This is the author's final accepted version.

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 O. Oyola[1,4], Alain Prince Okouga[5], Larson Boundenga[5], Eric Willaume[6], Barthélémy Ngoubangoye[5], Nancy Diamella Moukodoum[5], Christophe Paupy[2], Patrick Durand[2], Virginie Rougeron[2,5], Benjamin Ollomo[5], François Renaud[2], Chris Newbold[1,7], Matthew Berriman[1,*] & 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 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 deadly human malaria parasite, has been the subject of uncertainty and debate[1,2]. Recently it has become clear that *P. falciparum* is derived from a group of parasites infecting African Great Apes, known as the *Laverania* subgenus[2]. Until 2009, chimpanzee-infecting *P. reichenowi* was the only other species known in this subgenus, for which only one isolate was available[3]. It is now clear that there are a total of at least seven species in Great Apes that naturally infect chimpanzees (*P. gaboni*, *P. billcollinsi* and *P. reichenowi*), gorillas (*P. praefalciparum*, *P. blacklocki* and *P. adleri*)[4,5], or humans (*P. falciparum*) (Fig. 1a). Within this group, *P. falciparum* is the only parasite that has successfully adapted to humans after a transfer from gorillas and subsequently spread worldwide[2].

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

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

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

Genome sequencing from six *Laverania* species

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

Speciation history in the *Laverania* sub-genus

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

105 Since no data is available for the other species we have assumed hereon that these values generalise

106 across the subgenus. From Bayesian whole-genome estimates, the ancestor of all current day

107 parasites of this subgenus existed 0.7–1.2 million years ago, a time at which the subgenus divided

108 into two main clades, A (*P. adleri* and *P. gaboni*) and B that includes the remaining species (Fig.

109 1a). Our range of values is far more recent than previous estimates[3,11]. Following the Clade A/B

110 subdivision, several speciation events occurred leading either to new chimpanzee or gorilla

111 parasites. Interestingly, the divergence between *P. adleri* and *P. gaboni* in one lineage and *P.*

112 *reichenowi* and the ancestor of *P. praefalciparum/P. falciparum* in the other lineage occurred at

113 approximately the same time (140–230 thousand years ago; Fig. 1a, Table 2). Based on

114 coalescence estimates, *P. falciparum* begun to emerge in humans from *P. praefalciparum* around

115 40–60 thousand years ago (Fig. 1a), significantly later than the evolution of the first modern

116 humans and their spread throughout Africa[12]. Our analysis also indicates significant gene flow

117 between these two parasite species after their initial divergence (Supplementary Table 3).

118   *P. falciparum* has strikingly low diversity (π =0.0004), compared with the other *Laverania*

119 species (0.002–0.0049) (Supplementary Fig. 1). It has been proposed that *P. falciparum* arose from

120 a single transfer of *P. praefalciparum* into humans[6] and based in part on the paucity of neutral SNPs

121 within the genome, that *P. falciparum* emerged from a bottleneck of a single parasite around 10,000

122 years ago, after agriculture was established[6,8]. In light of our results, we estimate that the

123 *P. falciparum* population declined around 11,000 years ago and reached a minimum about 5,000

124 years ago (Fig. 1b) with an effective population size (*Ne*) of around 3,000 (Supplementary Note 1;

125 generally the census number of parasites is higher than *Ne*[13]). The hypothesis of a single progenitor

126 is also inconsistent with the observation of several ancient gene dimorphisms that have been

127 observed in *P. falciparum*. A previous analysis using *P. reichenowi* and *P. gaboni* sequence data,

128 provided some evidence that different dimorphic loci diverged at different points in the tree[14].

129 Looking at each of these *P. falciparum* loci across the *Laverania*, we found different patterns of

130 evolution at the *msp1*, *var1csa*, and *msp3* loci (Supplementary Fig. 3a). Most strikingly, a mutually

131 exclusive dimorphism (described as MAD20/K1[15]) in the central 70% of the *msp1* sequence, clearly

132 pre-dates the *P. falciparum–P. praefalciparum* common ancestor and dimorphism in *var1csa* (an

133 unusual *var* gene of unknown function that is transcribed late in the asexual cycle) occurred before

134 the split with *P. reichenowi*.

135   In contrast, the gene *eba-175* that encodes a parasite surface ligand involved in red blood cell

136 invasion contains a dimorphism that arose after the emergence of *P. falciparum* (Supplementary

137 Fig. 3b). The time to the most recent common ancestor of *eba-175* has been estimated as 130–140

138 thousand years in an analysis[16] that assumed *P. falciparum* and *P. reichenowi* diverged 6 million

139 years ago. However, based on our new estimate for *P. falciparum–P. reichenowi* divergence, we

140 recalibrated their estimate of the most recent common ancestor of the *eba-175* alleles to be around

141 4,000 years ago, which is in good agreement with our divergence time for *P. falciparum*

142 (Supplementary Note 1). The recent dimorphism cannot however explain the observation of an

143 ancient dimorphism near the human and ape loci for glycophorin[17] – an EBA-175 binding protein.

144 The formation and maintenance of all of these dimorphic loci has therefore been shaped by different

145 balancing selection pressures over time.

146

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

148

149   During its move away from gorillas, *P. falciparum* had to adapt to a new vertebrate host

150 (human) and new vector species (e.g. *Anopheles gambiae*)[18]. To infer *P. falciparum* specific

151 adaptive changes, we considered the *P. falciparum* / *P. praefalciparum* /*P. reichenowi* genome trio

152 and then applied two lineage based tests to find positive selection that occurred in the *P. falciparum*

153 branch (see methods). The two tests identified 172 genes (out of 4,826) with signatures of positive

154 selection in the human parasite species only (Supplementary Table 4). Two genes (*rop14* and

155 PF3D7_0609900) were significant in both tests. Among the 172 genes, almost half (n=82) encoded

156 proteins of unknown function. Analysis of those with functional annotation indicated that genes

157 involved in pathogenesis, entry into host, actin movement and organization and in drug response

158 were significantly over-represented. Other genes, expressed in different stages of the *P. falciparum*

159 life cycle (e.g. *sera4* and *emp3*, involved during the erythrocytic stages; *trsp* and *lisp*1, involved in

160 the hepatic stages; and *plp4*, *CelTOS* and *Cap380*, involved in the mosquito stages) also showed a

161 significant signal of adaptive evolution (Supplementary Table 4).

162

163 **Evolution through introgression, gene transfer and convergence**

164   Frequent mixed species infections in apes and mosquitoes[18] provide clear opportunities for

165 interspecific gene flow between these parasites. A recent study[6] reported a gene transfer event

166 between *P. adleri* and the ancestor of *P. falciparum* and *P. praefalciparum* of a region on

167 chromosome 4 including key genes involved in erythrocyte invasion (*rhr5* and *cyrpA*). Because

168 such events preserve the phylogenetic history of the genes involved, we systematically examined

169 the evidence for introgression or gene transfer events across the complete subgenus by testing the

170 congruence of each gene tree to the species tree for genes with one-to-one orthologues. Beyond the

171 region that includes *rh5* (Fig 2b, Supplementary Fig. 4a), few signals of gene flow between

172   parasites infecting the same host species were obtained (n=11) suggesting that these events were

173   rare or usually strongly deleterious (Supplementary Fig. 5).

174   The *Laverania* subgenus evolved to infect chimpanzees and gorillas but, on a genome-wide

175   scale, the convergent evolution of host-specific traits has not left a signature (Supplementary Note

176   2). We therefore examined each CDS independently and were able to identify genes with

177   differences fixed within specific hosts, falling into three categories: 53 in chimpanzee-infective

178   parasites, 49 in gorilla-infective and 12 with fixed traits in both host species (Fig. 2; Supplementary

179   Table 5a). For at least 67 genes, these differences were unlikely to have arisen by chance (p <0.05)

180   and GO term enrichment analysis revealed that several of these genes are involved in host invasion

181   and pathogenesis (Supplementary Table 5b) including *rh5* (which has a signal for convergent

182   evolution even when the introgressed tree topology is taken into consideration; Supplementary Fig.

183   4b). *Rh5* is the only gene identified in *P. falciparum* that is essential for erythrocyte recognition

184   during invasion, via binding to Basigin. *P. falciparum rh5* cannot bind to gorilla Basigin and binds

185   poorly to the chimpanzee protein[19]. We notice that one of the convergent sites is known to be a

186   binding site for the host receptor Basigin[20] (Supplementary Fig. 4b). The gene *eba-165* encodes a

187   member of the erythrocyte binding like (EBL) super family of proteins that are involved in

188   erythrocyte invasion. Although *eba-165* is a pseudogene in *P. falciparum*[21], it is not a pseudogene

189   in the other *Laverania* species and may therefore be involved in erythrocyte invasion, like other

190   EBL members. The protein has three convergent sites in gorillas. One falls inside the F2 region, a

191   domain involved in the interactions with erythrocyte receptors. The role of this protein and of these

192   convergent sites in the invasion of gorilla red cells remains to be determined. Finally, genes

193   involved in gamete fertility (the 6-cysteine protein P230) or implicated in *Plasmodium* invasion of

194   erythrocytes (doc2[22]) also displayed signals of convergent evolution. Twelve parasite coding

195   sequences had fixed differences at the same amino-acid position in chimpanzees and gorillas. Of

196   these P230 was the only one found with a position that was different and fixed across all three host

197   species. P230 is involved in gamete development and trans-specific reproductive barriers[23], possibly

198   through enabling male gametes to bind to erythrocytes prior to exflagellation[24]. Host-specific

199   residues observed in P230 might affect the efficiency of the binding to the erythrocyte receptors and

200   result from co-evolution between the parasite molecule and the host receptor.

201

202   **Subtelomeric gene families**

203   To date, the only in-depth data on the subtelomeric gene families of the *Laverania* have come

204   from *P. reichenowi* and *P. falciparum*. These important families are well represented in our

205   assemblies (Supplementary Tables 2 and 6A) and we provide a comprehensive picture of their

206   evolution.

207   Most gene families were likely present in the ancestor of all Laverania. The same general

208   pattern of one-to-one orthology throughout the subgenus indicates that many underwent gene

209   duplication early (e.g. FIKK) or prior (e.g. ETRAMP, PHIST and SURFIN) to the development of a

210   distinct Laverania lineage. Only a subset displayed contractions or expansions between specific

211   Laverania species (Fig. 3 and Supplementary Table 6a, Supplementary Figure 7). For these latter

212   families, Clade A and most species of Clade B clearly differ in their composition. *P. blacklocki*

213   (Clade B) is intermediate in its composition. Some gene families, like the group of exported

214   proteins *hyp4, hyp5*, *mc-2tm* and *EPF1*, have expanded only in *P. praefalciparum* and *P.*

215   *falciparum* (and even more in *P. falciparum* for *hyp4* and *hyp5*). Since all four are components of

216   Maurer's clefts, an organelle involved in protein export[25], some evolution of function in this

217   organelle may have been an important precursor to human infection. The family of acyl-CoA

218   synthetase genes, reported to be expanded and diversified in *P. falciparum*[26] is in fact expanded

219   across the *Laverania* and has four fewer copies in *P. falciparum* (Supplementary Fig. 6). Other

220   genes that show clade or group specific expansion include DBLmsp, glycophorin binding protein

221   and CLAG (Supplementary Fig. 7).

222   One striking inter-clade difference concerns the largest gene family that is likely common to all

223   other malaria species: the *Plasmodium* interspersed repeat family (*pir,* which includes the *rif* and

224   *stevor* families in *P. falciparum*) (Fig. 3, 4). This family has been proposed to be involved in

225   important functions such as antigenic variation, immune evasion, signalling, trafficking, red cell

226   rigidity and adhesion[27] and yet has expanded only in Clade B, after the *P. blacklocki* split (Fig 3).

227   The *rif* genes comprise a small conserved group and a much larger group of more diverse members

228   that contains just 13 genes from Clade A species and at least 180 members per Clade B species

229   (Fig. 4). There is however no evidence for host-specific adaptation in these sequences.

230   In contrast, a subset of *stevor* genes showed strong host-specific sequence diversification (Fig

231   4 and Supplementary Fig. 8). Based on full-length alignments, there is a deep phylogenetic split

232   between *stevor* genes but when only short conserved protein motifs are considered, a group of

233   Stevor proteins (*stevor II,* Fig 4a) forms a cluster comprising almost entirely of members from

234   gorilla-infecting species. Since *stevor* genes are known to be involved in host–parasite interactions

235   (such as binding to host glycophorin C in *P. falciparum*[28]), this host specific sequence may reflect

236   sequence differences in host-specific factors in gorillas.

237

238 **Evolution of *var* genes**

239 The *var* genes, crucial mediators of pathogenesis and the establishment of chronic infection through

240 cytoadherence and immune evasion, are the best studied *P. falciparum* multi-gene family and

241 unique to the *Laverania*[29]. They are two-exon genes and their products have three types of major

242 domain; exon 1 encodes Duffy Binding like (DBL) and Cysteine Rich Interdomain Regions (CIDR)

243 and exon 2 encodes Acidic Terminal Sequence (ATS)[30]. Similar to *P. falciparum*, our data are

244 consistent with all *Laverania* species having *var* genes (Fig. 3) that retain a two-exon structure and

245 are organized into subtelomeric or internal *var* gene clusters. There are however three notable

246 features of *var* evolution within the sub-genus.

247 First, there is a deep division in how the repertoire is organised between the major clades. The

248 *var* genes of Clade B parasites, with the exception of *P. blacklocki*, resemble those of *P. falciparum*

249 in terms of genomic organisation, domain types and numbers (Fig 5, Supplementary Table 7). In

250 contrast, the repertoires of Clade A parasites and *P. blacklocki* (treated as one group hereafter in

251 this section) differ in their domain composition, contain a novel CIDR-like domain (CIDRn, Fig 5a,

252 Supplementary Fig. 9) and have lower sequence diversity per domain but cluster into more sub-

253 groups than Clade B domains (Fig 5b, Supplementary Fig. 10). The paucity of domains similar to

254 those in *P. falciparum* (such as CIDRα) that are involved in cytoadherence to some specific and

255 common host receptors, means that if endothelial cytoadherence was important in Clade A, some

256 alternative receptors must have been utilised.

257 Second, in total there are 10 internal *var* gene clusters (confirmed by contiguous sequence data)

258 but 8 are oppositely oriented between the two clades (Supplementary Fig. 11, Supplementary Table

259 8). Clade B parasites also show a much greater number of associated GC-rich RNAs of unknown

260 Function (RUF) elements than Clade A (Supplementary Table 8).

261 Third, the ATS domains cluster tightly within Clade A. Within Clade B there is clear evidence

262 of species-specific diversification, except in *P. praefalciparum* and *P. falciparum* reflecting their

263 recent speciation. There is one intact ATS from *P. falciparum* as well as several pseudogenes that

264 cluster with Clade A (Fig 5c). Moreover, of seven internal *var* arrays (Supplementary Fig. 11) in

265 *P. falciparum*, containing a functional *var* gene, five terminate with one of these pseudogenes (on

266 the opposite DNA strand) suggesting that they may be remnants from ancient rearrangements. The

267 intact *P. falciparum* gene is *var2csa*, a *var*-like gene that is highly conserved between *P. falciparum*

268 isolates[31], involved both in cytoadherence in the placenta in primigravidae, and proposed to be a

269 central intermediate in *var* gene switching during antigenic variation[32]. We therefore propose

270   *var2csa* is a remnant of an ancient multigene family that has been maintained as a single complete

271   gene in *P. falciparum*, for the dual purposes of *var*-switching and placental cytoadherence.

272   There is other evidence of retention of ancient *var* gene sequence across the subgenus. First,

273   in Clade B we find a nearly full length *var* pseudogene that has highest similarity to *P. adleri* and *P.*

274   *gaboni var* genes, within an internal *var* cluster on chromosome 4 in *P. falciparum* and *P.*

275   *praefalciparum* but on the opposite strand to the other *var* genes. It is found in all *P. falciparum*

276   isolates, but not in *P. reichenowi*. Second, in *P. gaboni* and *P. adleri,* three genes have the N-

277   terminal DBLα/CIDRα architecture typical of Clade B genes and their domains cluster within Clade

278   B based on similarity (Fig. 5b, larger nodes). Directly adjacent to two of these *var* genes are two *rif*

279   pseudogenes that also show greatest similarity to those from Clade B. Last, we find a further nine

280   *rif* pseudogenes of Clade A parasites that cluster with Clade B *rif* genes (Fig. 4). If these

281   observations reflect retention of ancient copies, their high sequence conservation suggests that they

282   are under extremely unusual selection pressure. Alternatively, they may represent relics of gene

283   transfer between species that occurred after the Clade A/B split.

284

285   **Conclusion**

286   We have produced high quality genomes and used mutation rates and generation times, covering the

287   full range of most recent estimates, to calculate the date of speciation for all known members of the

288   *Laverania*, with only a small margin of error. In our analysis, we have shown that the successful

289   infection of humans by *P. falciparum* occurred quite recently and involved numerous parasites

290   rather than a single one as previously proposed. After the establishment in its new host, the parasite

291   population went through a bottleneck around 5,000 years ago during the period of rapid human

292   population expansion due to farming (Fig. 1b). We summarise the major genomic events during the

293   evolution of the *Laverania* in Fig. 6.

294   As a result of our analyses we propose the following series of events for the emergence of *P.*

295   *falciparum* as a major human pathogen. First, the crucial lateral transfer event of the *rh5* locus

296   between Clade A and B parasites may also have involved *var* and *rifin* genes in other parts of the

297   genome that, because of their orientation on the opposite strand, were not lost during later

298   recombination. Next, facilitatory mutations are likely to have occurred in *rh5* that in the first

299   instance allowed invasion of both gorilla and human red cells. Modern humans emerged more than

300   300,000 years ago[33] and existed as small isolated populations[12]. Our evidence suggests that *P.*

301   *falciparum* and *P. praefalciparum* started to diverge around 40,000-60,000 years ago. In the

302   following 40,000 years with low population densities in humans and gorillas there would have not

303   been high selection pressure to optimise infectivity in either the hosts or vectors, enabling at least

304   some movement of parasites between hosts. We find evidence for gene flow between lineages

305   throughout this period. The expansion of the human population with the advent of farming likely

306   led to strong evolutionary pressure for mosquito species (specifically *An. gambiae*) to feed

307   primarily on humans[34]. Therefore, the existing human infective (*P. falciparum)* genotypes would be

308   selected for human and appropriate vector success and the fittest would rapidly expand. Subsequent

309   rapid accumulation of mutations that favoured growth in humans, and in the anthropophilic vectors

310   such as *An. gambiae*, are likely to have occurred to increase human-specific reproductive success.

311   The resulting specific parasite genotypes that expanded (and appeared as an emergence from a

312   bottleneck), would have had a much lower probability of a direct transfer back to apes. With

313   experiments on gorillas and chimpanzees not possible it will be difficult directly to prove the

314   precise combination of different alleles that allowed the emergence of *P. falciparum*. However, for

315   the genes that we have implicated in this process, existing data (www.genedb.org, plasmodb.org)

316   suggest they are expressed throughout the life cycle but that only half have been characterised. This

317   opens up new opportunities for future studies on host specificity and host adaptation in

318   *Plasmodium*.

319

320 **Online Methods**

321 **Sample collection**

322   All but two infected blood samples from chimpanzees (*Pan troglodytes troglodytes*) and

323 gorillas (*Gorilla gorilla gorilla*) were obtained from the sanctuary "Parc de La Lékédi", Bakoumba

324 (Haut-Ogooué, Gabon), during routine sanitary controls of the animals. This park holds various

325 primate species, including gorillas, chimpanzees and monkeys (*Cercopithecinae*), that have been

326 orphaned due to bushmeat-poaching activities and have been confiscated by the Gabonese

327 Government, quarantined at the Centre International de Recherches Médicales de Franceville

328 (CIRMF, Gabon) and finally released into semi-free ranging enclosures in the sanctuary. Every six

329 months, chimpanzees (12 individuals) and gorillas (2 individuals) are anesthetized for medical

330 check up. Blood samples were collected from the animals during sanitary controls (July 2011,

331 September 2012, May 2013 and December 2013). Two additional infected blood samples were

332 obtained from gorilla orphans (GG05, GG06) seized by the Gabonese government in 2011 and 2013

333 and sent to the CIRMF for a quarantine before being released in a sanctuary. All animal work was

334 conducted according to relevant national and international guidelines. From each animal, 15 ml of

335 whole blood were collected in EDTA tubes. For all samples but three, white blood cell depletion

336 was performed on 10 ml of the freshly collected samples using cellulose columns as described in [35].

337 Remaining blood was subsequently used for DNA extraction and detection of *Plasmodium*

338 infections as described in Ollomo et al[3]. Overall, 15 blood samples from 7 chimpanzees and 4

339 gorillas were found to contain the *Laverania* samples used in the present study (Supplementary

340 Table 1).

341

342 **Ethical consideration**

343 The animal well-being was guaranteed by the veterinarians of the "Parc of la Lékédi" and the

344 CIRMF who proceeded to the sanitary controls and the blood sampling. Because these blood

345 samples were collected as part of the standard protocol for the sanitary controls (and not specifically

346 for our experiment), our study did not need the approval of an Institutional Animal Care or Use

347 Committee. Note also that our study did not involve randomisation nor blinding.

348

349 **Sample preparation**

350   Three methods were used for DNA amplification prior to sequencing (Supplementary Table 1).

351 For all but one sample, whole genome amplification (WGA) was performed with a REPLI-g Mini

352 Kit (Qiagen) following a modified protocol[36] to enrich genomic DNA. The genome of *P. blacklocki*

353   was generated using selective WGA (sWGA) as indicated in[37] using 20 primers, followed by a

354   WGA. Finally, for the PprfG03 (a *P. praefalciparum* isolate) and PadlG02 (a *P. adleri* isolate)

355   samples, we used a cell sorting approach[38].

356

357   **Sample sequencing**

358   All samples were first sequenced with Illumina. Amplification-free Illumina libraries of 400-

359   600 bp were prepared from the enriched genomic DNA[39] and run on MiSeq and HiSeq 2000 (v3

360   chemistry) Illumina machines.

361   After the Illumina sequencing, six samples with a combination of the least number of multiple

362   infections (see below) and the lowest level of host contamination were chosen for long read

363   sequencing, using Pacific Biosciences (PacBio). The DNA of the samples (after WGA) was size-

364   selected to 8 kb and sequenced with the C3/P5 chemistry. The number of SMRT cells (Pacific

365   Bioscience sequencing runs) used varied between samples (Supplementary Table 1).

366

367   **Genome assembly, genome QC, split of infection & annotation**

368   *Determination of multiple infections*

369   To initially quantify multiple infections and so allow samples to be selected for PacBio

370   sequencing from those comprising a low number of species, Illumina reads from each sample were

371   mapped against a concatenation of all available *Cox 3* and *CytB* genes of the *Laverania* from NCBI,

372   using SNP-o-Matic[40] (parameter chop=5) to position reads only where they aligned perfectly. SNP-

373   o-Matic returns all the positions of repetitive mapping reads. This output allowed us to count the

374   read depth of these two genes across all species and therefore determine the number and relative

375   amount of different malaria species per sample.

376

377   *Whole genome amplification (WGA) bias*

378   The uneven coverage that resulted from WGA bias, host contamination and multiple infections

379   presented a challenge for sequence assembly. To overcome the bias and the host contamination,

380   each DNA sample was sequenced deeper than normally necessary. Lower coverage of the

381   subtelomeres was obtained for the sWGA sample (*P. blacklocki*) meaning that the subtelomeres in

382   that assembly were not as complete as those in the assemblies for other species.

383

384   *Long reads (Pacific Bioscience) assemblies*

385   Six reference genomes were assembled using HGAP[41], with different settings for the genome

386   size parameter, ranging from 23 Mb (*P. reichenowi*) to 72 Mb (*P. billcollinsi*). This parameter

387   encodes how many long reads are corrected for use in the assembly and depends on the host

388   contamination and the amount of different isolates in the samples. The obtained contigs from

389   HGAP were ordered with ABACAS[42] against a *P. falciparum* 3D7 reference that has no

390   subtelomeric regions. Assembly errors and WGA artefacts were manually corrected using ACT[43].

391   After this step, three iterations of ICORN2[44] were run, followed by another ABACAS step,

392   allowing overlapping contigs to be merged (parameter: ABA_CHECK_OVERLAP=1). For the

393   PrG01, PgabG01 and PadlG01 assembly, we also ran PBjelly to close some of the sequencing

394   gaps[45].

395

396   *Host decontamination*

397   To detect and remove sequence data derived from host DNA, contigs were compared with the

398   chimpanzee or gorilla genomes using BLAST. Contigs were considered as host contamination if

399   more than 50% of their BLAST hits had higher than 95% identity to any of the great ape genomes.

400   Unordered contigs with a GC content >32% were searched against the non-redundant nucleotide

401   database, to detect and remove further contaminants.

402

403   *Resolving multiple infections*

404   The first assembled genome was a single *P. reichenowi* infection, PrG01. We detected low

405   levels of *P. vivax-like* and virus contamination (TT virus, AB038624.1), which were excluded. For

406   quality control, the assembly was compared against the existing PrCDC[9] reference genome. The

407   number of *Plasmodium* interspersed repeats (PIRs) was similar, and there were no breaks in

408   synteny. There were however significantly fewer sequencing gaps and 17 Rep20 regions could be

409   found (a known repeat close to the telomeres in *P. falciparum)*. Thus, the assembly of PacBio data

410   (PrG01; Supplementary Table 2) appears to be of higher quality than the existing *P. reichenowi*

411   PrCDC reference.

412   The *P. adleri* sample comprised a single infection. Because a large number of cycles of

413   amplification were used, a greater number of SMRT cells were sequenced (Supplementary Table 1)

414   to overcome the problem of uneven coverage resulting in under-represented regions. An estimated

415   genome size of 60 Mb was chosen for the HGAP analysis to ensure that all regions were covered.

416   PgabG01 was a *P. gaboni* isolate with a *P. vivax-like* co-infection. To detect contigs of *P.*

417   *vivax*, unordered contigs (those that could not be placed against Pf3D7 using ABACAS) were

418   searched against the protein sequences of *P. falciparum* 3D7 and the *P. vivax* PvP01 reference

419   genome using TBLASTx. For each contig, the relative number of genes hitting against the two

420   genomes was used to assign it to *P. gaboni* or *P. vivax*. In most cases, all genes for a given contig

421   consistently hit only one genome so that the attribution to either species was clear. Overall, 14 Mb

422   of *P. vivax-like* sequences were obtained that will be described elsewhere.

423   The *P. billcollinsi* genome (PbilcG01) was obtained from a co-infection with a *P. gaboni*

424   genome (PgabG02). Rather than ordering the contigs just against Pf3D7 with ABACAS, contigs

425   were ordered against a combined reference comprising *P. gaboni* (PgabG01) and the Pf3D7

426   (parameters: overlap 500 bp, identity 90%). The species designation of contigs was confirmed with

427   a TBLASTx searches of annotated genes against a combination of the proteomes of PgabG01 and

428   PrCDC. For subtelomeric gene families, contigs were attributed to species if the hit was significant

429   for one species, not the other. Some of the contigs could not be attributed unambiguously and were

430   discarded. Due to sequencing gaps, some of the core genes are missing from the final assembly.

431   The sample used to produce the *P. praefalciparum* genome (PprfG01) had a high level of host

432   contamination, a low level of co-infection with *P. adleri* and contained two distinct

433   *P. praefalciparum* genotypes. For the core genome, we used iCORN to select the dominant

434   genotype at each position. Where it was not possible to phase the genotypes, due to a lack of

435   variation, we assumed that they were identical. In the subtelomeres however, it was possible to

436   distinguish but not phase the two *P. praefalciparum* genotypes resulting in approximately twice the

437   number of *var* genes as seen in *P. falciparum*. Due to contamination of construction vectors (*E.*

438   *coli*) and host, 29 SMRT cells were sequenced and the HAGP parameter for the assembly size was

439   set to 60 Mb. Contigs were screened against *P. adleri* and *P. falciparum* to exclude a *P. adleri* co-

440   infection. All of the contigs that had a *P. falciparum* BLAST hit or had no clear hit (such as those

441   containing species-specific gene families) were attributed to the *P. praefalciparum* assembly. Last,

442   all samples (Supplementary Table 1) including five *P. falciparum* genomes were mapped against

443   the Pf3D7, *P. praefalciparum and P. adleri* assemblies. Contigs were excluded where more

444   normalized hits to the three *P. adleri* samples were found than to one of the two other *P.*

445   *praefalciparum* samples. Similarly, this method was used to eliminate the remote possibility that

446   any of the contigs in the *P. praefalciparum* assembly were in fact derived from *P. falciparum* co-

447   infection.

448   The *P. blacklocki* sample was from a single infection. Due to sWGA, the PacBio sequence data

449   covered regions not covered by Illumina but due to the bias of the primers, the subtelomeres were

450   not covered fully. However, the internal *var* gene clusters are all assembled. Some of the core genes

451   from this species are also missing.

452

453   *Annotation*

454   The genomes were annotated as described in[46]. In short, the annotation of *P. falciparum*

455   (version July 2015) was transferred with RATT[47] and new gene models were called with

456   Augustus[48]. Obvious structural errors in core genes were manually corrected in Artemis[49].

457

458   **Mapping - generation of further samples**

459   To generate the gene sequence for different samples, Illumina reads were mapped against a set

460   of reference genomes using BWA[50] and default parameters. For the gorilla samples, we mapped

461   against the combined PacBio reference genomes of *P. adleri*, *P. blacklocki* and *P. praefalciparum*

462   and for the chimpanzee samples, the combined references of *P. gaboni* (PgabG01), *P. billcollinsi*

463   and *P. reichenowi* (PrG01). SNPs with Phred score $\geq$ 100 were called using GATK

464   UnifiedGenotyper[51] v2.0.35 (parameters: -pnrm POOL -ploidy 2 -glm POOLBOTH). From these

465   SNP calls we constructed the new gene set, masking regions in genes with less than 10x coverage

466   of 'properly' (correct distance and orientation) mapped paired reads. To generate the sequences of

467   the other 13 isolates, homozygous SNP calls were obtained (consensus program from bcftools-

468   1.2[52]). We quality controlled the SNP calling by regenerating PrCDC and PgabG02 gene set from

469   PrG01 and PgabG01, respectively and confirmed that they were placed with nearly no differences

470   in a phylogenetic tree.

471

472   **Orthologous group determination and alignment**

473   Orthologous groups were identified using OrthoMCL v1.4[53] across: (i) the seven core

474   *Laverania* genomes; (ii) the seven core genomes, the *Laverania* isolates PgabG02, PrCDC and *P.*

475   *falciparum* IT, as well as two outgroup genomes *Plasmodium vivax* Sal1 and *Plasmodium knowlesi*

476   strain H; and (iii) just Pf3D7, PprfG01 and PrG01. *P. praefalciparum* II was excluded due to its

477   partial genome. From these groups, different complete sets of 1:1 orthologues were extracted:

478   (1) "Lav12sp" set of 3,369 orthologues across the seven core *Laverania* species, the PrCDC

479   and *P. falciparum* IT isolates, *P. vivax* and *P. knowlesi*

480   (2) "Lav25st" set of 424 1:1 orthologues from across the 25 *Laverania* isolates, including the

481   previously published *P. reichenowi* CDC and five *P. falciparum* isolates (3D7, IT, DD2,

482   HB3 and 7G8[9]).

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

484   (4) "Lav15st" set of 3,808 orthologues, with at least two representative sequences per species,

485   excluding *P. blacklocki* and the most divergent *P. praefalciparum* lineage Pprf3.

486   (5) "Lav3sp" set of 4,826 1:1 orthologues across all the *P. reichenowi*, *P. praefalciparum* and

487   *P. falciparum* isolates

488   The first two sets were used to reconstruct the species tree, the third one for the comparative

489   genomic analyses (introgression, convergence and gene family evolution), the fourth one for the

490   analyses of within species polymorphism and the fifth one for the analysis of *P. falciparum*

491   adaptive evolution.

492   To reduce the rate of false positives in the evolutionary analyses due to misalignments (e.g.[54]),

493   codon-based multiple alignments were performed using PRANK[55,56] with the -codon and +F

494   options, as it was shown to outperform other programs in the context of the detection of positive

495   selection[57,58]. Prior to aligning codons, low complexity regions were excluded in the nucleotide

496   sequences using dustmasker[59] and in amino acid sequences using segmasker[60] from NCBI-BLAST.

497   Poorly aligned regions were excluded using Gblocks[61], with default settings.

498

499   **Analysis of interspecific gene flow, introgression or gene transfer**

500   *Species-tree inference*

501   Two ML trees were performed using RAxMLv8.1.20[62] to illustrate the phylogenetic

502   relationships between the *Laverania* species and genotypes studied here using the "Lav12sp" and

503   the "Lav25st" set of orthologues. For each tree, multiple nucleotide alignments of each orthologous

504   group were conducted as described above. Trees were then constructed from the concatenated

505   alignments of the "Lav12sp" set of orthologues for the species tree and the "Lav25st" set for the

506   strain tree using RAxML and the following options "-m GTRGAMMA –f a -# 100". Trees were

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

508   clade for the genotype tree.

509

510   *Tree topology test*

511   Interspecific gene flow was investigated by testing congruence between each gene tree

512   topology and the species tree topology. We performed the Shimodaira-Hasegawa test (SH test[63])

513   using RAxMLv8.1.20 to test whether the phylogenetic tree for each gene significantly differed from

514   the *Laverania* species tree. Topology tests were based on multiple nucleotide alignments of the

515   4,350 "Lav7sp" set of orthologues. For each coding sequence, RAxML was called with the options

516   "-m GTRGAMMA –f h".

517

518   **Convergent evolution analyses**

519   *Genome-wide test of convergent evolution*

520   Convergent substitutions can occur by chance and the number of random convergent

521   substitutions between two lineages is correlated with the number of divergent substitutions

522   observed in these two lineages[64,65]. Excess of convergent substitutions in specific branch pairs can

523   thus be identified by analyzing the correlation between the number of convergent and divergent

524   substitutions between all the branch pairs in a phylogeny using orthogonal regression, and looking

525   for outlier branch pairs: branch pairs with a high positive residual show an excess of convergent

526   substitutions relatively to the number of divergent substitutions[64]. We used the software Grand-

527   Convergence (available at https://github.com/dekoning-lab/grand-conv) to estimate for each

528   chromosome the numbers of divergent and convergent substitutions between all branch pairs in the

529   *Laverania* tree and investigate whether branch pairs including *Laverania* species infecting the same

530   host species (gorilla or chimpanzee) presented an excess of convergence. Analyses were performed

531   under different models of amino-acid evolution: LG, WAG, JONES and DAYHOFF.

532

533   *Gene-based detection of convergent evolution throughout the Laverania*

534   For each orthologue of the "Lav7sp" set, the number and percentage of fixed amino acid

535   differences between parasites infecting the same host were calculated, *i.e.* the number of positions

536   showing the same amino acid within a host species but different amino acid between host species.

537   Alignments of all the available sequences ("Lav15st") from all the sequenced isolates were then

538   used to determine what number of host-specific differences were fixed within each host and each

539   species. To evaluate whether the observed number of host-specific fixed differences in an alignment

540   can be attributed to neutral evolution/purifying selection alone (with no positive selection), we used

541   a simulation-based approach. For each coding sequence, 1,000 sequences of the same size were

542   simulated, evolving along the same tree with the same specified branch lengths, substitution model,

543   codon frequencies and omega (d_N/d_S), using the program Evolver from PAML v4.8a[66]. The program

544   Codeml from PAML v4.8a[66] was first used to estimate the tree, the codon frequencies and the

545   average omega values for each of the coding sequences with fixed amino acid differences. For each

546   simulated dataset, the number of fixed amino-acid differences between the parasites infecting a

547   same host was estimated. The probability of observing *n* fixed differences was then computed as the

548   proportion of the simulated dataset of 1000 sequences that showed at least the same number of

549   fixed differences as observed in the real data.

550

551

552   **Tests for positive selection**

553   *Branch site tests*

554   To search for genes that have been subjected to positive selection in the *P. falciparum* lineage

555   alone after the divergence from *P. praefalciparum*, we used the updated Branch site test [67]

556   implemented in PAML v4.4c [66]. This test detects sites that have undergone positive selection in a

557   specific branch of the phylogenetic tree (foreground branch). The "Lav3sp" set of 4,826

558   orthologuous groups between *P. reichenowi*, *P. praefalciparum I* and *P. falciparum* was used for

559   the test. d_N/d_S ratio estimates per branch and genes were obtained using Codeml (PAML v4.4c)

560   with a *free-ratio* model of evolution. This identified 139 genes with a significant signal of positive

561   selection in *P. falciparum* only.

562   A Branch Site test was also applied, for each gene, on each terminal branch of the entire species

563   tree using the "Lav7sp" dataset. d_N/d_S ratio estimates per branch and genes were obtained using

564   Codeml (PAML v4.4c) with a *free-ratio* model of evolution, Figure 2.

565

566   *McDonald–Kreitman (MK) tests*

567   Selection in *P. falciparum* was also tested using McDonald–Kreitman (MK) tests [68] to compare

568   the polymorphism within species to the divergence between species, using *P. praefalciparum* as the

569   outgroup. Analyses were performed using the 4,826 "Lav3sp" set of orthologues. MK tests were

570   performed as described before[9]. Thirty-five genes had an MK ratio significantly higher than 1.

571

572   **Gene Ontology enrichment analyses**

573   Analysis of Gene Ontology (GO) term-enrichment was performed in R, using TopGO[69] with

574   default parameters. GO annotations from GeneDB were used but with unreviewed automated

575   annotations excluded.

576

577   **Gene family analyses**

578   To estimate the differential abundance of gene families across species, the Gene products and

579   the Pfam domains were counted and analysed by the variance of the occurrence. Unless otherwise

580   stated, trees were constructed using PhyML[70] (default parameters) or RAxML[62] (model estimated)

581   from alignments generated with Muscle[71] and trimmed with Gblocks[61] in Seaview[72] with default

582   values. Many of the findings were confirmed manually through ACT and bamview[49]. The analysis

583   of the *var* genes was performed on *var* genes larger than 2.5kb. Domains were called with the

584   HMMer models from varDom[73]. Distance matrices were generated based on BLASTp scores,

585   without filtering low complexity regions. Representation was done in R through the heatmap.2

586   program from gplot (see also Supplementary Note 3).

587

588   *Allelic dimorphisms*

589   For the analysis of dimorphism in *msp*, all sequences available for the *Laverania* were

590   downloaded from Uniprot[74]. Data were subsampled to obtain a similar number of sequences for

591   each group. Phylogenetic trees were constructed with PhyML[70], using default parameters and drawn

592   in Figtree. The *eba-175* alignment was visualized with Jalview[75].

593

594   **Divergence Dating**

595   Alignments of the *Laverania* included intergenic regions where possible. Assuming 402–681

596   mitotic events per year (Supplementary Note 1) and a mutation rate of 3.78E-10 for 4 mitotic

597   events[76,77](mutation rate from latter paper was taken from Pf3D7 line without drugs), equivalent to

598   around 0.9–1.5 mutations per genome per year. Although we observed similar mutation rates in

599   clinical samples (Supplementary Note 1), these estimates have potential errors and therefore we

600   report ratios of divergence times in the figures that are robust to errors in these parameters. For

601   coalescence based estimates of speciation times, G-Phocs[78] was used and multiple sequentially

602   Markovian coalescent (MSMC) on segregating sites[79] was used to estimate the *P. falciparum*

603   bottleneck.

604

620
621
622 **Data availability**
623 All sequences have been submitted to the European Nucleotide Archive. The accession numbers of
624 the raw reads, and assembly data can be found in Supplementary Table 9. The genomes are being
625 submitted to EBI, project ID PRJEB13584. The genomes are available from plasmodb.org and
626 from ftp://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/Laverania/.

627
628 **Competing financial interests.**
629 None
630
631 **Computer code**
632 Custom computer code is available on request.

633
634
635

636   **Legends**

637

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

647

648   **Figure 2. Overview of the analyses of core genes over all *Laverania* genomes.** (a) Summary of
649   evolution of core genes. From outer to inner track: scatterplot of branch site test for each genome
650   (see Supplementary Table 4 for *P. falciparum* data); per-species d_N/d_S values ($0.5 < d_N/d_S < 2$);
651   orthologues represented by vertical black lines under the chromosome track represent, with dots
652   representing *P. falciparum* 3D7 *var* genes on the forward (blue) or reverse strands (red), or *var*
653   pseudogenes (black); average of the relative polymorphism (π) across species, with the underlying π
654   for each species calculated from multiple strains ("Lav15st" dataset) and normalized by the average
655   for that species; signatures of convergent evolution based on host-specific fixed differences analysis
656   with the chromosome 4 region that includes the *Rh5* locus highlighted (black box). (b) Magnified
657   view of the *Rh5* region that is enriched with host specific fixed differences.   Convergent evolution
658   analysis was performed using orthologues conserved across the *Laverania*. Filled circles represent
659   the subset of differences that were fixed within all the isolates available ("Lav15st" set) and for
660   which we could reject neutral evolution (for the gene list see Supplementary Table 5).

661

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

666

667

668

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.

673

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 (******).

683

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*.

689
690

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. *Malaria journal* **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   Roy, S. W. The Plasmodium gaboni genome illuminates allelic dimorphism of immunologically important surface antigens in P. falciparum. *Infection, Genetics and Evolution* **36**, 441-449, doi:10.1016/j.meegid.2015.08.014 (2015).

15   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).

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

17   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).

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

19   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).

20   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).

21   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).

22   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).

23   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).

749   24   Eksi, S. *et al.* Malaria transmission-blocking antigen, Pfs230, mediates human red blood cell binding to
750   exflagellating male parasites and oocyst production. *Molecular Microbiology* **61**, 991-998, doi:10.1111/j.1365-
751   2958.2006.0528.x (2006).
752   25   Mundwiler-Pachlatko, E. & Beck, H. P. Maurer's clefts, the enigma of Plasmodium falciparum. *Proceedings of*
753   *the National Academy of Sciences of the United States of America* **110**, 19987-19994,
754   doi:10.1073/pnas.1309247110 (2013).
755   26   Bethke, L. L. *et al.* Duplication, gene conversion, and genetic diversity in the species-specific acyl-CoA
756   synthetase gene family of Plasmodium falciparum. *Mol Biochem Parasitol* **150**, 10-24,
757   doi:10.1016/j.molbiopara.2006.06.004 (2006).
758   27   Cunningham, D., Lawton, J., Jarra, W., Preiser, P. & Langhorne, J. The pir multigene family of Plasmodium:
759   antigenic variation and beyond. *Mol Biochem Parasitol* **170**, 65-73, doi:10.1016/j.molbiopara.2009.12.010
760   (2010).
761   28   Niang, M. *et al.* STEVOR is a Plasmodium falciparum erythrocyte binding protein that mediates merozoite
762   invasion and rosetting. *Cell Host Microbe* **16**, 81-93, doi:10.1016/j.chom.2014.06.004 (2014).
763   29   Kraemer, S. M. & Smith, J. D. A family affair: var genes, PfEMP1 binding, and malaria disease. *Curr Opin*
764   *Microbiol* **9**, 374-380, doi:10.1016/j.mib.2006.06.006 (2006).
765   30   Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**,
766   498-511, doi:10.1038/nature01097 (2002).
767   31   Bordbar, B. *et al.* Genetic diversity of VAR2CSA ID1-DBL2Xb in worldwide Plasmodium falciparum
768   populations: impact on vaccine design for placental malaria. *Infection, genetics and evolution : journal of*
769   *molecular epidemiology and evolutionary genetics in infectious diseases* **25**, 81-92,
770   doi:10.1016/j.meegid.2014.04.010 (2014).
771   32   Frank, M., Dzikowski, R., Amulic, B. & Deitsch, K. Variable switching rates of malaria virulence genes are
772   associated with chromosomal position. *Mol Microbiol* **64**, 1486-1498, doi:10.1111/j.1365-2958.2007.05736.x
773   (2007).
774   33   Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution.
775   *Nature reviews. Genetics* **13**, 745-753, doi:10.1038/nrg3295 (2012).
776   34   Carter, R. & Mendis, K. N. Evolutionary and historical aspects of the burden of malaria. *Clinical microbiology*
777   *reviews* **15**, 564-594, doi:10.1128/CMR.15.4.564-594.2002 (2002).
778   35   Auburn, S. *et al.* An effective method to purify plasmodium falciparum dna directly from clinical blood
779   samples for whole genome high-throughput sequencing. *PLoS ONE* **6**, doi:10.1371/journal.pone.0022213
780   (2011).
781   36   Oyola, S. O. *et al.* Optimized whole-genome amplification strategy for extremely AT-biased template. *DNA*
782   *research : an international journal for rapid publication of reports on genes and genomes* **21**, 661-671,
783   doi:10.1093/dnares/dsu028 (2014).
784   37   Oyola, S. O. *et al.* Whole genome sequencing of Plasmodium falciparum from dried blood spots using
785   selective whole genome amplification. *bioRxiv*, doi:10.1101/067546 (2016).
786   38   Boissiere, A. *et al.* Isolation of Plasmodium falciparum by flow-cytometry: implications for single-trophozoite
787   genotyping and parasite DNA purification for whole-genome high-throughput sequencing of archival samples.
788   *Malaria Journal* **11**, 163, doi:10.1186/1475-2875-11-163 (2012).
789   39   Quail, M. A. *et al.* Optimal enzymes for amplifying sequencing libraries. *Nat Methods* **9**, 10-11,
790   doi:10.1038/nmeth.1814 (2012).
791   40   Manske, H. & Kwiatkowski, D. SNP-o-matic. *Bioinformatics* **25**, 2434-2435,
792   doi:10.1093/bioinformatics/btp403 (2009).
793   41   Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.
794   *Nat Methods* **10**, 563-569, doi:10.1038/nmeth.2474 (2013).
795   42   Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: algorithm-based automatic
796   contiguation of assembled sequences. *Bioinformatics Vol. 25 No. 15*, 1968-1969 (2009).
797   43   Carver, T. *et al.* Artemis and ACT: viewing, annotation and comparing sequences stored in relational database.
798   *Bioinformatics* **24**, 2672-2676, doi:10.1093/bioinformatics/btn529 (2008).
799   44   Otto, T. D., Sanders, M., Berriman, M. & Newbold, C. Iterative Correction of Reference Nucleotides (iCORN)
800   using second generation sequencing technology. *Bioinformatics* **26**, 1704-1707,
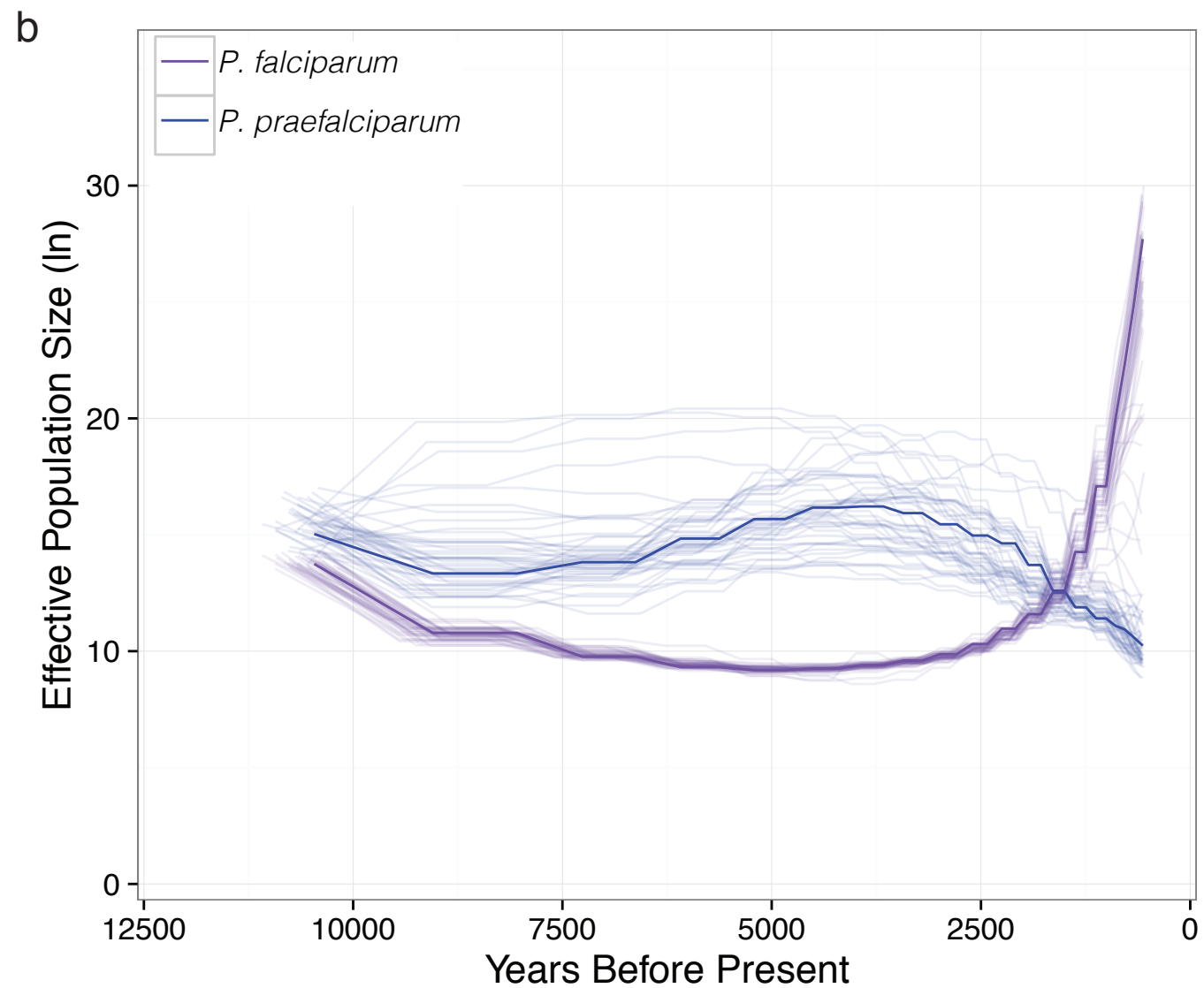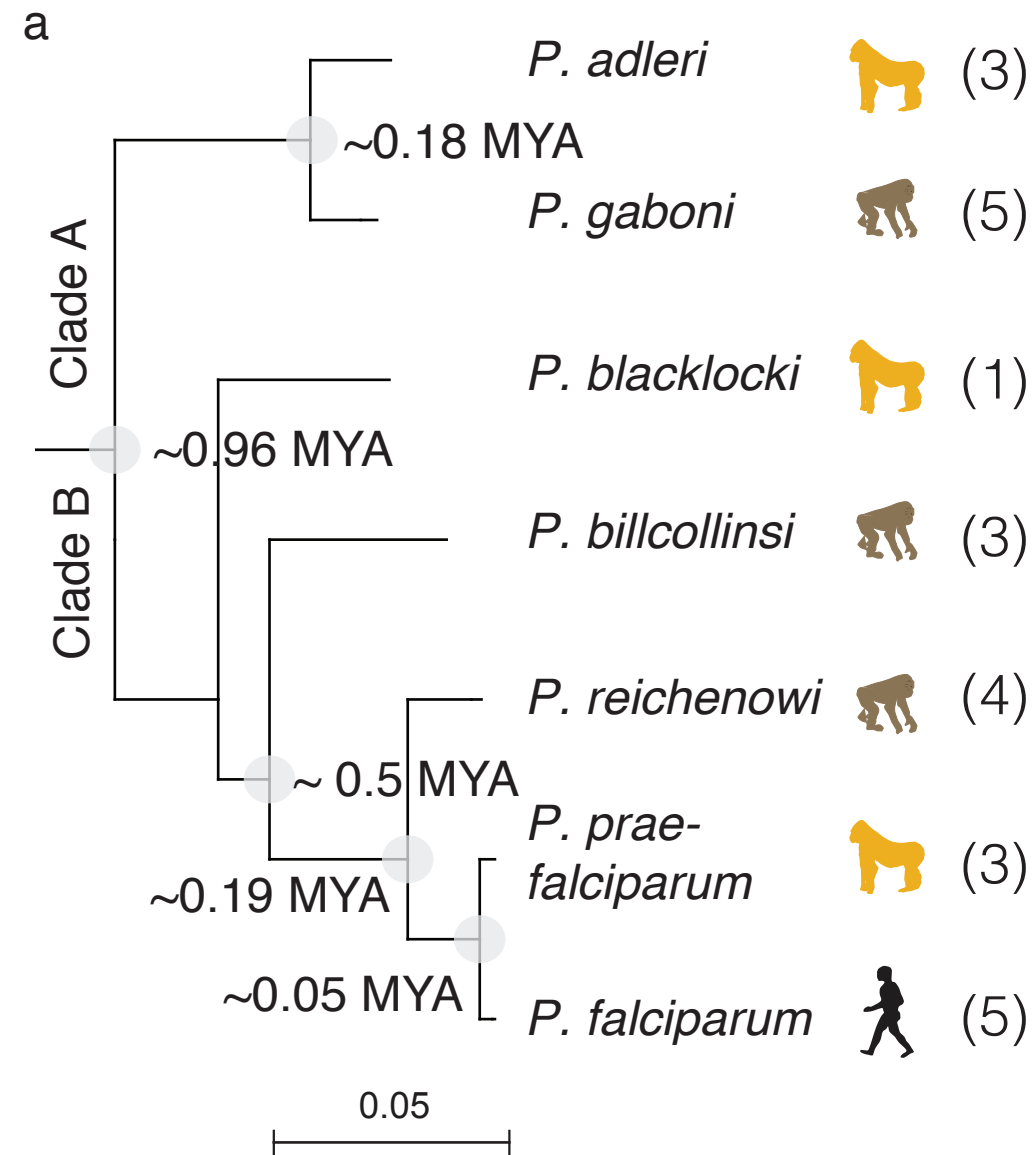801   doi:10.1093/bioinformatics/btq269 (2010).
802   45   English, A. C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing
803   technology. *PLoS One* **7**, e47768, doi:10.1371/journal.pone.0047768 (2012).
804   46   Otto, T. D. From sequence mapping to genome assemblies. *Methods Mol Biol* **1201**, 19-50, doi:10.1007/978-1-
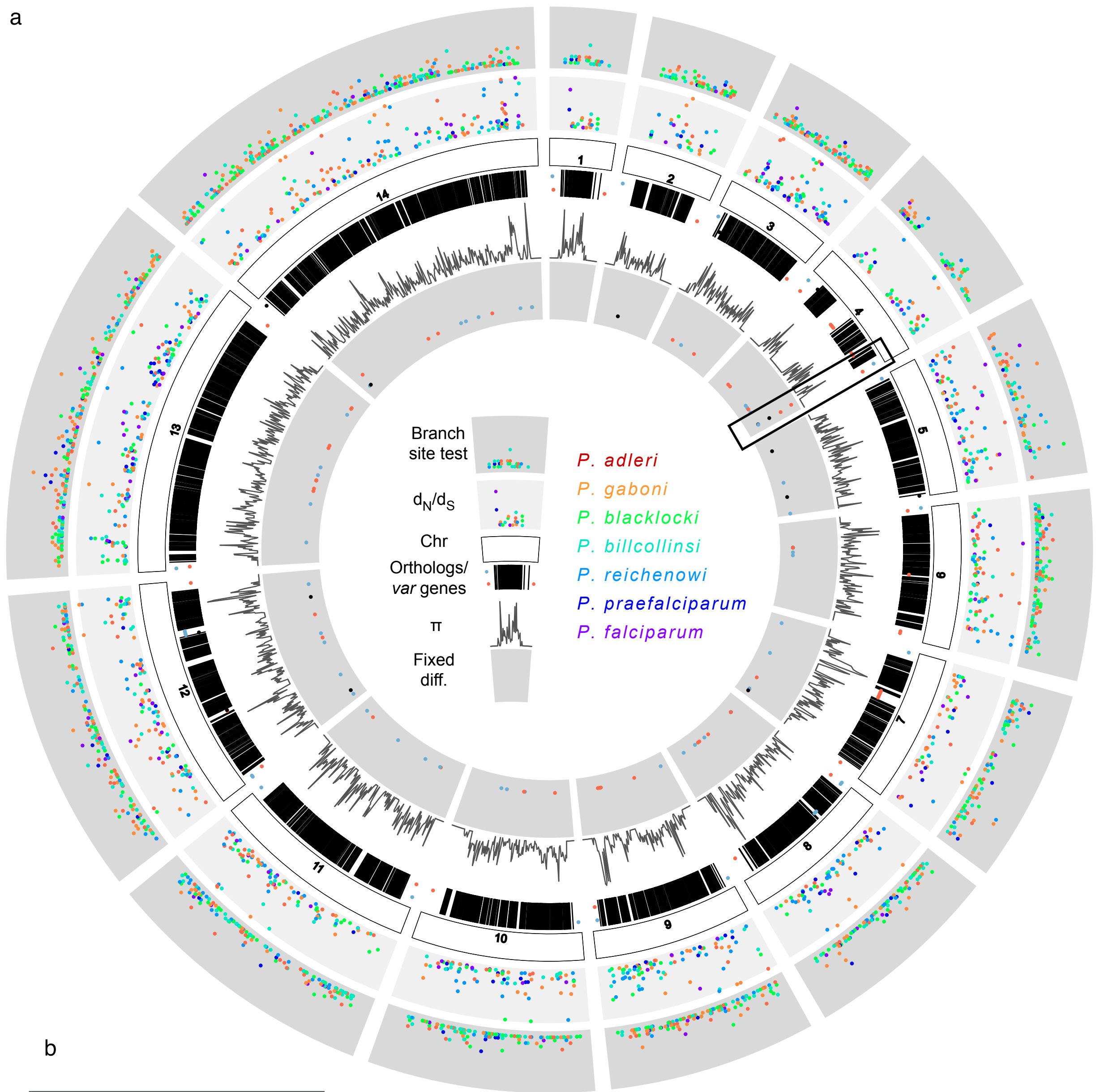805   4939-1438-8_2 (2015).
806   47   Otto, T. D., Dillon, G. P., Degrave, W. S. & Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucleic*
807   *Acids Research* **39**, e57, doi:10.1093/nar/gkq1268 (2011).

808   48   Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**,
809   W435-439, doi:10.1093/nar/gkl200 (2006).
810   49   Carver, T. *et al.* BamView: visualizing and interpretation of next-generation sequencing read. *Briefings in*
811   *bioinformatics*, doi:10.1093/bib/bbr073 (2013).
812   50   Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*
813   **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
814   51   McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation
815   DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
816   52   Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158,
817   doi:10.1093/bioinformatics/btr330 (2011).
818   53   Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes.
819   *Genome Res* **13**, 2178-2189, doi:10.1101/gr.1224503 (2003).
820   54   Jordan, G. & Goldman, N. The Effects of Alignment Error and Alignment Filtering on the Sitewise Detection
821   of Positive Selection. *Molecular biology and evolution* **29**, 1125-1139, doi:10.1093/molbev/msr272 (2012).
822   55   Loytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions.
823   *Proceedings of the National Academy of Sciences of the United States of America* **102**, 10557-10562,
824   doi:10.1073/pnas.0409137102 (2005).
825   56   Loytynoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and
826   evolutionary analysis. *Science* **320**, 1632-1635, doi:10.1126/science.1158395 (2008).
827   57   Fletcher, W. & Yang, Z. The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of
828   Positive Selection. *Molecular biology and evolution* **27**, 2257-2267, doi:10.1093/molbev/msq115 (2010).
829   58   Markova-Raina, P. & Petrov, D. High sensitivity to aligner and high rate of false positives in the estimates of
830   positive selection in the 12 Drosophila genomes. *Genome Research* **21**, 863-874, doi:10.1101/gr.115949.110
831   (2011).
832   59   Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A Fast and Symmetric DUST Implementation to
833   Mask Low-Complexity DNA Sequences. *Journal of Computational Biology* **13**, 1028-1040,
834   doi:10.1089/cmb.2006.13.1028 (2006).
835   60   Wootton, J. C. & Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases.
836   *Computers & Chemistry* **17**, 149-163, doi:10.1016/0097-8485(93)85006-X (1993).
837   61   Castresana, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic
838   Analysis. *Molecular biology and evolution* **17**, 540-552, doi:10.1093/oxfordjournals.molbev.a026334 (2000).
839   62   Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.
840   *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).
841   63   Shimodaira, H. & Hasegawa, M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic
842   Inference. *Molecular biology and evolution* **16**, 1114, doi:10.1093/oxfordjournals.molbev.a026201 (1999).
843   64   Castoe, T. A. *et al.* Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings*
844   *of the National Academy of Sciences of the United States of America* **106**, 8986-8991,
845   doi:10.1073/pnas.0900233106 (2009).
846   65   Thomas, G. W. C. & Hahn, M. W. Determining the Null Model for Detecting Adaptive Convergence from
847   Genomic Data: A Case Study using Echolocating Mammals. *Molecular biology and evolution* **32**, 1232-1236,
848   doi:10.1093/molbev/msv013 (2015).
849   66   Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular biology and evolution* **24**,
850   1586-1591, doi:10.1093/molbev/msm088 (2007).
851   67   Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an Improved Branch-Site Likelihood Method for Detecting
852   Positive Selection at the Molecular Level. *Molecular biology and evolution* **22**, 2472-2479,
853   doi:10.1093/molbev/msi237 (2005).
854   68   McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652-
855   654, doi:10.1038/351652a0 (1991).
856   69   Rahnenfuhrer, A. A. a. J. topGO: Enrichment analysis for Gene Ontology. *R package*,
857   doi:10.18129/B9.bioc.topGO (2010).
858   70   Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
859   performance of PhyML 3.0. *Syst Biol* **59**, 307-321, doi:10.1093/sysbio/syq010 (2010).
860   71   Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids*
861   *Res* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).
862   72   Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for
863   sequence alignment and phylogenetic tree building. *Molecular biology and evolution* **27**, 221-224,
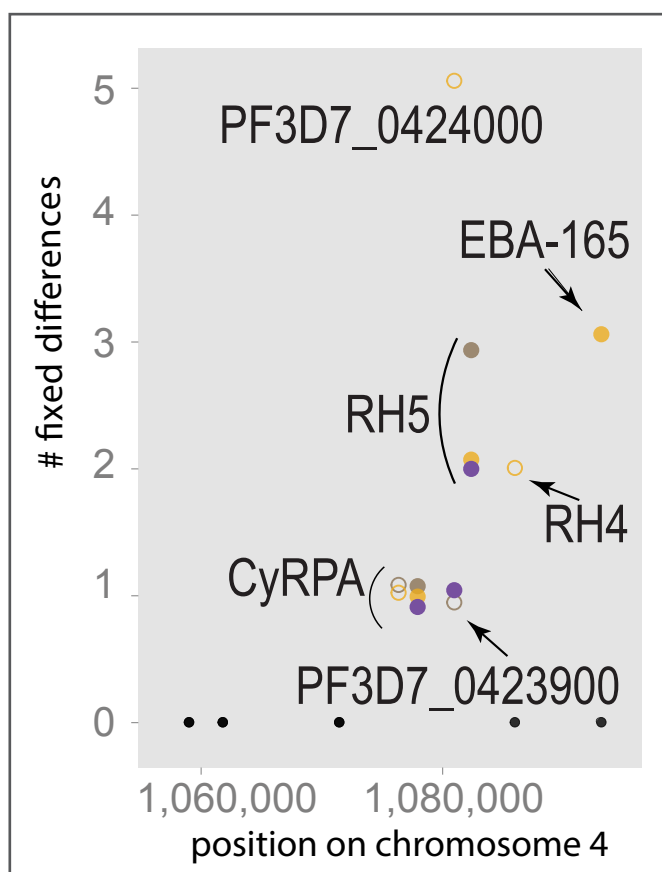864   doi:10.1093/molbev/msp259 (2010).

865   73   Rask, T. S., Hansen, D. A., Theander, T. G., Gorm Pedersen, A. & Lavstsen, T. Plasmodium falciparum
866   erythrocyte membrane protein 1 diversity in seven genomes--divide and conquer. *PLoS Comput Biol* **6**,
867   doi:10.1371/journal.pcbi.1000933 (2010).
868   74   UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212, doi:10.1093/nar/gku989
869   (2015).
870   75   Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple
871   sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191,
872   doi:10.1093/bioinformatics/btp033 (2009).
873   76   Claessens, A. *et al.* Generation of antigenic diversity in Plasmodium falciparum by structured rearrangement
874   of Var genes during mitosis. *PLoS Genet* **10**, e1004812, doi:10.1371/journal.pgen.1004812 (2014).
875   77   Bopp, S. E. *et al.* Mitotic evolution of Plasmodium falciparum shows a stable core genome but recombination
876   in antigen families. *PLoS Genet* **9**, e1003293, doi:10.1371/journal.pgen.1003293 (2013).
877   78   Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human
878   demography from individual genome sequences. *Nat Genet* **43**, 1031-1034, doi:10.1038/ng.937 (2011).
879   79   Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome
880   sequences. *Nat Genet* **46**, 919-925, doi:10.1038/ng.3015 (2014).

881

a

Clade A

Clade B

~0.18 MYA

~0.96 MYA

~0.5 MYA

~0.19 MYA

~0.05 MYA

P. adleri (3)

P. gaboni (5)

P. blacklocki (1)

P. billcollinsi (3)

P. reichenowi (4)

P. prae-falciparum (3)

P. falciparum (5)

0.05

b

P. falciparum
P. praefalciparum

Effective Population Size (ln)

Years Before Present

a

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

PF3D7_0424000
EBA-165
RH5
RH4
CyRPA
PF3D7_0423900

position on chromosome 4

fixed in great ape parasites
fixed in gorilla parasites
fixed in chimpanzee parasites

P. adleri   L F I P S Y N T Y D E N I T N D F
P. gaboni   L F M P S Y N S Y D E N I A N N F
P. blacklocki L F I P S N N T Y D E N V S N D F
P. billcolllinsi L F M P S N N I Y D E N M A N D F
P. reichenowi L F M P S N N I Y D E N V A N D F
P. praefalciparum L F I P S N K T Y D E N V D N D F
P. falciparum
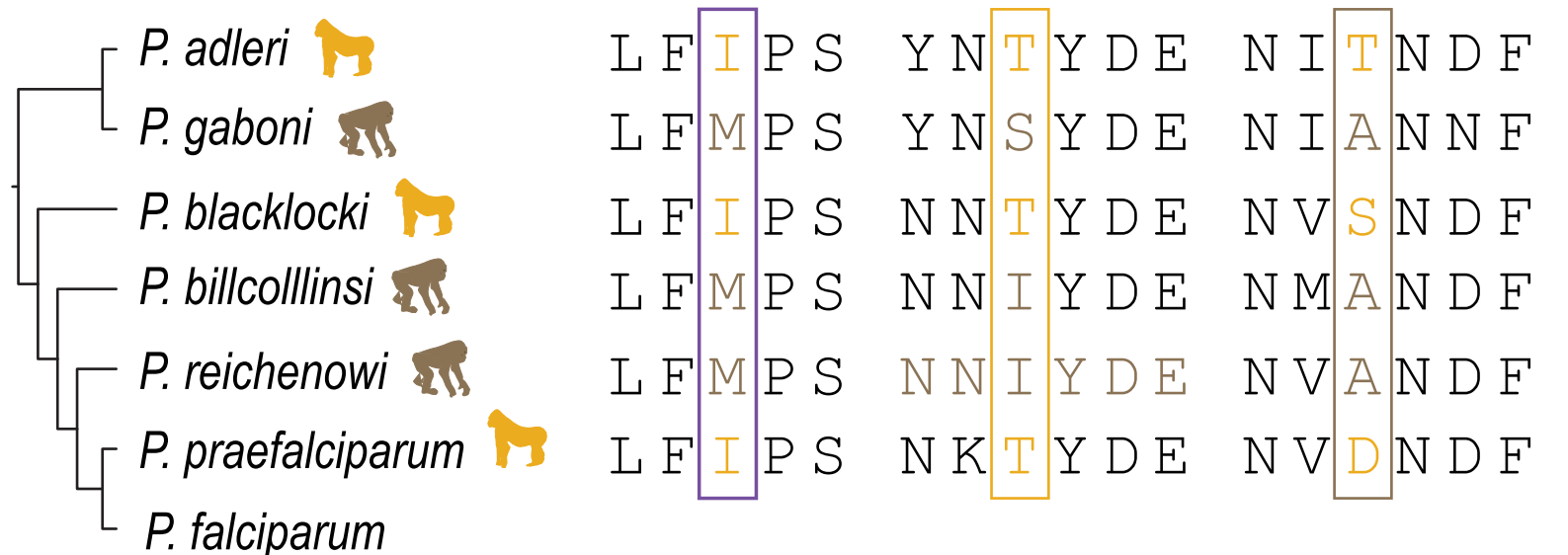
	P. falciparum	*P. praefalciparum*	*P. reichenowi*	*P. billcollinsi*	*P. blacklocki*	*P. gaboni*	*P. adleri*
var ≥ 2.5kb	67	112	92	35	43	61	58
rif > 250aa	183	582	440	280	13	12	19
stevor > 250aa	41	78	54	45	21	1	13
hyp4	9	2	1	1	0	0	0
hyp5	9	2	2	1	5	3	0
Maurer	13	12	5	4	7	0	0
EPF1	8	6	4	3	5	1	1
RESA-like	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

Rifin

Stevor II

Stevor I

Rifin B

Rifin A
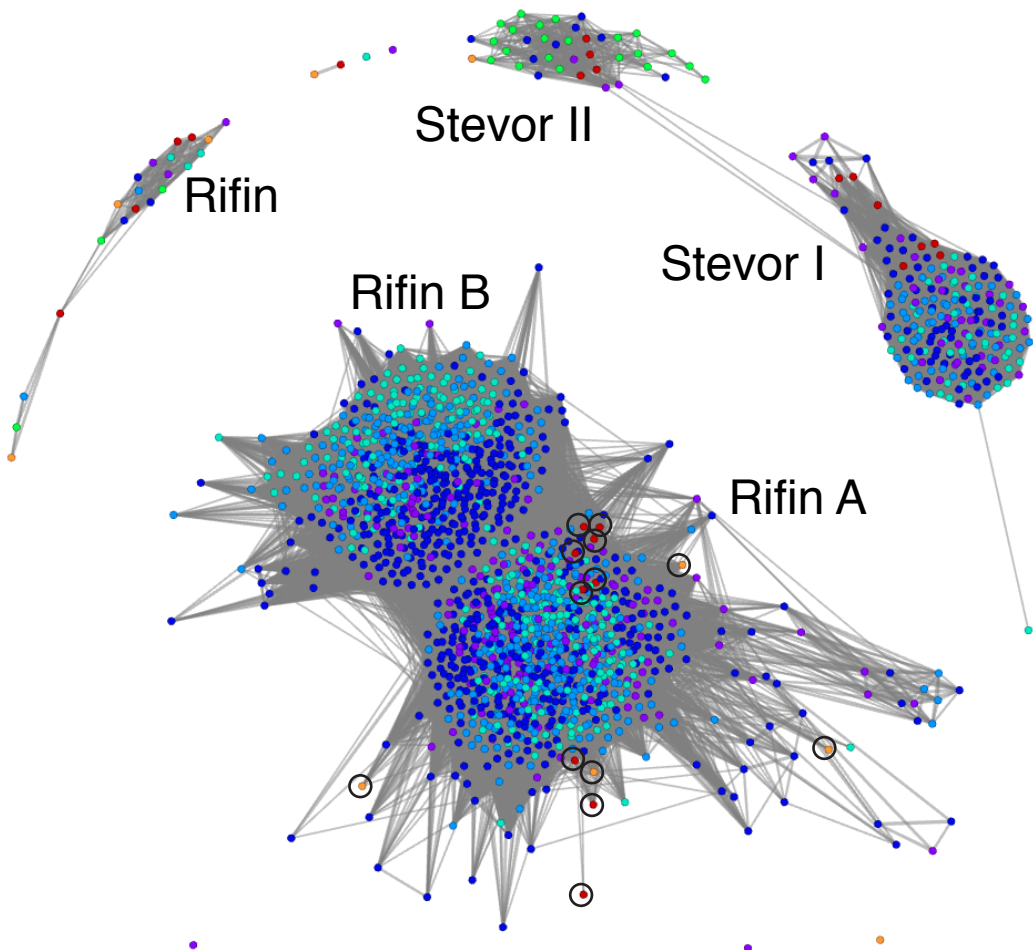
P. adleri *P. blacklocki* *P. reichenowi* *P. falciparum*

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

* Pfvar2csa /
** Clade A

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