The genome of the zoonotic malaria parasite *Plasmodium simium* reveals adaptions to host-switching

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

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

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

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

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

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

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

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

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

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

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

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

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

13) Malaria Unit, Department of Pathology, Institute of Tropical Medicine (NEKKEN), 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

Summary

Plasmodium simium, a malaria parasite of non-human primates in the Atlantic forest region of Brazil was recently shown to cause zoonotic infection in humans in the region. Phylogenetic analyses based on the whole genome sequences of six *P. simium* isolates infecting 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 from humans carrying *P. vivax*. We provide molecular evidence that the current zoonotic infections of people have likely resulted from multiple independent host switches, each seeded from a different monkey infection. Very low levels of genetic diversity within *P. simium* genomes and the absence of *P. simium*-*P. vivax* hybrids suggest that the *P. simium* population emerged recently and has subsequently experienced a period of independent evolution in Platyrrhini monkeys. We further find that Plasmodium Interspersed Repeat (PIR) genes, Plasmodium Helical Interspersed Subtelomeric (PHIST) genes and Tryptophan-Rich Antigens (TRAg) genes in *P. siumium* are genetically divergent from *P. vivax* 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 in *P. simium*. Genotyping of *P. simium* isolates from non-human primates (NHPs) and zoonotic

human infections showed that a precise deletion of 38 amino acids in DBP1 is exclusively present in all human infecting isolates, whereas non-human primate infecting isolates were polymorphic for the deletion. We speculate that these deletions in the parasite-encoded key erythrocyte invasion ligands and the additional rapid genetic changes have facilitated zoonotic transfer to humans. Non-human primate malaria parasites can be considered a reservoir of potential infectious human parasites that must be considered in any attempt of malaria elimination. The genome of *P. simium* will thus form an important basis for future functional characterizations on the mechanisms underlying malaria zoonosis.

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 latter three species are more commonly parasitic on non-human primates and have only relatively recently been shown to infect humans [1-3].

As interventions against human malaria parasites, particularly *P. falciparum* and *P. vivax,* continue to reduce their prevalence, the increasing importance of zoonotic malaria is becoming apparent. In countries currently moving towards the elimination of malaria, the presence of populations of potentially zoonotic parasites in non-human primates constitutes a significant obstacle.

The propensity of malaria parasites to switch hosts and the consequences of this for human health are underlined by the fact that both *P. vivax* and *P. falciparum* first arose as human pathogens as the result of host switches from great apes in Africa [4-6]. As contact between humans and the mosquitoes that feed on non-human primates increases due to habitat destruction and human encroachment into non-human primate habitats [7], there is increasing danger of zoonotic malaria transmission leading to the emergence of novel human malaria pathogens. Understanding how malaria parasites adapt to new hosts and new transmission environments allows assessment of the risks posed by novel zoonotic malaria outbreaks.

The clinical epidemiology of zoonotic malaria varies according to the parasite species involved and the demographics of the human-host population infected. Severe and lethal outcomes have been reported in people infected with *P. knowlesi* in Malaysia [8], whilst infection with *P. cynomolgi* in the same region appears to cause moderate/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], and this may be due to the relative virulence of the parasite strains circulating there and/or differences in the susceptibility of the local human populations. As the parasites of non-human primates have co-evolved with and adapted to their monkey hosts, it is impossible to predict their potential pathogenesis in zoonotic human infections. The virulence of *P. falciparum*, for example, has been attributed to its relatively recent emergence as a human pathogen [11], which appears to have occurred following a single host transfer from a gorilla in Africa [5].

Eighty nine percent of the malaria infections in Brazil are caused by *P. vivax,* with over 99% of these cases occurring in the Amazonian region. This region accounts for almost 60% of the area of Brazil, and is home to 13% of the population (https://www.ibge.gov.br/). Of the 0.4% of cases registered outside the Amazon, around 90% occur in the Atlantic Forest, a region of tropical forest that extends along the Atlantic coast of Brazil, and are caused by an apparently mild, vivax-like malaria parasite transmitted by *Anopheles (Kerteszia) cruzii*, a mosquito species that breeds in the leaf axils of bromeliad plants [12].

Following a malaria outbreak in the Atlantic Forest of Rio de Janeiro in 2015/2016, it was shown that these infections were caused by the non-human primate malaria parasite *P. simium* [1]. DNA samples collected both from humans and non-human primates (NHPs) in the same region shared identical mitochondrial genome sequences, distinct from *P. vivax* isolates from anywhere in the world and identical to that of a *P. simium* parasite isolated from a monkey in the same region in 1966, and to all isolates of *P. simium* recovered from NHPs since [13,14].

It was previously thought that *Plasmodium vivax* became a parasite of humans following a host switch from macaques in Southeast Asia, due to its close phylogenetic relationship with a clade of parasites infecting monkeys in this region and due to the high genetic diversity among *P. vivax* isolates from Southeast Asia [15]. We now know, however, that it became a human parasite following a host switch from great apes in Africa [6]. It is likely that it was introduced to the Americas by European colonisers following Columbus' journey to the New

World in 1492. Indeed, present-day *P. vivax* in South America is closely related to a strain of the parasite present, historically, in Spain [16]. The genetic diversity of extant *P. vivax* in the Americas suggests multiple post-Columbian colonising events associated with the passage of infected people from various regions throughout the world [17]. There is some evidence to suggest that *P. vivax* parasites may also have been introduced to South America in pre-Columbian times, and may have contributed to the extensive genetic diversity of the parasite on this continent [17].

Plasmodium simium, a parasite of various species of Platyrrhini monkeys whose range is restricted to the Atlantic Forest of south and southeast Brazil [18], is genetically and morphologically similar to *P. vivax* [1,19-22]. Based on this similarity, it appears likely that *P. simium* originated as a parasite of monkeys in Brazil following a host switch from humans carrying *P. vivax*. 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 that has occurred between *P. vivax* and *P. simium*, and whether adaptation to monkeys has led to the evolution of a parasite with clinical relevance to human health that differs from that of *P. vivax*.

It is unclear whether the current outbreak of *P. simium* in the human population of Rio de Janeiro was the result of a single transfer of the parasite from a monkey to a human and its subsequent transfers between people, or whether multiple independent host switches have occurred, each seeded from a different monkey infection. Furthermore, the degree and nature of adaptation to a non-human primate host and a sylvatic transmission cycle that has occurred in *P. simium* following its anthroponotic origin is of relevance to the understanding of how malaria parasites adapt to new hosts. It is also of interest to determine whether the current, human-infecting *P. simium* parasites have recently undergone changes at the genomic level that have allowed them to infect people in this region, as it has previously been suggested that *P. simium* has historically lacked the ability to infect man [23].

In order to resolve these questions, and so to better understand the epidemiology and natural history of this emerging zoonotic parasite, we analysed the whole genome sequences of *P. simium* parasites isolated from both humans and non-human primates in the Atlantic Forest region of Rio de Janeiro.

Results

Genome assembly and phylogeny

From a single *P. simium* sample collected from Rio de Janeiro state in 2016 [1] short read sequences were obtained and assembled into a draft genome (see Supplementary Materials). The assembled genome consists of 2,192 scaffolds over 1kb with a combined size of 29 Mb (Table S1). Two scaffolds corresponding to the apicoplast and mitochondrial organelles are also identified (Figure S1). Gene content analysis showed an annotation completeness comparable to previously published *Plasmodium* assemblies (Figure S35). A phylogenetic tree constructed from 3,181 of 1:1 orthologs of the annotated *P. simium* protein-coding genes with *Plasmodium 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 S2).

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 *P. simium* samples, two monkey *P. simium* samples, two *P. vivax* samples from Brazilian Amazon, and a range of *P. vivax* strains representing a global distribution retrieved from the literature [25]. Including only SNPs with a minimum depth of five reads, a total of 232,780 SNPs were initially called across 79 samples. Sixteen samples were subsequently removed from further analysis primarily due to low coverage resulting in a total of 63 samples (Table S2, Table S3). Few SNP loci are covered across all samples, and to enable diversity analysis, we restrict all further analysis to the 124,968 SNPs for which data is available from at least 55 samples (Figure S3).

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 S4). The latter observation suggesting that both human and monkey *P. simium* samples form a single population that is genetically differentiated from other American *P. vivax* populations. A similar pattern is observed when performing a multidimensional scaling analysis of the SNP data (Figure S5). 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, and the Asian and American samples were clearly separated (Figure 1A, a tree with sample IDs is available in Figure S6). *P. simium* samples clustered as a monophyletic group with Mexican vivax samples (Figure 1A), consistent with a recent American origin for *P. simium.*

To examine whether the *P. simium* isolates we obtained were part of a continuous population with local *P. vivax*, we examined population ancestry with the ADMIXTURE program [26] (Figure S7). This analysis is consistent with the PCA and MDS analysis (Figure S4 & Figure S5) and the phylogenetic analysis of segregating SNPs (Figure 1), showing that *P. simium* forms a genetically distinct population of *P. vivax*. The absence of *P. simium*-*P. vivax* hybrids (introgression events) suggests that *P. simium* has undergone a period of independent evolution in Platyrrhini monkeys.

P. simium genetic differentiation from _P. vivax_ is enhanced in host-parasite interacting genes

To characterise the *P. simium* population further, we estimated the nucleotide diversity in *P. simium* and *P. vivax* samples (see Materials and Methods). *P. simium* diversity (genome-median: 1.3×10^{-4}) is more than five times lower than the diversity observed when comparing all *P. vivax* samples (genome-median: 7.5×10^{-4}) (Figure 2). Diversity within coding sequences in *P. vivax* is consistent with previous reports [6]. The median nucleotide diversity between *P. simium* and *P. vivax* genomes of 8.4×10^{-4} and the low diversity within *P. simium* suggest that the strains we examined are part of a relatively recent or isolated population.

We then examined the population differentiation over the entire genome using $F_{ST,}$ a measure of the proportion of ancestry private to a population (F_{ST}=0 for completely intermixed populations, F_{ST}=1 for populations with completely independent ancestry). Although our analysis contains very few samples, F_{ST} estimates can be very accurate if multiple genomic sites are used [27]. Consistent with phylogenetic and admixture analysis, we observed a high level of differentiation between human *P. simium* and American *P. vivax* (F_{ST}=0.46). For comparison, the differentiation between vivax from America and vivax from Myanmar and Thailand (henceforth referred to as 'Asian vivax') is less than half of this (F_{ST}=0.22). To examine whether there were any signals of adaptive change in *P. simium* that may have

occurred during its adaptation in monkeys upon anthroponotic transfer, we calculated the fixation index for 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 4,341 *P. vivax* genes with at least one SNP in our data set, we examined the top-25% of the 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* orthologs instead – when available – gave similar results (not shown). We next tested if any of the gene families (Figure S8, Figure S9, 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; $p=3.5\times10^{-3}$, PHIST; $p=4.1\times10^{-4}$, MSP7; $p=0.034$, TRAg; $p=2.5\times10^{-3}$).

As these gene families are involved in parasite-host interactions, the observation of 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 vivax and a selection of Asian vivax isolates (Myanmar and Thailand samples only). Consistent with the phylogenetic analysis (Figure 1A) gene F_{ST} was slightly higher overall between simium and American vivax than between American and Asian vivax samples (Figure 3A). However, none of the gene families were overrepresented among genes with high F_{ST} (top-25%) between American and Asian vivax. To further examine if the elevated F_{ST} measures found for PIR, PHIST, MSP7, and TRAg genes are exclusive to the comparison between simium and American 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=0.12). Although the *P. simium* and the *P. vivax* P01 both genomes encode a high number of the gene family members, our analysis is restricted to the *P. vivax* genes for 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

across at least 80% of their gene length (Figure S10). Further, an even smaller number of these genes have detectable SNPs between simium and American vivax samples and are included in the analysis (numbers shown below Figure 3B).

To test if the sequence redundancy among gene family loci could result in spurious cross-mapping of short sequence reads we specifically tested the quality of SNPs in gene families, and SNPs residing in gene families showed no signs of decreased calling, mapping, or base qualities compared to other SNPs (Figure S11).

A phylogenetic analysis of PIR, PHIST and TRAg proteins harbouring genomic SNPs revealed no apparent association between certain protein phylogenetic sub-groups and high F_{ST} ratios (Figure S12-S14), consistent with a subtle signature of polygenic adaptation in these gene families.

When testing all exported genes and genes involved in invasion and exported genes (Table S8), the observed F_{ST} ratios were not significantly different from the background (Mann-Whitney U, p=0.5473). Hence, the differences in F_{ST} observed for PIR, PHIST and TRAg genes are not a general phenomenon amongst the genes known to be involved in interactions with the host and red cell invasion.

The observed skew towards higher F_{ST} values when comparing simium and American vivax (Figure 3A) could be a result of an inherent diversity between different American vivax populations potentially stemming from multiple introductions of *P. vivax* to the American continent [17]. To test if such founder effects and subsequent population bottlenecks could explain the observations, we repeated the F_{ST} analysis using only Mexican vivax samples as American representatives. Four Mexican samples (SRS693273, SRS694229, SRS694244, SRS694267) were used. These clustered close together in both the SNP phylogeny (Figure 1) and in the PCA and MDS plots (Figure S4 & Figure S5), and are assumed to share a recent evolutionary history. This analysis revealed the same pattern of elevated F_{ST} values between simium and Mexican vivax, and PIR genes did again display significantly higher F_{ST} ratios (Figure S15). Although PHIST and TRAg genes also showed higher F_{ST} ratios, these were no longer significant (Figure S15). We therefore conclude that the observed higher F_{ST} values between simium and American vivax PIR genes are not solely a result of diversity within American vivax populations, but rather appear specific to comparisons with *P. simium*.

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, respectively, 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 overall display markedly higher F_{ST} values between simium and vivax suggest 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 invasome components

In invading *P. vivax* merozoites, binding to host red blood cells is mediated by two gene families: Duffy Binding Proteins (DBPs) bind the Duffy Antigen Receptor for Chemokines (DARC) [32,33], which is present on both host normocytes and reticulocytes, whereas Reticulocyte Binding Proteins (RBPs) preferentially bind host reticulocytes [34-36]. Recently, the reported protein structure of *P. vivax* RBP2b revealed the evolutionary conservation of residues involved in the invasion complex formation [36]. Two DBPs, DBP1 and DBP2, are present in *P. vivax* P01 (Table S9). 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 S9).

The *P. vivax* DBP and RBP were used to search the *P. simium* proteins, resulting in the detection of the two DBP proteins and RBP1a, RBP1b, RBP2a, RBP2b, and RBP3 and failure to detect RBP2c and RBP2d (Figure 4; Table S9; Figure S16; Figure S17) across all sequenced *P. simium* samples. As in other *P. vivax* genomes, the *P. simium* RBP3 is a pseudogene [38], indicating that the pseudogenization event happened prior to the split between *P. vivax* and *P. simium*.

To determine whether the apparent absences of individual RBP genes in *P. simium* were due to incomplete genome assembly, we examined the coverage of *P. simium* reads mapped onto *P. vivax* RBP gene loci. As expected, no *P. simium* coverage was observed at the RBP2c, RBP2d, and RBP2e genes in *P. simium* samples, including the previously published CDC strain deposited in GenBank (accession ACB42432) [39] (Figure S18).

Coverage of mapped reads across invasome gene loci revealed no apparent elevated coverage in genes compared to their flanking genomic regions, which would otherwise be expected if the *P. simium* genome contained multiple (duplicated) copies of non-assembled invasion genes (Figure S19). Similarly, analysis of *P. simium* read mapping data using the DELLY software [40] showed no large genomic duplications and deletions events occurring at loci harbouring invasion genes (Table S10) although numerous short indels were detected within protein-coding genes (Table S11).

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

The simium assembly revealed that the invasion gene DBP1 contains a large deletion within its coding sequence (Figure 4) (a full alignment is provided in Figure S20). Intriguingly, the previously published *P. simium* CDC strain (originally isolated in 1966) DBP1 does not contain the deletion ('simium CDC' in Figure 4B). A haplotype network confirms that this previously published DBP1 gene is indeed a *P. simium* sequence (Figure S22), and the SNP analyses consistently assign the CDC strain to the simium cluster (Figure 1, Figure S4, and Figure S5). 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). This deletion is also present in most *P. vivax* isolates (Figure S23). Additional deletion patterns exist among isolates, and in a few cases multiple versions are detected within samples (Figure S23).

The presence of repetitive sequences within the DBP1 gene could potentially result in aberrant assembly across the DBP1 locus, which could appear as an apparent deletion in subsequent bioinformatic analysis. We tested this possibility and the DBP1 gene does not harbour any noticeable degree of repetitiveness (Figure S24). Several read mapping analyses confirmed that the *P. simium*-specific 115 bp deletion was not an assembly artefact (Figure S25-S27).

We next designed primers for PCR amplification of a genomic segment across the deleted region in the *P. simium* DBP1 gene and tested the occurrence of these deletion events in a range of *P. vivax* and *P. simium* field samples from Brazil. All *P. vivax* samples tested by PCR produced bands consistent with absence of the deletion whereas all samples from human-infecting *P. simium* produced bands consistent with the presence of the precise 115 bp deletion (Figure S28, top & middle). Interestingly, non-human primate (NHP)-infecting *P. simium* isolates were a mix of samples with and without deletions (Figure S28, bottom). 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 DBP1 allele required for zoonotic transfer to humans.

A large, additional deletion was observed in the *P. simium* RBP2a gene, the presence of which was also supported by read mapping and PCR analysis (Figure 3, Figure S29-S32).

Potential structural implications of the deletion in DBP1 and RBP2a

We next investigated if the observed deletions render DBP1 and RBP2a dysfunctional. DBP1 contains a large extracellular region, which includes the N-terminal DBL region which is mediating 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 DARC binding site (Figure S33). Based on our models, these substitutions are unlikely to negatively affect the association with DARC, supporting that the DBL domains of both *P. simium* would be capable of binding to human DARC. Hence, the human-infecting *P. simium* sequence encodes for a protein that retains the capacity to bind to human DARC, but would have the interacting domain positioned closer to the membrane than in the monkey-infecting CDC strain.

The deletion we detected in human-infecting RBP2a was more severe, resulting in the loss of 1003 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 does neither affect the transmembrane region, nor the receptor-binding domain, our analysis supports that the resulting truncated RBP2a protein can still associate with the human receptor, but that the binding event would occur closer to the plasmodium 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. The genome content confirmed the close phylogenetic relationship between *P. simium* and *P. vivax*, and further analyses on single nucleotide divergences 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 vivax populations, their genetic differentiation appears to be associated with host-switching between human and monkey.

Two proteins involved in host invasion, DBP1 and RBP2a, were found to harbour extensive deletions in *P. simium* compared to *P. vivax*. Interestingly, experimental analysis of *P. simium* samples revealed that isolates from human hosts all carried the DBP1 deletion, whereas isolates from non-human primates displayed both absence and presence of the deletion. This DBP1 deletion is not present in the *P. simium* isolated from a brown howler monkey in the 1960s, which was previously shown to be incapable of infecting humans [23], although some degree of laboratory adaptation of this parasite may have affected its genome. However, this deletion is also absent in *P. vivax*, so cannot in itself explain the ability of *P. simium* to infect

humans in the current outbreak. It is possible, however, that this deletion is required for *P. simium* to invade human red blood cells given the alterations that have occurred elsewhere in its invasome following adaptation to non-human primates since the split between *P. simium* and its human-infecting *P. vivax* ancestor.

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

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

Due to uncertainties regarding the number of individual genomes that were transferred during the original host switch from man 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]. However, our nuclear genome phylogeny is based on genome-wide SNPs, and so represents an "average" phylogeny across the genome. This cannot be considered to reflect a true history of parasite ancestry due to the effects of recombination, and it is possible that trees produced from individual genes might reveal different phylogenetic relationships.

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

The biological definition of a species is a group of organisms that can exchange genetic material and produce viable offspring. We have no way of knowing whether this is the case for *P. vivax* and *P. simium*, and genetic crossing experiments would be required to resolve this question. Our phylogenetic analysis, however, clearly shows *P. simium* forming a clade on its own within the broader diversity of *P. vivax*, and that 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* for long enough for significant genetic differentiation to occur (F_{ST} = 0.46), with some invasome 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. In this sense non-human primate malaria parasites can be considered a reservoir of potential infectious human parasites that must be considered in any attempt of malaria eradication. Little is known about the genetic basis for zoonosis, yet the presented genome sequence of *P. simium* suggests a deletion within the DBP1 gene as a possible facilitator of zoonotic transfer. The genome of *P. simium* will thus form an important basis for future functional characterizations on the mechanisms underlying malaria zoonosis.

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 described[1]. The genomic DNA for each sample was quantitated using the Qubit® 2.0 Fluorometer and was used for library preparation. The DNA for intact samples was sheared using a Covaris E220 DNA sonicator to fragments of 500bp. The DNA libraries

for intact samples were made using the TruSeq Nano DNA Library Prep kit (Illumina), whereas the DNA libraries for degraded samples were made using Ovation Ultralow Library System V2 kit (Nugen), according to the manufacturers' instructions. The amplified libraries were stored in -20 °C. The pooled libraries were sequenced in an Illumina HiSeq4000 instrument (2 x 150 bp PE reads) (Illumina). A PhiX control library was applied to the sequencing run as a base balanced sequence for the calibration of the instrument so that each base type is captured during the entire run. Raw sequence reads were submitted to FastQC v.0.11.5 and the quality score of the sequences generated was determined. Samples AF22, AF26, AF36 were additionally sequenced and scaffolded by PacBio RS II platform (Pacific Biosciences, California, US) using a SMRT library. Genomic DNA from the *P. vivax* samples was extracted from filter paper as previously described [47].

Illumina reads preparation and mapping

Fastqc v 0.11.6 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) was used to evaluate the quality of Illumina reads. Illumina adapters were removed, and reads were trimmed using the trimmomatic v0.33 [48] software with the following conditions:

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

To exclude human reads from our analysis, trimmed reads were mapped against the human reference genome (v. hg38) and the *Plasmodium vivax* strain P01 reference genome (v. 36) from PlasmoDB (www.plasmodb.org) with bowtie2 (v 2.3.3.1) [49]. Reads mapping against the human genome were removed from further analysis.

Genome Assembly

P. simium sample AF22 was selected for genome assembly based on read quality and coverage. After removal of human contaminants, Illumina reads were assembled into contigs using the Spades (v 3.70) assembler [50]. Contigs assembled into scaffolds running SSPACE (v 3.0) [51] for 15 rounds and gaps filled with Gapfiller (v 1.10) [52]. Scaffolds were subsequently corrected with Illumina reads using the Pilon (v 1.22) software [53]. Blobtools (v 1.0) (DOI: 10.5281/zenodo.845347) [54] was used to remove any residual contaminant scaffolds. Genome size and GC content is in line with that of *P. vivax* species (Table S1). Genomic scaffolds representing the mitochondrial and apicoplast genome were identified through blastn searches against the corresponding *P. falciparum* and *P. vivax* sequences (Figure S1). The *P. simium* mitochondrial genome was aligned against a range of previously published *P. vivax* and *P. simium* mitochondrial genomes [55,56]. A gap-filled region in the alignment where 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 S34).

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 lower 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 S35).

PlasmoDB Genome References and Annotations

Genome fasta files, as well as annotated protein and CDS files were obtained from PlasmoDB 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, 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 [63]. Approximately 88% of the predicted genes in *P. simium* have orthologs in the *P. vivax* P01 (Figure S36).

Indels in genes

Structural variations were detected using DELLY [40] (v 0.7.9). Coordinates of structural rearrangements their nearest genes are listed in Table S10. Shorter indels were detected from

608 soft-clipping information in read mapping (using the '-i' option in DELLY)(Table S11). Indels
609 in exons were further compared to indels present in the *P. simium* AF22 genome assembly,
610 suggesting a high false discovery rate of DELLY indels compared to assembly indels (Figure
611 S37).
612
613 **Protein phylogeny**
614 Protein sequences were aligned using mafft (v 7.222) [64] and alignments were subsequently
615 trimmed with trimAl (v 1.2rev59) [65] using the heuristic 'automated1' method to select the best
616 trimming procedure. Trimmed alignments were concatenated and a phylogenetic tree was
617 constructed using RAxML (v 8.2.3) [66] with the PROTGAMMALG model.
618
619 **SNP calling and analysis**
620 Short sequence reads from 15 simium samples (13 human and 2 monkey) and two vivax
621 samples, all from this study (Table S2), were aligned against a combined human (hg38) and
622 *P. vivax* (strain P01, version 39) genome using NextGenMap (v0.5.5) [67]. This was similarly
623 done for 30 previously published *P. vivax* strains [25] and the Sal1 reference. These data sets
624 were downloaded from ENA (https://www.ebi.ac.uk/ena) (Table S3).
625 Duplicate reads were removed using samtools (v 1.9) [68] and the filtered reads were realigned
626 using IndelRealigner from the GATK package (v 4.0.11) [69]. SNPs were called independently
627 with GATK HaplotypeCaller and freebayes (v 1.2.0) [70], keeping only SNPs with a QUAL
628 score above 30. The final SNP set were determined from the inter-section between GATK and
629 freebayes. Allele frequencies and mean coverage across SNP sites are shown in Figure S38.
630 PCA plot was constructed using plink (v 1.90) [71], and admixture analysis was done with
631 Admixture (v 1.3.0) [26]. F_{ST} values were estimated from nucleotide data with the PopGenome
632 R package [72,73] using the Weir & Cockerham method [74]. Non-synonymous and synonymous
633 SNPs were identified using snpeff [75].
634
635 **SNP phylogeny**
636 Alleles from SNP positions with data in 55 samples were retrieved, concatenated, and aligned
637 using mafft [64]. Tree was produced by PhyML [76,77] with the GTR substitution model selected
638 by SMS [78]. Branch support was evaluated with the Bayesian-like transformation of
639 approximate likelihood ratio test, aBayes [79]. Phylogenetic network was made in SplitsTree [80]
640 using the NeighborNet network [81].
641

Nucleotide diversity

Conventional tools calculating nucleotide diversity directly from the variant call files assumes that samples are aligned across the entire reference sequence. But as read coverage across the reference genome was highly uneven between samples (Figure S38), adjustment for this was required. Coverage across the reference genome was thus calculated for each sample using samtools mpileup (v 1.9) [68]. For each comparison between two samples, the nucleotide divergence was calculated as number of detected bi-allelic SNPs per nucleotide with read coverage of at least 5X in both samples.

Gene sequence deletions

Exploratory Neighbor-Joining phylogenies produced with CLUSTALW [82,83] and visualized with FigTree (https://github.com/rambaut/figtree/) after alignment with mafft [82]. Pacbio reads were aligned using Blasr (v 5.3.2) [84], short Illumina reads using NextGenMap (v0.5.5) [67]. Dotplots done with FlexiDot (v1.05) [85].

Gene families and groups

Exported gene sets were compiled from the literature [86-88]. Invasion genes were retrieved from [89]. Gene families were assessed in seven Plasmodium genomes (*P. simium*, *P. vivax* SalI, *P. vivax* P01, *P. vivax-like* Pvl01, *P. cynomolgi* M, *P. cynomolgi* B, and *P. knowlesi* H) using the following pipeline: For all genomes annotated genes were collected for each gene families. These 'seed' sequences were used to search all proteins from all genomes using BLASTP and best hits for all proteins were recorded. For each gene family 'seed' sequences were then aligned with mafft [64], trimmed with trimAl [65], and HMM models were then built using HMMer (http://hmmer.org/). For PIR/VIR and PHIST genes, models were built for each genome independently, for all other gene families a single model was built from all genomes. These models were then used to search all proteins in all genomes. All proteins with best BLASTP hit to a 'seed' sequence from a given genome were sorted according to their bit score. The lowest 5% of hits were discarded and remaining proteins with best hits to a 'seed' sequence were assigned one 'significant' hit. As all proteins were searched against 'seeds' from the six annotated genomes (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 consists of previously annotated genes and un-annotated genes with at least two 'significant' hits (either BLASTP or HMM).

676

677

PCR amplification of DBP1 and RBP2a genes

678 PCR primers were initially designed from alignments between *P. vivax* and *P. simium*
679 sequences and subsequent tested using Primer-BLAST [90] and PlasmoDB
680 (www.plasmodb.org). For DBP1, the reaction was performed in 10 μL volumes containing
681 0.5 μM of each oligonucleotide primer, 1 μL DNA and 5 μL of Master Mix 2x (Promega)
682 (0.3 units of Taq Polymerase, 200 μM each deoxyribonucleotide triphosphates and 1.5 mM
683 MgCl2). Samples were run with the following settings: 2 minutes of activation at 95°C,
684 followed by 35 cycles with 30 seconds denaturation at 95°C, 30 seconds annealing at 57°C
685 (ΔT=-0.2 °C from 2nd cycle) and 1 minute extension at 72°C, then 5 minutes final extension
686 at 72°C and hold in 4°C. For RBP2a PCR, the reaction was performed in 10 μL volumes
687 containing 0.5 μM of each oligonucleotide primer, 1 μL DNA, 0.1 μL PlatinumTaq DNA
688 Polymerase High Fidelity (Invitrogen, 5U/ μL), 0.2 mM each deoxyribonucleotide
689 triphosphates and 2 mM $MgSO_4$. The PCR assays were performed with the following cycling
690 parameters: an initial denaturation at 94°C for 1.5 min followed by 40 cycles of denaturation
691 at 94°C for 15 sec, annealing at 65°C for 30 sec (ΔT=-0.2 °C from 2nd cycle) and extension at
692 68°C for 3.5 min. The temperature was then reduced to 4 °C until the samples were taken. All
693 Genotyping assays were performed in the thermocycler Veriti 96 wells, Applied Biosystems,
694 and the amplified fragments were visualized by electrophoresis on agarose gels (2% for DBP1
695 and 1% for RBP2a) in 1x TAE buffer (40 mM Tris-acetate, 1 mM EDTA) with 5 μg/ mL
696 ethidium bromide (Invitrogen) in a horizontal system (Bio-Rad) at 100 V for 30 min. Gels
697 were examined with a UV transilluminator (UVP - Bio-Doc System).

698

699

700 To prevent cross-contamination, the DNA extraction and mix preparation were performed in
701 "parasite DNA-free rooms" distinct from each other. Furthermore, each of these separate
702 areas has different sets of pipettes and all procedures were performed using plugged pipette
703 tips. DNA extraction was performed twice on different days. Positive (DNA extracted from
704 blood from patients with known *P. vivax* infection) and negative (no DNA and DNA extracted
705 from individuals who have never traveled to malaria-endemic areas) controls were used in
706 each round of amplification. DNA extracted from blood of a patient with high parasitemia for
707 *P. vivax* and DNA of *P. simium* of a non-human primate with an acute infection and
708 parasitemia confirmed by optical microscopy served as positive controls in the PCR assays.
Primer sequences are provided in Figure S28 and S32.

Structural modelling of DBP1 and RBP2a genes

RaptorX [91] 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 [92], using the crystallographic structure of the DBL domain from *Plasmodium vivax* DBP bound to the ectodomain of the human DARC receptor (PDB ID 4nuv), with an identity of 98%, 96% and 96% for *P. vivax*, *P. simium* AF22 and *P. simium* CDC, respectively. QMEAN values were -2.27, -2.04 and -2.03, respectively. The homology model for the reticulocyte binding protein 2 (RBP2a) of *P. vivax* strain P01 was produced based on the cryoEM structure of the complex between the *P. vivax* RBP2b and the human transferrin receptor TfR1 (PDB ID 6d05)[36], with an identity of 31% and QMEAN value of -2.46. The visualization and structural analysis of the produced models was done with PyMOL (https://pymol.org/2/).

Data availability

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

Acknowledgements

We thank Prof. Xin-zhuan Su at the National Institute of Allergy and Infectious Diseases, NIH, for invaluable help in obtaining the parasite gDNA from the BEI Resources; Sidnei Silva and Graziela Zanini, for assistance on the parasitological diagnosis of the human samples; Aline Lavigne and Larissa Gomes for undertaking the PCR for *P vivax*; Alcides Pissinatti and Silvia Bahadian Moreira for the facilities provided at the Primate Centre of Rio de Janeiro; Orzinete Rodrigues Soares for non-human primates' blood slides'; Marcelo Quintela, Waldemir Paixão Vargas, Carlos Alberto C. da Silva, Alexandre B. de Souza, Vicente Klonowski, Romenique L. Araújo, Luis R. Nogueira, Fernando Barreto, Ana L. Quijada, Luiz P.P. Silva, Gelson Medeiros, Adilson B. Ramos, Marcilene B. Ramos, Carlos A.A. Júnior, Paulo G. Barbosa, Sérgio F. Fragoso, Adilson R. Silva, Cecília Cronemberger, Marcelo Rheingantz, Leonardo Nascimento and João Marins for the field support; *Grupo Técnico de Vigilância de Arboviroses* (GT-Arbo – Brazilian Ministry of Health) for field and

material supports; and Cassio Leonel Peterka from The Brazilian Ministry of Health for malaria epidemiological data. The following reagent was obtained through BEI Resources, NIAID, NIH: *Plasmodium simium*, Strain Howler, MRA-353, contributed by William E. Collins. The work was supported by the King Abdullah University of Science and Technology (KAUST) through the baseline fund BRF1020/01/01 to AP and BAS/1/1056-01-01 to STA, and the Award No. URF/1/1976-25 from the Office of Sponsored Research (OSR). The field work in the Atlantic Forest and laboratory analysis in Brazil received financial support from the Secretary for Health Surveillance of the Ministry of Health through the Global Fund (agreement IOC-005-Fio-13), *Programa Nacional de Excelência (PRONEX)* and contract 407873/2018-0 of the *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)*, the *Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig* CBB-APQ-02620-15*)* and the *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (Faperj)*, Brazil. CNPq supports CFAB, CTDR, MFFC, PB and RLO, with a research productivity fellowship. CTDR (CNE: E-26/202.921/2018), MFFC, PB and RLO are also supported by Faperj as *Cientistas do nosso estado*. AdP-C was supported by a postdoctoral fellowship from the Faperj and DAMA by a fellowship from the CGZV-SVS (Brazilian Ministry of Health) TED 49/2018 grant. SF was supported by a Wellcome Seed Award in Science to DCJ (208965/Z/17/Z).

Author contributions

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

References

1 Brasil, P. *et al.* Outbreak of human malaria caused by Plasmodium simium in the Atlantic Forest in Rio de Janeiro: a molecular epidemiological investigation. *Lancet Glob Health* **5**, e1038-e1046, doi:10.1016/S2214-109X(17)30333-9 (2017).

2 Cox-Singh, J. *et al.* Plasmodium knowlesi malaria in humans is widely distributed and potentially life threatening. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **46**, 165-171, doi:10.1086/524888 (2008).

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

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

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

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

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

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

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

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

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

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).

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).

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).

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

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

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

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

923 58 Leigh, J. W. & Bryant, D. POPART: full-feature software for haplotype network construction.
924 *Methods Ecol Evol* **6**, 1110-1116, doi:10.1111/2041-210x.12410 (2015).

925 59 Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management
926 tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491, doi:10.1186/1471-
927 2105-12-491 (2011).

928 60 Steinbiss, S. *et al.* Companion: a web server for annotation and analysis of parasite genomes.
929 *Nucleic acids research* **44**, W29-34, doi:10.1093/nar/gkw292 (2016).

930 61 Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene prediction
931 and phylogenomics. *Molecular biology and evolution*, doi:10.1093/molbev/msx319 (2017).

932 62 Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M.
933 BUSCO: assessing genome assembly and annotation completeness with single-copy
934 orthologs. *Bioinformatics* **31**, 3210-3212, doi:10.1093/bioinformatics/btv351 (2015).

935 63 Lechner, M. *et al.* Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC*
936 *Bioinformatics* **12**, 124, doi:10.1186/1471-2105-12-124 (2011).

937 64 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
938 improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780,
939 doi:10.1093/molbev/mst010 (2013).

940 65 Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated
941 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973,
942 doi:10.1093/bioinformatics/btp348 (2009).

943 66 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
944 phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).

945 67 Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read
946 mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790-2791,
947 doi:10.1093/bioinformatics/btt468 (2013).

948 68 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-
949 2079, doi:btp352 [pii]
950 10.1093/bioinformatics/btp352 (2009).

951 69 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing
952 next-generation DNA sequencing data. *Genome research* **20**, 1297-1303,
953 doi:10.1101/gr.107524.110 (2010).

954 70 Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing.
955 *arXiv*, 1207.3907 (2012).

956 71 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
957 datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

958 72 Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient
959 Swiss army knife for population genomic analyses in R. *Molecular biology and evolution* **31**,
960 1929-1936, doi:10.1093/molbev/msu136 (2014).

961 73 R Development Core Team. R: A language and environment for statistical computing. R
962 Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL
963 http://www.R-project.org. (2007).

964 74 Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population
965 Structure. *Evolution* **38**, 1358-1370, doi:10.1111/j.1558-5646.1984.tb05657.x (1984).

966 75 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide
967 polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2;
968 iso-3. *Fly (Austin)* **6**, 80-92, doi:10.4161/fly.19695 (2012).

969 76 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies:
970 assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321, doi:10.1093/sysbio/syq010
971 (2010).

972 77 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large
973 phylogenies by maximum likelihood. *Syst Biol* **52**, 696-704, doi:10.1080/10635150390235520
974 (2003).

975 78 Lefort, V., Longueville, J. E. & Gascuel, O. SMS: Smart Model Selection in PhyML.
976 *Molecular biology and evolution* **34**, 2422-2424, doi:10.1093/molbev/msx149 (2017).

977 79 Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C. & Gascuel, O. Survey of branch
978 support methods demonstrates accuracy, power, and robustness of fast likelihood-based
979 approximation schemes. *Syst Biol* **60**, 685-699, doi:10.1093/sysbio/syr041 (2011).

80 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).

81 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).

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

83 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).

84 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).

85 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).

86 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).

87 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).

88 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).

89 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).

90 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).

91 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).

92 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).

Figure Legends

Figure 1 - SNP phylogeny

A) Mid-point rooted maximum likelihood tree produced from 143,123 concatenated SNP positions with data from at least 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 shown below tree. *P. vivax* isolates are denoted as colored circles by their country of sample origin. A tree with specific sample IDs is available in Figure S7. B) Magnification of the *P. simium* clade (as in panel A). C) Map denoting the geographic location of *P. simium* samples.

Figure 2 - Nucleotide diversity

Box plot showing the pair-wise nucleotide diversity between human-infecting *P. simium* samples (left), *P. vivax* samples (middle), and between *P. simium* and *P. vivax* samples (right). Diversity is shown for entire genome (left-most plots, blue) and exonic regions only (right-most plots, orange). Individual values from pairwise comparsions 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; $p=8.79\times10^{-20}$, exons; $p=3.72\times10^{-20}$).

Figure 3 - F_{ST} ratios

A) Gene F_{ST} values were calculated between simium and American vivax samples (x-axis) and American and Asian vivax samples (y-axis). Each dot corresponds to a gene, and the distributions of the two FST measures are shown as bar charts above and to the right of the scatter plot. B) The ratio between F_{ST} values between i) simium and American vivax samples, and ii) American and Asian vivax samples were calculated for each gene (top). A pseudo count of one was added to all F_{ST} values. The distributions of log2-ratios are shown as violin plots (bottom) for all genes (grey),

PIR genes (red), PHIST genes (brown), TRAg genes (turquoise), invasion genes (yellow), and exported genes (dark blue). Only genes with SNP differences between the three populations are included in this analysis. F_{ST} values and ratios are provided in Table S7. P-values from Mann-Whitney U tests for differences in medians between PIR, PHIST, TRAg, invasion genes, exported genes, and all remaining genes are indicated on plot (e.g. PIR genes are compared to all non-PIR genes).

Figure 4 - Invasome deletions

A) Overview of the invasome gene groups, Reticulocyte Binding Proteins (RBPs) and Duffy Binding Proteins (DBPs) in *Plasmodium vivax* and *P. simium*. The *P. vivax* genome harbours two RBP2d genes, one of which is a pseudogene (Table S9). 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 modeled 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 S20. D) Similar to panel C) but for RBP2a. The complex between the reticulocyte binding domain and the human receptor was modeled 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 S21.

Figure 1

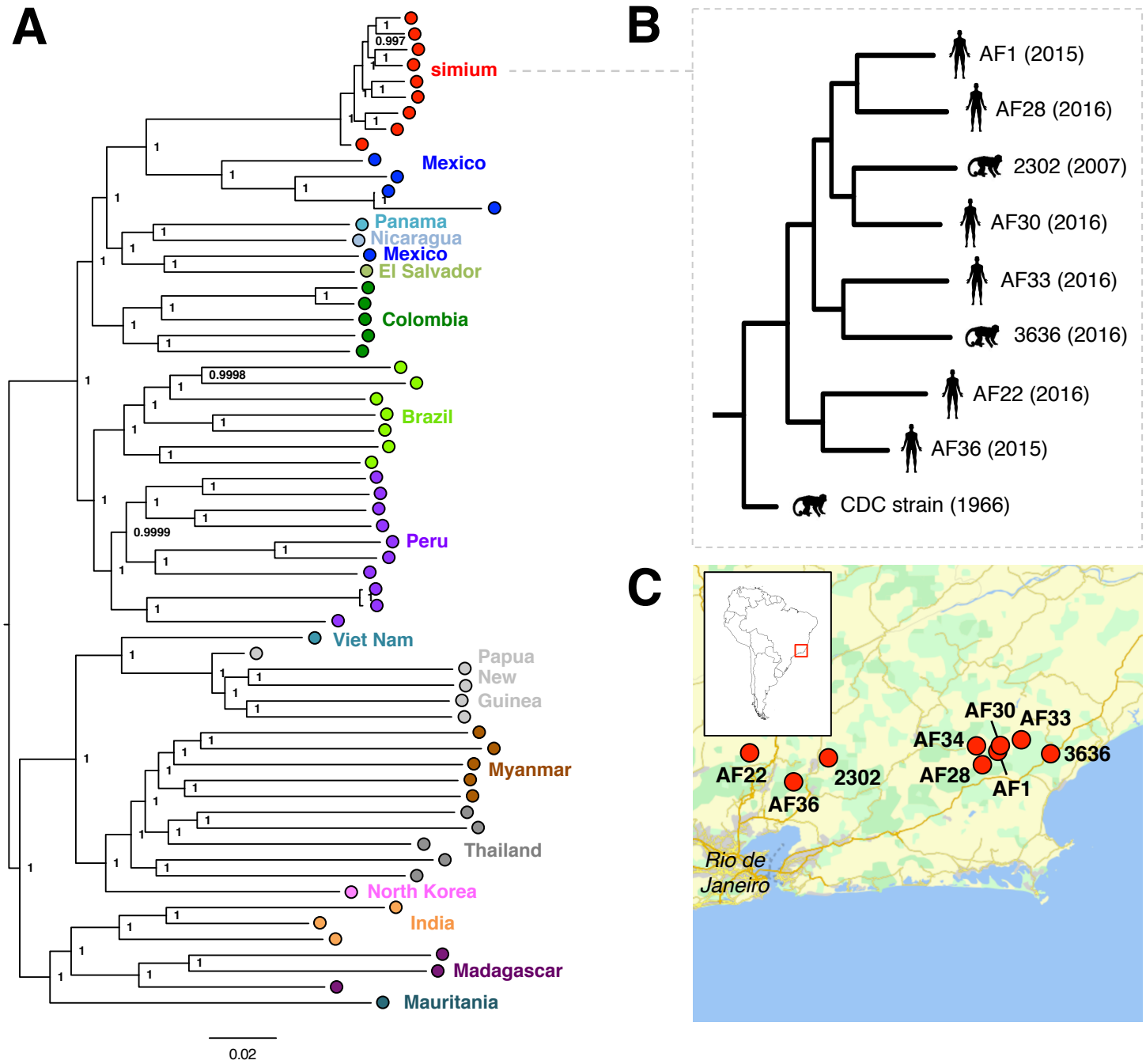

A
simium
Mexico
Panama
Nicaragua
Mexico
El Salvador
Colombia
Brazil
Peru
Viet Nam
Papua New Guinea
Myanmar
Thailand
North Korea
India
Madagascar
Mauritania
0.997
0.9998
0.9999
0.02
B
AF1 (2015)
AF28 (2016)
2302 (2007)
AF30 (2016)
AF33 (2016)
3636 (2016)
AF22 (2016)
AF36 (2015)
CDC strain (1966)
C
AF22
AF36
2302
AF30
AF33
AF34
3636
AF28
AF1
Rio de Janeiro

Figure 2

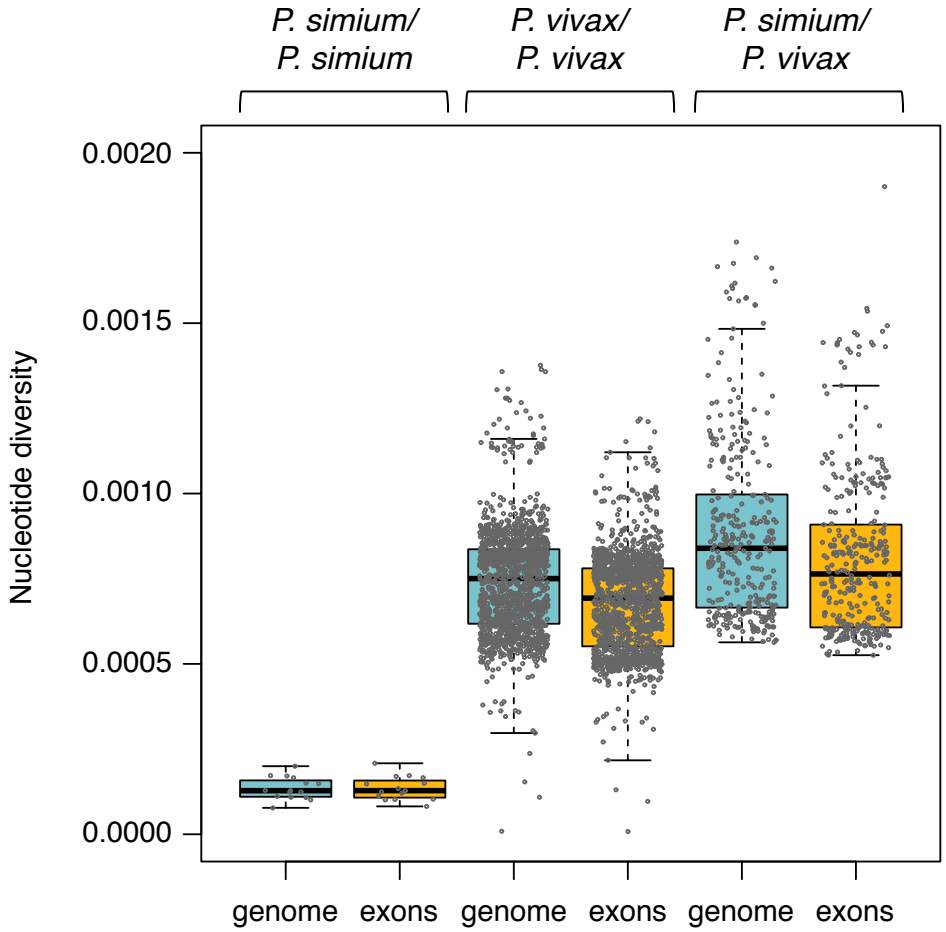

P. simium/
P. simium
P. vivax/
P. vivax
P. simium/
P. vivax
Nucleotide diversity
0.0020
0.0015
0.0010
0.0005
0.0000
genome
exons
genome
exons
genome
exons

Figure 3

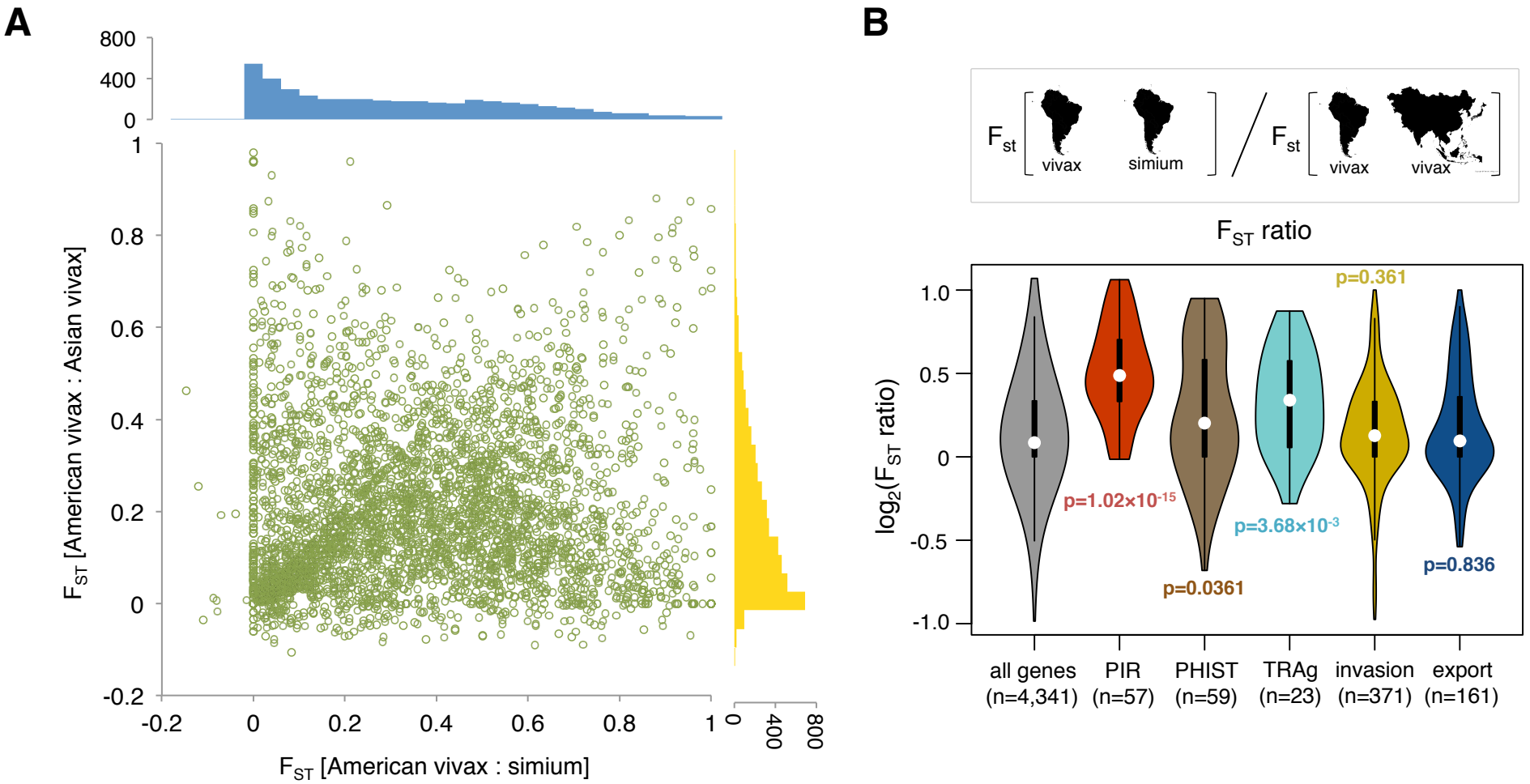
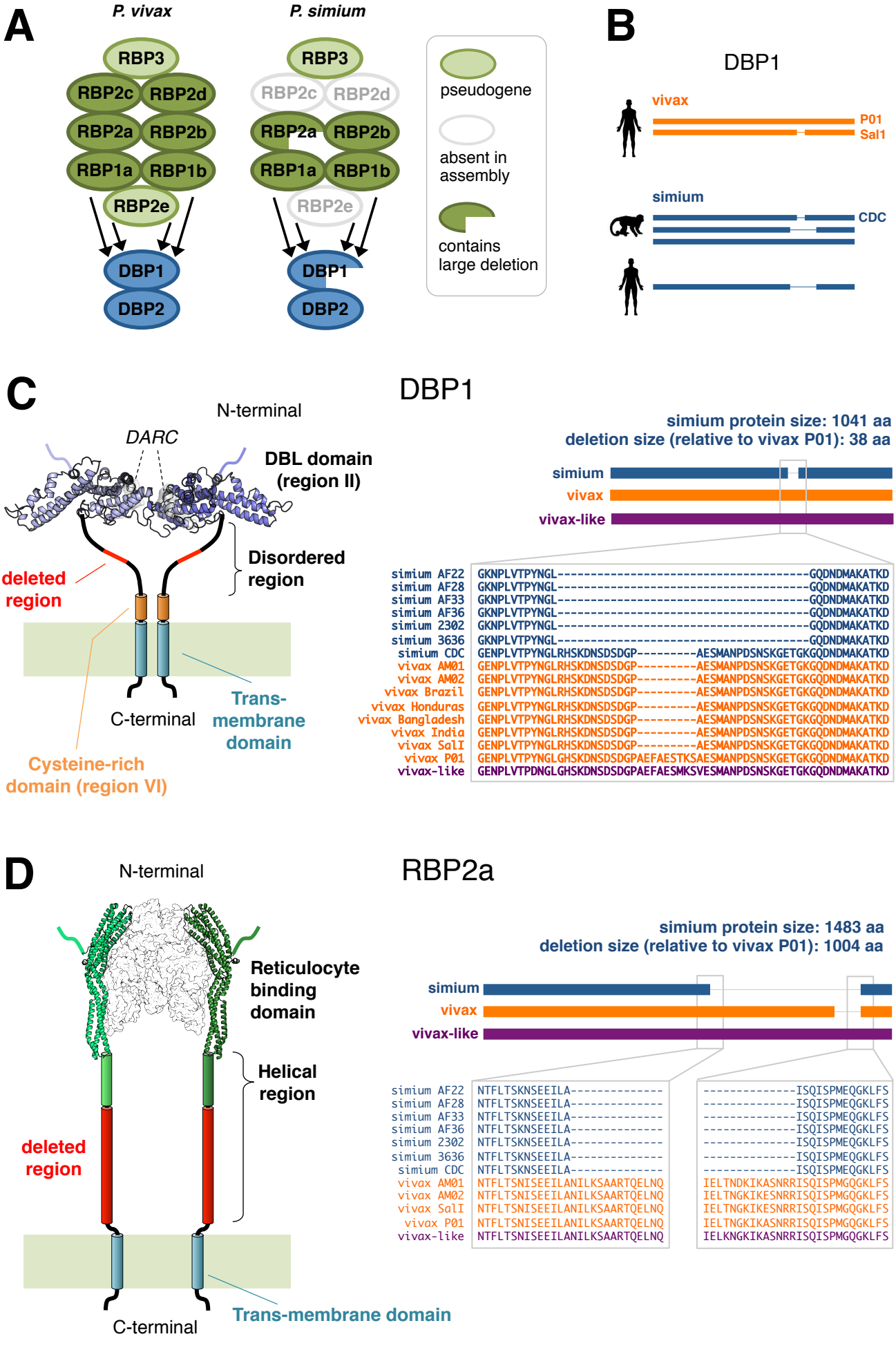

A
800
400
0
1
0.8
0.6
0.4
0.2
0
-0.2
F_{ST} [American vivax : Asian vivax]
-0.2
0
0.2
0.4
0.6
0.8
1
F_{ST} [American vivax : simium]
0
400
800
B
F_{st}
vivax
simium
F_{st}
vivax
vivax
F_{ST} ratio
$\log_2(F_{ST}$ ratio)
1.0
0.5
0
-0.5
-1.0
p=1.02×10-15
p=0.0361
p=3.68×10-3
p=0.361
p=0.836
all genes (n=4,341)
PIR (n=57)
PHIST (n=59)
TRAg (n=23)
invasion (n=371)
export (n=161)

Figure 4


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