bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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]

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

constructed from 3,181 of 1:1 orthologs of the annotated *P. simium* protein-coding genes with the orthologues of *P. vivax, P. cynomolgi, P. coatneyi, P. knowlesi, P. malariae, P. falciparum, P. reichenowi,* and *P. gallinaceum* confirmed that *P. simium* is very closely related to *P. vivax* (Figure S3). Analysis of gene families revealed a *P. simium* gene repertoire largely similar to *P. vivax* (Figure S4) and no indication of a recent expansion of the *P. simium* PIR gene family (Figure S5).

P. simium-P. vivax diversity analysis

To detect single nucleotide polymorphisms (SNPs) within the *P. vivax*/*P. simium* clade, short Illumina paired-end sequence reads were mapped onto the *P. vivax* P01 reference genome [24]. Reads were collected from eleven human-derived *P. simium* samples, two monkey *P. simium* samples, two *P. vivax* samples from the Brazilian Amazon, and a previously published *P. simium* CDC strain (originally isolated in 1966, see Methods). Data from a range of *P. vivax* strains representing a global distribution was retrieved from the literature [25]. Including only SNPs with a minimum depth of five reads, a total of 232,780 SNPs were called initially across 79 samples. Sixteen samples were subsequently removed, primarily due to low coverage, resulting in a total of 63 samples for further analysis (Table S2, Table S3). Since few SNP loci were covered across all samples, and to enable diversity analysis, we restricted all further analysis to the 124,968 SNPs for which data were available from at least 55 samples (Figure S6).

P. simium-P. vivax population analysis

A Principal Component Analysis (PCA) plot constructed from these genome-wide SNP loci showed a clear separation between American and Asian *P. vivax* samples as well as a distinct grouping of *P. simium* samples (Figure S7). The latter observation suggests that both human and NHP *P. simium* isolates represent a single population that is genetically different from American *P. vivax* populations. A similar pattern is observed by performing a multidimensional scaling analysis of the SNP data (Figure S8). To enable a phylogenetic approach, we constructed an alignment from the 124,968 SNP sites. In the resulting phylogenetic tree, *P. vivax* strains generally clustered according to their geographical origin, with clear separation of the Asian and American samples (Figure 1A, a tree with sample IDs is available in Figure S9). *Plasmodium simium* samples clustered as a monophyletic group with Mexican *P. vivax* samples (Figure 1A), consistent with a recent American origin for *P. simium*. This phylogeny, based on genome-wide SNPs represents an “average” phylogeny

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

all individual genes. Clearly, the small number of samples renders this analysis prone to false and incorrect signals, and F_{ST} values for individual genes should be interpreted with caution. Nevertheless, we attempted to look for general patterns in F_{ST} values across gene groups.

Amongst the 4,432 *P. vivax* genes with at least one SNP in our data set, we examined the top-25% of genes with highest F_{ST} values for enrichment in functional Gene Ontology (GO) terms or metabolic pathways. No GO terms or pathways were significant at the 0.05 level after Bonferroni correction (Table S4 & Table S5). Using the *P. falciparum* orthologues – when available – gave similar results (not shown). We next tested whether any gene families (Figure S4, Table S6) were associated with high F_{ST} values. Genes belonging to the Plasmodium Interspersed Repeat (PIR) family involved in antigenic variation [28], the Plasmodium Helical Interspersed Subtelomeric (PHIST) genes, a family of exported proteins [29], the merozoite surface proteins MSP7[30], and Tryptophan-rich antigens (TRAg) [31] were enriched among the genes with high F_{ST} values (binomial distribution, PIR; p=0.016, PHIST; $p=4.2\times10^{-4}$, MSP7; p=0.034, TRAg; $p=5.5\times10^{-5}$).

As these gene families are involved in parasite-host interactions the elevated F_{ST} values may simply reflect a general pattern of rapid genetic divergence in *Plasmodium* parasites. To test this, we repeated the F_{ST} analysis between American *P. vivax* and a selection of Asian *P. vivax* isolates (Myanmar and Thailand samples only). Consistent with the phylogenetic analysis (Figure 1A) gene F_{ST} was slightly higher overall between *P. simium* and American *P. vivax* than between American and Asian *P. vivax* samples (Figure 3A). However, none of the gene families were overrepresented among genes with high F_{ST} (top-25%) between American and Asian *P. vivax*.

To further examine whether the elevated F_{ST} measures found for PIR, PHIST, MSP7, and TRAg genes are exclusive to the comparison between *P. simium* and American *P. vivax*, we calculated the ratio between the two F_{ST} measurements ('simium versus American vivax' and ' American versus Asian vivax') (Figure 3B). The ratios for PIR, PHIST and TRAg genes were significantly higher than observed for the remaining genes (Figure 3C), whereas ratios for MSP7 genes were not (Mann-Whitney U, p=0.11). Although the *P. simium* and the *P. vivax* P01 genomes both encode a high number of the gene family members, our analysis is restricted to the *P. vivax* genes to which our *P. simium* short read sequences can map. For example, only 408 out of the 1209 *P. vivax* PIR genes have coverage from *P. simium* reads

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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*.

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

P. simium invadome components

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

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 4; Table S8; Figure S18; Figure S19) across all sequenced *P. simium* samples. As in *P. vivax* genomes, the *P. simium* RBP3 is a pseudogene [38], indicating that conversion to a pseudogene happened prior to the split between *P. vivax* and *P. simium*.

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 S20).

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 been expected if the *P. simium* genome contained multiple (duplicated) copies of non-assembled invasion genes (Figure S21).

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 S29, top & middle). Interestingly, NHP-infecting *P. simium* isolates contained a mix of sequences with and without deletions (Figure S29, bottom). PCR results are summarized in Table S9. 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 S30-S33). PCR-based genotyping results confirmed the presence of this deletion in all *P. simium* isolates irrespective of their host (Figure S34, Table S9).

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 [40] to *P. simium* assemblies. This set consisted of 244 indels in 222 genes (Table S10). 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 S35). 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 deletion 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* [41], followed by a largely disordered region and a cysteine-rich domain (Figure 4c). 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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DARC binding site (Figure S36). Based on our models, these substitutions are unlikely to negatively affect the association with DARC, supporting the idea that the DBL domains of both *P. simium* would be capable of binding to human DARC. Hence, the human-infecting *P. simium* DBL1 probably retains the capacity to bind to human DARC, but has the interacting domain positioned closer to the membrane than in the NHP-infecting CDC strain.

The deletion we detected in the *P. simium* RBP2a was larger, resulting in the loss of 1003 amino acid residues. These residues are predicted to form a mostly α-helical extracellular stem-like structure that positions the reticulocyte binding domain away from the membrane (Figure 4d). However, given that the deletion affects neither the transmembrane region, nor the receptor-binding domain, our analysis suggests the resulting truncated RBP2a protein could still associate with the human receptor, but that the binding event would occur closer to the merozoite membrane.

Discussion

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

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*

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

441 samples revealed that isolates from human hosts all carried the DBP1 deletion, whereas
442 isolates from NHPs displayed both absence and presence of the deletion. This DBP1 deletion
443 is not present in the *P. simium* isolated from a brown howler monkey in the 1960s, which was
444 previously shown to be incapable of infecting humans [23], although some degree of laboratory
445 adaptation of this parasite may have affected its genome. However, this deletion is also absent
446 in *P. vivax*, so cannot in itself explain the ability of *P. simium* to infect humans in the current
447 outbreak. It is possible, however, that this deletion is required for *P. simium* to invade human
448 red blood cells given the alterations that have occurred elsewhere in its invadome following
449 adaptation to non-human primates since the split between *P. simium* and its human-infecting
450 *P. vivax* ancestor.
451
452 Invadome proteins are obvious candidates for genetic factors underlying host-specificity; an
453 inactivating mutation in a *P. falciparum* erythrocyte binding antigen has recently been shown
454 to underlie host-specificity [43]. Traditionally, functional studies on invadome proteins have
455 focused on domains known to bind or interact directly with the host. Although the *P. simium*-
456 specific DBP1 and RBP2a deletions reported here do not cover known structural motifs, these
457 deletions could nevertheless affect host cell recognition as disordered protein regions have
458 known roles in cellular regulation and signal transduction [44]. Further, a shorter, less flexible
459 linker between the merozoite membrane and the receptor-binding DBP1 domain may favour a
460 more rigid and better oriented positioning of the dimeric DBP1, enhancing its capacity to
461 engage the human receptor.
462
463 Phylogenetic analyses of the *P. simium* clade gives the geographical location of its most
464 closely related *P. vivax* strain as Mexico, and not Brazil. In imported populations, the
465 relationship between geographical and genetic proximity may be weak. Multiple introductions
466 of diverse strains from founder populations may occur independently over large distances, so
467 that two closely related strains may be introduced in distantly located regions. It may be
468 postulated that there occurred the introduction of strains of *P. vivax* to Mexico from the Old
469 World that were closely related, due to similar regions of origin, to strains introduced to the
470 Atlantic Forest which then went on to become *P. simium* in New World monkeys. Strains
471 from a different point of origin were introduced to the Amazonian region of Brazil. This
472 hypothesis necessitates reproductive isolation of the *P. simium* clade from the Brazilian *P.*
473 *vivax* parasites following their initial introduction; an isolation that would be facilitated,
474 presumably, by their separate host ranges.

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 *Plasmodium simium* sample AF22 was selected for genome assembly based on read quality
567 and coverage. After removal of human contaminants, Illumina reads were assembled into
568 contigs using the Spades (v 3.70) assembler [50]. Contigs assembled into scaffolds running
569 SSPACE (v 3.0) [51] for 15 rounds and gaps filled with Gapfiller (v 1.10) [52]. Scaffolds were
570 subsequently corrected with Illumina reads using the Pilon (v 1.22) software [53]. Blobtools (v
571 1.0) (DOI: 10.5281/zenodo.845347) [54] was used to remove any residual contaminant
572 scaffolds. Genome size and GC content was in line with that of *P. vivax* (Table S1). Genomic
573 scaffolds representing the mitochondrial and apicoplast genome were identified through
574 blastn searches against the corresponding *P. falciparum* and *P. vivax* sequences (Figure S1).
575 The *P. simium* mitochondrial genome was aligned against a range of previously published *P.*
576 *vivax* and *P. simium* mitochondrial genomes [55,56]. A gap-filled region in the alignment where

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the distal parts of the *P. simium* scaffold were merged was manually deleted. A minimum spanning haplotype network was produced using PopART [57,58] confirming the authenticity of the *P. simium* mitochondrial genome (Figure S37).

Genome Annotation

Two approaches were used to annotate the reference *P. simium* AF22 genome. Firstly, the Maker pipeline (v 2.31.8) [59] was run for two rounds, using ESTs and protein evidence from *P. vivax* and *P. cynomolgi* strain B and *P. falciparum* to generate Augustus gene models. Secondly, a separate annotation was produced using the Companion web server [60]. Companion was run using the *P. vivax* P01 reference assembly and default parameters. Basic annotation statistics are provided in Table S1. The relatively low number of genes (5966) is due to the fragmented and incomplete nature of the *P. simium* assembly (Table S1). Gene content was estimated using BUSCO [61,62] (v3.0) revealing an annotation completeness comparable to other *Plasmodium* genome assemblies (Figure S2).

PlasmoDB Genome References and Annotations

Genome FASTA files, as well as annotated protein and CDS files were obtained from PlasmoDB [63] for the following species: *P. gallinaceum* 8A, *P. cynomolgi* B and M, *P knowlesi* H, *P. falciparum* 3D7, *P. reichenowi* G01, *P. malariae* UG01, *P. ovale curtisi* GH01, *P. coatneyi* Hackeri, *P. vivax* P01 and *P. vivax* SalI. For each species, PlasmoDB version 36 was used.

Orthologous group determination

Amino-acid sequences-based phylogenetic trees were prepared using protein sequences from the *P. simium* annotation, as well as the protein annotations from 10 malaria species downloaded from PlasmoDB: *P. vivax* P01, *P. cynomolgi* B, *P. knowlesi* H, *P. vivax*-like Pvl01, *P. coatneyi* Hackeri, *P. falciparum* 3D7, *P. gallinaceum* 8A, *P. malariae* UG01, *P. ovale* curtisi GH01, and *P. reichenowi* G01. *P. vivax*-like from PlasmoDB version 43, all other annotations from version 41. A total of 3181 1:1 orthologous genes were identified using the Proteinortho (v 6.0.3) software [64]. Approximately 88% of the predicted genes in *P. simium* have orthologues in *P. vivax* P01 (Figure S38).

Short indels in genes

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Shorter indels (<500 bp) were detected from soft-clipping information in read mapping using the '-i' option in DELLY [40] (v 0.7.9). Predicted indels in protein-coding genes were then compared to independent assemblies of *P. simium* samples AF22, AF28, AF33, AF36, 2302 and 3636 (Table S2) and indels present in both DELLY predictions and all assemblies were kept (Table S9). Indel boundaries shifted by a maximum of two amino acid positions were allowed between predictions and assemblies. As assemblies were not complete, data for any gene was only required to be present in five of the six *P. simium* assemblies used.

Protein phylogeny

Protein sequences were aligned using mafft (v 7.222) [65] and alignments were subsequently trimmed with trimAl (v 1.2rev59) [66] using the heuristic 'automated1' method to select the best trimming procedure. Trimmed alignments were concatenated and a phylogenetic tree was constructed using RAxML (v 8.2.3) [67] with the PROTGAMMALG model.

SNP calling and analysis

Short sequence reads from 15 *P. simium* samples (13 derived from humans and two from NHPs) and two *P. vivax* samples, all from this study (Table S2), were aligned against a combined human (hg38) and *P. vivax* (strain P01, version 39) genome using NextGenMap (v0.5.5) [68]. This was similarly done for 30 previously published *P. vivax* strains [25] and the Sal1 reference. These data sets were downloaded from ENA (https://www.ebi.ac.uk/ena) (Table S3). Duplicate reads were removed using samtools (v 1.9) [69] and the filtered reads were realigned using IndelRealigner from the GATK package (v 4.0.11) [70]. SNPs were called independently with GATK HaplotypeCaller and freebayes (v 1.2.0) [71], keeping only SNPs with a QUAL score above 30. The final SNP set was determined from the inter-section between GATK and freebayes. Allele frequencies and mean coverage across SNP sites are shown in Figure S39. A PCA plot was constructed using plink (v 1.90) [72], and admixture analysis was done with Admixture (v 1.3.0) [26]. F_{ST} values were estimated from nucleotide data with the PopGenome R package [73,74] using the Weir & Cockerham method [75]. Non-synonymous and synonymous SNPs were identified using snpeff [76].

SNP phylogeny

Alleles from SNP positions with data in 55 samples were retrieved, concatenated, and aligned using mafft [65]. Tree was produced by PhyML [77,78] with the GTR substitution model selected by SMS [79]. Branch support was evaluated with the Bayesian-like transformation of

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

PCR amplification of DBP1 and RBP2a genes

PCR primers were initially designed from alignments between *P. vivax* and *P. simium* sequences and tested using Primer-BLAST [91] and PlasmoDB (www.plasmodb.org). For DBP1, the reaction was performed in 10 µL volumes containing 0.5 µM of each oligonucleotide primer, 1 µL DNA and 5 μ L of Master Mix 2x (Promega) (0.3 units of Taq Polymerase, 200 µM each deoxyribonucleotide triphosphates and 1.5 mM MgCl2). Samples were run with the following settings: 2 minutes of activation at 95°C, followed by 35 cycles with 30 seconds denaturation at 95°C, 30 seconds annealing at 57°C (ΔT=-0.2 °C from 2nd cycle) and 1 minute extension at 72°C, then 5 minutes final extension at 72°C.

For the RBP2a PCR, the reaction was performed in 10 µL volumes containing 0.5 µM of each oligonucleotide primer, 1 µL DNA, 0.1 µL PlatinumTaq DNA Polymerase High Fidelity (Invitrogen, 5U/µL), 0.2 mM each deoxyribonucleotide triphosphates and 2 mM $MgSO_4$. The PCR assays were performed with the following cycling parameters: an initial denaturation at 94°C for 1.5 min followed by 40 cycles of denaturation at 94°C for 15 sec, annealing at 65°C for 30 sec (ΔT=-0.2 °C from 2nd cycle) and extension at 68°C for 3.5 min. All Genotyping assays were performed in the thermocycler Veriti 96 wells, Applied Biosystems, and the amplified fragments were visualized by electrophoresis on agarose gels (2% for DBP1 and 1% for RBP2a) in 1x TAE buffer (40 mM Tris-acetate, 1 mM EDTA) with 5 µg/ mL ethidium bromide (Invitrogen) in a horizontal system (Bio-Rad) at 100 V for 30 min. Gels were examined with a UV transilluminator (UVP - Bio-Doc System).

To prevent cross-contamination, the DNA extraction and mix preparation were performed in "parasite DNA-free rooms" distinct from each other. Furthermore, each of these separate areas has different sets of pipettes and all procedures were performed using filtered pipette tips. DNA extraction was performed twice on different days. Positive (DNA extracted from 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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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).

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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).

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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).

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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).

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure Legends

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 sample origin. A tree with specific sample IDs is available in Figure S9. B) Magnification of the *Plasmodium simium* clade (as in panel A). C) Map showing the geographic locations at which the *P. simium* samples were collected.

Figure 2 - Nucleotide diversity

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

Figure 3 - F_{ST} ratios

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

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

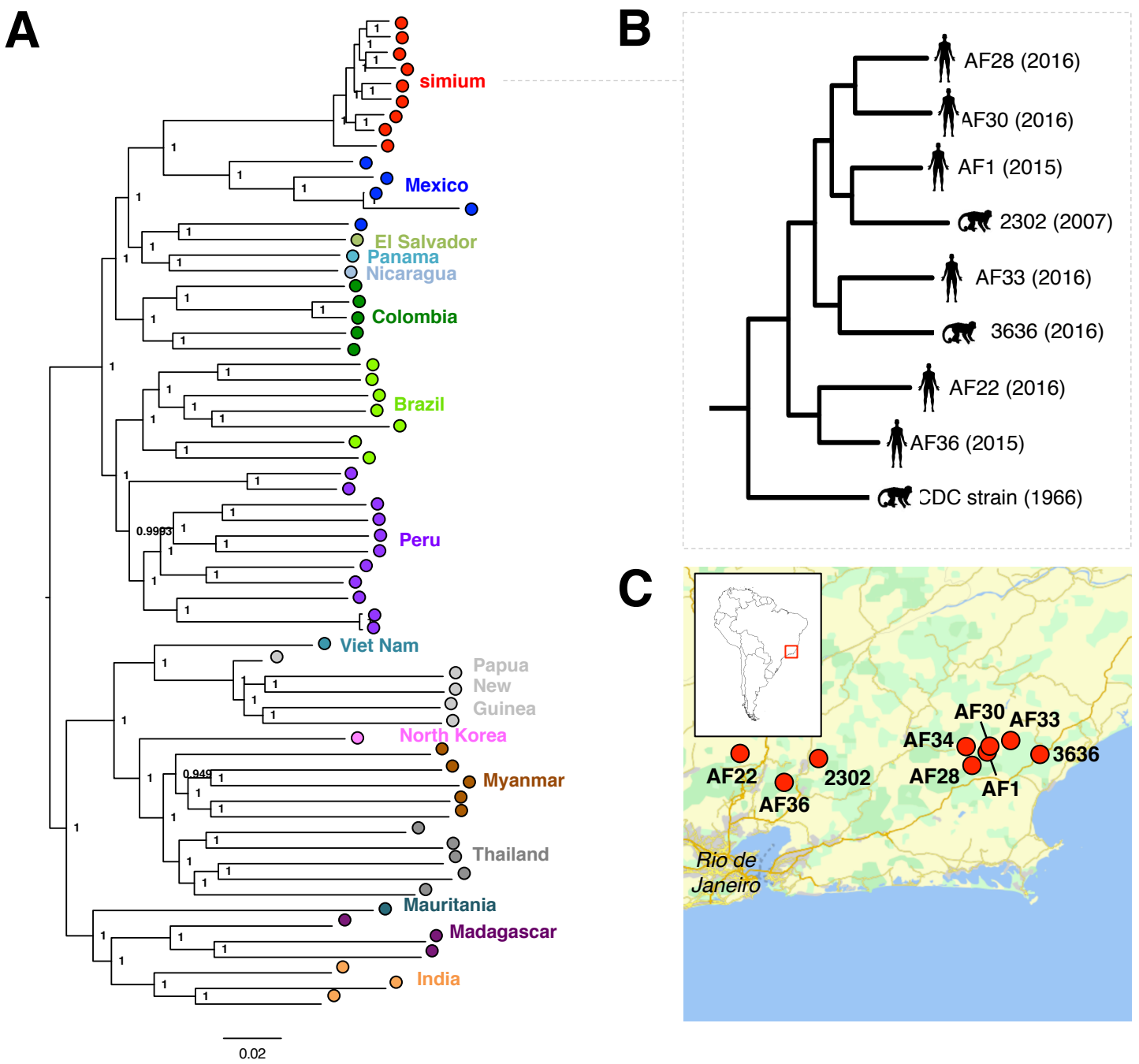

A
simium
Mexico
El Salvador
Panama
Nicaragua
Colombia
Brazil
Peru
Viet Nam
Papua New Guinea
North Korea
Myanmar
Thailand
Mauritania
Madagascar
India
0.9993
0.949
0.02
B
AF28 (2016)
AF30 (2016)
AF1 (2015)
2302 (2007)
AF33 (2016)
3636 (2016)
AF22 (2016)
AF36 (2015)
CDC strain (1966)
C
AF30
AF33
AF34
3636
AF22
2302
AF28
AF1
AF36
Rio de Janeiro

Figure 2

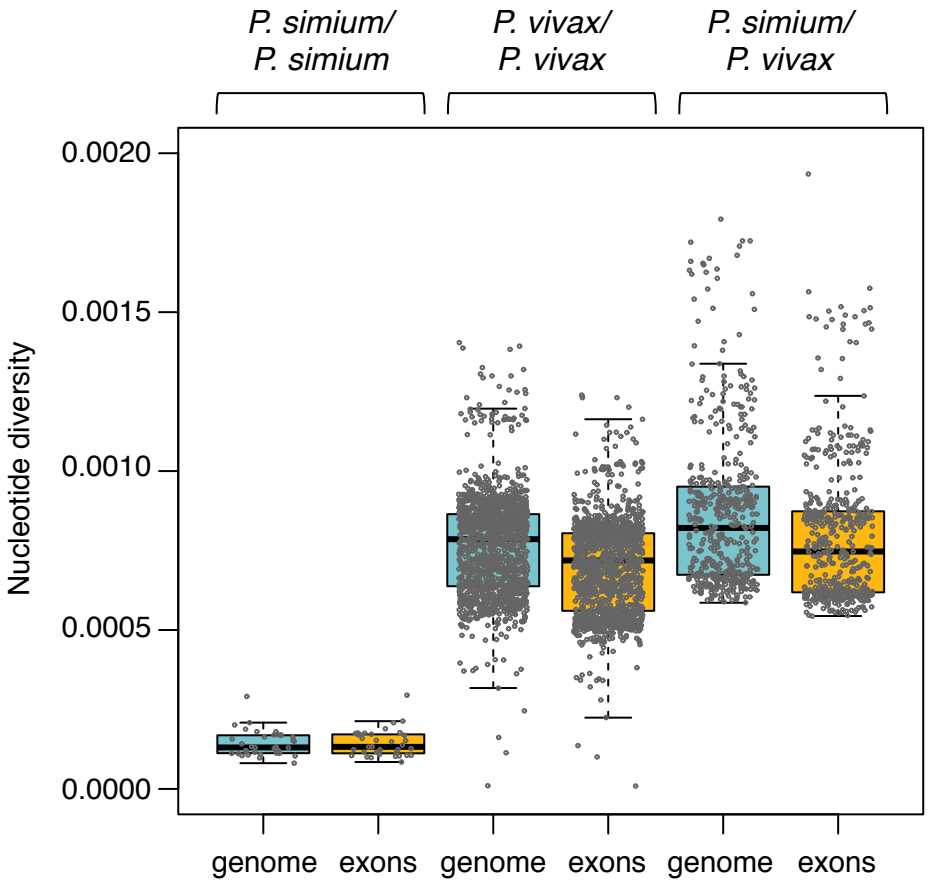
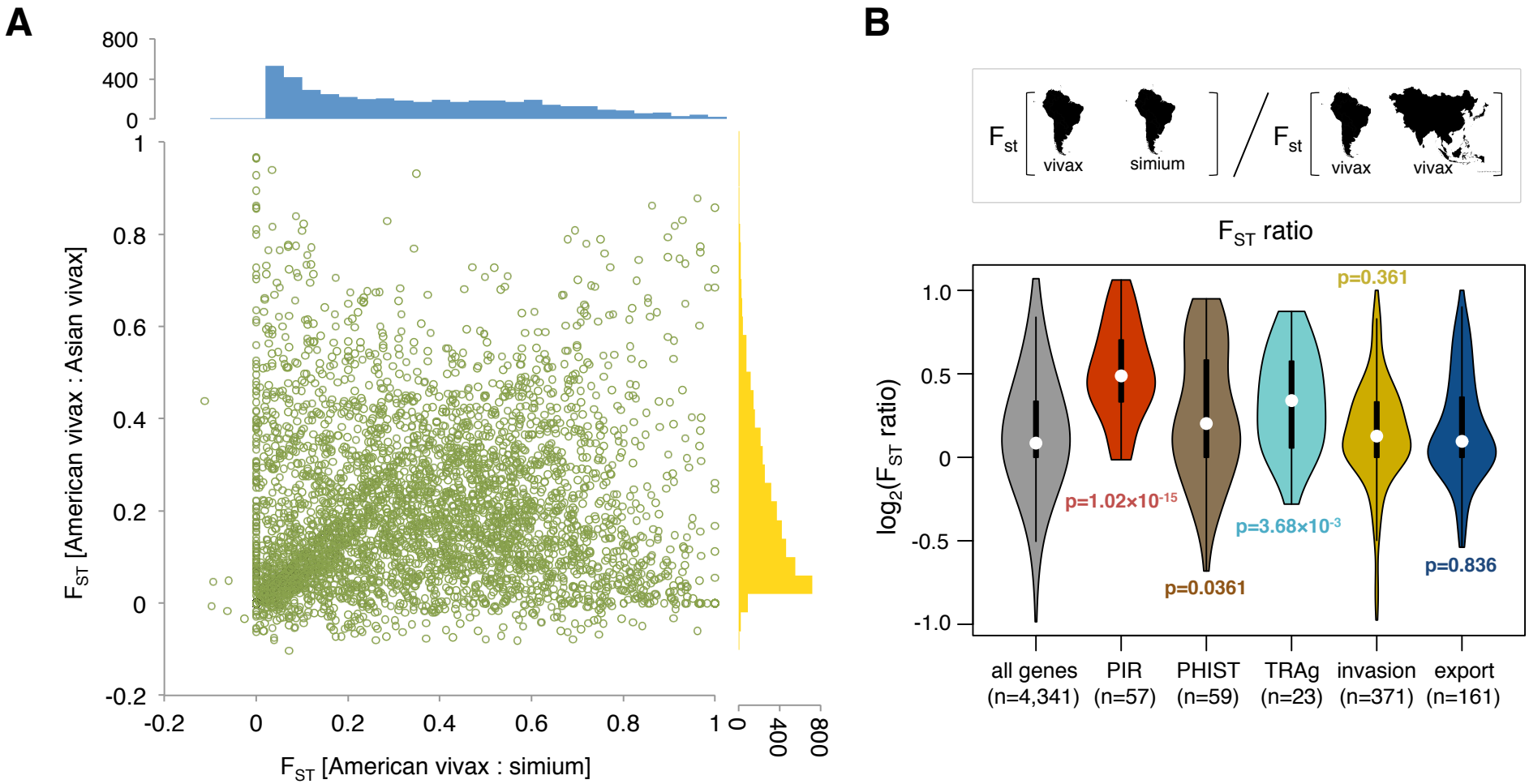

P. simium/
P. simium
P. vivax/
P. vivax
P. simium/
P. vivax
Nucleotide diversity
0.0020
0.0015
0.0010
0.0005
0.0000
genome
exons
genome
exons
genome
exons

Figure 3


A
800
400
0
1
0.8
0.6
0.4
0.2
0
-0.2
F_{ST} [American vivax : Asian vivax]
-0.2
0
0.2
0.4
0.6
0.8
1
F_{ST} [American vivax : simium]
0
400
800
B
F_{st}
vivax
simium
F_{st}
vivax
vivax
F_{ST} ratio
$\log_2(F_{ST}$ ratio)
1.0
0.5
0
-0.5
-1.0
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)

bioRxiv preprint doi: https://doi.org/10.1101/841171; this version posted December 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

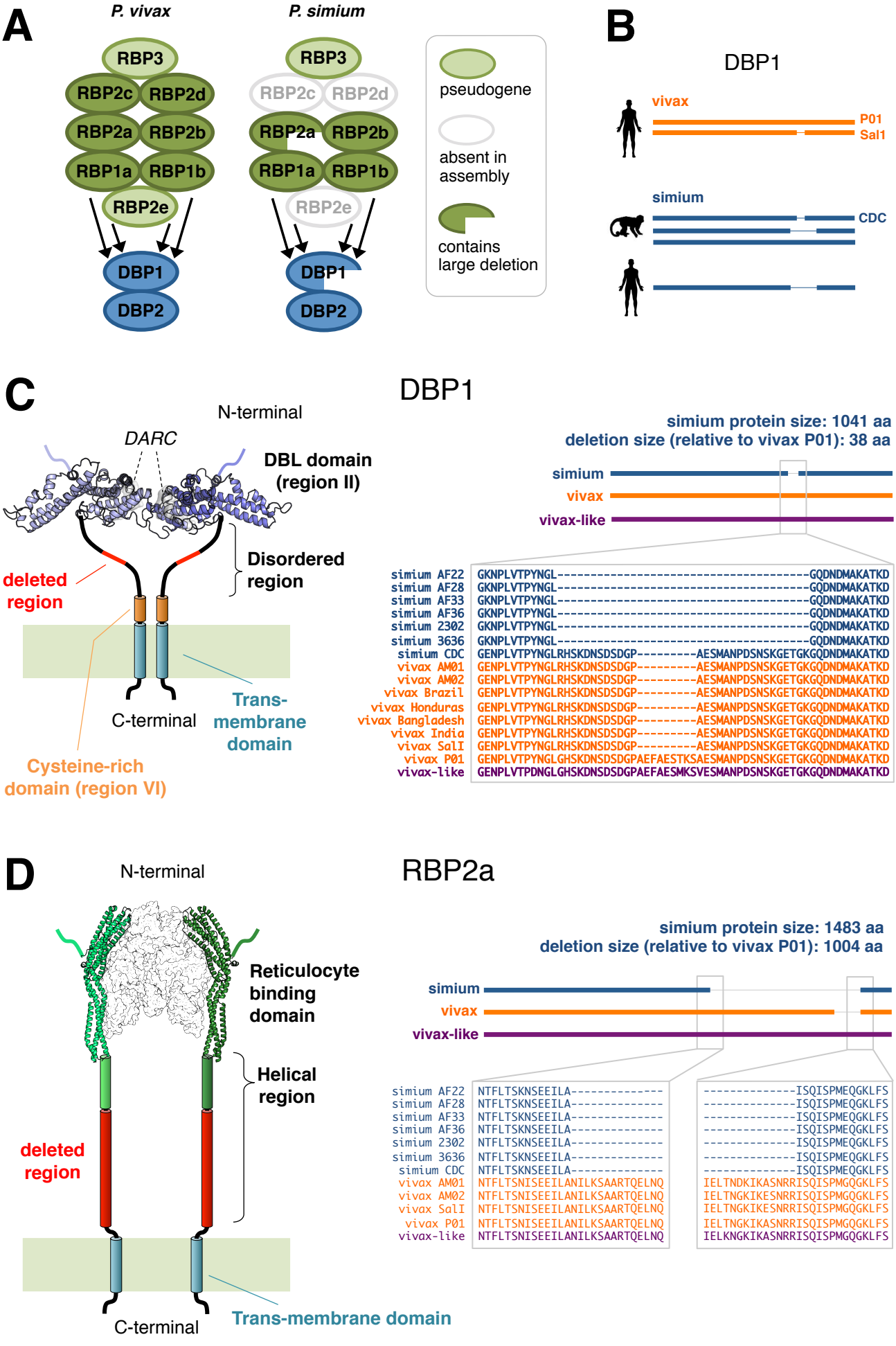



Figure 4