Title: The genome of the zoonotic malaria parasite *Plasmodium simium* reveals adaptations to host-switching

Tobias Mourier[1], Denise Anete Madureira de Alvarenga[2], Abhinav Kaushik[1], Anielle de Pina-Costa[3,4,5§], Olga Douvropoulou[1§], Qingtian Guan[1§], Francisco J. Guzmán-Vega[6§], Sarah Forrester[7], Filipe Vieira Santos de Abreu[3,8], Cesare Bianco Júnior[3,9], Julio Cesar de Souza Junior[10], Silvia Bahadian Moreira[11], Zelinda Maria Braga Hirano[10], Alcides Pissinatti[11], Maria de Fátima Ferreira-da-Cruz[3,9], Ricardo Lourenço de Oliveira[3,8], Stefan T. Arold[6,12], Daniel C. Jeffares[7], Patrícia Brasil[3,4], Cristiana Ferreira Alves de Brito[2], Richard Culleton[13], Cláudio Tadeu Daniel-Ribeiro[3,9*] & Arnab Pain[1,14,15*]

1) Pathogen Genomics Laboratory, Biological and Environmental Sciences and Engineering (BESE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

2) Grupo de Pesquisa em Biologia Molecular e Imunologia da Malária, Instituto René Rachou, Fundação Oswaldo Cruz (Fiocruz), Belo Horizonte/MG, 30190-009, Brazil

3) Centro de Pesquisa, Diagnóstico e Treinamento em Malária (CPD-Mal), Fiocruz, Rio de Janeiro/RJ, 21040-360, Brazil

4) Laboratório de Pesquisa Clínica em Doenças Febris Agudas, Instituto Nacional de Infectologia Evandro Chagas, Fiocruz, Rio de Janeiro/RJ, 21040-360, Brazil

5) Centro Universitário Serra dos Órgãos (UNIFESO), Teresópolis/RJ, 25964-004, Brazil

6) Computational Bioscience Research Center, Biological and Environmental Sciences and Engineering (BESE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

7) Department of Biology and York Biomedical Research Institute, University of York, Wentworth Way, York, YO10 5DD, U.K.

8) Laboratório de Mosquitos Transmissores de Hematozoários. Instituto Oswaldo Cruz (IOC), Fiocruz, Rio de Janeiro/RJ, 21040-360, Brazil

9) Laboratório de Pesquisa em Malária, IOC, Fiocruz, Rio de Janeiro/RJ, 21040-360, Brazil

10) Universidade Regional de Blumenau (FURB), Centro de Pesquisas Biológicas de Indaial (CEPESBI)/ Projeto bugio, Blumenau and Indaial, SC, Brazil.

11) Centro de Primatologia do Rio de Janeiro (CPRJ/Inea), 25940-000, Guapimirim, RJ, Brazil

12) Centre de Biochimie Structurale, CNRS, INSERM, Université de Montpellier, 34090 Montpellier, France

13) Division of Molecular Parasitology, Proteo-Science Center, Ehime University, Toon, Ehime 791-0295, Japan

14) Global Station for Zoonosis Control, Global Institution for Collaborative Research and Education (GI-CoRE), Hokkaido University, N20 W10 Kita-ku, Sapporo, Japan

15) Nuffield Division of Clinical Laboratory Sciences (NDCLS), University of Oxford, Headington, Oxford, OX3 9DU, UK

*) Corresponding authors. CTDR: malaria@fiocruz.br, AP: arnab.pain@kaust.edu.sa

§) Contributed equally, arranged alphabetically

46 **Abstract**

47

48 Background

49 *Plasmodium simium*, a malaria parasite of non-human primates (NHP) was recently shown to
50 cause zoonotic infections in humans in Brazil. We sequenced the *P. simium* genome to
51 investigate its evolutionary history and to identify any genetic adaptions that may underlie the
52 ability of this parasite to switch between host species.

53 Results

54 Phylogenetic analyses based on whole genome sequences of *P. simium* from humans and
55 NHPs reveals that *P. simium* is monophyletic within the broader diversity of South American
56 *Plasmodium vivax*, suggesting *P. simium* first infected NHPs as a result of a host-switch of *P.*
57 *vivax* from humans. The *P. simium* isolates show the closest relationship to Mexican *P. vivax*
58 isolates. Analysis of erythrocyte invasion genes reveals differences between *P. vivax* and *P.*
59 *simium*, including large deletions in the Duffy Binding Protein 1 (DBP1) and Reticulocyte
60 Binding Protein 2a genes of *P. simium*. Analysis of *P. simium* isolated from NHPs and
61 humans revealed a deletion of 38 amino acids in DBP1 present in all human-derived isolates,
62 whereas NHP isolates were multi-allelic.

63 Conclusions

64 Analysis of the *P. simium* genome confirmed a close phylogenetic relationship between *P.*
65 *simium* and *P. vivax*, and suggests a very recent American origin for *P. simium.* The presence
66 of the DBP1 deletion in all human-derived isolates tested suggests that this deletion, in

combination with other genetic changes in *P. simium*, may facilitate the invasion of human red blood cells and may explain, at least in part, the basis of the recent zoonotic infections.

Keywords

Plasmodium simium, *Plasmodium vivax*, malaria, zoonosis, comparative genomics

Background

There are currently eight species of malaria parasites known to cause disease in humans: *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae*, *Plasmodium ovale curtisi*, *Plasmodium ovale wallikeri*, *Plasmodium knowlesi*, *Plasmodium cynomolgi* and *Plasmodium simium*. The last three species are non-human primate (NHP) parasites that have recently been shown to infect humans [1-3].

As interventions against human parasites, particularly *P. falciparum* and *P. vivax,* reduce their prevalence, the importance of zoonotic malaria is becoming increasingly apparent. In countries moving towards malaria elimination, the presence of potentially zoonotic parasite populations in NHPs is a significant obstacle.

The propensity of malaria parasites to switch hosts and the consequences of this for human health are highlighted by the fact that both *P. vivax* and *P. falciparum* first arose as human pathogens following host switches from non-human great apes in Africa [4-6]. Contact between humans and the mosquitoes that feed on NHPs is increasing due to habitat destruction and human encroachment into NHP habitats [7]. Therefore, there is increasing danger of zoonotic transmission and the emergence of novel human malaria pathogens.

Understanding how malaria parasites adapt to new hosts and new transmission environments allows the risks posed by novel zoonotic malaria outbreaks to be assessed.

The clinical epidemiology of zoonotic malaria varies according to the parasite species and the demographics of the human host population. Severe and fatal outcomes for people infected with *P. knowlesi* in Malaysia have been reported [8], whilst *P. cynomolgi* infection in the same region causes moderate or mild clinical symptoms [9]. Interestingly, both *P. knowlesi* and *P. cynomolgi* infections in the Mekong region appear less virulent than in Malaysia, and are often asymptomatic [3, 10]. This may be due to differences in the relative virulence of the circulating parasite strains and/or in the susceptibility of the local human populations. As NHP parasites have adapted to co-evolved with their natural hosts, it is impossible to predict their virulence in zoonotic infections. For example, the pathogenicity of *P. falciparum* has been attributed to its relatively recent emergence as a human pathogen [11], which is thought to have occurred following a single transfer from a gorilla in Africa [5].

Eighty-nine percent of malaria infections in Brazil are caused by *P. vivax* and over 99.9% of these cases occur in the Amazonian region. This region accounts for almost 60% of the area of Brazil and is home to 13% of the human population (https://www.ibge.gov.br/). Around 90% of the malaria cases registered outside the Amazonian region occur in the Atlantic Forest, a region of tropical forest that extends along the Brazil's Atlantic coast. The infections are apparently mild and are caused by a vivax-like malaria parasite transmitted by *Anopheles (Kerteszia) cruzii*, a mosquito species that breeds in the leaf axils of bromeliad plants [12].

A malaria outbreak in the Atlantic Forest of Rio de Janeiro in 2015/2016 was shown to be caused by the NHP parasite *P. simium* [1]. Parasite DNA samples collected from both humans and NHPs in the same region had identical mitochondrial genome sequences, distinct from that of *P. vivax* isolates collected anywhere in the world but identical to that of a *P. simium* parasite isolated from a monkey in the same region in 1966, and to all subsequent *P. simium* isolates recovered from NHPs [13, 14].

Previously, it was thought that *P. vivax* became a human parasite following a host switch from macaques in Southeast Asia, due to its close phylogenetic relationship with a clade of parasites infecting macaques in this region, and due to the high genetic diversity among *P. vivax* isolates from Southeast Asia [15]. However, it is now considered likely that it became a human parasite following a host switch from non-human great apes in Africa [6]. It seems likely that *P. vivax* was introduced to the Americas by European colonisers following Columbus' journey to the New World towards the end of the 15th Century, since present-day South American *P. vivax* is closely related to a strain of the parasite present, historically, in Spain [16]. However, there is some evidence to suggest that *P. vivax* may also have been introduced to South America in pre-Columbian times [17], and this, together with the post-Columbian influx of infected people from various regions of the world associated with colonization, may have contributed to the extensive genetic diversity of present day *P. vivax* in Central and South America [17].

133 *Plasmodium simium*, a parasite of various species of Platyrrhini monkey whose range is
134 restricted to the Atlantic Forest from Southeast and South of Brazil [18], is genetically,
135 morphologically, and immunologically similar to *P. vivax* [1, 19-22]. Based on this
136 similarity, it is likely that the *P. simium* parasite of NHPs in Brazil originated from *P. vivax*
137 following a host switch from humans. The recent 2015/2016 outbreak of *P. simium* in the
138 local human population of Rio de Janeiro's Atlantic Forest raises questions about the degree
139 of divergence between *P. vivax* and *P. simium*, and whether adaptation to NHPs has led to the
140 evolution of a parasite with relevance to human health that differs from that of *P. vivax*.
141

142 It remains unproven whether the current outbreak of *P. simium* in the human population of
143 Rio de Janeiro was the result of a single parasite transfer from a NHP to a human and its
144 subsequent transfer between people, or whether multiple independent transfers have occurred
145 with different NHP-derived parasites. Furthermore, the nature and degree of adaptation to
146 NHP hosts and a sylvatic transmission cycle that has occurred in *P. simium* following its
147 anthroponotic origin is relevant to understanding how malaria parasites adapt to new hosts. It
148 is of interest to determine whether the current, human-infecting *P. simium* parasites have
149 undergone recent changes at the genomic level which allow them to infect humans, as it has
150 been previously suggested that *P. simium* lacks this ability [23].
151

152 In order to address these questions, and so to better understand the epidemiology and natural
153 history of this emerging zoonotic parasite, we analysed whole genome sequences of *P.*
154 *simium* parasites isolated from both humans and NHPs.

155

156

157

Results

158

159

Genome assembly and phylogeny

160 From a single *P. simium* sample collected from Rio de Janeiro state in 2016 [1] short read
161 DNA sequences were obtained and assembled into a draft genome (see Supplementary
162 Materials). The assembled genome consists of 2,192 scaffolds over 1kb in length with a
163 combined size of 29 Mb (Table S1). Two scaffolds corresponding to the apicoplast and
164 mitochondrial genomes were also identified (Figure S1). Gene content analysis showed a
165 completeness of annotation comparable to previously published *Plasmodium* assemblies
166 (Figure S2). Hence, although the fragmented nature of the chromosome assemblies prevents
167 precise assessment of synteny conservation with other *Plasmodium* species, the set of gene
168 components in the *P. simium* genome is relatively complete. A phylogenetic tree constructed
169 from 3,181 of 1:1 orthologs of the annotated *P. simium* protein-coding genes with the
170 orthologues of *P. vivax, P. cynomolgi, P. coatneyi, P. knowlesi, P. malariae, P. falciparum,*
171 *P. reichenowi,* and *P. gallinaceum* confirmed that *P. simium* is very closely related to *P.*
172 *vivax* (Figure S3). Analysis of gene families revealed a *P. simium* gene repertoire largely
173 similar to *P. vivax* (Figure S4; Table S2) and no indication of a recent expansion of the *P.*
174 *simium* PIR gene family (Figure S5).

175

176 ### *P. simium-P. vivax* diversity analysis

177 To detect single nucleotide polymorphisms (SNPs) within the *P. vivax*/*P. simium* clade, short
178 Illumina paired-end sequence reads were mapped onto the *P. vivax* P01 reference genome
179 [24]. Reads were collected from eleven human-derived *P. simium* samples, two monkey *P.*
180 *simium* samples, two *P. vivax* samples from the Brazilian Amazon, and a previously
181 published *P. simium* CDC strain (originally isolated in 1966, see Methods). Data from a
182 range of *P. vivax* strains representing a global distribution was retrieved from the literature
183 [25]. Including only SNPs with a minimum depth of five reads, a total of 232,780 SNPs were
184 called initially across 79 samples. Sixteen samples were subsequently removed, primarily due
185 to low coverage, resulting in a total of 63 samples for further analysis (Table S3, Table S4).
186 Since few SNP loci were covered across all samples, and to enable diversity analysis, we
187 restricted all further analysis to the 124,968 SNPs for which data were available from at least
188 55 samples (Figure S6).

189

190 *P. simium*-*P. vivax* population analysis

191 A Principal Component Analysis (PCA) plot constructed from these genome-wide SNP loci
192 showed a clear separation between American and Asian *P. vivax* samples as well as a distinct
193 grouping of *P. simium* samples (Figure S7). The latter observation suggests that both human
194 and NHP *P. simium* isolates represent a single population that is genetically different from
195 American *P. vivax* populations. A similar pattern is observed by performing a
196 multidimensional scaling analysis of the SNP data (Figure S8). To enable a phylogenetic
197 approach, we constructed an alignment from the 124,968 SNP sites. In the resulting
198 phylogenetic tree, *P. vivax* strains generally clustered according to their geographical origin,
199 with clear separation of the Asian and American samples (Figure 1A, a tree with sample IDs

200 is available in Figure S9). *Plasmodium simium* samples clustered as a monophyletic group
201 with Mexican *P. vivax* samples (Figure 1A), consistent with a recent American origin for *P.*
202 *simium.* This phylogeny, based on genome-wide SNPs represents an "average" phylogeny
203 across the genome and cannot be considered to reflect a true history of parasite ancestry due
204 to the effects of recombination. It is possible that trees produced from individual genes might
205 reveal different phylogenetic relationships. Yet, the association between *P. simium* samples
206 and *P. vivax* samples from Mexico is also observed in a phylogenetic network (Figure S10).

207

208 To examine whether the *P. simium* isolates we studied were part of a continuous population
209 with local *P. vivax* parasites, we examined population ancestry with the ADMIXTURE
210 program [26] (Figure S11). This analysis is consistent with the PCA and MDS analyses
211 (Figure S7 & Figure S8) and the phylogenetic analysis of segregating SNPs (Figure 1),
212 showing that *P. simium* forms a genetically distinct population separate from *P. vivax*. The
213 absence of *P. simium-P. vivax* hybrids (indicating genetic introgression events) indicates that
214 *P. simium* has undergone a period of independent evolution.

215

216 <u>Genetic diversity and population divergence</u>

217 To characterise the *P. simium* population further, we estimated the nucleotide diversity in *P.*
218 *simium* and *P. vivax* samples (see Materials and Methods). *Plasmodium simium* diversity
219 (genome-median: 1.3×10^{-4}) is approximately six times lower than the diversity observed
220 when comparing all *P. vivax* samples (genome-median: 7.9×10^{-4}) (Figure 2A). Diversity
221 measured within coding sequences in *P. vivax* is consistent with previous reports [6]. The

222 median nucleotide divergence between *P. simium* and *P. vivax* genomes of 8.7×10^{-4} and the
223 low diversity within *P. simium* suggest that the strains we examined are part of a relatively
224 recent or isolated population.

225 The difference in diversity within *P. simium* and *P. vivax* populations will influence relative
226 measures of population differentiation such as F_{ST} [27], and we therefore calculated D_{XY}, an
227 absolute measure of diversity that is independent of the levels of diversity within the two
228 populations being compared [27, 28]. When comparing D_{XY} in 10kb windows across the
229 genome, the diversity observed between *P. simium* samples and *P. vivax* samples from
230 America is significantly lower than the diversity between *P. simium* and *P. vivax* populations
231 from Asia and Africa (Figure 2B), supporting the origin of *P. simium* from American *P. vivax*
232 populations.

233 The vast majority of genes have an D_{XY} of zero, consistent with a recent split between *P.*
234 *simium* and *P. vivax* (Figure S12A). We noticed that genes showing the highest D_{XY} are from
235 multi-gene families (Figure S12B). Although polymorphic antigens are expected to display
236 high levels of diversity, this could also be explained by the uncertainties in SNP calling
237 among multi-gene families with high levels of sequence redundancy.

238

239 <u>*P. simium* invadome components</u>

240 Binding and red blood cell invasion is mediated by two key malaria gene families expressed
241 in merozoites: the Duffy Binding Proteins (DBPs) and the Reticulocyte Binding Proteins
242 (RBPs). In *P. vivax*, DBPs bind to the Duffy Antigen Receptor for Chemokines (DARC) [29,
243 30], which is present on both host normocytes and reticulocytes, and RBPs are known to

restrict *P. vivax* to binding and invading reticulocytes [31-33]. Two DBPs, DBP1 and DBP2, are present in *P. vivax* P01 (Table S5). Recently, the reported protein structure of *P. vivax* RBP2b revealed the conservation of residues involved in the invasion complex formation [33]. RBPs can be divided into three subfamilies, RBP1, RBP2, and RBP3 [34]. The *P. vivax* P01 genome encodes 11 RBPs (including the reticulocyte binding surface protein, RBSA), of which three are pseudogenes (Table S5).

The *P. vivax* DBP and RBP protein sequences were used to search for *P. simium* orthologues, resulting in the detection of the two DBP proteins and RBP1a, RBP1b, RBP2a, RBP2b, and RBP3, and a failure to detect RBP2c and RBP2d (Figure 3; Table S5; Figure S13; Figure S14) across all sequenced *P. simium* samples. As in *P. vivax* genomes, the *P. simium* RBP3 is a pseudogene [35], indicating that conversion to a pseudogene happened prior to the split between *P. vivax* and *P. simium*.

To determine whether the apparent absences of individual RBP genes in *P. simium* was due to incomplete genome assembly, we examined the coverage of *P. simium* reads mapped onto *P. vivax* RBP gene loci. No *P. simium* coverage was observed at the RBP2c, RBP2d, and RBP2e genes in *P. simium* samples, including the previously published CDC strain (Figure S15).

Coverage of mapped reads across invadome-associated gene loci revealed no apparent elevated coverage in genes compared to their flanking genomic regions, which would have

266 been expected if the *P. simium* genome contained multiple (duplicated) copies of non-
267 assembled invasion genes (Figure S16).

268

269 Structural variation in *P. simium* Duffy Binding Protein 1

270 The *P. simium* assembly revealed that the invasion gene DBP1 contains a large deletion
271 within its coding sequence (Figure 3; Figure S17). Intriguingly, the previously published *P.*
272 *simium* CDC strain DBP1 (GenBank accession: ACB42432) [36] does not contain this
273 deletion ('simium CDC' in Figure 3C). A haplotype network confirms that this previously
274 published DBP1 gene is indeed a *P. simium* sequence (Figure S18), and the SNP analyses
275 consistently assign the CDC strain to the *P. simium* cluster (Figure 1, Figure S7, and Figure
276 S8). Compared to the *P. vivax* P01 reference genome the SalI reference harbours a 27 base
277 pair deletion in DBP1, in contrast to the 115 bp deletion observed in all *P. simium* samples
278 isolated from humans (Figure 3). The 27 base pair deletion is also present in most *P. vivax*
279 isolates (Figure S19). Additional deletion patterns exist among isolates, and in a few cases
280 multiple versions are detected within samples (Figure S19).

281

282 The presence of repetitive sequences within the DBP1 gene could potentially result in
283 aberrant assembly across the DBP1 locus, which may appear as an apparent deletion in
284 subsequent bioinformatic analysis. We tested this possibility and showed that the DBP1 gene
285 does not harbour any noticeable degree of repetitiveness at the nucleotide level (Figure S20).
286 Several read mapping analyses confirmed that the *P. simium*-specific 115 bp deletion was not
287 an assembly artefact (Figure S21-S23).

We next designed primers for PCR amplification of a nucleotide sequence that spans the deleted region in the *P. simium* DBP1 gene and tested the occurrence of these deletion events in a range of *P. vivax* and *P. simium* field samples from Brazil. All *P. vivax* samples tested by PCR produced bands consistent with absence of the deletion whereas all samples from human-infecting *P. simium* produced bands consistent with the presence of the precise 115 bp deletion (Figure S24, top & middle). Interestingly, NHP-infecting *P. simium* isolates contained a mix of sequences with and without deletions (Figure S24, bottom). PCR results are summarized in Table S6. Under the simplified assumption that the sampled NHP *P. simium* samples reflect the DBP1 allele frequency in the entire *P. simium* population (4 with deletion in DBP1, 8 without the deletion; Table S6), the probability of randomly sampling 15 human *P. simium* samples all having the DBP1 deletion is 6.97×10^{-8} using a binomial distribution. If the *P. simium*-specific deletion in DBP1 is a prerequisite for the ability to infect humans, this suggests that only a subset of NHP-infecting *P. simium* parasites currently possess the ability to infect humans.

A large, 3012 bp additional deletion was observed in the *P. simium* RBP2a gene, the presence of which was also supported by read mapping (Figure 3, Figure S25-S28). PCR-based genotyping results confirmed the presence of this deletion in all *P. simium* isolates irrespective of their host (Figure S29, Table S6).

To test for the presence of insertions and deletions (indels) in other protein-coding genes, we curated a set of indels by comparing short indels predicted by the DELLY software [37] to *P. simium* assemblies. This set consisted of 244 indels in 222 genes (Table S7). Indel sizes were almost exclusively integers of three (hence conserving reading frames) and indels were predominantly found in low-complexity regions and in genes encoding long proteins (Figure S30). Notably, short indels were recorded in 138 genes with known functional annotations with an overrepresentation of DNA-binding function (GO:0005488 'binding', p=0.0011; GO:0043565 'sequence-specific DNA binding', p=0.0062).

Potential structural implications of the deletions in DBP1 and RBP2a

We next investigated whether the observed deletions render DBP1 and RBP2a non-functional. DBP1 contains a large extracellular region, which includes the N-terminal DBL region which mediates the association with DARC in *P. vivax* [38], followed by a largely disordered region and a cysteine-rich domain (Figure 3C). DBP1 has a single-pass transmembrane helix and a short cytoplasmic tail. The deletion observed in the human-infecting *P. simium* only affects the disordered region, leaving the flanking domains intact. We produced homology models of the DBL domains from the *P. vivax* strain P01, the human-infecting *P. simium* strain AF22, and the *P. simium* CDC strain, based on the crystal structure of the > 96% identical DBL domain of *P. vivax* bound to DARC (PDB ID 4nuv). Whereas no significant substitutions were found in the DBL domain between both *P. simium* sequences, our analysis showed that residue substitutions between *P. simium* and *P. vivax* DBL domains cluster in proximity of the DARC binding site (Figure S31). Based on our models, these substitutions are unlikely to negatively affect the association with DARC,

332 supporting the idea that the DBL domains of both *P. simium* would be capable of binding to
333 human DARC. Hence, the human-infecting *P. simium* DBL1 probably retains the capacity to
334 bind to human DARC, but has the interacting domain positioned closer to the membrane than
335 in the NHP-infecting CDC strain.

336

337 The deletion we detected in the *P. simium* RBP2a was larger, resulting in the loss of 1003
338 amino acid residues. These residues are predicted to form a mostly α-helical extracellular
339 stem-like structure that positions the reticulocyte binding domain away from the membrane
340 (Figure 3D). However, given that the deletion affects neither the transmembrane region, nor
341 the receptor-binding domain, our analysis suggests the resulting truncated RBP2a protein
342 could still associate with the human receptor, but that the binding event would occur closer to
343 the merozoite membrane.

344

345

346 **Discussion**

347

348 We present the genome of *Plasmodium simium*, the eighth malaria parasite species known to
349 infect humans in nature. In recent evolutionary time, *P. simium* has undergone both
350 anthroponosis and zoonosis making it unique for the study of the genetics underlying host-
351 switching in malaria parasites. Analysis of its genome confirmed a close phylogenetic
352 relationship between *P. simium* and *P. vivax*, and further analyses on single nucleotide
353 divergence support a very recent American origin for *P. simium.*

354

355 Two proteins involved in host invasion, DBP1 and RBP2a, were found to harbour extensive
356 deletions in *P. simium* compared to *P. vivax*. Interestingly, experimental analysis of *P.*
357 *simium* samples revealed that isolates from human hosts all carried the DBP1 deletion,
358 whereas isolates from NHPs displayed both absence and presence of the deletion. This DBP1
359 deletion is not present in the *P. simium* isolated from a brown howler monkey in the 1960s,
360 which was previously shown to be incapable of infecting humans [23], although some degree
361 of laboratory adaptation of this parasite may have affected its genome. However, this deletion
362 is also absent in *P. vivax*, so cannot in itself explain the ability of *P. simium* to infect humans
363 in the current outbreak. It is possible, however, that this deletion is required for *P. simium* to
364 invade human red blood cells given the alterations that have occurred elsewhere in its
365 invadome following adaptation to non-human primates since the split between *P. simium* and
366 its human-infecting *P. vivax* ancestor (Figure 3B).

367 Our data is consistent with the hypothesis that the DBP1 deletion is required for efficient
368 invasion of human RBCs by the *P. simium* parasites present in the Atlantic Forest. However,
369 another study from Espírito Santo, suggests that humans may be infected with *P. simium*
370 strains that carry DBP1 without this deletion [39]. This discrepancy may be caused by genetic
371 differences in the *P. simium* strains circulating in each region. Further, the distinction
372 between *P. simium* and *P. vivax* infections based on epidemiological data may be confounded
373 by human migration between the Amazonas and Espírito Santo [40].

374

375 Invadome proteins are obvious candidates for genetic factors underlying host-specificity; an
376 inactivating mutation in a *P. falciparum* erythrocyte binding antigen has recently been shown

377 to underlie host-specificity [41]. Traditionally, functional studies on invadome proteins have
378 focused on domains known to bind or interact directly with the host. Although the *P. simium*-
379 specific DBP1 and RBP2a deletions reported here do not cover known structural motifs,
380 these deletions could nevertheless affect host cell recognition as disordered protein regions
381 have known roles in cellular regulation and signal transduction [42]. Further, a shorter, less
382 flexible linker between the merozoite membrane and the receptor-binding DBP1 domain may
383 favour a more rigid and better oriented positioning of the dimeric DBP1, enhancing its
384 capacity to engage the human receptor.
385

386 Phylogenetic analyses of the *P. simium* clade give the geographical location of its most
387 closely related *P. vivax* strain as Mexico, and not Brazil. In imported populations, the
388 relationship between geographical and genetic proximity may be weak. Multiple
389 introductions of diverse strains from founder populations may occur independently over large
390 distances, so that two closely related strains may be introduced in distantly located regions. It
391 may be postulated that there occurred the introduction of strains of *P. vivax* to Mexico from
392 the Old World that were closely related (due to similar regions of origin) to strains introduced
393 to the Atlantic Forest which then went on to become *P. simium* in New World monkeys.
394 Strains from a different point of origin were introduced to the Amazonian region of Brazil.
395 This hypothesis necessitates reproductive isolation of the *P. simium* clade from the Brazilian
396 *P. vivax* parasites following their initial introduction; an isolation that would be facilitated,
397 presumably, by their separate host ranges or via adaptation to different vectors.
398

Due to uncertainties regarding the number of individual genomes that were transferred during the original host switch from human to NHPs that resulted in the formation of the *P. simium* clade, it is impossible to perform dating analyses to determine a time for the split between *P. vivax* and *P. simium* with which we can be confident. The phylogeny shown in Figure 1 is consistent with the hypothesis that all present-day *P. vivax*/*P. simium* originated from a now extinct, or as yet unsampled, Old World population. The most parsimonious explanation for this is that today's New World *P. vivax/P. simium* originated from European *P. vivax*, which was itself a remnant of the original Eurasian/African *P. vivax* driven to extinction (or near extinction) in Africa by the evolution of the Duffy negative condition in the local human populations, and from Europe by malaria eradication programmes in the latter half of the twentieth century. This hypothesis is supported by the evidence of a close relationship between historical Spanish *P. vivax* and South American strains of the parasite [16], and by previous analyses of the mitochondrial genome [43]. Therefore, we postulate that the host switch between humans and non-human primates that eventually led to establishment of *P. simium* in howler monkeys must have occurred subsequent to the European colonisation of the Americas, within the last 600 years.

We find no evidence from the nuclear genome, the mitochondrial genome or the apicoplast genome that any of the *P. vivax /P. simium* strains from the New World considered in our analyses are more closely related to Old World parasites than they are to each other, as previously contended [44].

421 Given the limited genetic diversity amongst the *P. simium* isolates considered here compared
422 to that of *P. vivax*, it is almost certain that the original host switch occurred from humans to
423 NHPs, and not the other way around [22]. Similarly, the larger amount of genetic diversity in
424 the current NHP-infecting *P. simium* compared to those *P. simium* strains isolated from
425 humans (as indicated by the higher degree of DBP1 polymorphism in the NHP-infecting *P.*
426 *simium* compared to the strains infecting humans), suggests that humans are being infected
427 from a pool of NHP parasites in a true zoonotic manner, as opposed to the sharing of a
428 common parasite pool between humans and NHPs.
429

430 The biological definition of a species is a group of organisms that can exchange genetic
431 material and produce viable offspring. We have no way of knowing whether this is the case
432 for *P. vivax* and *P. simium*, and genetic crossing experiments would be required to resolve
433 this question. Our phylogenetic analyses, however, clearly show *P. simium* forming a clade
434 on its own within the broader diversity of *P. vivax*, and this strongly suggests, given what we
435 know about its biology, that allopatric speciation has been/is occurring.
436 *Plasmodium simium* is currently recognised as a species separate from *P. vivax*; it has been
437 well characterised and described in the literature, and there is a type specimen available, with
438 which all the strains sequenced here cluster in one monophyletic group. Therefore, we cannot
439 at present overturn the species status of *P. simium* in the absence of conclusive proof from
440 crossing experiments.
441

442 **Conclusions**

443

444 The recent outbreak of human malaria in the Atlantic Forest of Rio de Janeiro underlines the
445 impact of zoonotic events on human health. Non-human primate malaria parasites must be
446 considered a reservoir of potential infectious human parasites relevant to any malaria
447 eradication strategy. Little is known about the genetic basis for zoonoses, yet the genome
448 sequence of *P. simium* suggests a deletion within the DBP1 gene is a possible facilitator of
449 zoonotic transfer. The genome of *P. simium* will thus form an important resource for the
450 future functional characterizations of the mechanisms underlying zoonotic malaria.

451

452

453 **Methods**

454

455 Sample Collection and Preparation

456 Human and primate samples of *P. simium* were collected and prepared as part of a previous
457 study [1, 14]. Additionally, two *P. vivax* samples from the Amazon area of Brazil were also
458 collected from human patients (Table S3). All participants provided informed written
459 consent. The *P. simium* CDC (Howler) strain (Catalog No. MRA-353) from ATCC was
460 obtained *via* the BEI Resources Repository in NIAID-NIH (https://www.beiresources.org/).

461

462 DNA extraction and sequencing

463 DNA was extracted as previously described[1]. Genomic DNA for each sample was
464 quantified using the Qubit® 2.0 Fluorometer and was used for library preparation. DNA for
465 intact samples was sheared using a Covaris E220 DNA sonicator to fragments of 500bp. The
466 DNA libraries for intact samples were made using the TruSeq Nano DNA Library Prep kit
467 (Illumina), whereas the DNA libraries for degraded samples were made using Ovation
468 Ultralow Library System V2 kit (Nugen), according to the manufacturers' instructions. The
469 amplified libraries were stored at -20 °C. The pooled libraries were sequenced in an Illumina
470 HiSeq4000 instrument (2 x 150 bp PE reads) (Illumina). A PhiX control library was applied
471 to the sequencing run as a base balanced sequence for the calibration of the instrument so that
472 each base type is captured during the entire run. Samples AF22, AF26, AF36 were
473 additionally sequenced and scaffolded by PacBio RS II platform (Pacific Biosciences,
474 California, US) using a SMRT library. Genomic DNA from the *P. vivax* samples was
475 extracted from filter paper as previously described [45].

476

477 <u>Illumina reads preparation and mapping</u>

478 FastQC v 0.11.6 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) was used to
479 evaluate the quality of Illumina reads. Illumina adapters were removed, and reads were
480 trimmed using the trimmomatic v0.33 [46] software with the following conditions:

481 *LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:36*

482 To exclude human reads from our analysis, trimmed reads were mapped against the human
483 reference genome (v. hg38) and the *Plasmodium vivax* strain P01 reference genome (v. 36)

484 from PlasmoDB (www.plasmodb.org) with bowtie2 (v 2.3.3.1) [47]. Reads mapping against
485 the human genome were removed from further analysis.

486

487 Genome Assembly

488 *Plasmodium simium* sample AF22 was selected for genome assembly based on read quality
489 and coverage. After removal of human contaminants, Illumina reads were assembled into
490 contigs using the Spades (v 3.70) assembler [48]. Contigs assembled into scaffolds running
491 SSPACE (v 3.0) [49] for 15 rounds and gaps filled with Gapfiller (v 1.10) [50]. Scaffolds
492 were subsequently corrected with Illumina reads using the Pilon (v 1.22) software [51].
493 Blobtools (v 1.0) (DOI: 10.5281/zenodo.845347) [52] was used to remove any residual
494 contaminant scaffolds. Genome size and GC content was in line with that of *P. vivax* (Table
495 S1). Genomic scaffolds representing the mitochondrial and apicoplast genome were
496 identified through blastn searches against the corresponding *P. falciparum* and *P. vivax*
497 sequences (Figure S1). The *P. simium* mitochondrial genome was aligned against a range of
498 previously published *P. vivax* and *P. simium* mitochondrial genomes [53, 54]. A gap-filled
499 region in the alignment where the distal parts of the *P. simium* scaffold were merged was
500 manually deleted. A minimum spanning haplotype network was produced using PopART
501 [55, 56] confirming the authenticity of the *P. simium* mitochondrial genome (Figure S32).

502

503 Genome Annotation

504 Two approaches were used to annotate the reference *P. simium* AF22 genome. Firstly, the
505 Maker pipeline (v 2.31.8) [57] was run for two rounds, using ESTs and protein evidence from

506 *P. vivax* and *P. cynomolgi* strain B and *P. falciparum* to generate Augustus gene models.
507 Secondly, a separate annotation was produced using the Companion web server [58].
508 Companion was run using the *P. vivax* P01 reference assembly and default parameters. Basic
509 annotation statistics are provided in Table S1. The relatively low number of genes (5966) is
510 due to the fragmented and incomplete nature of the *P. simium* assembly (Table S1). Gene
511 content was estimated using BUSCO [59, 60] (v3.0) revealing a gene annotation
512 completeness comparable to other *Plasmodium* genome assemblies (Figure S2).

513

514 PlasmoDB Genome References and Annotations

515 Genome FASTA files, as well as annotated protein and CDS files were obtained from
516 PlasmoDB [61] for the following species: *P. gallinaceum* 8A, *P. cynomolgi* B and M, *P*
517 *knowlesi* H, *P. falciparum* 3D7, *P. reichenowi* G01, *P. malariae* UG01, *P. ovale curtisi*
518 GH01, *P. coatneyi* Hackeri, *P. vivax* P01 and *P. vivax* SalI. For each species, PlasmoDB
519 version 36 was used.

520

521 Orthologous group determination

522 Amino-acid sequences-based phylogenetic trees were prepared using protein sequences from
523 the *P. simium* annotation, as well as the protein annotations from 10 malaria species
524 downloaded from PlasmoDB: *P. vivax* P01, *P. cynomolgi* B, *P. knowlesi* H, *P. vivax*-like
525 Pvl01, *P. coatneyi* Hackeri, *P. falciparum* 3D7, *P. gallinaceum* 8A, *P. malariae* UG01, *P.*
526 *ovale* curtisi GH01, and *P. reichenowi* G01. *P. vivax*-like from PlasmoDB version 43, all
527 other annotations from version 41. A total of 3181 1:1 orthologous genes were identified

528 using the Proteinortho (v 6.0.3) software [62]. Approximately 88% of the predicted genes in
529 *P. simium* have orthologues in *P. vivax* P01 (Figure S33).

530

531 Short indels in genes

532 Shorter indels (<500 bp) were detected from soft-clipping information in read mapping using
533 the '-i' option in DELLY [37] (v 0.7.9). Predicted indels in protein-coding genes were then
534 compared to independent assemblies of *P. simium* samples AF22, AF28, AF33, AF36, 2302
535 and 3636 (Table S3) and indels present in both DELLY predictions and all assemblies were
536 kept (Table S7). Indel boundaries shifted by a maximum of two amino acid positions were
537 allowed between predictions and assemblies. As assemblies were not complete, data for any
538 gene was only required to be present in five of the six *P. simium* assemblies used.

539

540 Protein phylogeny

541 Protein sequences were aligned using mafft (v 7.222) [63] and alignments were subsequently
542 trimmed with trimAl (v 1.2rev59) [64] using the heuristic 'automated1' method to select the
543 best trimming procedure. Trimmed alignments were concatenated and a phylogenetic tree
544 was constructed using RAxML (v 8.2.3) [65] with the PROTGAMMALG model.

545

546 SNP calling and analysis

547 Short sequence reads from 15 *P. simium* samples (13 derived from humans and two from
548 NHPs) and two *P. vivax* samples, all from this study (Table S3), were aligned against a
549 combined human (hg38) and *P. vivax* (strain P01, version 39) genome using NextGenMap

550 (v0.5.5) [66]. This was similarly done for 30 previously published *P. vivax* strains [25] and
551 the Sal1 reference. These data sets were downloaded from ENA (https://www.ebi.ac.uk/ena)
552 (Table S4). Duplicate reads were removed using samtools (v 1.9) [67] and the filtered reads
553 were realigned using IndelRealigner from the GATK package (v 4.0.11) [68]. SNPs were
554 called independently with GATK HaplotypeCaller and freebayes (v 1.2.0) [69], keeping only
555 SNPs with a QUAL score above 30. The final SNP set was determined from the inter-section
556 between GATK and freebayes. Allele frequencies and mean coverage across SNP sites are
557 shown in Figure S34. A PCA plot was constructed using plink (v 1.90) [70], and admixture
558 analysis was done with Admixture (v 1.3.0) [26].

559

560 SNP phylogeny

561 Alleles from SNP positions with data in 55 samples were retrieved, concatenated, and aligned
562 using mafft [63]. Tree was produced by PhyML [71, 72] with the GTR substitution model
563 selected by SMS [73]. Branch support was evaluated with the Bayesian-like transformation
564 of approximate likelihood ratio test, aBayes [74]. A phylogenetic network was made in
565 SplitsTree [75] using the NeighborNet network [76].

566

567 Nucleotide diversity

568 Conventional tools calculating nucleotide diversity directly from the variant call files assume
569 that samples are aligned across the entire reference sequence. But as read coverage across the
570 reference genome was highly uneven between samples (Figure S34), adjustment for this was
571 required. Coverage across the reference genome was thus calculated for each sample using

572 samtools mpileup (v 1.9) [67]. For each comparison between two samples, the nucleotide
573 divergence was calculated as number of detected bi-allelic SNPs per nucleotide with read
574 coverage of at least 5X in both samples.

575

576 Population divergence

577 To account for the differences in read coverage between samples we used the pixy software
578 [77], which produces unbiased estimates of D_{XY} in the presence of missing data. Population
579 divergence (D_{XY}) was calculated from this all-site VCF using pixy version 1.0.4.beta1, either
580 in 10 kb windows or in gene coordinate windows.

581

582 Gene sequence deletions

583 Exploratory Neighbour-Joining phylogenies were produced with CLUSTALW [78, 79] and
584 visualised with FigTree (https://github.com/rambaut/figtree/) after alignment with mafft [78].
585 Pacbio reads were aligned using Blasr (v 5.3.2) [80], short Illumina reads using NextGenMap
586 (v0.5.5) [66]. Dotplots were produced with FlexiDot (v1.05) [81].

587

588 Gene families and groups

589 Exported protein-coding gene sets were compiled from the literature [82-84]. Invasion genes
590 were retrieved from Hu *et al*. (2010) [85]. Gene families were assessed in seven *Plasmodium*
591 genomes (*P. simium*, *P. vivax* SalI, *P. vivax* P01, *P. vivax-like* Pvl01, *P. cynomolgi* M, *P.*
592 *cynomolgi* B, and *P. knowlesi* H) using the following pipeline: for all genomes, annotated
593 genes were collected for each gene family. These 'seed' sequences were used to search all

594 proteins from all genomes using BLASTP and best hits for all proteins were recorded. For
595 each gene family 'seed' sequences were then aligned with mafft [63], trimmed with trimAl
596 [64], and HMM models were then built using HMMer (http://hmmer.org/).
597
598 For PIR/VIR and PHIST genes, models were built for each genome independently, for all
599 other gene families a single model was built from all genomes. These models were then used
600 to search all proteins in all genomes. All proteins with best BLASTP hit to a 'seed' sequence
601 from a given genome were sorted according to their bit score. The lowest 5% of hits were
602 discarded and remaining proteins with best hits to a 'seed' sequence were assigned one
603 'significant' hit. As all proteins were searched against 'seeds' from the six annotated genomes
604 (*P. simium* excluded), a maximum of six 'significant' BLAST hits could be obtained.
605 Similarly, for each HMM model the bottom 25% hits were discarded and remaining hits were
606 considered 'significant'. The final set of gene families consisted of previously annotated genes
607 and un-annotated genes with at least two 'significant' hits (either BLASTP or HMM).
608
609 PCR amplification of DBP1 and RBP2a genes
610 PCR primers were initially designed from alignments between *P. vivax* and *P. simium*
611 sequences and tested using Primer-BLAST [86] and PlasmoDB (www.plasmodb.org). For
612 DBP1, the reaction was performed in 10 µL volumes containing 0.5 µM of each
613 oligonucleotide primer, 1 µL DNA and 5 µL of Master Mix 2x (Promega) (0.3 units of Taq
614 Polymerase, 200 µM each deoxyribonucleotide triphosphates and 1.5 mM MgCl2). Samples
615 were run with the following settings: 2 minutes of activation at 95°C, followed by 35 cycles

616 with 30 seconds denaturation at 95°C, 30 seconds annealing at 57°C (ΔT=-0.2 °C from 2nd
617 cycle) and 1 minute extension at 72°C, then 5 minutes final extension at 72°C.
618
619 For the RBP2a PCR, the reaction was performed in 10 µL volumes containing 0.5 µM of
620 each oligonucleotide primer, 1 µL DNA, 0.1 µL PlatinumTaq DNA Polymerase High Fidelity
621 (Invitrogen, 5U/µL), 0.2 mM each deoxyribonucleotide triphosphates and 2 mM $MgSO_4$. The
622 PCR assays were performed with the following cycling parameters: an initial denaturation at
623 94°C for 1.5 min followed by 40 cycles of denaturation at 94°C for 15 sec, annealing at 65°C
624 for 30 sec (ΔT=-0.2 °C from 2nd cycle) and extension at 68°C for 3.5 min. All Genotyping
625 assays were performed in the thermocycler Veriti 96 wells, Applied Biosystems, and the
626 amplified fragments were visualized by electrophoresis on agarose gels (2% for DBP1 and
627 1% for RBP2a) in 1x TAE buffer (40 mM Tris-acetate, 1 mM EDTA) with 5 µg/ mL
628 ethidium bromide (Invitrogen) in a horizontal system (Bio-Rad) at 100 V for 30 min. Gels
629 were examined with a UV transilluminator (UVP - Bio-Doc System).
630
631 To prevent cross-contamination, the DNA extraction and mix preparation were performed in
632 "parasite DNA-free rooms" distinct from each other. Furthermore, each of these separate
633 areas has different sets of pipettes and all procedures were performed using filtered pipette
634 tips. DNA extraction was performed twice on different days. Positive (DNA extracted from
635 blood from patients with known *P. vivax* infection) and negative (no DNA and DNA
636 extracted from individuals who have never traveled to malaria-endemic areas) controls were
637 used in each round of amplification. DNA extracted from blood of a patient with high
638 parasitemia for *P. vivax* and DNA of *P. simium* of a non-human primate with an acute

infection and parasitemia confirmed by optical microscopy served as positive controls in the PCR assays. Primer sequences are provided in Figure S24 and S29.

Structural modelling of DBP1 and RBP2a genes

RaptorX [87] was used for prediction of secondary structure and protein disorder. Homology models for the DBL domain of *P. vivax* P01 strain, *P. simium* AF22, and the previously published CDC *P. simium* strain were produced by SWISS-MODEL [88], using the crystallographic structure of the DBL domain from *Plasmodium vivax* DBP bound to the ectodomain of the human DARC receptor (PDB ID 4nuv), with an identity of 98%, 96% and 96% for *P. vivax*, *P. simium* AF22 and *P. simium* CDC, respectively. QMEAN values were -2.27, -2.04 and -2.03, respectively. The homology model for the reticulocyte binding protein 2 (RBP2a) of *P. vivax* strain P01 was produced based on the cryoEM structure of the complex between the *P. vivax* RBP2b and the human transferrin receptor TfR1 (PDB ID 6d05)[33], with an identity of 31% and QMEAN value of -2.46. The visualization and structural analysis of the produced models was done with PyMOL (https://pymol.org/2/).

Declarations

Ethics approval and consent to participate:

This study was conducted under the approved project by the IRB commitees in Fiocruz (CAAE Fiocruz: 88554718.0.0000.5262) and in KAUST (19IBEC12).

Consent for publication

Not applicable

Availability of data and materials

The reference genome assembly and short sequence reads have been uploaded to European Nucleotide Archive (https://www.ebi.ac.uk/ena/) under the Study accession number PRJEB34061.

Competing interests

The authors declare that they have no competing interests

Funding

The work was supported financially by the King Abdullah University of Science and Technology (KAUST) through the baseline fund BRF1020/01/01 to AP and BAS/1/1056-01-01 to STA, and the Award No. URF/1/1976-25 from the Office of Sponsored Research (OSR). The field work in the Atlantic Forest and laboratory analysis in Brazil received financial support from the Secretary for Health Surveillance of the Ministry of Health through the Global Fund (agreement IOC-005-Fio-13), *Programa Nacional de Excelência (PRONEX)* and contract 407873/2018-0 of the *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)*, the *Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig* CBB-APQ-02620-15*)* and the *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (Faperj)*, Brazil. CNPq supports CFAB, CTDR, MFFC, PB and RLO, with a research productivity fellowship. CTDR (CNE: E-26/202.921/2018), MFFC, PB and RLO are also supported by Faperj as *Cientistas do nosso*

estado. AdP-C was supported by a postdoctoral fellowship from the Faperj and DAMA by a fellowship from the CGZV-SVS (Brazilian Ministry of Health) TED 49/2018 grant. SF was supported by a Wellcome Seed Award in Science to DCJ (208965/Z/17/Z).

Authors' contributions

CTDR, PB, AdPC, RLdO, RC and AP conceived the study. CTDR, PB, CFAdB, MdFFdC, RC and AP supervised and co-ordinated the study. AdPC, FVSdA, DAMA, CBJ, JCdSJ and ZMBH collected materials. OD, QG, AdPC, CFAdB, MdFFdC, FVSdA, DAMA, CBJ, JCsSJ and ZMBH conducted wet-lab experiments. TM, AK, SF, DCJ, FJGV, SA, CTDR, PB, RLdO, CFAdB, MdFFdC, FVSdA, DAMA, CBJ, JCdSJ, ZMBH, RC and AP analysed and interpreted the data. TM, RC, AP, CTDR, PB, CFAdB and RLdO drafted and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Prof. Xin-zhuan Su at the National Institute of Allergy and Infectious Diseases, NIH, for invaluable help in obtaining the parasite gDNA from the BEI Resources; Sidnei Silva and Graziela Zanini, for assistance on the parasitological diagnosis of the human samples; Aline Lavigne and Larissa Gomes for undertaking the PCR for *P vivax*; Alcides Pissinatti and Silvia Bahadian Moreira for the facilities provided at the Primate Centre of Rio de Janeiro; Orzinete Rodrigues Soares for non-human primates' blood slides'; Marcelo Quintela, Waldemir Paixão Vargas, Carlos Alberto C. da Silva, Alexandre B. de Souza, Vicente Klonowski, Romenique L. Araújo, Luis R. Nogueira, Fernando Barreto, Ana L.

Quijada, Luiz P.P. Silva, Gelson Medeiros, Adilson B. Ramos, Marcilene B. Ramos, Carlos A.A. Júnior, Paulo G. Barbosa, Sérgio F. Fragoso, Adilson R. Silva, Cecília Cronemberger, Marcelo Rheingantz, Leonardo Nascimento and João Marins for the field support; *Grupo Técnico de Vigilância de Arboviroses* (GT-Arbo – Brazilian Ministry of Health) for field and material supports; and Cassio Leonel Peterka from The Brazilian Ministry of Health for malaria epidemiological data. The following reagent was obtained through BEI Resources, NIAID, NIH: *Plasmodium simium*, Strain Howler, MRA-353, contributed by William E. Collins. We thank Kieran Samuk for assistance with pixy, Richard Carter for invaluable discussion and comment on previous versions of this manuscript, and Tony Holder for critical reading of the manuscript.

References

1. Brasil P, Zalis MG, de Pina-Costa A, Siqueira AM, Junior CB, Silva S, Areas ALL, Pelajo-Machado M, de Alvarenga DAM, da Silva Santelli ACF *et al*: **Outbreak of human malaria caused by Plasmodium simium in the Atlantic Forest in Rio de Janeiro: a molecular epidemiological investigation**. *Lancet Glob Health* 2017, **5**(10):e1038-e1046.

2. Cox-Singh J, Davis TME, Lee K-S, Shamsul SSG, Matusop A, Ratnam S, Rahman HA, Conway DJ, Singh B: **Plasmodium knowlesi malaria in humans is widely distributed and potentially life threatening**. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2008, **46**(2):165-171.

3. Imwong M, Madmanee W, Suwannasin K, Kunasol C, Peto TJ, Tripura R, von Seidlein L, Nguon C, Davoeung C, Day NPJ *et al*: **Asymptomatic Natural Human Infections With the Simian Malaria Parasites Plasmodium cynomolgi and Plasmodium knowlesi**. *J Infect Dis* 2019, **219**(5):695-702.

4. Loy DE, Liu W, Li Y, Learn GH, Plenderleith LJ, Sundararaman SA, Sharp PM, Hahn BH: **Out of Africa: origins and evolution of the human malaria parasites Plasmodium falciparum and Plasmodium vivax**. *Int J Parasitol* 2017, **47**(2-3):87-97.

733 5. Liu W, Li Y, Learn GH, Rudicell RS, Robertson JD, Keele BF, Ndjango JB, Sanz CM,
734 Morgan DB, Locatelli S *et al*: **Origin of the human malaria parasite Plasmodium**
735 **falciparum in gorillas**. *Nature* 2010, **467**(7314):420-425.

736 6. Liu W, Li Y, Shaw KS, Learn GH, Plenderleith LJ, Malenke JA, Sundararaman SA, Ramirez
737 MA, Crystal PA, Smith AG *et al*: **African origin of the malaria parasite Plasmodium**
738 **vivax**. *Nature communications* 2014, **5**:3346.

739 7. Marques GR, Condino ML, Serpa LL, Cursino TV: **[Epidemiological aspects of**
740 **autochthonous malaria in the Atlantic forest area of the northern coast of the State of**
741 **Sao Paulo, 1985-2006]**. *Rev Soc Bras Med Trop* 2008, **41**(4):386-389.

742 8. Cox-Singh J, Davis TM, Lee KS, Shamsul SS, Matusop A, Ratnam S, Rahman HA, Conway
743 DJ, Singh B: **Plasmodium knowlesi malaria in humans is widely distributed and**
744 **potentially life threatening**. *Clin Infect Dis* 2008, **46**(2):165-171.

745 9. Ta TH, Hisam S, Lanza M, Jiram AI, Ismail N, Rubio JM: **First case of a naturally**
746 **acquired human infection with Plasmodium cynomolgi**. *Malar J* 2014, **13**:68.

747 10. Marchand RP, Culleton R, Maeno Y, Quang NT, Nakazawa S: **Co-infections of Plasmodium**
748 **knowlesi, P. falciparum, and P. vivax among Humans and Anopheles dirus Mosquitoes,**
749 **Southern Vietnam**. *Emerg Infect Dis* 2011, **17**(7):1232-1239.

750 11. Otto TD, Gilabert A, Crellen T, Bohme U, Arnathau C, Sanders M, Oyola SO, Okouga AP,
751 Boundenga L, Willaume E *et al*: **Genomes of all known members of a Plasmodium**
752 **subgenus reveal paths to virulent human malaria**. *Nat Microbiol* 2018, **3**(6):687-697.

753 12. de Pina-Costa A, Brasil P, Di Santi SM, de Araujo MP, Suarez-Mutis MC, Santelli AC,
754 Oliveira-Ferreira J, Lourenco-de-Oliveira R, Daniel-Ribeiro CT: **Malaria in Brazil: what**
755 **happens outside the Amazonian endemic region**. *Mem Inst Oswaldo Cruz* 2014,
756 **109**(5):618-633.

757 13. Collins WE, Contacos PG, Guinn EG: **Observations on the sporogonic cycle and**
758 **transmission of Plasmodium simium Da Fonseca**. *J Parasitol* 1969, **55**(4):814-816.

759 14. de Alvarenga DAM, Culleton R, de Pina-Costa A, Rodrigues DF, Bianco C, Jr., Silva S,
760 Nunes AJD, de Souza JC, Jr., Hirano ZMB, Moreira SB *et al*: **An assay for the**
761 **identification of Plasmodium simium infection for diagnosis of zoonotic malaria in the**
762 **Brazilian Atlantic Forest**. *Scientific reports* 2018, **8**(1):86.

763 15. Escalante AA, Cornejo OE, Freeland DE, Poe AC, Durrego E, Collins WE, Lal AA: **A**
764 **monkey's tale: the origin of Plasmodium vivax as a human malaria parasite**. *Proc Natl*
765 *Acad Sci U S A* 2005, **102**(6):1980-1985.

766 16. Gelabert P, Sandoval-Velasco M, Olalde I, Fregel R, Rieux A, Escosa R, Aranda C,
767 Paaijmans K, Mueller I, Gilbert MT *et al*: **Mitochondrial DNA from the eradicated**
768 **European Plasmodium vivax and P. falciparum from 70-year-old slides from the Ebro**
769 **Delta in Spain**. *Proceedings of the National Academy of Sciences of the United States of*
770 *America* 2016, **113**(41):11495-11500.

771 17. Rodrigues PT, Valdivia HO, de Oliveira TC, Alves JMP, Duarte A, Cerutti-Junior C, Buery
772 JC, Brito CFA, de Souza JC, Jr., Hirano ZMB *et al*: **Human migration and the spread of**
773 **malaria parasites to the New World**. *Scientific reports* 2018, **8**(1):1993.

774 18. Deane LM: **Simian malaria in Brazil**. *Mem Inst Oswaldo Cruz* 1992, **87 Suppl 3**:1-20.

775 19. Fonseca F: **[Plasmodium of a primate of Brazil]**. *Mem Inst Oswaldo Cruz* 1951, **49**:543-
776 553.

777 20. Leclerc MC, Durand P, Gauthier C, Patot S, Billotte N, Menegon M, Severini C, Ayala FJ,
778 Renaud F: **Meager genetic variability of the human malaria agent Plasmodium vivax**.
779 *Proceedings of the National Academy of Sciences of the United States of America* 2004,
780 **101**(40):14455-14460.

781 21. Duarte AM, Porto MA, Curado I, Malafronte RS, Hoffmann EH, de Oliveira SG, da Silva
782 AM, Kloetzel JK, Gomes Ade C: **Widespread occurrence of antibodies against**
783 **circumsporozoite protein and against blood forms of Plasmodium vivax, P. falciparum**
784 **and P. malariae in Brazilian wild monkeys**. *J Med Primatol* 2006, **35**(2):87-96.

785 22. Tazi L, Ayala FJ: **Unresolved direction of host transfer of Plasmodium vivax v. P. simium**
786 **and P. malariae v. P. brasilianum**. *Infect Genet Evol* 2011, **11**(1):209-221.

787 23. Coatney GR, Collins WE, Warren M, Contacos PG: **Plasmodium simium**. In: *The Primate*
788 *Malarias*. National Institutes of Health; 1971.

789 24. Auburn S, Bohme U, Steinbiss S, Trimarsanto H, Hostetler J, Sanders M, Gao Q, Nosten F,
790 Newbold CI, Berriman M *et al*: **A new Plasmodium vivax reference sequence with**
791 **improved assembly of the subtelomeres reveals an abundance of pir genes**. *Wellcome*
792 *Open Res* 2016, **1**:4.

793 25. Hupalo DN, Luo Z, Melnikov A, Sutton PL, Rogov P, Escalante A, Vallejo AF, Herrera S,
794 Arevalo-Herrera M, Fan Q *et al*: **Population genomics studies identify signatures of global**
795 **dispersal and drug resistance in Plasmodium vivax**. *Nat Genet* 2016, **48**(8):953-958.

796 26. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in**
797 **unrelated individuals**. *Genome research* 2009, **19**(9):1655-1664.

798 27. Cruickshank TE, Hahn MW: **Reanalysis suggests that genomic islands of speciation are**
799 **due to reduced diversity, not reduced gene flow**. *Mol Ecol* 2014, **23**(13):3133-3157.

800 28. Nei M, Li WH: **Mathematical model for studying genetic variation in terms of restriction**
801 **endonucleases**. *Proceedings of the National Academy of Sciences of the United States of*
802 *America* 1979, **76**(10):5269-5273.

803 29. Kanjee U, Rangel GW, Clark MA, Duraisingh MT: **Molecular and cellular interactions**
804 **defining the tropism of Plasmodium vivax for reticulocytes**. *Curr Opin Microbiol* 2018,
805 **46**:109-115.

806 30. Miller LH, McAuliffe FM, Mason SJ: **Erythrocyte receptors for malaria merozoites**. *Am J*
807 *Trop Med Hyg* 1977, **26**(6 Pt 2):204-208.

808 31. Iyer J, Gruner AC, Renia L, Snounou G, Preiser PR: **Invasion of host cells by malaria**
809 **parasites: a tale of two protein families**. *Mol Microbiol* 2007, **65**(2):231-249.

810 32. Chan LJ, Dietrich MH, Nguitragool W, Tham WH: **Plasmodium vivax Reticulocyte**
811 **Binding Proteins for invasion into reticulocytes**. *Cell Microbiol* 2019:e13110.

812 33. Gruszczyk J, Huang RK, Chan LJ, Menant S, Hong C, Murphy JM, Mok YF, Griffin MDW,
813 Pearson RD, Wong W *et al*: **Cryo-EM structure of an essential Plasmodium vivax**
814 **invasion complex**. *Nature* 2018, **559**(7712):135-139.

815 34. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, Crabtree J, Angiuoli SV,
816 Merino EF, Amedeo P *et al*: **Comparative genomics of the neglected human malaria**
817 **parasite Plasmodium vivax**. *Nature* 2008, **455**(7214):757-763.

818 35. Gilabert A, Otto TD, Rutledge GG, Franzon B, Ollomo B, Arnathau C, Durand P,
819 Moukodoum ND, Okouga AP, Ngoubangoye B *et al*: **Plasmodium vivax-like genome**
820 **sequences shed new insights into Plasmodium vivax biology and evolution**. *PLoS Biol*
821 2018, **16**(8):e2006035.

822 36. Ntumngia FB, McHenry AM, Barnwell JW, Cole-Tobian J, King CL, Adams JH: **Genetic**
823 **variation among Plasmodium vivax isolates adapted to non-human primates and the**
824 **implication for vaccine development**. *Am J Trop Med Hyg* 2009, **80**(2):218-227.

825 37. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO: **DELLY: structural**
826 **variant discovery by integrated paired-end and split-read analysis**. *Bioinformatics* 2012,
827 **28**(18):i333-i339.

828 38. Batchelor JD, Zahm JA, Tolia NH: **Dimerization of Plasmodium vivax DBP is induced**
829 **upon receptor binding and drives recognition of DARC**. *Nat Struct Mol Biol* 2011,
830 **18**(8):908-914.

831 39. de Oliveira TC, Rodrigues PT, Early AM, Duarte A, Buery JC, Bueno MG, Catao-Dias JL,
832 Cerutti C, Jr., Rona LDP, Neafsey DE *et al*: **Plasmodium simium: population genomics**
833 **reveals the origin of a reverse zoonosis**. *J Infect Dis* 2021.

834 40. Meneguzzi VC, Santos CB, Pinto Ide S, Feitoza LR, Feitoza HN, Falqueto A: **Use of**
835 **geoprocessing to define malaria risk areas and evaluation of the vectorial importance of**
836 **anopheline mosquitoes (Diptera: Culicidae) in Espirito Santo, Brazil**. *Mem Inst Oswaldo*
837 *Cruz* 2009, **104**(4):570-575.

838 41. Proto WR, Siegel SV, Dankwa S, Liu W, Kemp A, Marsden S, Zenonos ZA, Unwin S, Sharp
839 PM, Wright GJ *et al*: **Adaptation of Plasmodium falciparum to humans involved the loss**
840 **of an ape-specific erythrocyte invasion ligand**. *Nature communications* 2019, **10**(1):4512.

841 42. Wright PE, Dyson HJ: **Intrinsically disordered proteins in cellular signalling and**
842 **regulation**. *Nature reviews Molecular cell biology* 2015, **16**(1):18-29.

843 43. Culleton R, Carter R: **African Plasmodium vivax: distribution and origins**. *Int J Parasitol*
844 2012, **42**(12):1091-1097.

845 44. Li J, Collins WE, Wirtz RA, Rathore D, Lal A, McCutchan TF: **Geographic subdivision of**
846 **the range of the malaria parasite Plasmodium vivax**. *Emerging infectious diseases* 2001,
847 **7**(1):35-42.

848 45. Choi EH, Lee SK, Ihm C, Sohn YH: **Rapid DNA extraction from dried blood spots on**
849 **filter paper: potential applications in biobanking**. *Osong Public Health Res Perspect* 2014,
850 **5**(6):351-357.

851 46. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence**
852 **data**. *Bioinformatics* 2014, **30**(15):2114-2120.

853 47. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature methods*
854 2012, **9**(4):357-359.

855 48. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
856 Nikolenko SI, Pham S, Prjibelski AD *et al*: **SPAdes: a new genome assembly algorithm**
857 **and its applications to single-cell sequencing**. *J Comput Biol* 2012, **19**(5):455-477.

858 49. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled**
859 **contigs using SSPACE**. *Bioinformatics* 2011, **27**(4):578-579.

860 50. Boetzer M, Pirovano W: **Toward almost closed genomes with GapFiller**. *Genome biology*
861 2012, **13**(6):R56.

862 51. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q,
863 Wortman J, Young SK *et al*: **Pilon: an integrated tool for comprehensive microbial**
864 **variant detection and genome assembly improvement**. *PloS one* 2014, **9**(11):e112963.

865 52. Laetsch D, Blaxter M: **BlobTools: Interrogation of genome assemblies [version 1;**
866 **referees: 2 approved with reservations]**. *F1000Research* 2017, **6**(1287).

867 53. Jongwutiwes S, Putaporntip C, Iwasaki T, Ferreira MU, Kanbara H, Hughes AL:
868 **Mitochondrial genome sequences support ancient population expansion in Plasmodium**
869 **vivax**. *Molecular biology and evolution* 2005, **22**(8):1733-1739.

870 54. Rodrigues PT, Alves JM, Santamaria AM, Calzada JE, Xayavong M, Parise M, da Silva AJ,
871 Ferreira MU: **Using mitochondrial genome sequences to track the origin of imported**
872 **Plasmodium vivax infections diagnosed in the United States**. *Am J Trop Med Hyg* 2014,
873 **90**(6):1102-1108.

874 55. Bandelt HJ, Forster P, Rohl A: **Median-joining networks for inferring intraspecific**
875 **phylogenies**. *Molecular biology and evolution* 1999, **16**(1):37-48.

876 56. Leigh JW, Bryant D: **POPART: full-feature software for haplotype network construction**.
877 *Methods Ecol Evol* 2015, **6**(9):1110-1116.

878 57. Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database**
879 **management tool for second-generation genome projects**. *BMC Bioinformatics* 2011,
880 **12**(1):491.

881 58. Steinbiss S, Silva-Franco F, Brunk B, Foth B, Hertz-Fowler C, Berriman M, Otto TD:
882 **Companion: a web server for annotation and analysis of parasite genomes**. *Nucleic acids*
883 *research* 2016, **44**(W1):W29-34.

884 59. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva
885 EV, Zdobnov EM: **BUSCO applications from quality assessments to gene prediction and**
886 **phylogenomics**. *Molecular biology and evolution* 2017.

887 60. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing**
888 **genome assembly and annotation completeness with single-copy orthologs**.
889 *Bioinformatics* 2015, **31**(19):3210-3212.

890 61. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A,
891 Grant G, Harb OS *et al*: **PlasmoDB: a functional genomic database for malaria parasites**.
892 *Nucleic acids research* 2009, **37**(Database issue):D539-543.

893 62. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ: **Proteinortho: detection**
894 **of (co-)orthologs in large-scale analysis**. *BMC Bioinformatics* 2011, **12**:124.

895 63. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7:**
896 **improvements in performance and usability**. *Molecular biology and evolution* 2013,
897 **30**(4):772-780.

898 64. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T: **trimAl: a tool for automated**
899 **alignment trimming in large-scale phylogenetic analyses**. *Bioinformatics* 2009,
900 **25**(15):1972-1973.

901 65. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of**
902 **large phylogenies**. *Bioinformatics* 2014, **30**(9):1312-1313.

903 66. Sedlazeck FJ, Rescheneder P, von Haeseler A: **NextGenMap: fast and accurate read**
904 **mapping in highly polymorphic genomes**. *Bioinformatics* 2013, **29**(21):2790-2791.

905 67. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin
906 R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009,
907 **25**(16):2078-2079.

908 68. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,
909 Altshuler D, Gabriel S, Daly M *et al*: **The Genome Analysis Toolkit: a MapReduce**
910 **framework for analyzing next-generation DNA sequencing data**. *Genome research* 2010,
911 **20**(9):1297-1303.

912 69. Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing**.
913 *arXiv* 2012:1207.3907.

914 70. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation**
915 **PLINK: rising to the challenge of larger and richer datasets**. *Gigascience* 2015, **4**:7.

916 71. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms**
917 **and methods to estimate maximum-likelihood phylogenies: assessing the performance of**
918 **PhyML 3.0**. *Syst Biol* 2010, **59**(3):307-321.

919 72. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large**
920 **phylogenies by maximum likelihood**. *Syst Biol* 2003, **52**(5):696-704.

921 73. Lefort V, Longueville JE, Gascuel O: **SMS: Smart Model Selection in PhyML**. *Molecular*
922 *biology and evolution* 2017, **34**(9):2422-2424.

923 74. Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O: **Survey of branch support**
924 **methods demonstrates accuracy, power, and robustness of fast likelihood-based**
925 **approximation schemes**. *Syst Biol* 2011, **60**(5):685-699.

926 75. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies**.
927 *Molecular biology and evolution* 2006, **23**(2):254-267.

928 76. Bryant D, Moulton V: **Neighbor-net: an agglomerative method for the construction of**
929 **phylogenetic networks**. *Molecular biology and evolution* 2004, **21**(2):255-265.

930 77. Korunes KL, Samuk K: **pixy: Unbiased estimation of nucleotide diversity and divergence**
931 **in the presence of missing data**. *Mol Ecol Resour* 2021, **21**(4):1359-1368.

932 78. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin
933 F, Wallace IM, Wilm A, Lopez R *et al*: **Clustal W and Clustal X version 2.0**.
934 *Bioinformatics* 2007, **23**(21):2947-2948.

935 79. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of**
936 **progressive multiple sequence alignment through sequence weighting, position-specific**
937 **gap penalties and weight matrix choice**. *Nucleic acids research* 1994, **22**(22):4673-4680.

938 80. Chaisson MJ, Tesler G: **Mapping single molecule sequencing reads using basic local**
939 **alignment with successive refinement (BLASR): application and theory**. *BMC*
940 *Bioinformatics* 2012, **13**:238.

941 81. Seibt KM, Schmidt T, Heitkam T: **FlexiDot: highly customizable, ambiguity-aware**
942 **dotplots for visual sequence analyses**. *Bioinformatics* 2018, **34**(20):3575-3577.

943 82. van Ooij C, Tamez P, Bhattacharjee S, Hiller NL, Harrison T, Liolios K, Kooij T, Ramesar J,
944 Balu B, Adams J *et al*: **The malaria secretome: from algorithms to essential function in**
945 **blood stage infection**. *PLoS pathogens* 2008, **4**(6):e1000084.

946 83. Boddey JA, Carvalho TG, Hodder AN, Sargeant TJ, Sleebs BE, Marapana D, Lopaticki S,
947 Nebl T, Cowman AF: **Role of plasmepsin V in export of diverse protein families from the**
948 **Plasmodium falciparum exportome**. *Traffic* 2013, **14**(5):532-550.

949 84. Schulze J, Kwiatkowski M, Borner J, Schluter H, Bruchhaus I, Burmester T, Spielmann T,
950 Pick C: **The Plasmodium falciparum exportome contains non-canonical PEXEL/HT**
951 **proteins**. *Mol Microbiol* 2015, **97**(2):301-314.

952 85. Hu G, Cabrera A, Kono M, Mok S, Chaal BK, Haase S, Engelberg K, Cheemadan S,
953 Spielmann T, Preiser PR *et al*: **Transcriptional profiling of growth perturbations of the**
954 **human malaria parasite Plasmodium falciparum**. *Nat Biotechnol* 2010, **28**(1):91-98.

955 86. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL: **Primer-BLAST: a tool**
956 **to design target-specific primers for polymerase chain reaction**. *BMC Bioinformatics*
957 2012, **13**:134.

958 87. Wang S, Li W, Liu S, Xu J: **RaptorX-Property: a web server for protein structure**
959 **property prediction**. *Nucleic acids research* 2016, **44**(W1):W430-435.

960 88. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer
961 TAP, Rempfer C, Bordoli L *et al*: **SWISS-MODEL: homology modelling of protein**
962 **structures and complexes**. *Nucleic acids research* 2018, **46**(W1):W296-W303.
963

964

965

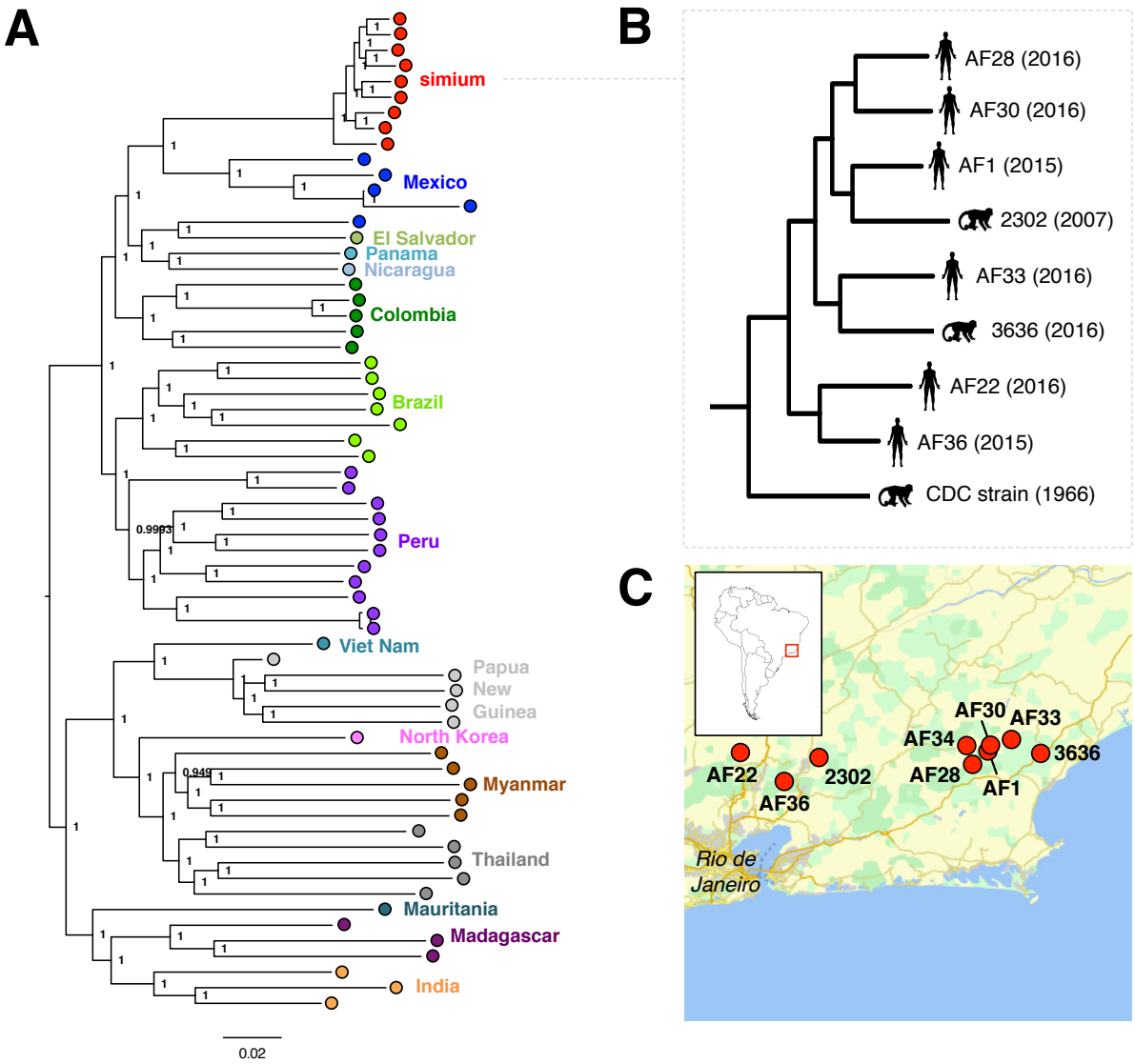

Figure 1. SNP phylogeny

A) Mid-point rooted maximum likelihood tree produced from 143,123 concatenated SNP positions with data from a minimum of 55 samples. The tree was produced using PhyML with the GTR evolutionary model. Branch support was evaluated with the Bayesian-like transformation of approximate likelihood ratio test (aBayes). Genetic distance is shown below the tree. *Plasmodium vivax* isolates are denoted as coloured circles by their country of

974 sample origin. A tree with specific sample IDs is available in Figure S9B) Magnification of
975 the *Plasmodium simium* clade (as in panel A). C) Map showing the geographic locations at
976 which the *P. simium* samples were collected.
977

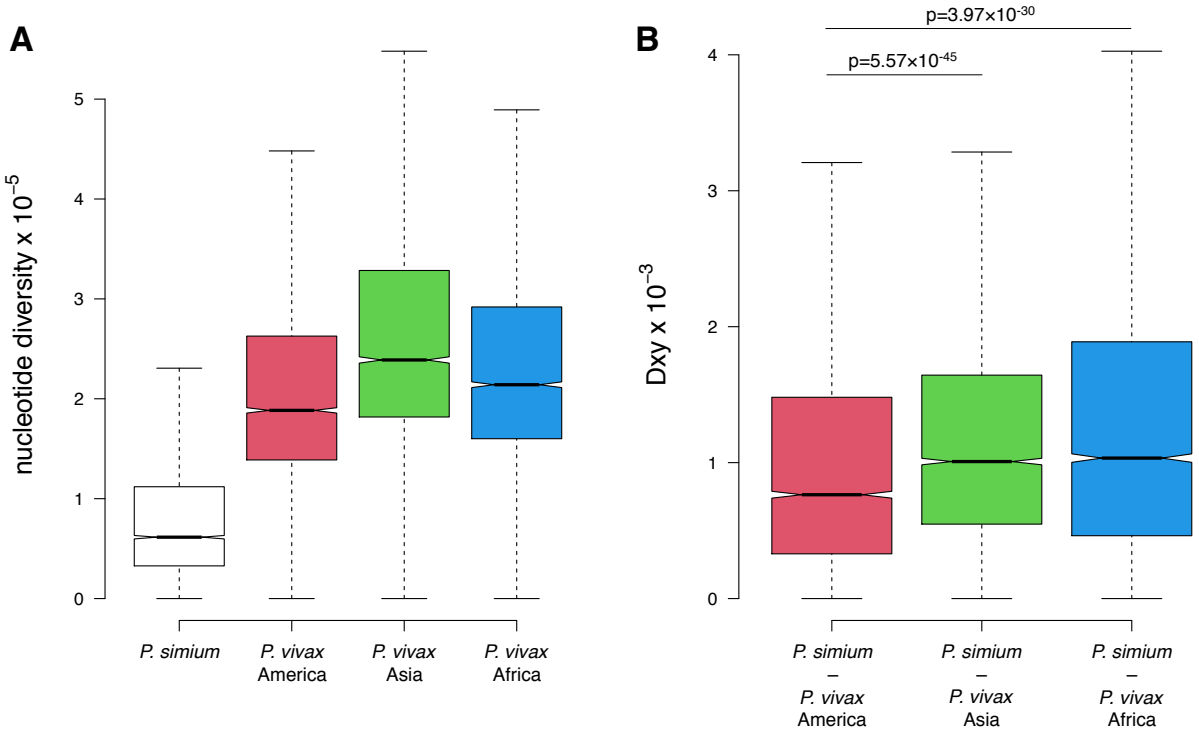

Figure 2. Nucleotide diversity and genetic distance between populations

A) Diversity within populations. Box plot showing the nucleotide diversity in 10 kb windows between *Plasmodium simium* samples (left), and *Plasmodium vivax* samples from America, Asia, and Africa. B) Genetic distance between populations. Box plot showing D_{XY} in 10 kb windows comparing *Plasmodium simium* samples to P. vivax samples from America, Asia, and Africa. P-values from one-sided Mann-Whitney tests for difference in medians are shown above boxes. For both plots, boxes denote 25th and 75th percentiles with all outliers removed.

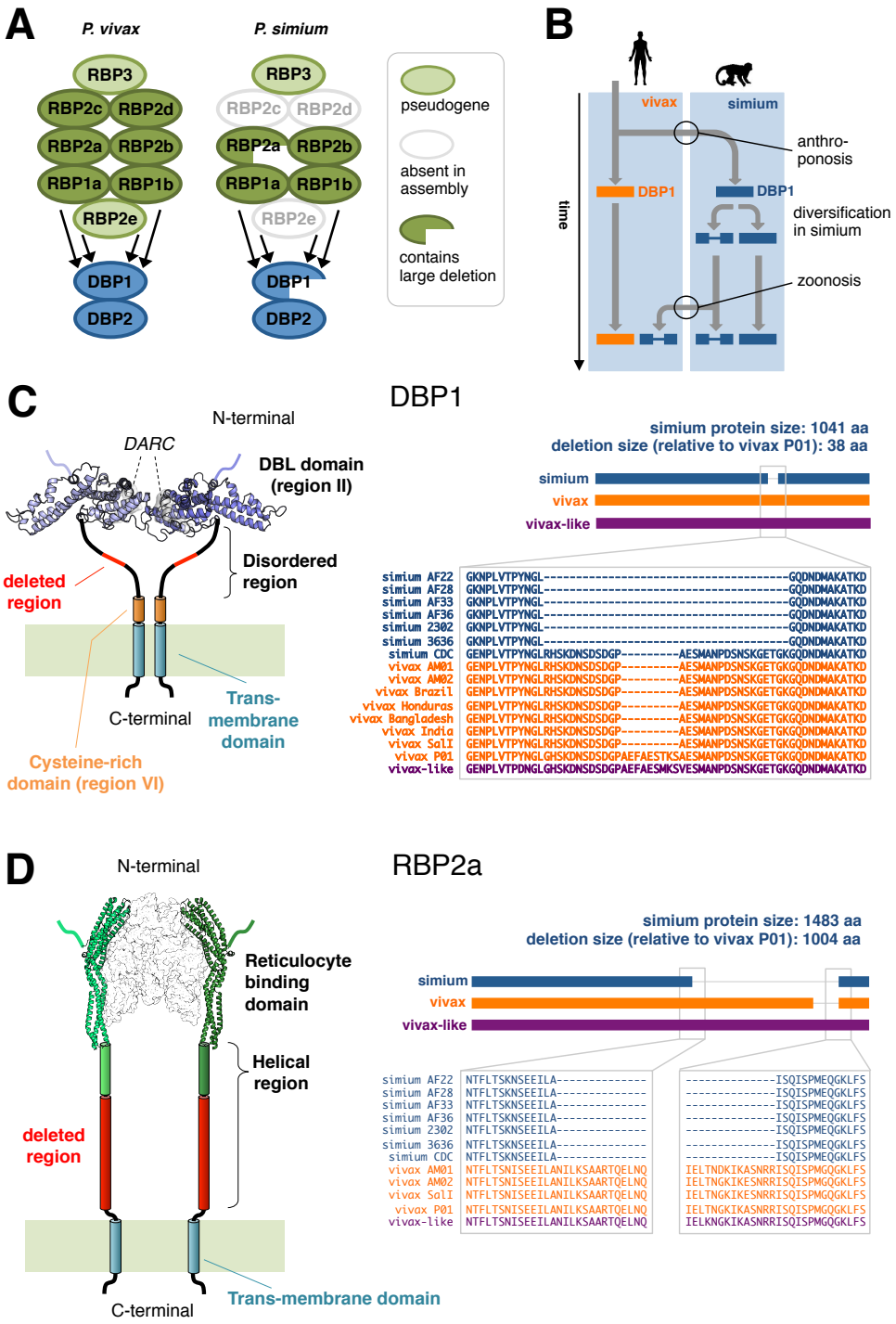

Figure 3. Invadome deletions

A) Overview of the invadome gene groups, Reticulocyte Binding Proteins (RBPs) and Duffy Binding Proteins (DBPs) in *Plasmodium vivax* and *Plasmodium simium*. The *P. vivax*

993 genome harbours two RBP2d genes, one of which is a pseudogene (Table S5). B) Schematic
994 depiction of the hypothesized scenario in which the DBP1 deletion – along with other
995 accumulated genetic changes in *P. simium* – is a prerequisite for the recently observed
996 zoonosis.

997 C) Left: Structural rendering of DBP1, showing known structural domains and motifs. The
998 two fragment molecules from the human DARC receptor are shown in grey. The 3-
999 dimensional structure of the DBL-DARC complex was modelled based on the *P. vivax*
1000 crystallographic model (PDB 4nuv). The region deleted in sequences from human-infecting
1001 *P. simium*, as compared to *P. vivax* P01, is highlighted in red. Right: Details of DBP1 protein
1002 alignments. A full alignment is available in Figure S20.

1003 D) Similar to panel C) but for RBP2a. The complex between the reticulocyte binding domain
1004 and the human receptor was modelled based on the cryoEM structure of the complex between
1005 the *P. vivax* RBP2b and the human transferrin receptor TfR1 (PDB 6d05). A full alignment is
1006 available in Figure S28.
1007

1008
1009