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, 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) Malaria Unit, Department of Pathology, Institute of Tropical Medicine, Nagasaki University, 1-12-4 Sakamoto, Nagasaki, 852-8523, 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

§) Contributed equally, arranged alphabetically
†) Corresponding authors. CTDR: malaria@fiocruz.br, AP: arnab.pain@kaust.edu.sa

Summary

Plasmodium simium, a malaria parasite of non-human primates in the Atlantic forest region of Brazil was recently shown to cause zoonotic infections in humans. Phylogenetic analyses based on the whole genome sequences of six *P. simium* isolates from humans and two isolates from brown howler monkeys revealed that *P. simium* is monophyletic within the broader diversity of South American *Plasmodium vivax*, consistent with the hypothesis that *P. simium* first infected non-human primates as a result of a host-switch of *P. vivax* from humans. Very low levels of genetic diversity within *P. simium* and the absence of *P. simium-P. vivax* hybrids suggest that the *P. simium* population emerged recently with a subsequent period of independent evolution in Platyrrhini monkeys. We find that Plasmodium Interspersed Repeat (PIR) genes, Plasmodium Helical Interspersed Subtelomeric (PHIST) genes and Tryptophan-Rich Antigen (TRAg) genes in *P. simium* are divergent from *P. vivax* orthologues and are enriched for non-synonymous single nucleotide polymorphisms, consistent with the rapid evolution of these genes. Analysis of genes involved in erythrocyte invasion revealed several notable differences between *P. vivax* and *P. simium*, including large deletions within the coding region of the Duffy Binding Protein 1 (DBP1) and Reticulocyte Binding Protein 2a (RBP2a) genes of *P. simium*. Sequence analysis of *P. simium* isolates from non-human primates (NHPs) and zoonotic human infections revealed a deletion of 38 amino acids in DBP1 present in all human-derived isolates, whereas NHP isolates were multi-allelic at this locus. We speculate that these deletions in key erythrocyte invasion ligands along with other

significant genetic changes may have facilitated zoonotic transfer to humans. NHPs are a reservoir of parasites potentially infectious to humans that must be considered in malaria eradication efforts. The *P. simium* genome is an important resource for understanding the mechanisms of malaria parasite zoonoses.

Introduction

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% 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/). Of the 0.4% of malaria cases registered outside the Amazonian region, around 90% 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, we now know 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 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]

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

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

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

Results

Genome assembly and phylogeny

From a single *P. simium* sample collected from Rio de Janeiro state in 2016 [1] short read DNA sequences were obtained and assembled into a draft genome (see Supplementary Materials). The assembled genome consists of 2,192 scaffolds over 1kb in length with a combined size of 29 Mb (Table S1). Two scaffolds corresponding to the apicoplast and mitochondrial genomes were also identified (Figure S1). Gene content analysis showed a completeness of annotation comparable to previously published *Plasmodium* assemblies (Figure S2). A phylogenetic tree

170
171 constructed from 3,181 of 1:1 orthologs of the annotated *P. simium* protein-coding genes with
172 the orthologues of *P. vivax, P. cynomolgi, P. coatneyi, P. knowlesi, P. malariae, P.*
173 *falciparum, P. reichenowi,* and *P. gallinaceum* confirmed that *P. simium* is very closely
174 related to *P. vivax* (Figure S3). Analysis of gene families revealed a *P. simium* gene repertoire
175 largely similar to *P. vivax* (Figure S4) and no indication of a recent expansion of the *P.*
simium PIR gene family (Figure S5).
176
177 ***P. simium-P. vivax* diversity analysis**
178 To detect single nucleotide polymorphisms (SNPs) within the *P. vivax*/*P. simium* clade, short
179 Illumina paired-end sequence reads were mapped onto the *P. vivax* P01 reference genome [24].
180 Reads were collected from eleven human-derived *P. simium* samples, two monkey *P. simium*
181 samples, two *P. vivax* samples from the Brazilian Amazon, and a previously published *P.*
182 *simium* CDC strain (originally isolated in 1966, see Methods). Data from a range of *P. vivax*
183 strains representing a global distribution was retrieved from the literature [25]. Including only
184 SNPs with a minimum depth of five reads, a total of 232,780 SNPs were called initially across
185 79 samples. Sixteen samples were subsequently removed, primarily due to low coverage,
186 resulting in a total of 63 samples for further analysis (Table S2, Table S3). Since few SNP loci
187 were covered across all samples, and to enable diversity analysis, we restricted all further
188 analysis to the 124,968 SNPs for which data were available from at least 55 samples (Figure
189 S6).
190
191 ***P. simium-P. vivax* population analysis**
192 A Principal Component Analysis (PCA) plot constructed from these genome-wide SNP loci
193 showed a clear separation between American and Asian *P. vivax* samples as well as a distinct
194 grouping of *P. simium* samples (Figure S7). The latter observation suggests that both human
195 and NHP *P. simium* isolates represent a single population that is genetically different from
196 American *P. vivax* populations. A similar pattern is observed by performing a
197 multidimensional scaling analysis of the SNP data (Figure S8). To enable a phylogenetic
198 approach, we constructed an alignment from the 124,968 SNP sites. In the resulting
199 phylogenetic tree, *P. vivax* strains generally clustered according to their geographical origin,
200 with clear separation of the Asian and American samples (Figure 1A, a tree with sample IDs
201 is available in Figure S9). *Plasmodium simium* samples clustered as a monophyletic group
202 with Mexican *P. vivax* samples (Figure 1A), consistent with a recent American origin for *P.*
203 *simium.* This phylogeny, based on genome-wide SNPs represents an "average" phylogeny

204 across the genome and cannot be considered to reflect a true history of parasite ancestry due
205 to the effects of recombination. It is possible that trees produced from individual genes might
206 reveal different phylogenetic relationships. Yet, the association between *P. simium* samples
207 and *P. vivax* samples from Mexico is also observed in a phylogenetic network (Figure S10).
208
209 To examine whether the *P. simium* isolates we studied were part of a continuous population
210 with local *P. vivax* parasites, we examined population ancestry with the ADMIXTURE
211 program [26] (Figure S11). This analysis is consistent with the PCA and MDS analyses (Figure
212 S7 & Figure S8) and the phylogenetic analysis of segregating SNPs (Figure 1), showing that
213 *P. simium* forms a genetically distinct population separate from *P. vivax*. The absence of *P.*
214 *simium-P. vivax* hybrids (indicating genetic introgression events) indicates that *P. simium* has
215 undergone a period of independent evolution.
216
217 ***P. simium* genetic differentiation from *P. vivax* is enhanced in host-parasite interacting**
218 **genes**
219 To characterise the *P. simium* population further, we estimated the nucleotide diversity in *P.*
220 *simium* and *P. vivax* samples (see Materials and Methods). *Plasmodium simium* diversity
221 (genome-median: 1.3×10^{-4}) is approximately six times lower than the diversity observed when
222 comparing all *P. vivax* samples (genome-median: 7.9×10^{-4}) (Figure 2). Diversity measured
223 within coding sequences in *P. vivax* is consistent with previous reports [6]. The median
224 nucleotide diversity between *P. simium* and *P. vivax* genomes of 8.7×10^{-4} and the low
225 diversity within *P. simium* suggest that the strains we examined are part of a relatively recent
226 or isolated population.
227
228 We then examined the population differentiation over the entire genome using F_{ST}, a measure
229 of the proportion of ancestry private to a population (F_{ST} = 0 for completely intermixed
230 populations, F_{ST} = 1 for populations with completely independent ancestry). Although our
231 analysis contains few samples, F_{ST} estimates can be very accurate if multiple genomic sites
232 are used [27]. Consistent with phylogenetic and admixture analyses, we observed a high level of
233 differentiation between human *P. simium* and American *P. vivax* (F_{ST}=0.47). For comparison,
234 the differentiation between *P. vivax* from America and *P. vivax* from Myanmar and Thailand
235 (henceforth referred to as 'American vivax' and 'Asian vivax', respectively) is less than half of
236 this (F_{ST}=0.21). To examine whether there were any signals of adaptive changes in *P. simium*
237 that may have occurred in transfer and growth in NHPs, we calculated the fixation index for

238 all individual genes. Clearly, the small number of samples renders this analysis prone to false
239 and incorrect signals, and F_{ST} values for individual genes should be interpreted with caution.
240 Nevertheless, we attempted to look for general patterns in F_{ST} values across gene groups.
241
242 Amongst the 4,432 *P. vivax* genes with at least one SNP in our data set, we examined the top-
243 25% of genes with highest F_{ST} values for enrichment in functional Gene Ontology (GO) terms
244 or metabolic pathways. No GO terms or pathways were significant at the 0.05 level after
245 Bonferroni correction (Table S4 & Table S5). Using the *P. falciparum* orthologues – when
246 available – gave similar results (not shown). We next tested whether any gene families
247 (Figure S4, Table S6) were associated with high F_{ST} values. Genes belonging to the
248 Plasmodium Interspersed Repeat (PIR) family involved in antigenic variation [28], the
249 Plasmodium Helical Interspersed Subtelomeric (PHIST) genes, a family of exported proteins
250 [29], the merozoite surface proteins MSP7[30], and Tryptophan-rich antigens (TRAg) [31] were
251 enriched among the genes with high F_{ST} values (binomial distribution, PIR; p=0.016, PHIST;
252 $p=4.2\times10^{-4}$, MSP7; p=0.034, TRAg; $p=5.5\times10^{-5}$).
253
254 As these gene families are involved in parasite-host interactions the elevated F_{ST} values may
255 simply reflect a general pattern of rapid genetic divergence in *Plasmodium* parasites. To test
256 this, we repeated the F_{ST} analysis between American *P. vivax* and a selection of Asian *P.*
257 *vivax* isolates (Myanmar and Thailand samples only). Consistent with the phylogenetic
258 analysis (Figure 1A) gene F_{ST} was slightly higher overall between *P. simium* and American *P.*
259 *vivax* than between American and Asian *P. vivax* samples (Figure 3A). However, none of the
260 gene families were overrepresented among genes with high F_{ST} (top-25%) between American
261 and Asian *P. vivax*.
262
263 To further examine whether the elevated F_{ST} measures found for PIR, PHIST, MSP7, and
264 TRAg genes are exclusive to the comparison between *P. simium* and American *P. vivax*, we
265 calculated the ratio between the two F_{ST} measurements ('simium versus American vivax' and '
266 American versus Asian vivax') (Figure 3B). The ratios for PIR, PHIST and TRAg genes were
267 significantly higher than observed for the remaining genes (Figure 3C), whereas ratios for
268 MSP7 genes were not (Mann-Whitney U, p=0.11). Although the *P. simium* and the *P. vivax*
269 P01 genomes both encode a high number of the gene family members, our analysis is
270 restricted to the *P. vivax* genes to which our *P. simium* short read sequences can map. For
271 example, only 408 out of the 1209 *P. vivax* PIR genes have coverage from *P. simium* reads

272 across at least 80% of their length (Figure S12). Further, an even smaller number of these
273 genes have detectable SNPs between *P. simium* and American *P. vivax* samples and are
274 included in the analysis (numbers shown below Figure 3B).
275
276 To test whether the sequence redundancy among gene family loci could result in spurious
277 cross-mapping of short sequence reads we specifically tested the quality of SNPs in gene
278 families, and SNPs residing in gene families showed no signs of decreased calling, mapping,
279 or base qualities compared to other SNPs (Figure S13).
280
281 A phylogenetic analysis of PIR, PHIST and TRAg proteins harbouring genomic SNPs
282 revealed no apparent association between certain protein phylogenetic sub-groups and high
283 F_{ST} ratios (Figure S14-S16), consistent with a subtle signature of polygenic adaptations in
284 these gene families.
285
286 When testing all genes involved in invasion and exported protein genes (Table S7), the
287 observed F_{ST} ratios were not significantly different from the background (Figure 3B). Hence,
288 the differences in F_{ST} observed for PIR, PHIST and TRAg genes are not a general
289 phenomenon amongst the genes known to be involved in interactions with the host.
290
291 The observed skew towards higher F_{ST} values when comparing *P. simium* and American *P.*
292 *vivax* (Figure 3A) could result from inherent diversity between different American *P. vivax*
293 populations, potentially stemming from multiple introductions of *P. vivax* to the American
294 continent [17]. To test whether such founder effects and subsequent population bottlenecks
295 could explain the observations, we repeated the F_{ST} analysis using only Mexican vivax
296 samples as American representatives. Four Mexican samples (SRS693273, SRS694229,
297 SRS694244, SRS694267) were used. These clustered together in both the SNP phylogeny
298 (Figure 1) and in the PCA and MDS plots (Figure S7 & Figure S8), and are assumed to share
299 a recent evolutionary history. This analysis revealed the same pattern of elevated F_{ST} values
300 between *P. simium* and Mexican *P. vivax*, and PIR genes again displayed significantly higher
301 F_{ST} ratios (Figure S17). Although PHIST and TRAg genes also showed higher F_{ST} ratios,
302 these were no longer significant (Figure S17). We therefore conclude that the observed higher
303 F_{ST} values between *P. simium* and American *P. vivax* PIR genes are not solely a result of
304 diversity within American *P. vivax* populations, but rather appear specific to comparisons
305 with *P. simium*.

306

307 Adaptive changes in PIR genes would be expected to produce stronger genetic divergence in
308 non-synonymous codon positions. To examine this, we divided genic SNPs into synonymous
309 and non-synonymous changes. In PIR genes, there are 353 non-synonymous and 185
310 synonymous SNPs (non-synonymous to synonymous SNP ratio = 1:1.91). Similarly, in
311 PHIST and TRAg genes we find 220 and 103 non-synonymous SNPs, and 67 and 41
312 synonymous SNPs, respectively (PHIST ratio = 1:3.28, TRAg ratio = 1:2.51). In all other
313 genes, the ratio between non-synonymous and synonymous SNPs is 1:1.49. Hence, the
314 proportion of non-synonymous SNPs in PIR, PHIST and TRAg genes is significantly higher
315 than in all other genes (chi-square, PIR; $p = 0.0073$, PHIST; $p = 9.4\times10^{-9}$, TRAg; $p = 0.0054$).

316

317 Our finding that PIR, PHIST and TRAg genes display overall markedly higher F_{ST} values
318 between *P. simium* and *P. vivax* suggests that these gene groups are enriched for private
319 alleles consistent with natural selection acting upon these genes subsequent to the split
320 between *P. simium* and *P. vivax*.

321

322 ***P. simium* invadome components**

323 Binding and red blood cell invasion is mediated by two key malaria gene families expressed
324 in merozoites: the Duffy Binding Proteins (DBPs) and the Reticulocyte Binding Proteins
325 (RBPs). In *P. vivax*, DBPs bind to the Duffy Antigen Receptor for Chemokines (DARC) [32,33],
326 which is present on both host normocytes and reticulocytes, and RBPs are known to restrict *P.*
327 *vivax* to binding and invading reticulocytes [34-36]. Two DBPs, DBP1 and DBP2, are present in
328 *P. vivax* P01 (Table S8). Recently, the reported protein structure of *P. vivax* RBP2b revealed
329 the evolutionary conservation of residues involved in the invasion complex formation [36].
330 RBPs can be divided into three subfamilies, RBP1, RBP2, and RBP3 [37]. The *P. vivax* P01
331 genome encodes 11 RBPs (including the reticulocyte binding surface protein, RBSA), of
332 which three are pseudogenes (Table S8).

333

334 The *P. vivax* DBP and RBP protein sequences were used to search for *P. simium* orthologues,
335 resulting in the detection of the two DBP proteins and RBP1a, RBP1b, RBP2a, RBP2b, and
336 RBP3, and a failure to detect RBP2c and RBP2d (Figure 4; Table S8; Figure S18; Figure S19)
337 across all sequenced *P. simium* samples. As in *P. vivax* genomes, the *P. simium* RBP3 is a
338 pseudogene [38], indicating that conversion to a pseudogene happened prior to the split between
339 *P. vivax* and *P. simium*.

340

341 To determine whether the apparent absences of individual RBP genes in *P. simium* was due to
342 incomplete genome assembly, we examined the coverage of *P. simium* reads mapped onto *P.*
343 *vivax* RBP gene loci. No *P. simium* coverage was observed at the RBP2c, RBP2d, and RBP2e
344 genes in *P. simium* samples, including the previously published CDC strain (Figure S20).
345

346 Coverage of mapped reads across invadome-associated gene loci revealed no apparent
347 elevated coverage in genes compared to their flanking genomic regions, which would have
348 been expected if the *P. simium* genome contained multiple (duplicated) copies of non-
349 assembled invasion genes (Figure S21).
350

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

352 The *P. simium* assembly revealed that the invasion gene DBP1 contains a large deletion
353 within its coding sequence (Figure 4 and Figure S22). Intriguingly, the previously published
354 *P. simium* CDC strain DBP1 (GenBank accession: ACB42432) [39] does not contain this
355 deletion ('simium CDC' in Figure 4B). A haplotype network confirms that this previously
356 published DBP1 gene is indeed a *P. simium* sequence (Figure S23), and the SNP analyses
357 consistently assign the CDC strain to the *P. simium* cluster (Figure 1, Figure S7, and Figure
358 S8). Compared to the *P. vivax* P01 reference genome the SalI reference harbours a 27 base
359 pair deletion in DBP1, in contrast to the 115 bp deletion observed in all *P. simium* samples
360 isolated from humans (Figure 4). The 27 base pair deletion is also present in most *P. vivax*
361 isolates (Figure S24). Additional deletion patterns exist among isolates, and in a few cases
362 multiple versions are detected within samples (Figure S24).
363

364 The presence of repetitive sequences within the DBP1 gene could potentially result in
365 aberrant assembly across the DBP1 locus, which may appear as an apparent deletion in
366 subsequent bioinformatic analysis. We tested this possibility and showed that the DBP1 gene
367 does not harbour any noticeable degree of repetitiveness (Figure S25). Several read mapping
368 analyses confirmed that the *P. simium*-specific 115 bp deletion was not an assembly artefact
369 (Figure S26-S28).
370

371 We next designed primers for PCR amplification of a nucleotide sequence that spans the
372 deleted region in the *P. simium* DBP1 gene and tested the occurrence of these deletion events
373 in a range of *P. vivax* and *P. simium* field samples from Brazil. All *P. vivax* samples tested by

374 PCR produced bands consistent with absence of the deletion whereas all samples from
375 human-infecting *P. simium* produced bands consistent with the presence of the precise 115 bp
376 deletion (Figure S29, top & middle). Interestingly, NHP-infecting *P. simium* isolates
377 contained a mix of sequences with and without deletions (Figure S29, bottom). PCR results
378 are summarized in Table S9. If the *P. simium*-specific deletion in DBP1 is a prerequisite for
379 the ability to infect humans, this suggests that only a subset of NHP-infecting *P. simium*
380 parasites currently possess the ability to infect humans.
381
382 A large, 3012 bp additional deletion was observed in the *P. simium* RBP2a gene, the presence
383 of which was also supported by read mapping (Figure 3, Figure S30-S33). PCR-based
384 genotyping results confirmed the presence of this deletion in all *P. simium* isolates
385 irrespective of their host (Figure S34, Table S9).
386
387
388 To test for the presence of insertions and deletions (indels) in other protein-coding genes, we
389 curated a set of indels by comparing short indels predicted by the DELLY software [40] to *P.*
390 *simium* assemblies. This set consisted of 244 indels in 222 genes (Table S10). Indel sizes were
391 almost exclusively integers of three (hence conserving reading frames) and indels were
392 predominantly found in low-complexity regions and in genes encoding long proteins (Figure
393 S35). Notably, short indels were recorded in 138 genes with known functional annotations
394 with an overrepresentation of DNA-binding function (GO:0005488 'binding', p=0.0011;
395 GO:0043565 'sequence-specific DNA binding', p=0.0062).
396

Potential structural implications of the deletion in DBP1 and RBP2a

397 We next investigated whether the observed deletions render DBP1 and RBP2a non-functional.
398 DBP1 contains a large extracellular region, which includes the N-terminal DBL region which
399 mediates the association with DARC in *P. vivax* [41], followed by a largely disordered region
400 and a cysteine-rich domain (Figure 4c). DBP1 has a single-pass transmembrane helix and a
401 short cytoplasmic tail. The deletion observed in the human-infecting *P. simium* only affects
402 the disordered region, leaving the flanking domains intact. We produced homology models of
403 the DBL domains from the *P. vivax* strain P01, the human-infecting *P. simium* strain AF22,
404 and the *P. simium* CDC strain, based on the crystal structure of the > 96% identical DBL
405 domain of *P. vivax* bound to DARC (PDB ID 4nuv). Whereas no significant substitutions
406 were found in the DBL domain between both *P. simium* sequences, our analysis showed that
407 residue substitutions between *P. simium* and *P. vivax* DBL domains cluster in proximity of the

408 DARC binding site (Figure S36). Based on our models, these substitutions are unlikely to
409 negatively affect the association with DARC, supporting the idea that the DBL domains of
410 both *P. simium* would be capable of binding to human DARC. Hence, the human-infecting *P.*
411 *simium* DBL1 probably retains the capacity to bind to human DARC, but has the interacting
412 domain positioned closer to the membrane than in the NHP-infecting CDC strain.
413
414 The deletion we detected in the *P. simium* RBP2a was larger, resulting in the loss of 1003
415 amino acid residues. These residues are predicted to form a mostly α-helical extracellular
416 stem-like structure that positions the reticulocyte binding domain away from the membrane
417 (Figure 4d). However, given that the deletion affects neither the transmembrane region, nor
418 the receptor-binding domain, our analysis suggests the resulting truncated RBP2a protein
419 could still associate with the human receptor, but that the binding event would occur closer to
420 the merozoite membrane.
421
422
423

Discussion

424
425 We present the genome of *Plasmodium simium*, the eighth malaria parasite species known to
426 infect humans in nature. In recent evolutionary time, *P. simium* has undergone both
427 anthroponosis and zoonosis making it unique for the study of the genetics underlying host-
428 switching in malaria parasites. Analysis of its genome confirmed a close phylogenetic
429 relationship between *P. simium* and *P. vivax*, and further analyses on single nucleotide
430 divergence support a very recent American origin for *P. simium.* This recent split between *P.*
431 *vivax* and *P. simium* precludes detection of genes under positive evolution [42], and we have
432 instead performed a general analysis of population differentiation between extant *P. simium*
433 and *P. vivax* isolates using F_{ST}. We find that members of three gene families involved in
434 antigenic variation, PIR, PHIST and TRAg, show significantly elevated F_{ST} levels between *P.*
435 *simium* and *P. vivax*. As higher F_{ST} values amongst these genes are not observed between
436 global *P. vivax* populations, their genetic differentiation appears to be associated with host-
437 switching between humans and non-human primates (NHPs).
438
439
440 Two proteins involved in host invasion, DBP1 and RBP2a, were found to harbour extensive
deletions in *P. simium* compared to *P. vivax*. Interestingly, experimental analysis of *P. simium*

samples revealed that isolates from human hosts all carried the DBP1 deletion, whereas isolates from NHPs displayed both absence and presence of the deletion. This DBP1 deletion is not present in the *P. simium* isolated from a brown howler monkey in the 1960s, which was previously shown to be incapable of infecting humans [23], although some degree of laboratory adaptation of this parasite may have affected its genome. However, this deletion is also absent in *P. vivax*, so cannot in itself explain the ability of *P. simium* to infect humans in the current outbreak. It is possible, however, that this deletion is required for *P. simium* to invade human red blood cells given the alterations that have occurred elsewhere in its invadome following adaptation to non-human primates since the split between *P. simium* and its human-infecting *P. vivax* ancestor.

Invadome proteins are obvious candidates for genetic factors underlying host-specificity; an inactivating mutation in a *P. falciparum* erythrocyte binding antigen has recently been shown to underlie host-specificity [43]. Traditionally, functional studies on invadome proteins have focused on domains known to bind or interact directly with the host. Although the *P. simium*-specific DBP1 and RBP2a deletions reported here do not cover known structural motifs, these deletions could nevertheless affect host cell recognition as disordered protein regions have known roles in cellular regulation and signal transduction [44]. Further, a shorter, less flexible linker between the merozoite membrane and the receptor-binding DBP1 domain may favour a more rigid and better oriented positioning of the dimeric DBP1, enhancing its capacity to engage the human receptor.

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

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 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 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 [45]. 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 [46].

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

The biological definition of a species is a group of organisms that can exchange genetic material and produce viable offspring. We have no way of knowing whether this is the case for *P. vivax* and *P. simium*, and genetic crossing experiments would be required to resolve this

question. Our phylogenetic analyses, however, clearly show *P. simium* forming a clade on its own within the broader diversity of *P. vivax*, and this strongly suggests, given what we know about its biology, that allopatric speciation has been/is occurring. *Plasmodium simium* appears to have been reproductively isolated from other strains of *P. vivax* long enough for significant genetic differentiation to occur (F_{ST} = 0.47), with some invadome genes showing even higher genetic differentiation.

Plasmodium simium is currently recognised as a species separate from *P. vivax*; it has been well characterised and described in the literature, and there is a type specimen available, with which all the strains sequenced here cluster in one monophyletic group. Therefore, we cannot at present overturn the species status of *P. simium* in the absence of conclusive proof from crossing experiments.

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

Methods

Sample Collection and Preparation

Human and primate samples of *P. simium* were collected and prepared as part of a previous study [1,14]. Additionally, two *P. vivax* samples from the Amazon area of Brazil were also collected from human patients (Table S2). All participants provided informed written consent. The *P. simium* CDC (Howler) strain (Catalog No. MRA-353) from ATCC was obtained *via* the BEI Resources Repository in NIAID-NIH (https://www.beiresources.org/).

DNA extraction and sequencing

DNA was extracted as previously described[1]. Genomic DNA for each sample was quantified using the Qubit® 2.0 Fluorometer and was used for library preparation. DNA for intact

543 samples was sheared using a Covaris E220 DNA sonicator to fragments of 500bp. The DNA
544 libraries for intact samples were made using the TruSeq Nano DNA Library Prep kit
545 (Illumina), whereas the DNA libraries for degraded samples were made using Ovation
546 Ultralow Library System V2 kit (Nugen), according to the manufacturers' instructions. The
547 amplified libraries were stored at -20 °C. The pooled libraries were sequenced in an Illumina
548 HiSeq4000 instrument (2 x 150 bp PE reads) (Illumina). A PhiX control library was applied
549 to the sequencing run as a base balanced sequence for the calibration of the instrument so that
550 each base type is captured during the entire run. Samples AF22, AF26, AF36 were
551 additionally sequenced and scaffolded by PacBio RS II platform (Pacific Biosciences,
552 California, US) using a SMRT library. Genomic DNA from the *P. vivax* samples was
553 extracted from filter paper as previously described [47].
554
555 **Illumina reads preparation and mapping**
556 FastQC v 0.11.6 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) was used to
557 evaluate the quality of Illumina reads. Illumina adapters were removed, and reads were
558 trimmed using the trimmomatic v0.33 [48] software with the following conditions:
559 *LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:36*
560 To exclude human reads from our analysis, trimmed reads were mapped against the human
561 reference genome (v. hg38) and the *Plasmodium vivax* strain P01 reference genome (v. 36)
562 from PlasmoDB (www.plasmodb.org) with bowtie2 (v 2.3.3.1) [49]. Reads mapping against the
563 human genome were removed from further analysis.
564
565 **Genome Assembly**
566
567 *Plasmodium simium* sample AF22 was selected for genome assembly based on read quality
568 and coverage. After removal of human contaminants, Illumina reads were assembled into
569 contigs using the Spades (v 3.70) assembler [50]. Contigs assembled into scaffolds running
570 SSPACE (v 3.0) [51] for 15 rounds and gaps filled with Gapfiller (v 1.10) [52]. Scaffolds were
571 subsequently corrected with Illumina reads using the Pilon (v 1.22) software [53]. Blobtools (v
572 1.0) (DOI: 10.5281/zenodo.845347) [54] was used to remove any residual contaminant
573 scaffolds. Genome size and GC content was in line with that of *P. vivax* (Table S1). Genomic
574 scaffolds representing the mitochondrial and apicoplast genome were identified through
575 blastn searches against the corresponding *P. falciparum* and *P. vivax* sequences (Figure S1).
576 The *P. simium* mitochondrial genome was aligned against a range of previously published *P.*
vivax and *P. simium* mitochondrial genomes [55,56]. A gap-filled region in the alignment where

577
578 the distal parts of the *P. simium* scaffold were merged was manually deleted. A minimum
579 spanning haplotype network was produced using PopART [57,58] confirming the authenticity of the *P. simium* mitochondrial genome (Figure S37).
580
581 **Genome Annotation**
582 Two approaches were used to annotate the reference *P. simium* AF22 genome. Firstly, the
583 Maker pipeline (v 2.31.8) [59] was run for two rounds, using ESTs and protein evidence from *P.*
584 *vivax* and *P. cynomolgi* strain B and *P. falciparum* to generate Augustus gene models.
585 Secondly, a separate annotation was produced using the Companion web server [60].
586 Companion was run using the *P. vivax* P01 reference assembly and default parameters. Basic
587 annotation statistics are provided in Table S1. The relatively low number of genes (5966) is
588 due to the fragmented and incomplete nature of the *P. simium* assembly (Table S1). Gene
589 content was estimated using BUSCO [61,62] (v3.0) revealing an annotation completeness
590 comparable to other *Plasmodium* genome assemblies (Figure S2).
591
592 **PlasmoDB Genome References and Annotations**
593 Genome FASTA files, as well as annotated protein and CDS files were obtained from
594 PlasmoDB [63] for the following species: *P. gallinaceum* 8A, *P. cynomolgi* B and M, *P*
595 *knowlesi* H, *P. falciparum* 3D7, *P. reichenowi* G01, *P. malariae* UG01, *P. ovale curtisi*
596 GH01, *P. coatneyi* Hackeri, *P. vivax* P01 and *P. vivax* SalI. For each species, PlasmoDB
597 version 36 was used.
598
599 **Orthologous group determination**
600 Amino-acid sequences-based phylogenetic trees were prepared using protein sequences from
601 the *P. simium* annotation, as well as the protein annotations from 10 malaria species
602 downloaded from PlasmoDB: *P. vivax* P01, *P. cynomolgi* B, *P. knowlesi* H, *P. vivax*-like
603 Pvl01, *P. coatneyi* Hackeri, *P. falciparum* 3D7, *P. gallinaceum* 8A, *P. malariae* UG01, *P.*
604 *ovale* curtisi GH01, and *P. reichenowi* G01. *P. vivax*-like from PlasmoDB version 43, all
605 other annotations from version 41. A total of 3181 1:1 orthologous genes were identified
606 using the Proteinortho (v 6.0.3) software [64]. Approximately 88% of the predicted genes in *P.*
607 *simium* have orthologues in *P. vivax* P01 (Figure S38).
608
609 **Short indels in genes**

610 Shorter indels (<500 bp) were detected from soft-clipping information in read mapping using
611 the '-i' option in DELLY [40] (v 0.7.9). Predicted indels in protein-coding genes were then
612 compared to independent assemblies of *P. simium* samples AF22, AF28, AF33, AF36, 2302
613 and 3636 (Table S2) and indels present in both DELLY predictions and all assemblies were
614 kept (Table S9). Indel boundaries shifted by a maximum of two amino acid positions were
615 allowed between predictions and assemblies. As assemblies were not complete, data for any
616 gene was only required to be present in five of the six *P. simium* assemblies used.
617
618 **Protein phylogeny**
619 Protein sequences were aligned using mafft (v 7.222) [65] and alignments were subsequently
620 trimmed with trimAl (v 1.2rev59) [66] using the heuristic 'automated1' method to select the best
621 trimming procedure. Trimmed alignments were concatenated and a phylogenetic tree was
622 constructed using RAxML (v 8.2.3) [67] with the PROTGAMMALG model.
623
624 **SNP calling and analysis**
625 Short sequence reads from 15 *P. simium* samples (13 derived from humans and two from
626 NHPs) and two *P. vivax* samples, all from this study (Table S2), were aligned against a
627 combined human (hg38) and *P. vivax* (strain P01, version 39) genome using NextGenMap
628 (v0.5.5) [68]. This was similarly done for 30 previously published *P. vivax* strains [25] and the
629 Sal1 reference. These data sets were downloaded from ENA (https://www.ebi.ac.uk/ena)
630 (Table S3). Duplicate reads were removed using samtools (v 1.9) [69] and the filtered reads
631 were realigned using IndelRealigner from the GATK package (v 4.0.11) [70]. SNPs were called
632 independently with GATK HaplotypeCaller and freebayes (v 1.2.0) [71], keeping only SNPs
633 with a QUAL score above 30. The final SNP set was determined from the inter-section
634 between GATK and freebayes. Allele frequencies and mean coverage across SNP sites are
635 shown in Figure S39. A PCA plot was constructed using plink (v 1.90) [72], and admixture
636 analysis was done with Admixture (v 1.3.0) [26]. F_{ST} values were estimated from nucleotide
637 data with the PopGenome R package [73,74] using the Weir & Cockerham method [75]. Non-
638 synonymous and synonymous SNPs were identified using snpeff [76].
639
640 **SNP phylogeny**
641 Alleles from SNP positions with data in 55 samples were retrieved, concatenated, and aligned
642 using mafft [65]. Tree was produced by PhyML [77,78] with the GTR substitution model selected
643 by SMS [79]. Branch support was evaluated with the Bayesian-like transformation of

644 approximate likelihood ratio test, aBayes [80]. A Phylogenetic network was made in SplitsTree
645 [81] using the NeighborNet network [82].
646
647 **Nucleotide diversity**
648 Conventional tools calculating nucleotide diversity directly from the variant call files assume
649 that samples are aligned across the entire reference sequence. But as read coverage across the
650 reference genome was highly uneven between samples (Figure S39), adjustment for this was
651 required. Coverage across the reference genome was thus calculated for each sample using
652 samtools mpileup (v 1.9) [69]. For each comparison between two samples, the nucleotide
653 divergence was calculated as number of detected bi-allelic SNPs per nucleotide with read
654 coverage of at least 5X in both samples.
655
656 **Gene sequence deletions**
657 Exploratory Neighbour-Joining phylogenies were produced with CLUSTALW [83,84] and
658 visualised with FigTree (https://github.com/rambaut/figtree/) after alignment with mafft [83].
659 Pacbio reads were aligned using Blasr (v 5.3.2) [85], short Illumina reads using NextGenMap
660 (v0.5.5) [68]. Dotplots were produced with FlexiDot (v1.05) [86].
661
662 **Gene families and groups**
663 Exported protein-coding gene sets were compiled from the literature [87-89]. Invasion genes
664 were retrieved from Hu *et al*. (2010) [90]. Gene families were assessed in seven *Plasmodium*
665 genomes (*P. simium*, *P. vivax* SalI, *P. vivax* P01, *P. vivax-like* Pvl01, *P. cynomolgi* M, *P.*
666 *cynomolgi* B, and *P. knowlesi* H) using the following pipeline: for all genomes, annotated
667 genes were collected for each gene family. These 'seed' sequences were used to search all
668 proteins from all genomes using BLASTP and best hits for all proteins were recorded. For
669 each gene family 'seed' sequences were then aligned with mafft [65], trimmed with trimAl [66],
670 and HMM models were then built using HMMer (http://hmmer.org/).
671
672 For PIR/VIR and PHIST genes, models were built for each genome independently, for all
673 other gene families a single model was built from all genomes. These models were then used
674 to search all proteins in all genomes. All proteins with best BLASTP hit to a 'seed' sequence
675 from a given genome were sorted according to their bit score. The lowest 5% of hits were
676 discarded and remaining proteins with best hits to a 'seed' sequence were assigned one
677 'significant' hit. As all proteins were searched against 'seeds' from the six annotated genomes

678 (*P. simium* excluded), a maximum of six 'significant' BLAST hits could be obtained.
679 Similarly, for each HMM model the bottom 25% hits were discarded and remaining hits were
680 considered 'significant'. The final set of gene families consisted of previously annotated genes
681 and un-annotated genes with at least two 'significant' hits (either BLASTP or HMM).
682
683

PCR amplification of DBP1 and RBP2a genes

684 PCR primers were initially designed from alignments between *P. vivax* and *P. simium*
685 sequences and tested using Primer-BLAST [91] and PlasmoDB (www.plasmodb.org). For
686 DBP1, the reaction was performed in 10 µL volumes containing 0.5 µM of each
687 oligonucleotide primer, 1 µL DNA and 5 µL of Master Mix 2x (Promega) (0.3 units of Taq
688 Polymerase, 200 µM each deoxyribonucleotide triphosphates and 1.5 mM MgCl2). Samples
689 were run with the following settings: 2 minutes of activation at 95°C, followed by 35 cycles
690 with 30 seconds denaturation at 95°C, 30 seconds annealing at 57°C (ΔT=-0.2 °C from 2nd
691 cycle) and 1 minute extension at 72°C, then 5 minutes final extension at 72°C.
692
693 For the RBP2a PCR, the reaction was performed in 10 µL volumes containing 0.5 µM of each
694 oligonucleotide primer, 1 µL DNA, 0.1 µL PlatinumTaq DNA Polymerase High Fidelity
695 (Invitrogen, 5U/µL), 0.2 mM each deoxyribonucleotide triphosphates and 2 mM $MgSO_4$. The
696 PCR assays were performed with the following cycling parameters: an initial denaturation at
697 94°C for 1.5 min followed by 40 cycles of denaturation at 94°C for 15 sec, annealing at 65°C
698 for 30 sec (ΔT=-0.2 °C from 2nd cycle) and extension at 68°C for 3.5 min. All Genotyping
699 assays were performed in the thermocycler Veriti 96 wells, Applied Biosystems, and the
700 amplified fragments were visualized by electrophoresis on agarose gels (2% for DBP1 and
701 1% for RBP2a) in 1x TAE buffer (40 mM Tris-acetate, 1 mM EDTA) with 5 µg/ mL ethidium
702 bromide (Invitrogen) in a horizontal system (Bio-Rad) at 100 V for 30 min. Gels were
703 examined with a UV transilluminator (UVP - Bio-Doc System).
704
705
706 To prevent cross-contamination, the DNA extraction and mix preparation were performed in
707 "parasite DNA-free rooms" distinct from each other. Furthermore, each of these separate
708 areas has different sets of pipettes and all procedures were performed using filtered pipette
709 tips. DNA extraction was performed twice on different days. Positive (DNA extracted from
710 blood from patients with known *P. vivax* infection) and negative (no DNA and DNA extracted
from individuals who have never traveled to malaria-endemic areas) controls were used in

each round of amplification. DNA extracted from blood of a patient with high 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 S29 and S34.

Structural modelling of DBP1 and RBP2a genes

RaptorX [92] 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 [93], 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)[36], 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/).

Data availability

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.

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 Richard Carter for invaluable discussion and comment on previous versions of this manuscript, and Tony Holder for critical reading of the manuscript.

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).

Author 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.

References

1 Brasil, P. *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* **5**, e1038-e1046, doi:10.1016/S2214-109X(17)30333-9 (2017).

2 Cox-Singh, J. *et al.* 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* **46**, 165-171, doi:10.1086/524888 (2008).

3 Imwong, M. *et al.* Asymptomatic Natural Human Infections With the Simian Malaria Parasites Plasmodium cynomolgi and Plasmodium knowlesi. *J Infect Dis* **219**, 695-702, doi:10.1093/infdis/jiy519 (2019).

4 Loy, D. E. *et al.* Out of Africa: origins and evolution of the human malaria parasites Plasmodium falciparum and Plasmodium vivax. *Int J Parasitol* **47**, 87-97, doi:10.1016/j.ijpara.2016.05.008 (2017).

5 Liu, W. *et al.* Origin of the human malaria parasite Plasmodium falciparum in gorillas. *Nature* **467**, 420-425, doi:10.1038/nature09442 (2010).

6 Liu, W. *et al.* African origin of the malaria parasite Plasmodium vivax. *Nature communications* **5**, 3346, doi:10.1038/ncomms4346 (2014).

7 Marques, G. R., Condino, M. L., Serpa, L. L. & Cursino, T. V. [Epidemiological aspects of autochthonous malaria in the Atlantic forest area of the northern coast of the State of Sao Paulo, 1985-2006]. *Rev Soc Bras Med Trop* **41**, 386-389 (2008).

8 Cox-Singh, J. *et al.* Plasmodium knowlesi malaria in humans is widely distributed and potentially life threatening. *Clin Infect Dis* **46**, 165-171, doi:10.1086/524888 (2008).

9 Ta, T. H. *et al.* First case of a naturally acquired human infection with Plasmodium cynomolgi. *Malar J* **13**, 68, doi:10.1186/1475-2875-13-68 (2014).

10 Marchand, R. P., Culleton, R., Maeno, Y., Quang, N. T. & Nakazawa, S. Co-infections of Plasmodium knowlesi, P. falciparum, and P. vivax among Humans and Anopheles dirus Mosquitoes, Southern Vietnam. *Emerg Infect Dis* **17**, 1232-1239, doi:10.3201/eid1707.101551 (2011).

11 Otto, T. D. *et al.* Genomes of all known members of a Plasmodium subgenus reveal paths to virulent human malaria. *Nature microbiology* **3**, 687-697, doi:10.1038/s41564-018-0162-2 (2018).

816 12 de Pina-Costa, A. *et al.* Malaria in Brazil: what happens outside the Amazonian
817 endemic region. *Mem Inst Oswaldo Cruz* **109**, 618-633, doi:10.1590/0074-
818 0276140228 (2014).

819 13 Collins, W. E., Contacos, P. G. & Guinn, E. G. Observations on the sporogonic cycle
820 and transmission of Plasmodium simium Da Fonseca. *J Parasitol* **55**, 814-816 (1969).

821 14 de Alvarenga, D. A. M. *et al.* An assay for the identification of Plasmodium simium
822 infection for diagnosis of zoonotic malaria in the Brazilian Atlantic Forest. *Scientific*
823 *reports* **8**, 86, doi:10.1038/s41598-017-18216-x (2018).

824 15 Escalante, A. A. *et al.* A monkey's tale: the origin of Plasmodium vivax as a human
825 malaria parasite. *Proc Natl Acad Sci U S A* **102**, 1980-1985,
826 doi:10.1073/pnas.0409652102 (2005).

827 16 Gelabert, P. *et al.* Mitochondrial DNA from the eradicated European Plasmodium
828 vivax and P. falciparum from 70-year-old slides from the Ebro Delta in Spain.
829 *Proceedings of the National Academy of Sciences of the United States of America*
830 **113**, 11495-11500, doi:10.1073/pnas.1611017113 (2016).

831 17 Rodrigues, P. T. *et al.* Human migration and the spread of malaria parasites to the
832 New World. *Scientific reports* **8**, 1993, doi:10.1038/s41598-018-19554-0 (2018).

833 18 Deane, L. M. Simian malaria in Brazil. *Mem Inst Oswaldo Cruz* **87 Suppl 3**, 1-20
834 (1992).

835 19 Fonseca, F. [Plasmodium of a primate of Brazil]. *Mem Inst Oswaldo Cruz* **49**, 543-
836 553 (1951).

837 20 Leclerc, M. C. *et al.* Meager genetic variability of the human malaria agent
838 Plasmodium vivax. *Proceedings of the National Academy of Sciences of the United*
839 *States of America* **101**, 14455-14460, doi:10.1073/pnas.0405186101 (2004).

840 21 Duarte, A. M. *et al.* Widespread occurrence of antibodies against circumsporozoite
841 protein and against blood forms of Plasmodium vivax, P. falciparum and P. malariae
842 in Brazilian wild monkeys. *J Med Primatol* **35**, 87-96, doi:10.1111/j.1600-
843 0684.2006.00148.x (2006).

844 22 Tazi, L. & Ayala, F. J. Unresolved direction of host transfer of Plasmodium vivax v.
845 P. simium and P. malariae v. P. brasilianum. *Infect Genet Evol* **11**, 209-221,
846 doi:10.1016/j.meegid.2010.08.007 (2011).

847 23 Coatney, G. R., Collins, W. E., Warren, M. & Contacos, P. G. in *The Primate*
848 *Malarias* (National Institutes of Health, 1971).

849 24 Auburn, S. *et al.* A new Plasmodium vivax reference sequence with improved
850 assembly of the subtelomeres reveals an abundance of pir genes. *Wellcome Open Res*
851 **1**, 4, doi:10.12688/wellcomeopenres.9876.1 (2016).

852 25 Hupalo, D. N. *et al.* Population genomics studies identify signatures of global
853 dispersal and drug resistance in Plasmodium vivax. *Nat Genet* **48**, 953-958,
854 doi:10.1038/ng.3588 (2016).

855 26 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry
856 in unrelated individuals. *Genome research* **19**, 1655-1664,
857 doi:10.1101/gr.094052.109 (2009).

858 27 Willing, E. M., Dreyer, C. & van Oosterhout, C. Estimates of genetic differentiation
859 measured by F(ST) do not necessarily require large sample sizes when using many
860 SNP markers. *PloS one* **7**, e42649, doi:10.1371/journal.pone.0042649 (2012).

861 28 Cunningham, D., Lawton, J., Jarra, W., Preiser, P. & Langhorne, J. The pir multigene
862 family of Plasmodium: antigenic variation and beyond. *Mol Biochem Parasitol* **170**,
863 65-73, doi:10.1016/j.molbiopara.2009.12.010 (2010).

864 29 Warncke, J. D., Vakonakis, I. & Beck, H. P. Plasmodium Helical Interspersed
865 Subtelomeric (PHIST) Proteins, at the Center of Host Cell Remodeling. *Microbiol*
866 *Mol Biol Rev* **80**, 905-927, doi:10.1128/MMBR.00014-16 (2016).

867 30 Castillo, A. I., Andreina Pacheco, M. & Escalante, A. A. Evolution of the merozoite
868 surface protein 7 (msp7) family in Plasmodium vivax and P. falciparum: A
869 comparative approach. *Infect Genet Evol* **50**, 7-19, doi:10.1016/j.meegid.2017.01.024
870 (2017).

871 31 Ntumngia, F. B. *et al.* Characterisation of a tryptophan-rich Plasmodium falciparum
872 antigen associated with merozoites. *Mol Biochem Parasitol* **137**, 349-353,
873 doi:10.1016/j.molbiopara.2004.06.008 (2004).

874 32 Kanjee, U., Rangel, G. W., Clark, M. A. & Duraisingh, M. T. Molecular and cellular
875 interactions defining the tropism of Plasmodium vivax for reticulocytes. *Curr Opin*
876 *Microbiol* **46**, 109-115, doi:10.1016/j.mib.2018.10.002 (2018).

877 33 Miller, L. H., McAuliffe, F. M. & Mason, S. J. Erythrocyte receptors for malaria
878 merozoites. *Am J Trop Med Hyg* **26**, 204-208, doi:10.4269/ajtmh.1977.26.204 (1977).

879 34 Iyer, J., Gruner, A. C., Renia, L., Snounou, G. & Preiser, P. R. Invasion of host cells
880 by malaria parasites: a tale of two protein families. *Mol Microbiol* **65**, 231-249,
881 doi:10.1111/j.1365-2958.2007.05791.x (2007).

882 35 Chan, L. J., Dietrich, M. H., Nguitragool, W. & Tham, W. H. Plasmodium vivax
883 Reticulocyte Binding Proteins for invasion into reticulocytes. *Cell Microbiol*, e13110,
884 doi:10.1111/cmi.13110 (2019).

885 36 Gruszczyk, J. *et al.* Cryo-EM structure of an essential Plasmodium vivax invasion
886 complex. *Nature* **559**, 135-139, doi:10.1038/s41586-018-0249-1 (2018).

887 37 Carlton, J. M. *et al.* Comparative genomics of the neglected human malaria parasite
888 Plasmodium vivax. *Nature* **455**, 757-763, doi:10.1038/nature07327 (2008).

889 38 Gilabert, A. *et al.* Plasmodium vivax-like genome sequences shed new insights into
890 Plasmodium vivax biology and evolution. *PLoS Biol* **16**, e2006035,
891 doi:10.1371/journal.pbio.2006035 (2018).

892 39 Ntumngia, F. B. *et al.* Genetic variation among Plasmodium vivax isolates adapted to
893 non-human primates and the implication for vaccine development. *Am J Trop Med*
894 *Hyg* **80**, 218-227 (2009).

895 40 Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and
896 split-read analysis. *Bioinformatics* **28**, i333-i339, doi:10.1093/bioinformatics/bts378
897 (2012).

898 41 Batchelor, J. D., Zahm, J. A. & Tolia, N. H. Dimerization of Plasmodium vivax DBP
899 is induced upon receptor binding and drives recognition of DARC. *Nat Struct Mol*
900 *Biol* **18**, 908-914, doi:10.1038/nsmb.2088 (2011).

901 42 Jeffares, D. C., Tomiczek, B., Sojo, V. & dos Reis, M. A beginners guide to
902 estimating the non-synonymous to synonymous rate ratio of all protein-coding genes
903 in a genome. *Methods Mol Biol* **1201**, 65-90, doi:10.1007/978-1-4939-1438-8_4
904 (2015).

905 43 Proto, W. R. *et al.* Adaptation of Plasmodium falciparum to humans involved the loss
906 of an ape-specific erythrocyte invasion ligand. *Nature communications* **10**, 4512,
907 doi:10.1038/s41467-019-12294-3 (2019).

908 44 Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling
909 and regulation. *Nature reviews. Molecular cell biology* **16**, 18-29,
910 doi:10.1038/nrm3920 (2015).

911 45 Culleton, R. & Carter, R. African Plasmodium vivax: distribution and origins. *Int J*
912 *Parasitol* **42**, 1091-1097, doi:10.1016/j.ijpara.2012.08.005 (2012).

913 46 Li, J. *et al.* Geographic subdivision of the range of the malaria parasite Plasmodium
914 vivax. *Emerging infectious diseases* **7**, 35-42, doi:10.3201/eid0701.010105 (2001).

915 47 Choi, E. H., Lee, S. K., Ihm, C. & Sohn, Y. H. Rapid DNA extraction from dried
916 blood spots on filter paper: potential applications in biobanking. *Osong Public Health*
917 *Res Perspect* **5**, 351-357, doi:10.1016/j.phrp.2014.09.005 (2014).

918 48 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
919 sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170
920 (2014).

921 49 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature*
922 *methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

923 50 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications
924 to single-cell sequencing. *J Comput Biol* **19**, 455-477, doi:10.1089/cmb.2012.0021
925 (2012).

926 51 Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-
927 assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579,
928 doi:10.1093/bioinformatics/btq683 (2011).

929 52 Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome*
930 *biology* **13**, R56, doi:10.1186/gb-2012-13-6-r56 (2012).

931 53 Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant
932 detection and genome assembly improvement. *PloS one* **9**, e112963,
933 doi:10.1371/journal.pone.0112963 (2014).

934 54 Laetsch, D. & Blaxter, M. BlobTools: Interrogation of genome assemblies [version 1;
935 referees: 2 approved with reservations]. *F1000Research* **6**,
936 doi:10.12688/f1000research.12232.1 (2017).

937 55 Jongwutiwes, S. *et al.* Mitochondrial genome sequences support ancient population
938 expansion in Plasmodium vivax. *Molecular biology and evolution* **22**, 1733-1739,
939 doi:10.1093/molbev/msi168 (2005).

940 56 Rodrigues, P. T. *et al.* Using mitochondrial genome sequences to track the origin of
941 imported Plasmodium vivax infections diagnosed in the United States. *Am J Trop*
942 *Med Hyg* **90**, 1102-1108, doi:10.4269/ajtmh.13-0588 (2014).

943 57 Bandelt, H. J., Forster, P. & Rohl, A. Median-joining networks for inferring
944 intraspecific phylogenies. *Molecular biology and evolution* **16**, 37-48,
945 doi:10.1093/oxfordjournals.molbev.a026036 (1999).

946 58 Leigh, J. W. & Bryant, D. POPART: full-feature software for haplotype network
947 construction. *Methods Ecol Evol* **6**, 1110-1116, doi:10.1111/2041-210x.12410
948 (2015).

949 59 Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database
950 management tool for second-generation genome projects. *BMC Bioinformatics* **12**,
951 491, doi:10.1186/1471-2105-12-491 (2011).

952 60 Steinbiss, S. *et al.* Companion: a web server for annotation and analysis of parasite
953 genomes. *Nucleic acids research* **44**, W29-34, doi:10.1093/nar/gkw292 (2016).

954 61 Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene
955 prediction and phylogenomics. *Molecular biology and evolution*,
956 doi:10.1093/molbev/msx319 (2017).

957 62 Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M.
958 BUSCO: assessing genome assembly and annotation completeness with single-copy
959 orthologs. *Bioinformatics* **31**, 3210-3212, doi:10.1093/bioinformatics/btv351 (2015).

960 63 Aurrecoechea, C. *et al.* PlasmoDB: a functional genomic database for malaria
961 parasites. *Nucleic acids research* **37**, D539-543, doi:gkn814 [pii]
962 10.1093/nar/gkn814 (2009).

963 64 Lechner, M. *et al.* Proteinortho: detection of (co-)orthologs in large-scale analysis.
964 *BMC Bioinformatics* **12**, 124, doi:10.1186/1471-2105-12-124 (2011).

965 65 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version
966 7: improvements in performance and usability. *Molecular biology and evolution* **30**,
967 772-780, doi:10.1093/molbev/mst010 (2013).

968 66 Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for
969 automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*
970 **25**, 1972-1973, doi:10.1093/bioinformatics/btp348 (2009).

971 67 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis
972 of large phylogenies. *Bioinformatics* **30**, 1312-1313,
973 doi:10.1093/bioinformatics/btu033 (2014).

68 Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790-2791, doi:10.1093/bioinformatics/btt468 (2013).

69 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:btp352 [pii] 10.1093/bioinformatics/btp352 (2009).

70 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

71 Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv*, 1207.3907 (2012).

72 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

73 Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Molecular biology and evolution* **31**, 1929-1936, doi:10.1093/molbev/msu136 (2014).

74 R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. (2007).

75 Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358-1370, doi:10.1111/j.1558-5646.1984.tb05657.x (1984).

76 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92, doi:10.4161/fly.19695 (2012).

77 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321, doi:10.1093/sysbio/syq010 (2010).

78 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696-704, doi:10.1080/10635150390235520 (2003).

79 Lefort, V., Longueville, J. E. & Gascuel, O. SMS: Smart Model Selection in PhyML. *Molecular biology and evolution* **34**, 2422-2424, doi:10.1093/molbev/msx149 (2017).

80 Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C. & Gascuel, O. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst Biol* **60**, 685-699, doi:10.1093/sysbio/syr041 (2011).

81 Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution* **23**, 254-267, doi:10.1093/molbev/msj030 (2006).

82 Bryant, D. & Moulton, V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution* **21**, 255-265, doi:10.1093/molbev/msh018 (2004).

83 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).

84 Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* **22**, 4673-4680 (1994).

85 Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238, doi:10.1186/1471-2105-13-238 (2012).

86 Seibt, K. M., Schmidt, T. & Heitkam, T. FlexiDot: highly customizable, ambiguity-aware dotplots for visual sequence analyses. *Bioinformatics* **34**, 3575-3577, doi:10.1093/bioinformatics/bty395 (2018).

87 van Ooij, C. *et al.* The malaria secretome: from algorithms to essential function in blood stage infection. *PLoS pathogens* **4**, e1000084, doi:10.1371/journal.ppat.1000084 (2008).

88 Boddey, J. A. *et al.* Role of plasmepsin V in export of diverse protein families from the Plasmodium falciparum exportome. *Traffic* **14**, 532-550, doi:10.1111/tra.12053 (2013).

89 Schulze, J. *et al.* The Plasmodium falciparum exportome contains non-canonical PEXEL/HT proteins. *Mol Microbiol* **97**, 301-314, doi:10.1111/mmi.13024 (2015).

90 Hu, G. *et al.* Transcriptional profiling of growth perturbations of the human malaria parasite Plasmodium falciparum. *Nat Biotechnol* **28**, 91-98, doi:10.1038/nbt.1597 (2010).

91 Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134, doi:10.1186/1471-2105-13-134 (2012).

92 Wang, S., Li, W., Liu, S. & Xu, J. RaptorX-Property: a web server for protein structure property prediction. *Nucleic acids research* **44**, W430-435, doi:10.1093/nar/gkw306 (2016).

93 Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research* **46**, W296-W303, doi:10.1093/nar/gky427 (2018).

1054 **Figure Legends**

1055

1056

1057 **Figure 1 - SNP phylogeny**

1058 A) Mid-point rooted maximum likelihood tree produced from 143,123 concatenated
1059 SNP positions with data from a minimum of 55 samples. The tree was produced using
1060 PhyML with the GTR evolutionary model. Branch support was evaluated with the
1061 Bayesian-like transformation of approximate likelihood ratio test (aBayes). Genetic
1062 distance is shown below the tree. *Plasmodium vivax* isolates are denoted as coloured
1063 circles by their country of sample origin. A tree with specific sample IDs is available
1064 in Figure S9. B) Magnification of the *Plasmodium simium* clade (as in panel A). C)
1065 Map showing the geographic locations at which the *P. simium* samples were
1066 collected.

1067

1068 **Figure 2 - Nucleotide diversity**

1069 Box plot showing the pair-wise nucleotide diversity between human-infecting
1070 *Plasmodium simium* samples (left), *Plasmodium vivax* samples (middle), and between
1071 *P. simium* and *P. vivax* samples (right). Nucleotide diversity (see Methods) is shown
1072 for entire genome (left-most plots, blue) and exonic regions only (right-most plots,
1073 orange). Individual values from pairwise comparisons are shown as grey dots, boxes
1074 denote 25th and 75th percentiles, and whiskers an additional 1.5 interquartile lengths.
1075 The observed nucleotide diversity between *P. simium* and *P. vivax* samples is
1076 significantly higher than between *P. vivax* samples (Mann-Whitney U, genome;
1077 $p=4.60\times10^{-21}$, exons; $p=1.78\times10^{-20}$).

1078

1079 **Figure 3 - F_{ST} ratios**

1080 A) Gene F_{ST} values were calculated between *Plasmodium simium* and American
1081 *Plasmodium vivax* samples (x-axis) and American and Asian *P. vivax* samples (y-
1082 axis). Each dot corresponds to a gene, and the distributions of the two F_{ST} measures
1083 are shown as bar charts above and to the right of the scatter plot. B) The ratio between
1084 F_{ST} values between i) *P. simium* and American *P. vivax* samples, and ii) American
1085 and Asian *P. vivax* samples were calculated for each gene (top). A pseudo count of
1086 one was added to all F_{ST} values. The distributions of $\log_2$-ratios are shown as violin
1087 plots (bottom) for all genes (grey), PIR genes (red), PHIST genes (brown), TRAg

genes (turquoise), invasion genes (yellow), and exported genes (dark blue). Only genes with SNP differences between the three populations are included in this analysis. F_{ST} values and ratios are provided in Table S11. P-values from Mann-Whitney U tests for differences in medians between PIR, PHIST, TRAg, invasion genes, exported genes, and all remaining genes (e.g. PIR genes are compared to all non-PIR genes) are indicated on plot.

Figure 4 - 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* genome harbours two RBP2d genes, one of which is a pseudogene (Table S8).
B) Schematic depiction of samples with and without the deletion found in DBP1.
C) Left: Structural rendering of DBP1, showing known structural domains and motifs. The two fragment molecules from the human DARC receptor are shown in grey. The 3-dimensional structure of the DBL-DARC complex was modelled based on the *P. vivax* crystallographic model (PDB 4nuv). The region deleted in sequences from human-infecting *P. simium*, as compared to *P. vivax* P01, is highlighted in red. Right: Details of DBP1 protein alignments. A full alignment is available in Figure S22.
D) Similar to panel C) but for RBP2a. The complex between the reticulocyte binding domain and the human receptor was modelled based on the cryoEM structure of the complex between the *P. vivax* RBP2b and the human transferrin receptor TfR1 (PDB 6d05). A full alignment is available in Figure S33.

Figure 1

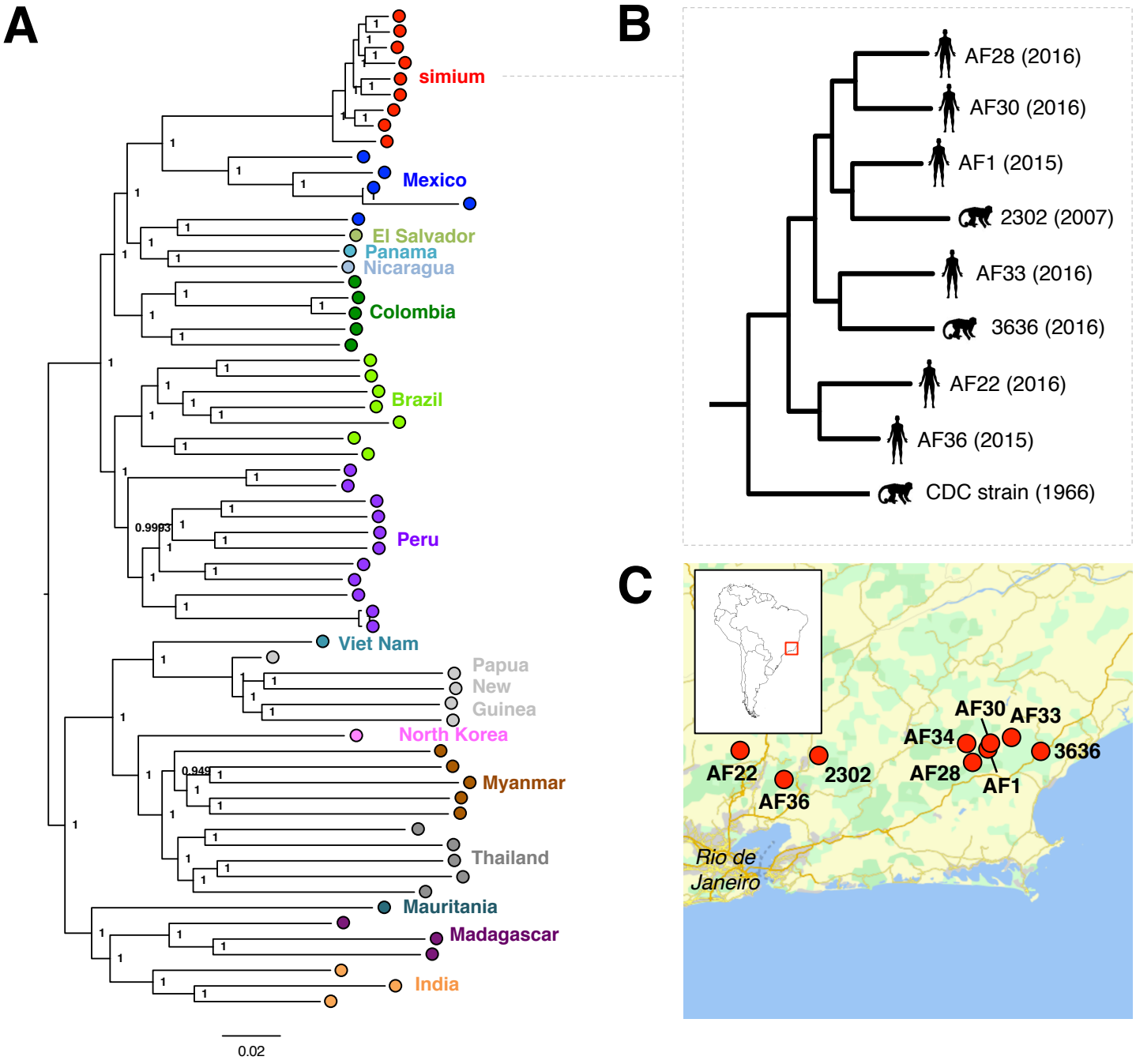

A
simium
Mexico
El Salvador
Panama
Nicaragua
Colombia
Brazil
Peru
Viet Nam
Papua
New
Guinea
North Korea
Myanmar
Thailand
Mauritania
Madagascar
India
0.02
B
AF28 (2016)
AF30 (2016)
AF1 (2015)
2302 (2007)
AF33 (2016)
3636 (2016)
AF22 (2016)
AF36 (2015)
CDC strain (1966)
C
AF22
AF36
2302
AF30
AF33
AF34
AF28
AF1
3636
Rio de
Janeiro

Figure 2

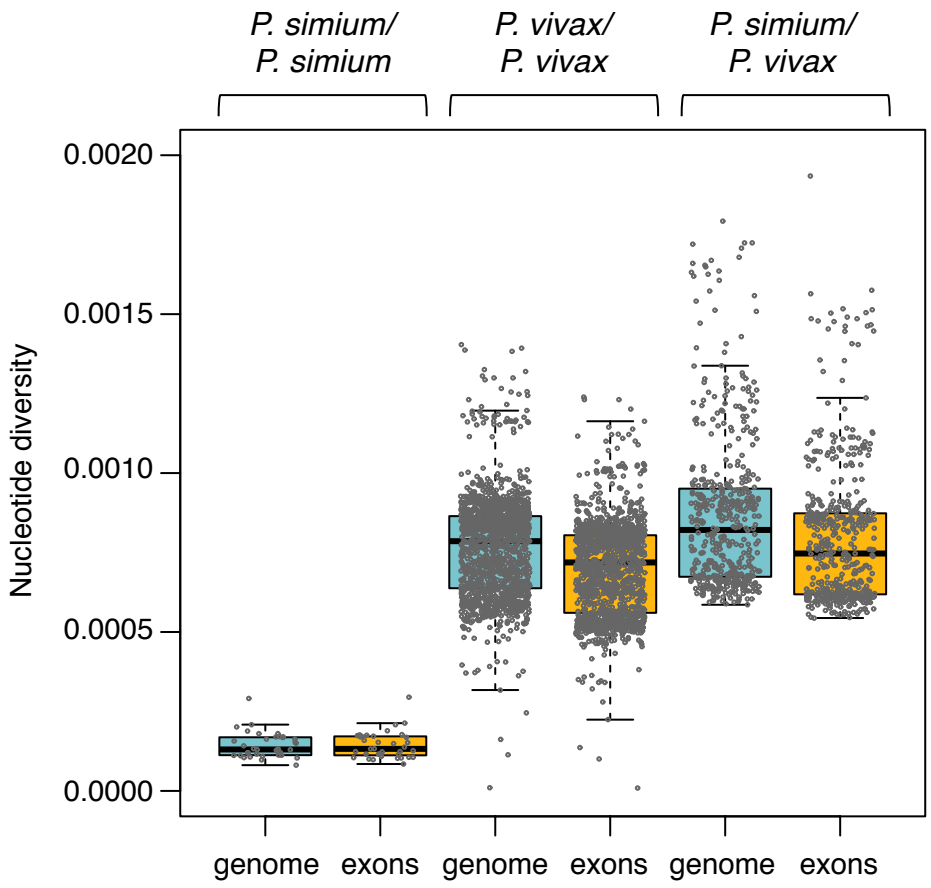

P. simium/
P. simium
P. vivax/
P. vivax
P. simium/
P. vivax
Nucleotide diversity
0.0020
0.0015
0.0010
0.0005
0.0000
genome
exons
genome
exons
genome
exons

Figure 3

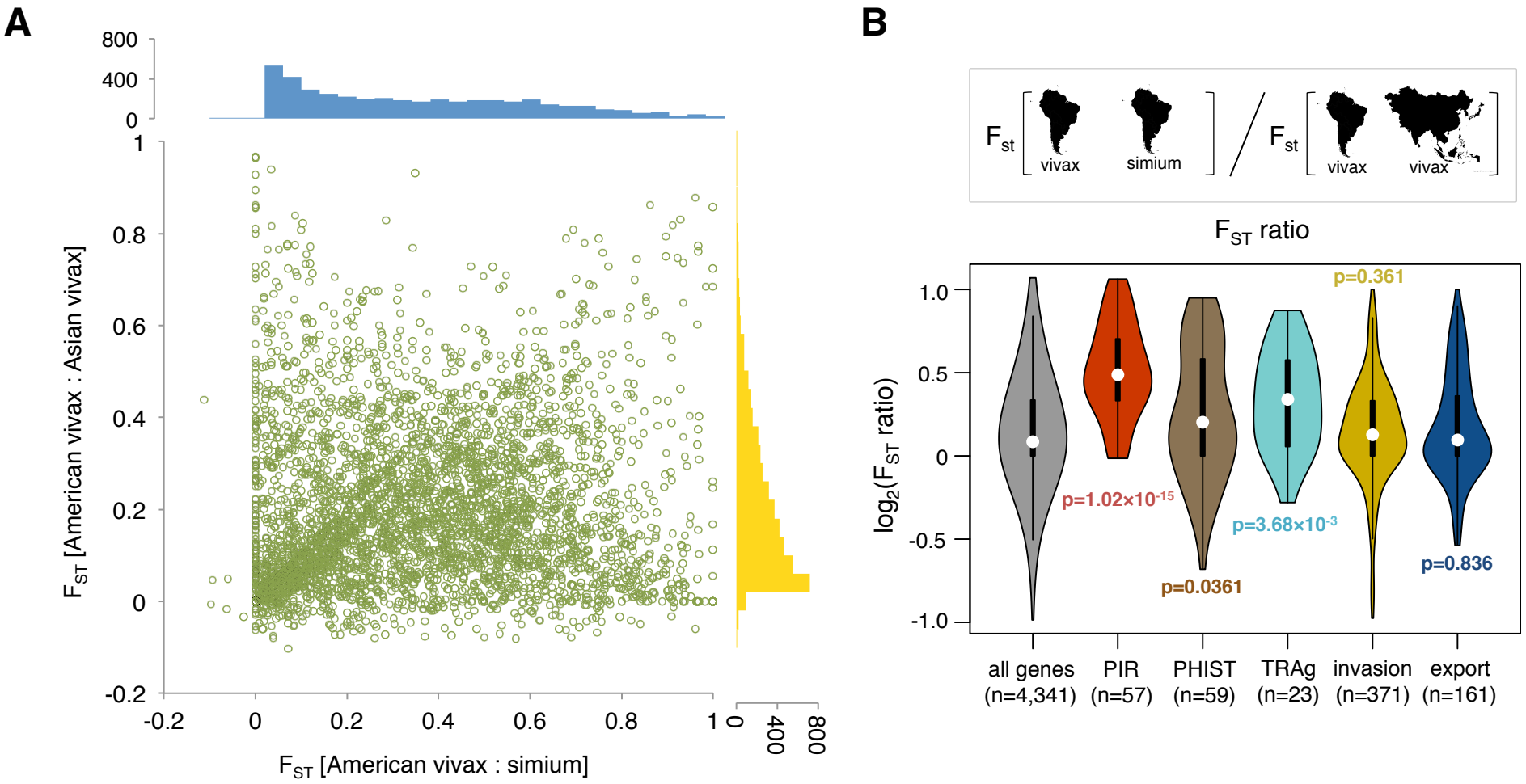
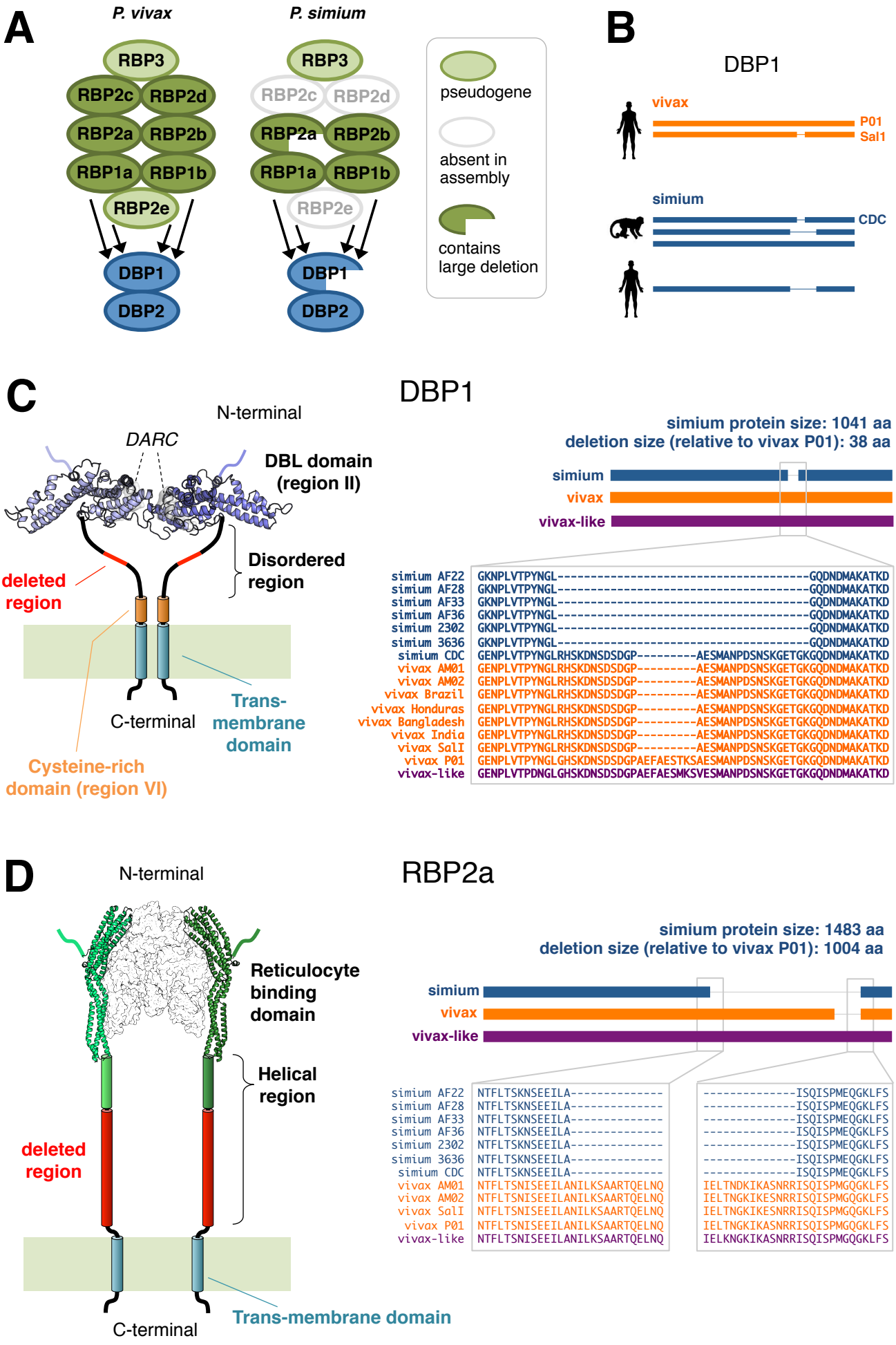

A
800
400
0
1
0.8
0.6
0.4
0.2
0
-0.2
-0.2
0
0.2
0.4
0.6
0.8
1
0
400
800
F_ST [American vivax : Asian vivax]
F_ST [American vivax : simium]
B
F_st
vivax
simium
F_st
vivax
vivax
F_ST ratio
1.0
0.5
0
-0.5
-1.0
log2(F_ST ratio)
p=1.02×10^-15
p=0.0361
p=3.68×10^-3
p=0.361
p=0.836
all genes (n=4,341)
PIR (n=57)
PHIST (n=59)
TRAg (n=23)
invasion (n=371)
export (n=161)

Figure 4


A
P. vivax
P. simium
RBP3
RBP2c
RBP2d
RBP2a
RBP2b
RBP1a
RBP1b
RBP2e
DBP1
DBP2
pseudogene
absent in assembly
contains large deletion
B
DBP1
vivax
P01
Sal1
simium
CDC
C
DBP1
N-terminal
DARC
DBL domain (region II)
Disordered region
deleted region
Trans-membrane domain
C-terminal
Cysteine-rich domain (region VI)
simium protein size: 1041 aa
deletion size (relative to vivax P01): 38 aa
simium
vivax
vivax-like
simium AF22 GKNPLVTPYNGL----------GQDNDMAKATKD
simium AF28 GKNPLVTPYNGL----------GQDNDMAKATKD
simium AF33 GKNPLVTPYNGL----------GQDNDMAKATKD
simium AF36 GKNPLVTPYNGL----------GQDNDMAKATKD
simium 2302 GKNPLVTPYNGL----------GQDNDMAKATKD
simium 3636 GKNPLVTPYNGL----------GQDNDMAKATKD
simium CDC GENPLVTPYNGLRHSKDNSDSDGP---------AESMANPDSNSKGETGKGQDNDMAKATKD
vivax AM01 GENPLVTPYNGLRHSKDNSDSDGP---------AESMANPDSNSKGETGKGQDNDMAKATKD
vivax AM02 GENPLVTPYNGLRHSKDNSDSDGP---------AESMANPDSNSKGETGKGQDNDMAKATKD
vivax Brazil GENPLVTPYNGLRHSKDNSDSDGP---------AESMANPDSNSKGETGKGQDNDMAKATKD
vivax Honduras GENPLVTPYNGLRHSKDNSDSDGP---------AESMANPDSNSKGETGKGQDNDMAKATKD
vivax Bangladesh GENPLVTPYNGLRHSKDNSDSDGP---------AESMANPDSNSKGETGKGQDNDMAKATKD
vivax India GENPLVTPYNGLRHSKDNSDSDGP---------AESMANPDSNSKGETGKGQDNDMAKATKD
vivax SalI GENPLVTPYNGLRHSKDNSDSDGP---------AESMANPDSNSKGETGKGQDNDMAKATKD
vivax P01 GENPLVTPYNGLGHSKDNSDSDGPAEFAESTKSAESMANPDSNSKGETGKGQDNDMAKATKD
vivax-like GENPLVTPDNGLGHSKDNSDSDGPAEFAESMKSVESMANPDSNSKGETGKGQDNDMAKATKD
D
N-terminal
RBP2a
Reticulocyte binding domain
Helical region
deleted region
C-terminal
Trans-membrane domain
simium protein size: 1483 aa
deletion size (relative to vivax P01): 1004 aa
simium
vivax
vivax-like
simium AF22 NTFLTSKNSEEILA---------- ----------ISQISPMEQGKLFS
simium AF28 NTFLTSKNSEEILA---------- ----------ISQISPMEQGKLFS
simium AF33 NTFLTSKNSEEILA---------- ----------ISQISPMEQGKLFS
simium AF36 NTFLTSKNSEEILA---------- ----------ISQISPMEQGKLFS
simium 2302 NTFLTSKNSEEILA---------- ----------ISQISPMEQGKLFS
simium 3636 NTFLTSKNSEEILA---------- ----------ISQISPMEQGKLFS
simium CDC NTFLTSKNSEEILA---------- ----------ISQISPMEQGKLFS
vivax AM01 NTFLTSNISEEILANILKSAARTQELNQ IELTNDKIKASNRRISQISPMGQGKLFS
vivax AM02 NTFLTSNISEEILANILKSAARTQELNQ IELTNGKIKESNRRISQISPMGQGKLFS
vivax SalI NTFLTSNISEEILANILKSAARTQELNQ IELTNGKIKESNRRISQISPMGQGKLFS
vivax P01 NTFLTSNISEEILANILKSAARTQELNQ IELTNGKIKASNRRISQISPMGQGKLFS
vivax-like NTFLTSNISEEILANILKSAARTQELNQ IELKNGKIKASNRRISQISPMGQGKLFS