairs can
518 thus be identified by analyzing the correlation between the number of convergent and divergent
519 substitutions between all the branch pairs in a phylogeny using orthogonal regression, and looking

520 for outlier branch pairs: branch pairs with a high positive residual show an excess of convergent
521 substitutions relatively to the number of divergent substitutions[65]. We used the software Grand-
522 Convergence (available at https://github.com/dekoning-lab/grand-conv) to estimate for each
523 chromosome the numbers of divergent and convergent substitutions between all branch pairs in the
524 *Laverania* tree and investigate whether branch pairs including *Laverania* species infecting the same
525 host species (gorilla or chimpanzee) presented an excess of convergence. Analyses were performed
526 under different models of amino-acid evolution: LG, WAG, JONES and DAYHOFF.
527
528 *Gene-based detection of convergent evolution throughout the Laverania*
529 For each orthologue of the "Lav7sp" set, the number and percentage of fixed amino acid
530 differences between parasites infecting the same host were calculated, *i.e.* the number of positions
531 showing the same amino acid within a host species but different amino acid between host species.
532 Alignments of all the available sequences ("Lav15st") from all the sequenced isolates were then
533 used to determine what number of host-specific differences were fixed within each host and each
534 species. To evaluate whether the observed number of host-specific fixed differences in an alignment
535 can be attributed to neutral evolution/purifying selection alone (with no positive selection), we used
536 a simulation-based approach. For each coding sequence, 1,000 sequences of the same size were
537 simulated, evolving along the same tree with the same specified branch lengths, substitution model,
538 codon frequencies and omega (d_N/d_S), using the program Evolver from PAML v4.8a[67]. The
539 program Codeml from PAML v4.8a[67] was first used to estimate the tree, the codon frequencies and
540 the average omega values for each of the coding sequences with fixed amino acid differences. For
541 each simulated dataset, the number of fixed amino-acid differences between the parasites infecting
542 a same host was estimated. The probability of observing *n* fixed differences was then computed as
543 the proportion of the simulated dataset of 1000 sequences that showed at least the same number of
544 fixed differences as observed in the real data.
545
546
547 **Tests for positive selection**
548 *Branch site tests*
549 To search for genes that have been subjected to positive selection in the *P. falciparum* lineage
550 alone after the divergence from *P. praefalciparum*, we used an updated branch site test [68]
551 implemented in PAML v4.4c [67]. This test detects sites that have undergone positive selection in a
552 specific branch of the phylogenetic tree (foreground branch). The "Lav3sp" set of 4,826

553 orthologous groups between *P. reichenowi*, *P. praefalciparum I* and *P. falciparum* was used for the
554 test. d_N/d_S ratio estimates per branch and gene were obtained using Codeml (PAML v4.4c) with a
555 *free-ratio* model of evolution. This identified 139 genes with a significant signal of positive
556 selection in *P. falciparum* only.
557 Using the "Lav7sp" dataset, a branch site test was also applied for each gene (Figure 2),
558 comparing the sequence from an individual species to the inferred sequence of its closest branch
559 point on the tree (e.g. *P. falciparum* compared with the sequence at the *P. falciparum–*
560 *P. praefalciparum* split).
561
562 *McDonald–Kreitman (MK) tests*
563 Selection in *P. falciparum* was also tested using McDonald–Kreitman (MK) tests [69] to compare
564 the polymorphism within species to the divergence between species, using *P. praefalciparum* as the
565 outgroup. Analyses were performed using the 4,826 "Lav3sp" set of orthologues. MK tests were
566 performed as described before[9]. Thirty-five genes had an MK ratio considerably higher than 1.
567
568 **Gene Ontology enrichment analyses**
569 Analysis of Gene Ontology (GO) term-enrichment was performed in R, using TopGO[70] with
570 default parameters. GO annotations from GeneDB were used but with unreviewed automated
571 annotations excluded.
572
573 **Gene family analyses**
574 To estimate the differential abundance of gene families across species, the Gene products and
575 the Pfam domains were counted and analysed by the variance of the occurrence. Unless otherwise
576 stated, trees were constructed using PhyML[71] (default parameters) or RAxML[63] (model estimated)
577 from alignments generated with Muscle[72] and trimmed with Gblocks[62] in Seaview[73] with default
578 values. Many of the findings were confirmed manually through ACT and bamview[50]. The analysis
579 of the *var* genes was performed on *var* genes larger than 2.5kb. Domains were called with the
580 HMMer models from varDom[74]. Distance matrices were generated based on BLASTp scores,
581 without filtering low complexity regions. Representation was done in R through the heatmap.2
582 program from gplot (see also Supplementary Note 3).
583
584 *Allelic dimorphisms*

585 For the analysis of dimorphism in *msp*, all sequences available for the *Laverania* were
586 downloaded from Uniprot[75]. Data were subsampled to obtain a similar number of sequences for
587 each group. Phylogenetic trees were constructed with PhyML[71], using default parameters and drawn
588 in Figtree. The *eba-175* alignment was visualized with Jalview[76].
589
590
591
592 **Divergence Dating**
593 Alignments of the *Laverania* included intergenic regions where possible. Assuming 402–681
594 mitotic events per year (Supplementary Note 1) and a mutation rate of 3.78E-10 for 4 mitotic
595 events[77,78](mutation rate from latter paper was taken from Pf3D7 line without drugs), equivalent to
596 around 0.9–1.5 mutations per genome per year. Although we observed similar mutation rates in
597 clinical samples (Supplementary Note 1), these estimates have potential errors and therefore we
598 report ratios of divergence times in the figures that are robust to errors in these parameters. For
599 coalescence based estimates of speciation times, G-Phocs[79] was used and multiple sequentially
600 Markovian coalescent (MSMC) on segregating sites[80] was used to estimate the *P. falciparum*
601 bottleneck.
602

603
604 **Author Contributions:** TDO**,** BO, FR, CN, MB, FP designed the study. CA, APO, LB, EW, BN,
605 ND, CP, PD, VR, FP collected and assessed samples. CA performed the WGA and cell sorting on
606 one sample. SO performed the WGA on the samples; MS organised the sequencing. TDO did
607 assembly and annotation. UB did manual gene curation; AG, FP performed the evolutionary
608 analyses on core genomes. TDO, CN, MB performed the analyses of gene families
609 and dimorphisms. TC performed the dating analyses. TDO, AG, CN, MB, FP wrote the manuscript.
610 All authors read and approved the paper.

611

612 **Acknowledgments:**

613 This work was funded by ANR ORIGIN JCJC 2012, LMI ZOFAC, CNRS, CIRMF, IRD and the
614 Wellcome Trust (grants WT 098051 and WT 206194 to the Sanger Institute, 104792/Z/14/Z to
615 CN). TC holds a MRC DTP Studentship. We thank Gavin Rutledge for performing the sWGA and
616 Julian Rayner and Francisco J. Ayala for helpful discussion. We thank the PlasmoDB team for
617 promptly making these data available.

618

619

620 **Data availability**

621 All sequences have been submitted to the European Nucleotide Archive. The accession numbers of
622 the raw reads, and assembly data can be found in Supplementary Table 8. The genomes are being
623 submitted to EBI, project ID PRJEB13584. The genomes are available from plasmodb.org and
624 from ftp://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/Laverania/.

625

626 **Competing financial interests.**

627 None
628

629 **Computer code**

630 Custom computer code is available on request.

631

632
633

634 **Figures**

635

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

645

646 **Figure 2. Overview of the analyses of core genes over all *Laverania* genomes.** (a) Summary of
647 evolution of core genes. From outer to inner track: scatterplot of branch site test for each genome
648 (see Supplementary Table 3 for *P. falciparum* data); per-species d_N/d_S values ($0.5 < d_N/d_S < 2$);
649 orthologues represented by vertical black lines under the chromosome track represent, with dots
650 representing *P. falciparum* 3D7 *var* genes on the forward (blue) or reverse strands (red), or *var*
651 pseudogenes (black); average of the relative polymorphism (π) across species, with the underlying π
652 for each species calculated from multiple strains ("Lav25st" dataset) and normalized by the average
653 for that species; signatures of convergent evolution based on host-specific fixed differences analysis
654 with the chromosome 4 region that includes the *Rh5* locus highlighted (black box). (b) Magnified
655 view of the *Rh5* region that is enriched with host specific fixed differences. Convergent evolution
656 analysis was performed using orthologues conserved across the *Laverania*. Filled circles represent
657 the subset of differences that were fixed within all the isolates available ("Lav15st" set) and for
658 which we could reject neutral evolution (for the gene list see Supplementary Table 4).
659

660 **Figure 3. Gene families in the *Laverania.*** Distribution of major multigene families including *var*
661 and those that show significant copy number variation among lineages. Data from
662 *P. praefalciparum* include the subtelomeric gene families from the two infecting genotypes.
663 Assembly of *P. billcollinsi* is incomplete in the subtelomeres.
664

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.

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*.

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

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

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

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

Population	Status	Coalescence, years (x 10^5)			Effective Population size, Ne (x 10^5)			Migration target	Probability of Parasite Migration		
		CDS	Intergenic	Intergenic no UTR	CDS	Intergenic	Intergenic no UTR		CDS	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	0.605 (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	-	-
All species	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 2: Estimated dates of speciation (time to coalescence), effective population size (Ne) and migration rates for present and ancestral *Laverania* populations. Estimated dates of speciation (time to coalescence), effective population size (Ne) and migration rates for Laverania populations, including inferred ancestors. Values inferred using GPhoCS coalescence model. The results are presented from sequence alignments of protein coding (CDS) (genic) regions as well as apparent intergenic regions inferred in the presence or absence of 0.5 kb flanking (UTR) sequences. The coalescence model parameters have been scaled based on an assumption of 401–681 mitotic events per year (Supplementary Note 1). The number in parentheses is the range based on the prediction (95% confidence interval). Values with (*) are linear estimates, as good alignments could not be obtained between those species due to the low GC content. The migration rate refers to the probability of a parasite from the source lineage migrating into the target over the time period. The data from these estimates are presented in Fig. 1.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

16 Tanabe, K., Mackay, M., Goman, M. & Scaife, J. G. Allelic dimorphism in a surface antigen gene of the malaria parasite Plasmodium falciparum. *Journal of molecular biology* **195**, 273-287, doi:10.1016/0022-2836(87)90649-8 (1987).

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

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

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

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

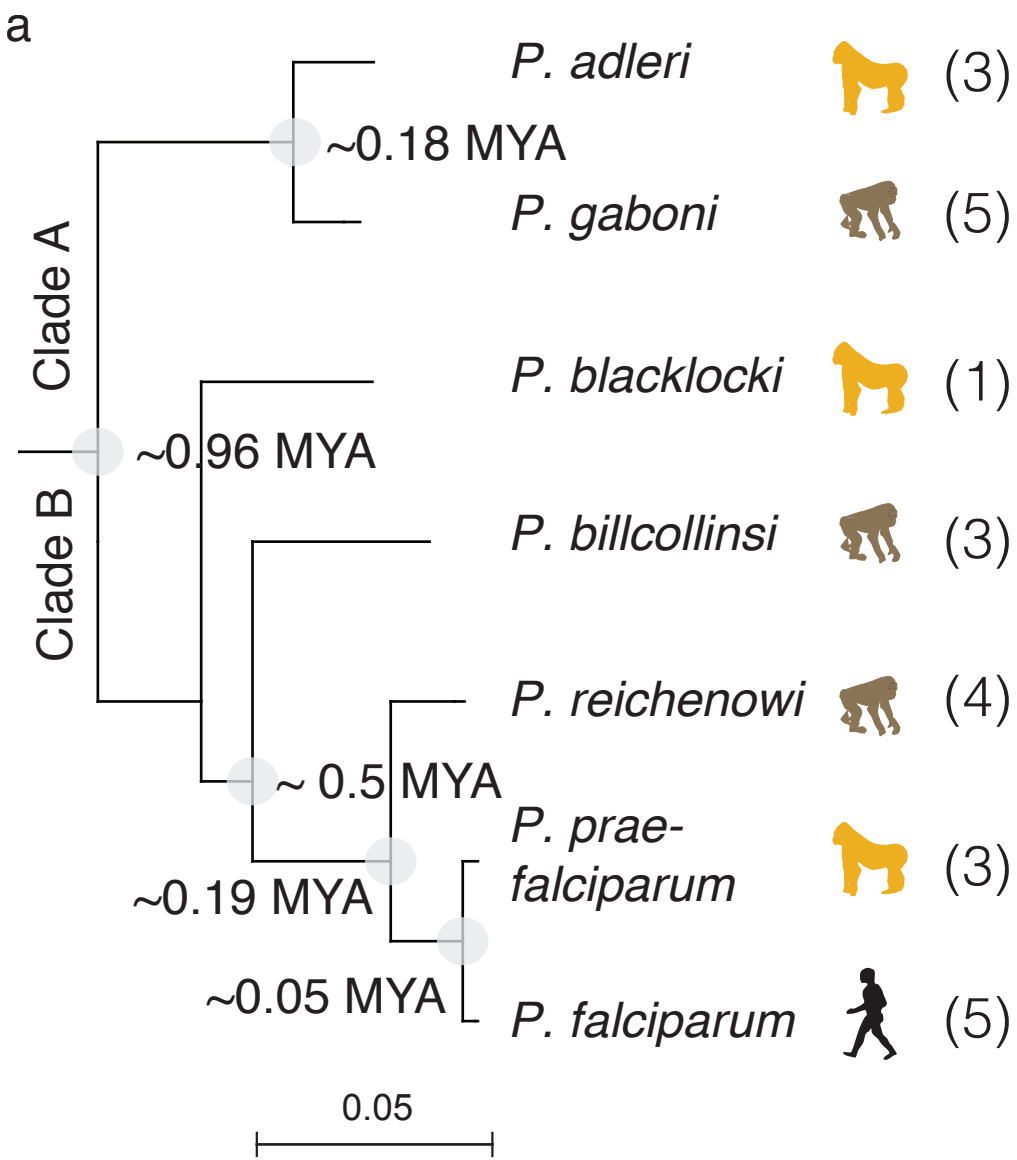
21 Wright, K. E. *et al.* Structure of malaria invasion protein RH5 with erythrocyte basigin and blocking antibodies. *Nature* **515**, 427-+, doi:10.1038/nature13715 (2014).

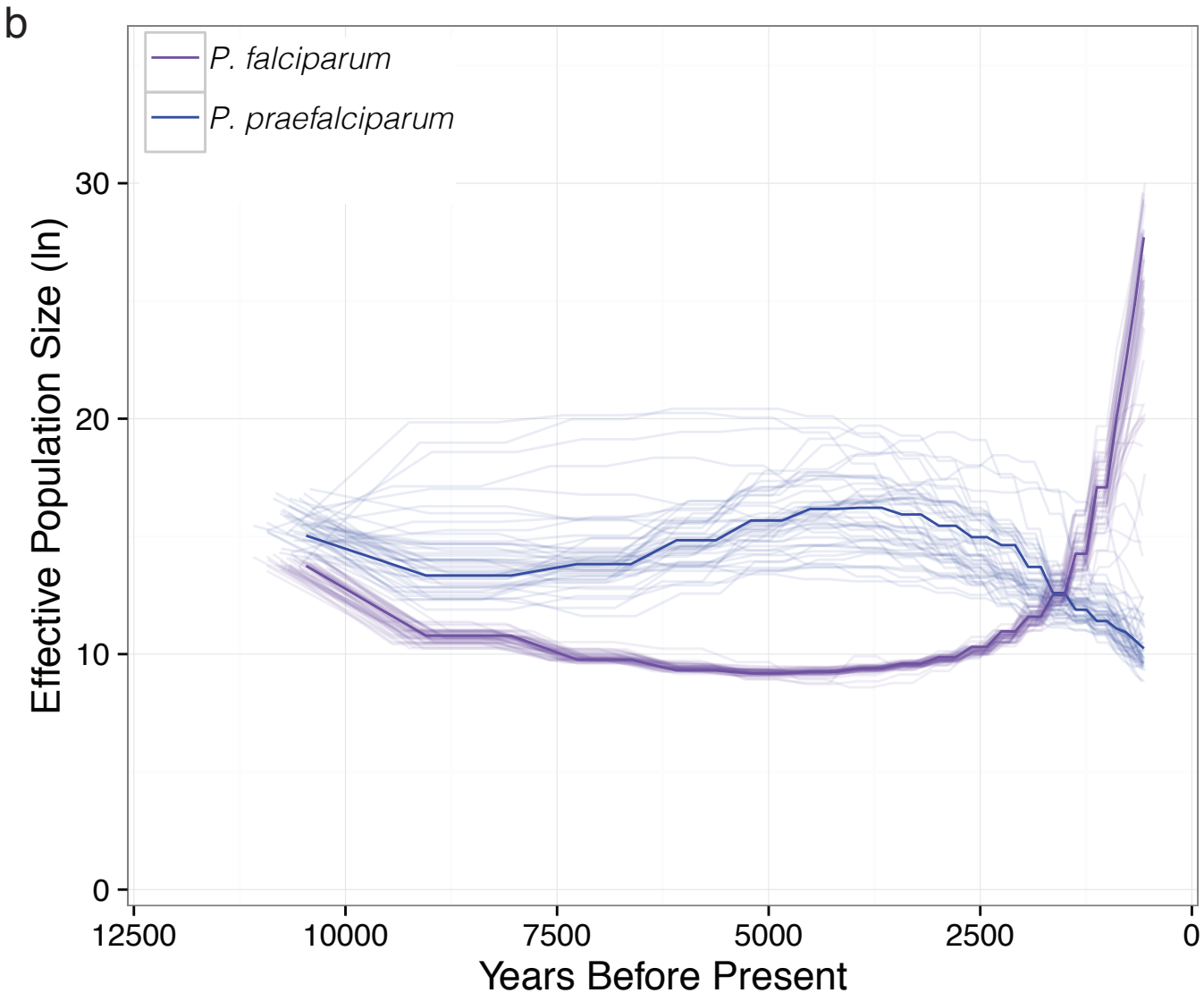
22 Triglia, T., Thompson, J. K. & Cowman, A. F. An EBA175 homologue which is transcribed but not translated in erythrocytic stages of Plasmodium falciparum. *Mol Biochem Parasitol* **116**, 55-63, doi:10.1016/S0166-6851(01)00303-6 (2001).

23 Farrell, A. *et al.* A DOC2 protein identified by mutational profiling is essential for apicomplexan parasite exocytosis. *Science* **335**, 218-221, doi:10.1126/science.1210829 (2012).
24 Ramiro, R. S. *et al.* Hybridization and pre-zygotic reproductive barriers in Plasmodium. *Proceedings of the Royal Society B-Biological Sciences* **282**, doi:10.1098/rspb.2014.3027 (2015).
25 Eksi, S. *et al.* Malaria transmission-blocking antigen, Pfs230, mediates human red blood cell binding to exflagellating male parasites and oocyst production. *Molecular Microbiology* **61**, 991-998, doi:10.1111/j.1365-2958.2006.0528.x (2006).
26 Mundwiler-Pachlatko, E. & Beck, H. P. Maurer's clefts, the enigma of Plasmodium falciparum. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 19987-19994, doi:10.1073/pnas.1309247110 (2013).
27 Bethke, L. L. *et al.* Duplication, gene conversion, and genetic diversity in the species-specific acyl-CoA synthetase gene family of Plasmodium falciparum. *Mol Biochem Parasitol* **150**, 10-24, doi:10.1016/j.molbiopara.2006.06.004 (2006).
28 Cunningham, D., Lawton, J., Jarra, W., Preiser, P. & Langhorne, J. The pir multigene family of Plasmodium: antigenic variation and beyond. *Mol Biochem Parasitol* **170**, 65-73, doi:10.1016/j.molbiopara.2009.12.010 (2010).
29 Niang, M. *et al.* STEVOR is a Plasmodium falciparum erythrocyte binding protein that mediates merozoite invasion and rosetting. *Cell Host Microbe* **16**, 81-93, doi:10.1016/j.chom.2014.06.004 (2014).
30 Kraemer, S. M. & Smith, J. D. A family affair: var genes, PfEMP1 binding, and malaria disease. *Curr Opin Microbiol* **9**, 374-380, doi:10.1016/j.mib.2006.06.006 (2006).
31 Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511, doi:10.1038/nature01097 (2002).
32 Bordbar, B. *et al.* Genetic diversity of VAR2CSA ID1-DBL2Xb in worldwide Plasmodium falciparum populations: impact on vaccine design for placental malaria. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* **25**, 81-92, doi:10.1016/j.meegid.2014.04.010 (2014).
33 Frank, M., Dzikowski, R., Amulic, B. & Deitsch, K. Variable switching rates of malaria virulence genes are associated with chromosomal position. *Mol Microbiol* **64**, 1486-1498, doi:10.1111/j.1365-2958.2007.05736.x (2007).
34 Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nature reviews. Genetics* **13**, 745-753, doi:10.1038/nrg3295 (2012).
35 Carter, R. & Mendis, K. N. Evolutionary and historical aspects of the burden of malaria. *Clinical microbiology reviews* **15**, 564-594, doi:10.1128/CMR.15.4.564-594.2002 (2002).
36 Auburn, S. *et al.* An effective method to purify plasmodium falciparum dna directly from clinical blood samples for whole genome high-throughput sequencing. *PLoS ONE* **6**, doi:10.1371/journal.pone.0022213 (2011).
37 Oyola, S. O. *et al.* Optimized whole-genome amplification strategy for extremely AT-biased template. *DNA research : an international journal for rapid publication of reports on genes and genomes* **21**, 661-671, doi:10.1093/dnares/dsu028 (2014).
38 Oyola, S. O. *et al.* Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. *bioRxiv*, doi:10.1101/067546 (2016).
39 Boissiere, A. *et al.* Isolation of Plasmodium falciparum by flow-cytometry: implications for single-trophozoite genotyping and parasite DNA purification for whole-genome high-throughput sequencing of archival samples. *Malaria Journal* **11**, 163, doi:10.1186/1475-2875-11-163 (2012).
40 Quail, M. A. *et al.* Optimal enzymes for amplifying sequencing libraries. *Nat Methods* **9**, 10-11, doi:10.1038/nmeth.1814 (2012).
41 Manske, H. & Kwiatkowski, D. SNP-o-matic. *Bioinformatics* **25**, 2434-2435, doi:10.1093/bioinformatics/btp403 (2009).
42 Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**, 563-569, doi:10.1038/nmeth.2474 (2013).
43 Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics Vol. 25 No. 15*, 1968-1969 (2009).
44 Carver, T. *et al.* Artemis and ACT: viewing, annotation and comparing sequences stored in relational database. *Bioinformatics* **24**, 2672-2676, doi:10.1093/bioinformatics/btn529 (2008).
45 Otto, T. D., Sanders, M., Berriman, M. & Newbold, C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26**, 1704-1707, doi:10.1093/bioinformatics/btq269 (2010).
46 English, A. C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**, e47768, doi:10.1371/journal.pone.0047768 (2012).

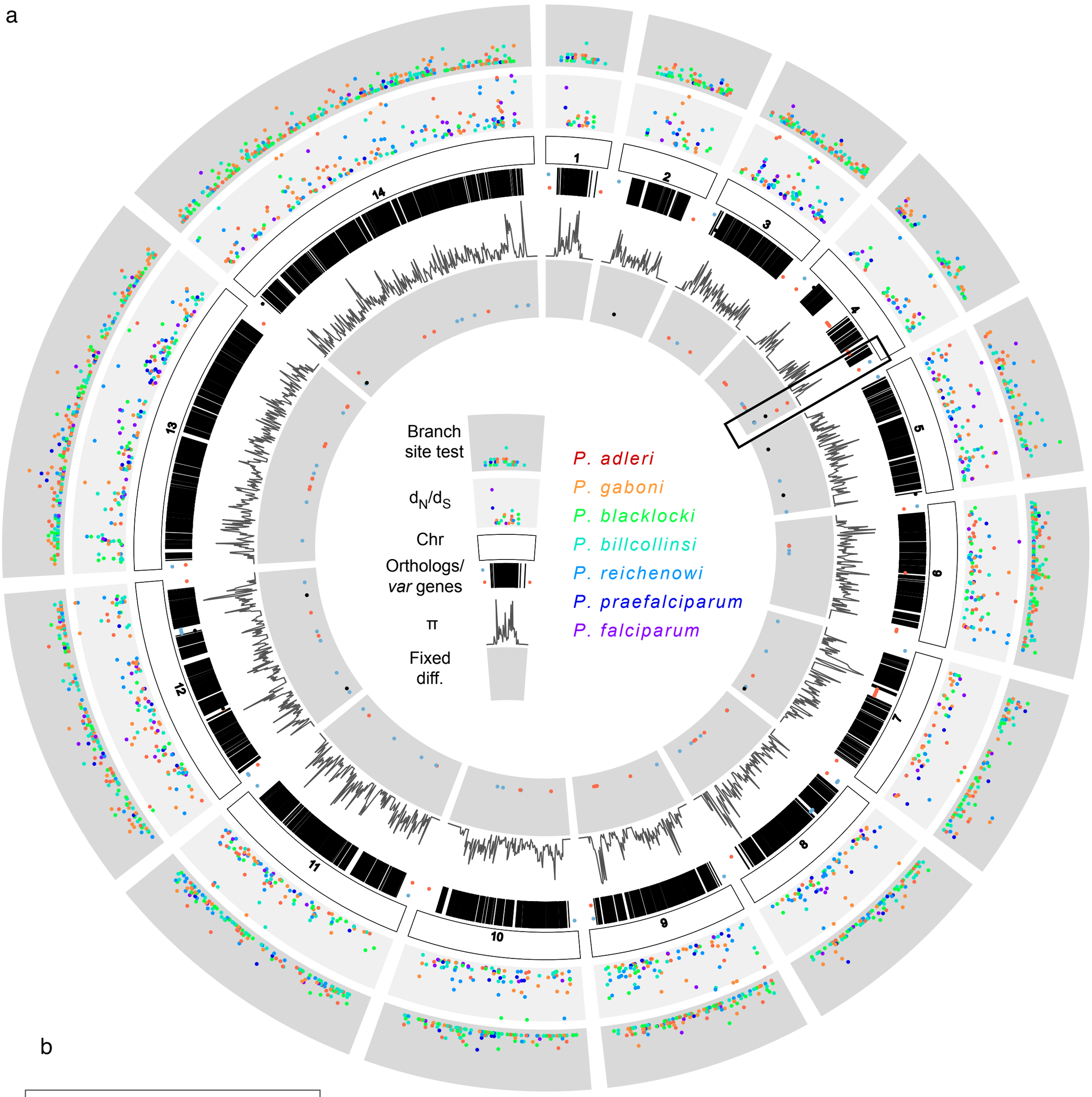
47 Otto, T. D. From sequence mapping to genome assemblies. *Methods Mol Biol* **1201**, 19-50, doi:10.1007/978-1-4939-1438-8_2 (2015).
48 Otto, T. D., Dillon, G. P., Degrave, W. S. & Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Research* **39**, e57, doi:10.1093/nar/gkq1268 (2011).
49 Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435-439, doi:10.1093/nar/gkl200 (2006).
50 Carver, T. *et al.* BamView: visualizing and interpretation of next-generation sequencing read. *Briefings in bioinformatics*, doi:10.1093/bib/bbr073 (2013).
51 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
52 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
53 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).
54 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189, doi:10.1101/gr.1224503 (2003).
55 Jordan, G. & Goldman, N. The Effects of Alignment Error and Alignment Filtering on the Sitewise Detection of Positive Selection. *Molecular biology and evolution* **29**, 1125-1139, doi:10.1093/molbev/msr272 (2012).
56 Loytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 10557-10562, doi:10.1073/pnas.0409137102 (2005).
57 Loytynoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632-1635, doi:10.1126/science.1158395 (2008).
58 Fletcher, W. & Yang, Z. The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection. *Molecular biology and evolution* **27**, 2257-2267, doi:10.1093/molbev/msq115 (2010).
59 Markova-Raina, P. & Petrov, D. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Research* **21**, 863-874, doi:10.1101/gr.115949.110 (2011).
60 Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. *Journal of Computational Biology* **13**, 1028-1040, doi:10.1089/cmb.2006.13.1028 (2006).
61 Wootton, J. C. & Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry* **17**, 149-163, doi:10.1016/0097-8485(93)85006-X (1993).
62 Castresana, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular biology and evolution* **17**, 540-552, doi:10.1093/oxfordjournals.molbev.a026334 (2000).
63 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).
64 Shimodaira, H. & Hasegawa, M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Molecular biology and evolution* **16**, 1114, doi:10.1093/oxfordjournals.molbev.a026201 (1999).
65 Castoe, T. A. *et al.* Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 8986-8991, doi:10.1073/pnas.0900233106 (2009).
66 Thomas, G. W. C. & Hahn, M. W. Determining the Null Model for Detecting Adaptive Convergence from Genomic Data: A Case Study using Echolocating Mammals. *Molecular biology and evolution* **32**, 1232-1236, doi:10.1093/molbev/msv013 (2015).
67 Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular biology and evolution* **24**, 1586-1591, doi:10.1093/molbev/msm088 (2007).
68 Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Molecular biology and evolution* **22**, 2472-2479, doi:10.1093/molbev/msi237 (2005).
69 McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652-654, doi:10.1038/351652a0 (1991).
70 Rahnenfuhrer, A. A. a. J. topGO: Enrichment analysis for Gene Ontology. *R package*, doi:10.18129/B9.bioc.topGO (2010).
71 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321, doi:10.1093/sysbio/syq010 (2010).
72 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).

73 Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution* **27**, 221-224, doi:10.1093/molbev/msp259 (2010).
74 Rask, T. S., Hansen, D. A., Theander, T. G., Gorm Pedersen, A. & Lavstsen, T. Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes--divide and conquer. *PLoS Comput Biol* **6**, doi:10.1371/journal.pcbi.1000933 (2010).
75 UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212, doi:10.1093/nar/gku989 (2015).
76 Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191, doi:10.1093/bioinformatics/btp033 (2009).
77 Claessens, A. *et al.* Generation of antigenic diversity in Plasmodium falciparum by structured rearrangement of Var genes during mitosis. *PLoS Genet* **10**, e1004812, doi:10.1371/journal.pgen.1004812 (2014).
78 Bopp, S. E. *et al.* Mitotic evolution of Plasmodium falciparum shows a stable core genome but recombination in antigen families. *PLoS Genet* **9**, e1003293, doi:10.1371/journal.pgen.1003293 (2013).
79 Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet* **43**, 1031-1034, doi:10.1038/ng.937 (2011).
80 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925, doi:10.1038/ng.3015 (2014).

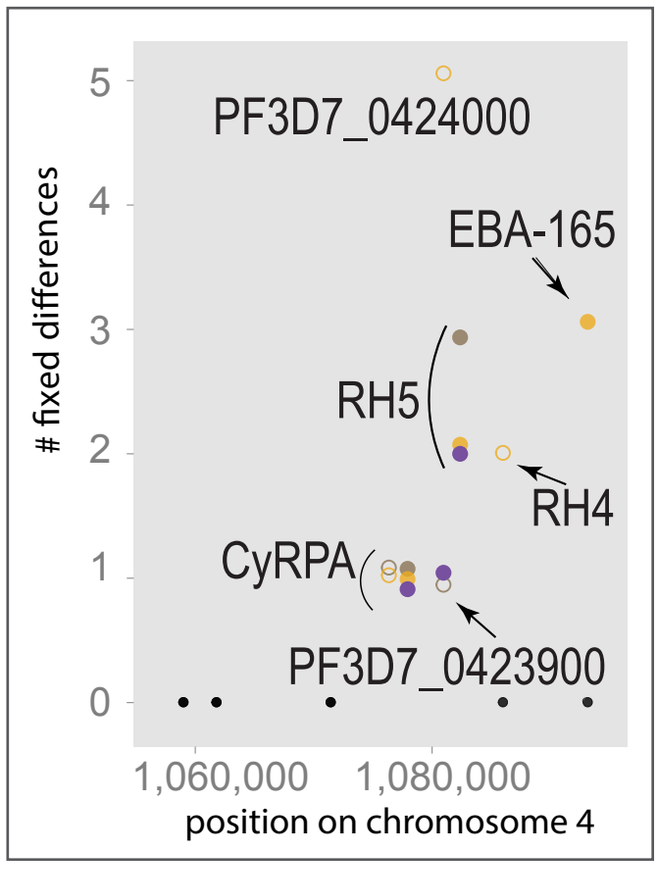

a
P. adleri
(3)
~0.18 MYA
P. gaboni
(5)
Clade A
P. blacklocki
(1)
~0.96 MYA
Clade B
P. billcollinsi
(3)
P. reichenowi
(4)
~ 0.5 MYA
P. prae-
falciparum
(3)
~0.19 MYA
~0.05 MYA
P. falciparum
(5)
0.05

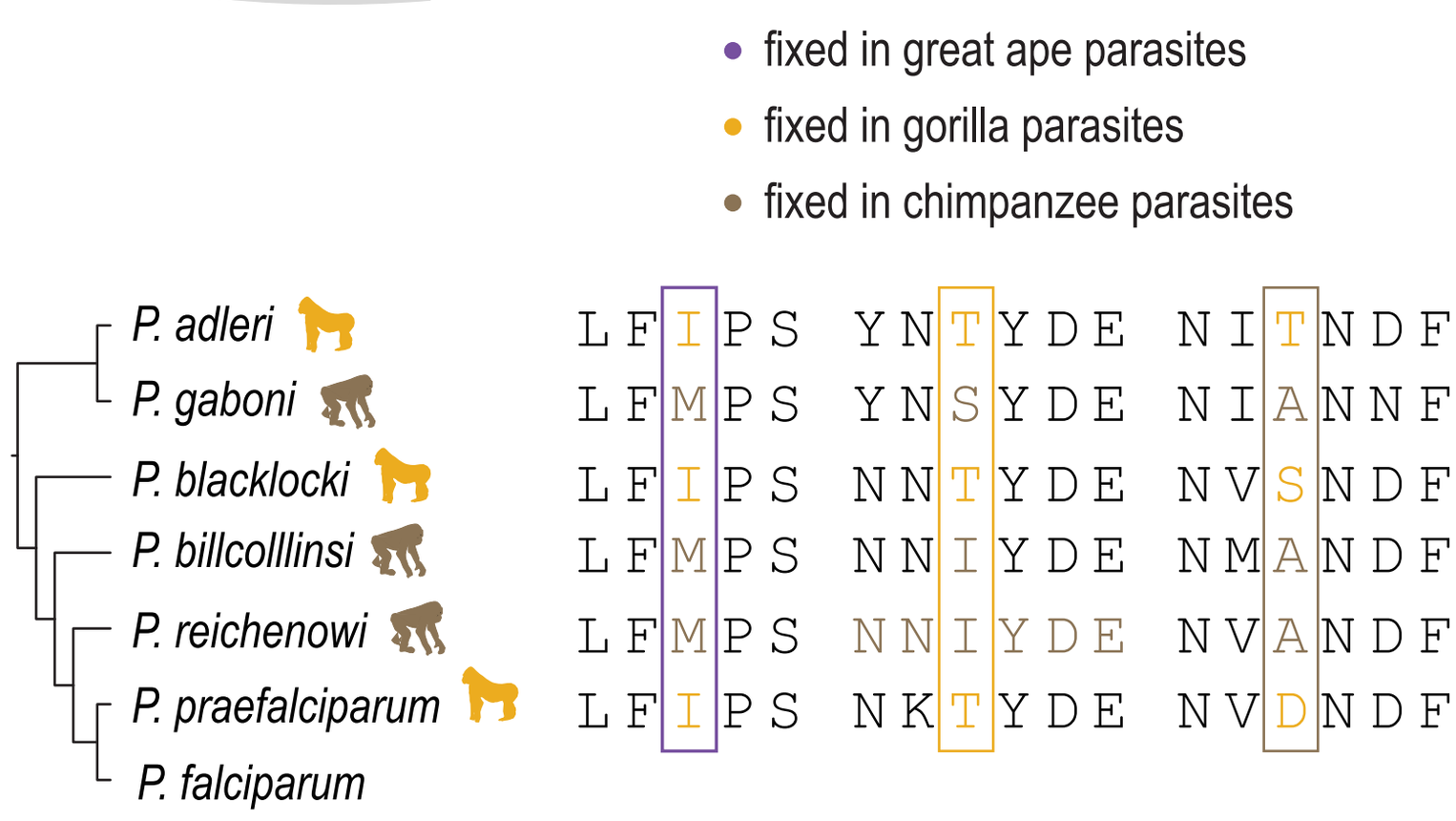

b
P. falciparum
P. praefalciparum
Effective Population Size (ln)
30
20
10
0
12500
10000
7500
5000
2500
0
Years Before Present

a

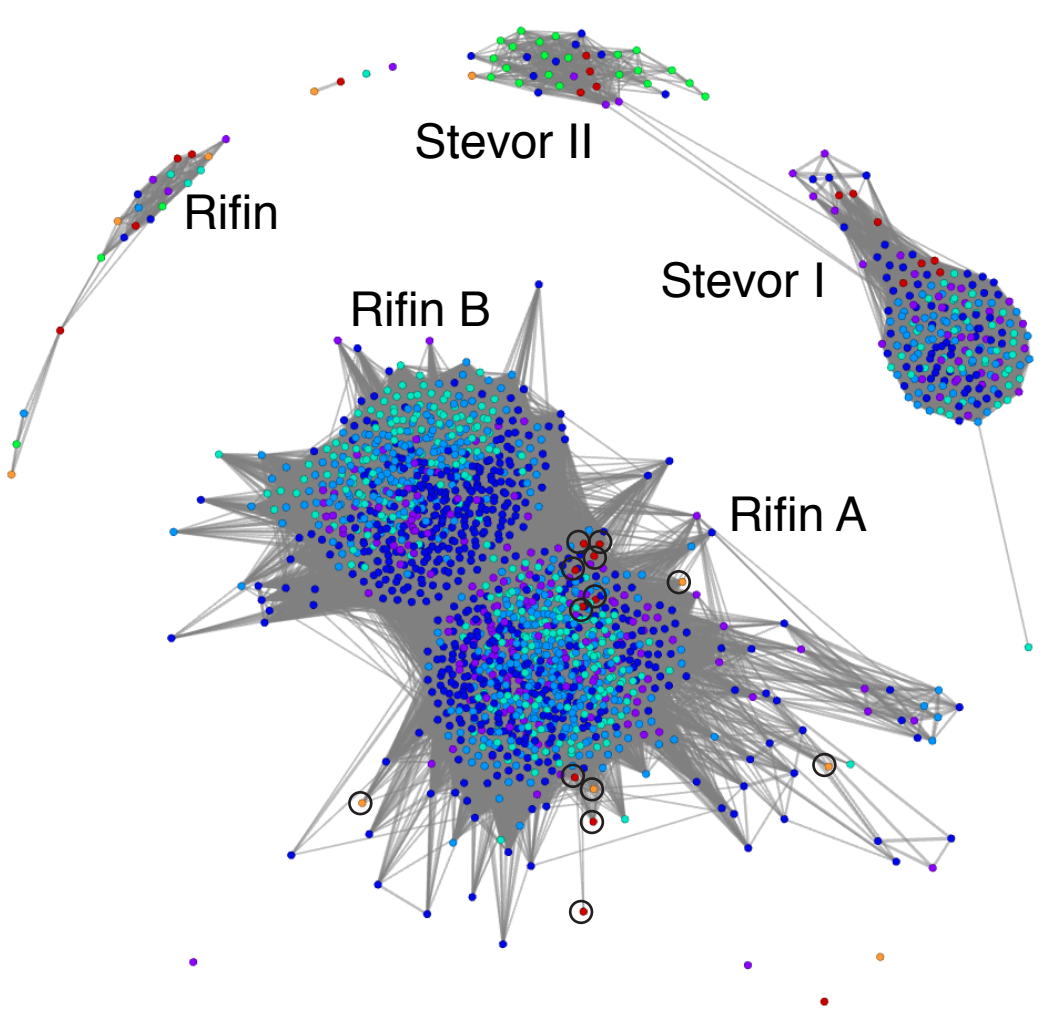

1
2
3
4
5
6
7
8
9
10
11
12
13
14
Branch site test
d_N/d_S
Chr
Orthologs/ *var* genes
π
Fixed diff.
P. adleri
P. gaboni
P. blacklocki
P. billcollinsi
P. reichenowi
P. praefalciparum
P. falciparum

b


fixed differences
5
4
3
2
1
0
PF3D7_0424000
EBA-165
RH5
RH4
CyRPA
PF3D7_0423900
1,060,000
1,080,000
position on chromosome 4


fixed in great ape parasites
fixed in gorilla parasites
fixed in chimpanzee parasites
P. adleri LFIPS YNTYDE NITNDF
P. gaboni LFMPS YNSYDE NIANNF
P. blacklocki LFIPS NNTYDE NVSNDF
P. billcolllinsi LFMPS NNIYDE NMANDF
P. reichenowi LFMPS NNIYDE NVANDF
P. praefalciparum LFIPS NKTYDE NVDNDF
P. falciparum

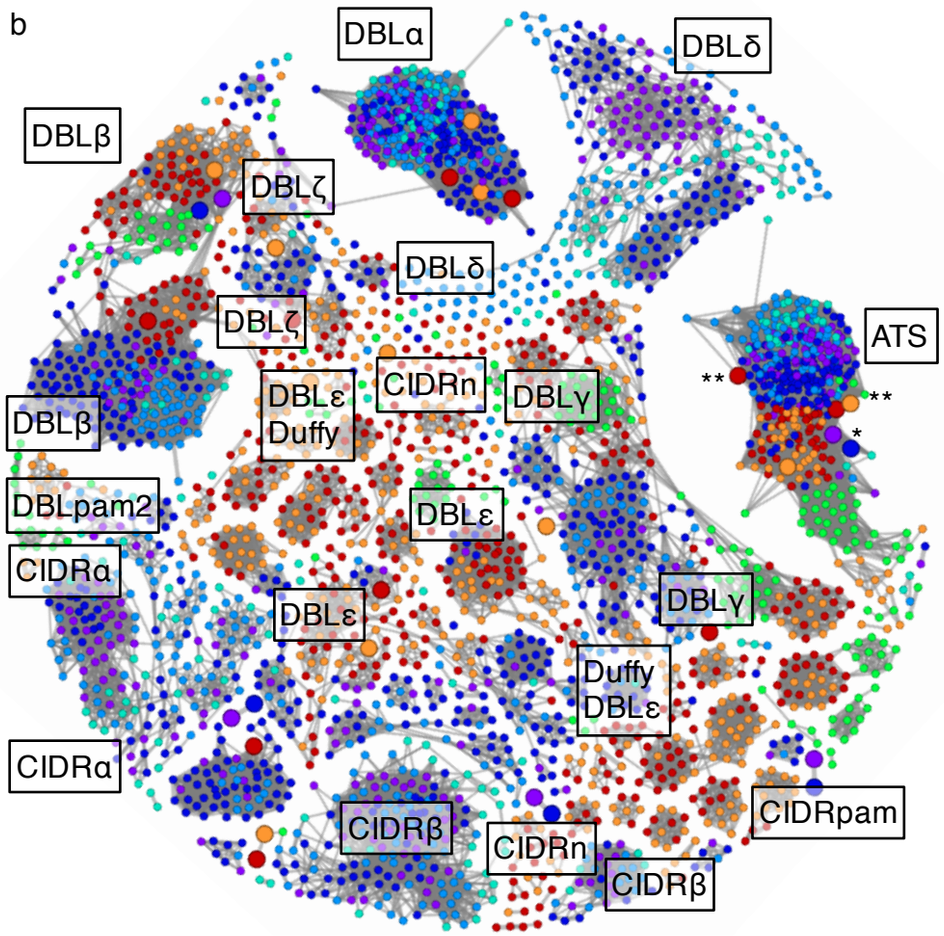
	P. falciparum	*P. praefalciparum*	*P. reichenowi*	*P. billcollinsi*	*P. blacklocki*	*P. gaboni*	*P. adleri*
var ≥ 2.5kb	67	112	92	35	43	61	58
rif > 250aa	183	582	440	280	13	12	19
stevor > 250aa	41	78	54	45	21	1	13
hyp4	9	2	1	1	0	0	0
hyp5	9	2	2	1	5	3	0
Maurer	13	12	5	4	7	0	0
EPF1	8	6	4	3	5	1	1
RESA-like	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


Stevor II
Rifin
Stevor I
Rifin B
Rifin A
P. adleri
P. blacklocki
P. reichenowi
P. falciparum
P. gaboni
P. billcollinsi
P. praefalciparum

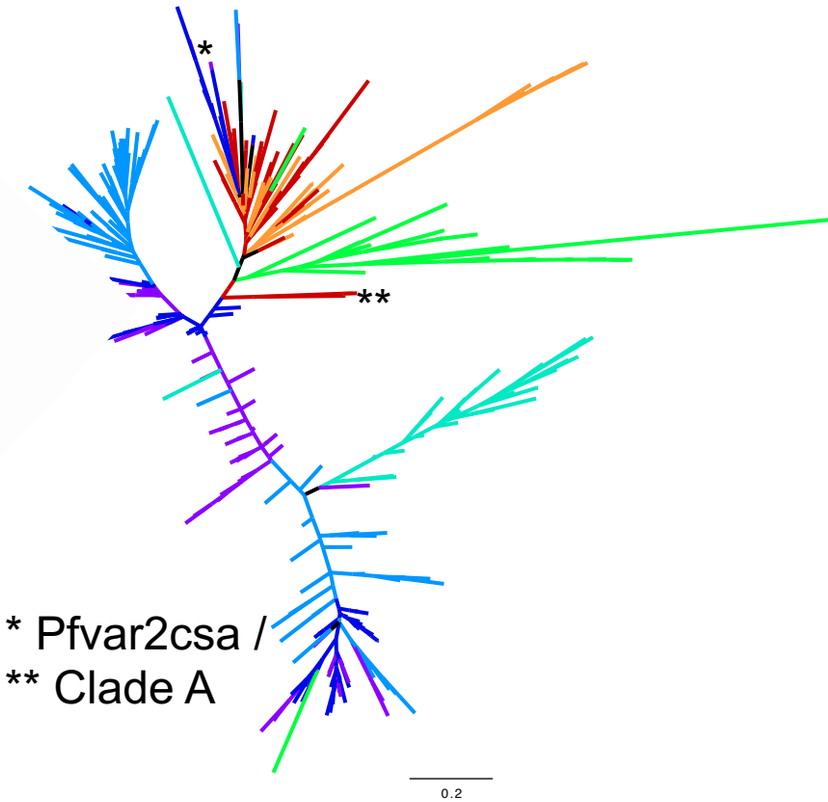
a

	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

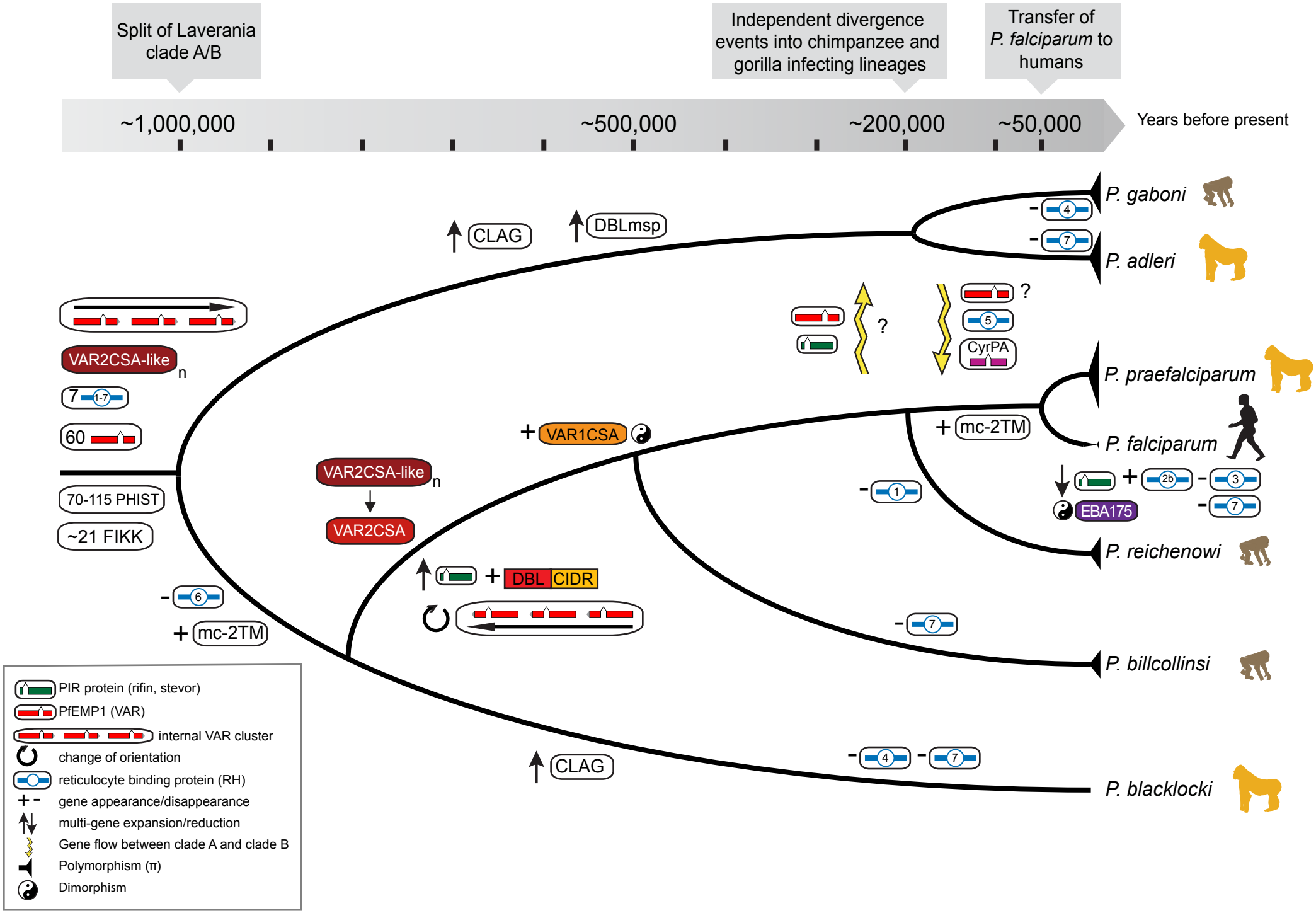
b



c



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


Split of Laverania clade A/B
Independent divergence events into chimpanzee and gorilla infecting lineages
Transfer of P. falciparum to humans
~1,000,000
~500,000
~200,000
~50,000
Years before present
P. gaboni
P. adleri
P. praefalciparum
P. falciparum
P. reichenowi
P. billcollinsi
P. blacklocki
CLAG
DBLmsp
VAR2CSA-like
n
7
1-7
60
70-115 PHIST
~21 FIKK
VAR2CSA
VAR1CSA
mc-2TM
DBL
CIDR
CyrPA
EBA175
PIR protein (rifin, stevor)
PfEMP1 (VAR)
internal VAR cluster
change of orientation
reticulocyte binding protein (RH)
gene appearance/disappearance
multi-gene expansion/reduction
Gene flow between clade A and clade B
Polymorphism (π)
Dimorphism