Title: *Plasmodium vivax-like* genome sequences shed new insights into *Plasmodium vivax* biology and evolution

Authors: Aude Gilabert[1,#], Thomas D. Otto[2,3,#], Gavin G. Rutledge[2], Blaise Franzon[1], Benjamin Ollomo[4], Céline Arnathau[1], Patrick Durand[1], Nancy D. Moukodoum[4], Alain-Prince Okouga[4], Barthélémy Ngoubangoye[4], Boris Makanga[4], Larson Boundenga[4], Francisco J. Ayala[5], Christophe Paupy[1,4], François Renaud[1], Franck Prugnolle[1,4,*], Virginie Rougeron[1,4,*]

Affiliations:

[1]Laboratoire MIVEGEC (Université de Montpellier-CNRS-IRD), 34394 Montpellier, FRANCE

[2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, UK

[3] Institute of Infection, Immunity and Inflammation, University of Glasgow, College of Medical, Veterinary and Life Sciences, Sir Graeme Davies Building, 120 University Place, Glasgow G12 8TA, UK

[4]Centre International de Recherches Médicales de Franceville, B.P. 769, Franceville, GABON

[5]Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697

[#]Contributed equally to this work

[*]Corresponding authors: Rougeron Virginie; Laboratoire MIVEGEC (UM-CNRS-IRD), 34394 Montpellier, France; rougeron.virginie@gmail.com / virginie.rougeron@ird.fr; Prugnolle Franck; Laboratoire MIVEGEC (UM-CNRS-IRD), 34394 Montpellier, France; franck.prugnolle@ird.fr

Abstract

Plasmodium vivax is responsible of the majority of malaria infections outside Africa. Its closer genetic relative, *Plasmodium vivax-like*, was discovered in African great apes and suggested to have given rise to *P. vivax* in humans*.* We generated two newly *P. vivax-like* reference genomes and 9 additional *P. vivax-like* genotypes, to unravel the evolutionary history of *P. vivax*. We showed a clear separation between the two clades, a higher genetic diversity of *P. vivax-like* parasites in comparison to the *P. vivax* ones, and the potential existence of two sub-clades of *P. vivax-like*. We dated the relative split between *P. vivax* and *P. vivax-like* as three times shorter than the split between *P. ovale wallikeri* and *P. ovale curtesi* and 1.5 times longer than the split between *Plasmodium malariae.* The sequencing of the *P. vivax-like* genomes is an undeniable advance in the understanding of *P. vivax* biology, evolution and emergence in human populations.

Main text.

Plasmodium vivax, the most prevalent human malaria parasite outside Africa, is responsible for severe and incapacitating clinical symptoms in humans[1]. Traditionally, *P. vivax* has been neglected because of its lower mortality in comparison to *Plasmodium falciparum*[2,3]. Its ability to produce a dormant liver-stage form (hypnozoite), responsible of relapsing infections, makes it a challenging public health issue for malaria elimination. The recent emergence of antimalarial drug resistance[4] as well as the discovery of severe and even fatal human cases[2,5,6] has renewed the interest for this enigmatic species, including its evolutionary history and its origin in humans.

Earlier studies placed the origin of *P. vivax* in humans in Southeast Asia ("Out of Asia" hypothesis) based on its phylogenetic position in a clade of parasites infecting Asian monkeys[7]. At that time, the closest known relative of *P. vivax* was considered to be *Plasmodium cynomolgi*, an Asian monkey parasite[8]. However, this hypothesis was recently challenged with the discovery of another *Plasmodium* species, genetically closer to *P. vivax* than *P. cynomolgi,* circulating in African great apes (chimpanzees and gorillas)[9,10]. This new lineage (hereafter named *Plasmodium vivax-like*) was considered by certain authors to have given rise to *P. vivax* in humans following a transfer of parasites from African apes[10], but this "Out of Africa" hypothesis still remains debated. Moreover, a transfer of *P. vivax-like* parasites has been documented, thus making possible the release of new strains in new hosts species, specifically in human populations [9]. In such a context, it now seems fundamental to characterize the genome of the closest ape-relative to the human *P. vivax* parasite in order to get a better understanding of the evolution of this parasite and also to identify the key genetic changes explaining the emergence of *P. vivax* in human populations.

Here we report the analysis of two reference and nine further genotypes of *P. vivax-like* parasites*.* We compare these genome sequences to those of several worldwide *P. vivax* isolates, and *P. cynomolgi* and *Plasmodium knowlesi* reference genomes. Our analyses show that the genomes of *P. vivax* and *P. vivax-like* are highly similar and co-linear within the core regions. Phylogenetic analyses clearly show that *P. vivax-like* parasites form a genetically distinct clade from *P. vivax*. Concerning the relative divergence dating, we estimate that both species diverged relatively from each other three times more recently than *Plasmodium ovale wallikeri and Plasmodium ovale curtisi*, and 1.5 times earlier than *Plasmodium malariae*, a human *Plasmodium*, and *Plasmodium* *malariae-like*, its ape relative. Similar to other ape-infective *Plasmodium* species, *P. vivax-like* exhibits far higher levels of genetic diversity than its human-infective relative, *P. vivax*. Finally, our genome-wide analyses provide new insights into the adaptive evolution of *P. vivax*. Indeed, we identify several key genes that exhibit signatures of positive selection exclusively in *P. vivax*, and show that some gene families important for red blood cell invasion have undergone species-specific evolution in the human parasite.

Genome Assemblies

Eleven *P. vivax-like* genotypes were obtained from two different kinds of samples: ten infected chimpanzee blood samples collected during successive routine sanitary controls of chimpanzees living in the Park of La Lékédi (a sanctuary in Gabon) and one infected *Anopheles* mosquito collected during an entomological survey realized in the same park[11] (Supplementary Table 1). For blood samples, white blood cells were depleted using the CF11 method[12] to reduce the amount of host DNA. After DNA extraction, samples were subjected to whole genome amplification (WGA) in order to obtain sufficient parasite DNA for library preparation. Sequencing was performed using short read Illumina technology. For one sample (Pvl06), long read sequencing (PacBio technology) was performed in order to get a better coverage of regions containing subtelomeric gene families.

Among the eleven samples, ten presented mixed infections with other *Plasmodium* species (Supplementary Table 1). Four samples containing *P. gaboni* or *P. malariae-like* co-infections were used in other studies (see Supplementary Table 1)[13,14]. In order to obtain the *P. vivax-like* genotypes, sequencing reads were extracted based on their similarity to the reference genome sequence of *P. vivax,* PvP01[15]. Sequencing reads from two samples, one obtained using Illumina sequencing, Pvl01, and another using PacBio technology, Pvl06, were used to perform *de novo* genome assemblies and annotated to produce reference genomes for *P. vivax-like* (Supplementary Table 2). Of the two assemblies, Pvl01 is of considerably higher quality with 4,570 orthologues to the PvP01 reference genome compared to 2,925 for Pvl06 (Table 1). Both assemblies consist of 14 supercontigs (corresponding to the 14 *P. vivax* chromosomes) and, respectively for Pvl01 and Pvl06, 226 and 370 unassigned contigs, comprising a total of 27.5Mb and 18.8Mb in size for Pvl01 and Pvl06 respectively. After annotation with Companion [16], these two genomes contained 5,532 and 4,953 annotated genes (Table 1).

The genome sequences obtained from the other samples were used for SNP calling, and population genetic and phylogenetic analyses.

Gene synteny and gene composition

Comparing the *P. vivax-like* reference genomes with those of *P. vivax* (PvP01 and SalI)[2,15], *P. cynomolgi* (B strain)[8] and *P. knowlesi* (H strain)[17] reveals several similarities, including a similar GC content, extensive collinearity and conservation of gene content/organization (Table 1). The *P. vivax-like* core genome sequences are completely syntenic to the *P. vivax* PvP01 reference genome sequence (Supplementary Figures 1 and 2).

For *Plasmodium* parasites, most species-specific genes are part of large gene families, such as *var* genes in *P. falciparum* or *pir* genes that currently are present in all *Plasmodium* genomes studied[18,19]. Table 2 provides a summarized view of gene content and copy number of the main multigene families in *P. vivax-like* in comparison to *P. vivax*, *P. knowlesi* and *P. cynomolgi*. Despite only partial coverage of the subtelomeric regions of our reference genomes (Supplementary Figures 1 and 2), at least one copy of each major gene family was detected (Table 2). In comparison to *P. vivax*, as expected because of the partial subtelomeric sequencing coverage, the number of copies in each family was generally lower or equal in *P. vivax-like*, except for the cyto-adherence linked asexual (CLAG) gene family. The CLAG family is an essential gene family in host-parasite interactions, playing a role in merozoite invasion, parasitophorous vacuole formation, cyto-adherence and in the uptake of ions and nutrients from the host plasma[20,21]. In *P. vivax-like*, the CLAG gene family is composed of one extra copy (n=4, 2 CLAG genes on chromosome 14, one CLAG gene on chromosome 8 and one CLAG gene on chromosome 7) compared to the *P. vivax* references PvP01 and Sal I (n=3, one CLAG gene on each of the following chromosomes 7, 8 and 14). Even if this needs to be confirmed in the future by full-genome sequencing of more *P. vivax-like* samples, this could suggest that *P. vivax* potentially lost one CLAG gene (the one situated on chromosome 14) during its adaptation to the human host. Another interesting observation is the expansion of most gene families involve in red blood cell invasion or cyto-adherence (such as PHIST, ETRAMP, PST-A, Pv-fam-e, Pv-fam-a, RBP) in the human *P. vivax* lineage in comparison to *P. vivax-like*, *P. cynomolgi* and *P. knowlesi*, which suggests that the gene duplications in the ancestral lineage of *P. vivax* could be an adaptation to humans (Table 2).

During the *Plasmodium* life cycle, host erythrocyte invasion is mediated by specific interactions between parasite ligands and host erythrocyte receptors. Two major gene families are involved in erythrocyte invasion: the Duffy-binding proteins (DBP) and the reticulocyte-binding proteins (RBP)[22]. For *P. vivax*, two DBP genes (DBP1 on chromosome 6 and DBP2 on chromosome 1) exist and seem to be essential to red blood cell invasion, as demonstrated by their inability to infect individuals not expressing the Duffy receptor on the surface of their red blood cells (i.e. Duffy-positive individuals)[23–25]. In the reference genome of *P. vivax-like* (Pvl01), we observe the DBP1 gene as for the *P. vivax* genome PvP01 (Table 2) and also in the other *Plasmodium* species, however we did not observe the DBP2 one.

Knowing that gorillas and chimpanzees are today all described as Duffy positive[10], we could hypothesize that *P. vivax-like* parasites infect only Duffy positive hosts. This would be in accordance with the fact that the only described transfer of *P. vivax-like* to humans was in a Caucasian Duffy positive individual[9] and that no transfers of *P. vivax-like* were recorded in Central African Duffy negative populations despite the fact that they live in close proximity with infected ape populations[26].

RBP genes encode a merozoite surface protein family present across all *Plasmodium* species and known to be involved in erythrocyte invasion and host specificity[22]. Comparison of the organization and characteristics of this gene family between *P. vivax*, *P. vivax-like*, *P. knowlesi* and *P. cynomolgi* (Table 2), reveals that: 1) all gene classes (RBP1, RBP2 and RBP3) are ancestral to the divergence of these species; 2) the expansion of RBP2, a class of genes that confer the ability to infect reticulocytes (immature red blood cells), was likely ancestral to the *P. vivax*/*P. vivax-like* lineage and 3) RBP3, which is a class of genes supposed to confer the ability to infect specifically normocytes (mature red blood cells), are functional in all species except in *P. vivax* (where the gene is pseudogenised), suggesting that *P. vivax* specifically lost the ability to infect this category of erythrocytes during its adaptation to humans (Figure 1).

Phylogenetic relationships to other *Plasmodium* species and divergence time

Conservation of the gene content between *P. vivax-like* with the other primate-infective *Plasmodium* species has enabled us to reconstruct with confidence the relationships between the different species and to estimate the age of the different speciation events. This analysis confirmed the position of *P. vivax-like* as the closest sister lineage of *P. vivax* (Figure 2).

Regarding the divergence time, with the fix point considered as the relative distance of the split of the speciation of the two *P. ovale* genomes and the *P. malariae-like* and *P. malariae*[14], relative split times between *Plasmodium* species were estimated using the method of Silva et al.[27]. Assuming consistent mutation rates and generation times across the branches, we observed that the time of the split between *P. vivax* and *P. vivax-like* is about three times more recently than than the split between *P. ovale wallikeri* and *P. ovale curtisi* and 1.5 times earlier than *Plasmodium malariae* and *Plasmodium malariae-like*. It has to be noticed that the split between these two last species was estimated to be five times earlier than the one between *P. ovale wallikeri* and *P. ovale curtesi*, which is consistent with the study of Rutledge et al.[14]. However, it has to be confirmed with other dating methods, because the GC content could bias this estimation. Indeed, the models used in our study assume a strict molecular clock, which would not apply to all *Plasmodium* species, specifically for *P. falciparum* because of its extreme GC content in comparison to other *Plasmodium* species.

Relationships to worldwide human *P. vivax* isolates

To analyse the relationship between our 11 *P. vivax-like* isolates with human *P. vivax*, we completed our dataset with 19 published human *P. vivax* genomes[28] (Table 1). All sequencing reads were aligned against the PvP01 reference genome[15] and SNPs were called and filtered as described in the Materials and Methods section. Maximum-likelihood phylogenetic trees were then produced based on 100,616 SNPs. Our results clearly demonstrate the presence of a significantly distinct clade (a bootstrap value of 100) composed of *P. vivax-like* strains on one side and human *P. vivax* isolates on the other side (Figure 3), which is in disagreement with previous results suggesting that human strains formed a monophyletic clade within the radiation of ape *P. vivax-like* parasites[10].

One explanation for this difference with previous results could be due to a phenomenon called Incomplete Lineage Sorting (ILS). ILS is the discordance observed between some gene trees and the species or population tree due to the coalescence of gene copies in an ancestral species or population[29]. Such a phenomenon is often observed when species or population divergence is recent, which is the case for *P.*

vivax/*P. vivax-like*[30,31]. ILS may thus result in the wrong conclusion of *P. vivax* and *P. vivax-like* populations being intermixed and *P. vivax* diversity being included in the diversity of *P. vivax-like*. In our study, the use of a lot more genetic information localized throughout the genome, both in genic and intergenic regions, allows us to reduce this effect of ILS and reflects a more accurate picture of the genetic relationship between the different parasite species. Another explanation to these contradictory results would rely on the analysis of *P. vivax-like* samples collected only in Gabon or Cameroon, which limit the access to the full genetic diversity of these parasites. Clearly for now, the origin, the direction of the transfer and the evolutionary history of these parasites are still unclear and need addition of more *P. vivax-like* samples from other locations to be elucidated.

Our results also show that *P. vivax-like* is composed of two distinct lineages: the one including the two reference genomes (Pvl01 and Pvl06) and seven other isolates that will hereafter be referred as *P. vivax-like 1* and another one including two isolates (Pvl09 and Pvl10) (referred as *P. vivax-like 2*) (Figure 3). These two lineages may reflect an ancient split within *P. vivax-like* or be the consequence of a recent introgression or hybridization event between *P. vivax-like* and *P. vivax* in Africa. Further analyses including sequencing of more African *P. vivax* populations, from different geographic areas, should be done to disentangle these two hypotheses.

Previous studies highlighted the high genetic diversity of *P. vivax-like* populations in comparison to *P. vivax* worldwide[15]. In this genome-wide analysis of the nucleotide diversity π[32], we confirm that *P. vivax-like* populations are significantly more diverse than *P. vivax* populations (P<0.001, Wilcoxon test), with *P. vivax-like* samples showing nearly ten times higher nucleotide diversity ($\pi_{P.vivax}$ = 0.0012; $\pi_{P.vivax\text{-}like}$ = 0.0096). This suggests that African great apes *P. vivax-like* parasites are probably more ancient that the human *P. vivax* strains and that the human *P. vivax* species as other human *Plasmodium* went through a bottleneck and only recently underwent population expansion.

P. vivax specific adaptive evolution

Comparison of the *P. vivax* genome to its closest sister lineage (*P. vivax-like*) and to the other primate *Plasmodium* provides a unique opportunity to identify *P. vivax* specific adaptations to humans. We applied a branch-site test of positive selection to detect events of positive selection that exclusively occurred in the *P. vivax* lineage.

Within the reference genome *P. vivax-like* (Pvl01), 418 genes exhibited significant signals of positive selection (Supplementary Table 3). In the human *P. vivax* genome PvP01, the test allowed the identification of 255 genes showing significant signals of positive selection (Supplementary Table 4). Among these genes presenting a significant dN/dS ratio, 71 were shared between *P. vivax* and *P. vivax-like*, including 56 encoding for proteins with unknown function, and 15 encode proteins that are involved either in energy metabolism regulation (n = 9), in chromatid segregation (n = 2) or cellular-based movement (n = 4).

We then took into consideration the genes detected under positive selection in *P. falciparum*[13] and compared them to those obtain in *P. vivax*. We identified of a subset of 10 genes under positive selection in the human *P. vivax* and *P. falciparum* parasites (P-value<0.05). Among these 10 genes, five are coding for conserved *Plasmodium* proteins with unknown function and three for proteins involved in either transcription or transduction. Interestingly, the two remaining genes under positive selection in these two human *Plasmodium* parasites code for the oocysts capsule protein, which is essential for malaria parasite survival in *Anopheles*' midgut, and for the rhoptry protein ROP14, involved in the protein maturation and the host cell invasion. These results suggest that these proteins could be essential for infection of humans or their vectors and future studies should focus on the involvement of these proteins in human parasite transmission and infection.

Conclusion.

In summary, we assembled the first *P. vivax-like* reference genomes, the closest sister clade to human *P. vivax*, which is an indispensable step in the development of a model system for a better understanding of this enigmatic species. We validated that *P. vivax-like* parasites form a genetically distinct clade from *P. vivax*. Concerning the relative divergence dating, we estimated that the divergence between both species occurred three times more recently than the split between *P. ovale wallikeri* and *P. ovale curtisi*, and 1.5 times earlier than the split between *Plasmodium malariae*. Our genome-wide analyses provided new insights into the adaptive evolution of *P. vivax*. Indeed, we identified several key genes that exhibit signatures of positive selection exclusively in the human *P. vivax* parasites, and show that some gene families important for red blood cell invasion have undergone species-specific evolution in the human parasite, such as for instance RBPs. Are these genes the keys of the emergence of *P. vivax* in the human populations? This pending question will need to be answered through functional studies associated to deeper whole genome analyses. To conclude, this study provides the foundation for further investigations into *Plasmodium vivax* parasite's traits of public health importance, such as features involved in host-parasite interactions, host specificity, and species-specific adaptations.

Material and methods

***P. vivax-like* sample collection and preparation.** *P. vivax-like* samples were identified by molecular diagnostic testing during a continuous survey of great ape *Plasmodium* infections carried out in the Park of La Lékédi, in Gabon, by the Centre International de Recherches Médicales de Franceville (CIRMF)[9]. In parallel, a survey of *Anopheles* mosquitoes circulating in the same area (Park of La Lékédi, Gabon) was conducted in order to identify potential vectors of ape *Plasmodium*[11]. Specifically, mosquitoes were trapped with CDC light traps in the forest of the Park of La Lékédi in Gabon. *Anopheles* specimen were retrieved and identified using a taxonomic key[33] before proceeding to dissection of isolate abdomen. Samples were then stored at -20°C until transportation to the CIRMF, Gabon, where they were stored at -80 °C until processed. Blood samples of great apes were treated using leukocyte depletion by CF11 cellulose column filtration[34]. *P. vivax-like* samples were identified either by amplifying and sequencing the *Plasmodium Cytochrome b* (*Cytb*) gene as described in Ollomo et al. or directly from samples already studied for other *Plasmodium* species[29,38]. This allowed the detection of 11 *P. vivax-like* samples, 10 from chimpanzees and 1 from an *Anopheles moucheti* mosquito. Most of them were co-infected with other *Plasmodium* species, and/or probably with multiple *P. vivax-like* isolates (see below and Supplementary Table 1). The identification of intraspecific *P. vivax-like* co-infections was made by analyzing the distribution of the reference allele frequency[35].

Ethical approval. These investigations were approved by the Government of the Republic of Gabon and by the Animal Life Administration of Libreville, Gabon (no. CITES 00956). All animal work was conducted according to relevant national and international guidelines.

Genome sequencing. DNA was extracted using Qiagen Midi extraction kits (Qiagen) following manufacturer's recommendation, and then enriched through a whole genome amplification step (WGA[36]). The Illumina isolates were sequenced using Illumina Standard libraries of 200-300bp fragments and amplification-free libraries of 400-600bp fragments were prepared and sequenced on the Illumina HiSeq 2500 and the MiSeq v2 according to the manufacturer's standard protocol (Supplementary Table 1). The Pvl06 isolate was sequenced using Pacific Biosciences with the C3/P5 chemistry after a size selection of 8 kb fragments. Raw sequence data are deposited in the European Nucleotide Archive. The accession numbers can be found in Supplementary Table 1.

Assembly of *P. vivax-like* genomes. Two *P. vivax-like* genomes (Pvl01 and Pvl06) were assembled from a co-infection with a *P. malariae-like* and a *P. reichenowi* (PmlGA01 sample in Rutledge et al. 2017)[14] for Pvl01 and from a co-infection with *P. gaboni* for Pvl06 (PGABG01 sample in Otto et al.)[37]. Briefly, the genome assembly of the Illumina sequenced sample Pvl01 was performed using MaSuRCA[38] and the assembled contigs belonging to *P. vivax-like* were extracted using a BLAST search against the *P. vivax* P01 reference genome (PvP01 genome; Auburn et al. 2016; http://www.genedb.org/Homepage/PvivaxP01)[15]. The draft assembly was further improved by iterative uses of SSPACE[39], GapFiller[40] and IMAGE[41]. The 3,540 contigs resulting from these analyses were then ordered against PvP01 genome and the *P. gaboni* and *P. reichenowi* reference genomes[13] to separate possible co-infections with a parasite species of chimpanzees from the *Laverania* subgenus using ABACAS2[42]. The genome assembly was further improved and annotated using the Companion web server[16]. BLAST searches of the unassembled contigs against the two reference genomes were performed before running Companion to keep the contigs with the best BLAST hits against PvP01 only. The PacBio assembly of Pvl06 was performed using Hierarchical Genome Assembly Process HGAP[43].

Read mapping and alignment. Nine additional *P. vivax-like* samples were sequenced for population genomics and polymorphism analyses (see Supplementary Table 1). The dataset was completed with 19 globally sampled *P. vivax* isolates[28] for human *vs.* great apes parasite comparisons, and the Asian parasite *P. cynomolgi* strain B was used as the root for phylogenetic inferences[8]. The 11 newly generated *P. vivax-like* samples, together with the already published 19 *P. vivax* samples and the reference strain *P. cynomolgi*[8] Illumina reads were mapped against the PvP01 reference genome using BWA[44] with default parameters. We then used Samtools to only keep properly paired reads and to remove PCR duplicates[45].

Gene family search. For all *P. vivax-like* Pvl01 and Pvl06, *P. vivax* PvP01 and SalI, *P. cynomolgi* B strain and *P. knowlesi* H strain genomes obtained, gene variants were detected and counted using Geneious software[46].

Orthologous group determination and alignment. Orthologous groups across (1) *P. vivax* PvP01, *P. vivax-like* Pvl01, *P. cynomolgi* B strain[8] and *P. knowlesi* H strain[17] reference genomes and (2) the 13 *Plasmodium* reference genomes used for the phylogeny (the seven *Laverania* genomes *P. falciparum*[47], *P. praefalciparum, P. reichenowi, P. billcollinsi, P. blacklocki, P. gaboni* and *P. adleri*[13], *P. cynomolgi* B strain and *P. knowlesi* H, *P. vivax* PvP01, *P. vivax-like* Pvl01 (this study), and *P. malariae* and *P. malariae-like* [14] were identified using OrthoMCL v2.09[48,49]. From those, we extracted different sets of one-to-one orthologues for the subsequent analyses: a set of 4,056 genes that included the one-to-one orthologues among the four restricted species, *P. vivax*, *P. vivax-like, P. cynomolgi* and *P. knowlesi*, and a set of 2,352 among the 13 *Plasmodium* species considered here for the interspecies phylogenetic analysis.

Amino acid sequences of the one-to-one orthologues were aligned using MUSCLE[50]. Prior to aligning codon sequences, we removed the low complexity regions identified on a nucleotide level using dustmasker[51] and then in amino acid sequences using segmasker [52] from ncbi-blast. After MUSCLE alignments[50], we finally excluded poorly aligned codon regions using Gblocks default parameters[53].

SNP discovery and annotation. SNPs were called independently for all 11 *P. vivax-like* and 19 *P. vivax* samples by first mapping the samples against the *P. vivax* PvP01 reference genome using SMALT and then calling SNPs using Samtools mpileup v. 0.1.9 (parameters –q 20 -Q 20 -C 50) followed by bcftools (call -c -V indels). SNPs were filtered using VCFTools (--minDP 5 –max-missing 1).

Divergence dating. To estimate the dates of speciation, we used 12 *Plasmodium* genomes: the here generated *P. vivax-like* Pvl01, the *P. vivax* PvP01[15], *P. cynomolgi* M Version 2[54], *P. coatneyi* PcyM[54], *P. knowlesi* H strain[17], *P. falciparum* 3D7[47] *P. reichenowi* PrCDC[37], *P. gallinaceum*[55], and *P. ovale wallikeri*, *P. ovale curtisi*, *P. malariae* and *P. malariae-like*[14]. From the proteins of the 12 genomes, low complexity regions were excluded with SEG filter, using default parameters[56]. After an all-against-all BLASTp (parameter Evalue 1e-6), OrthoMCL v.1.4[49] (using default parameters) was run. For each of the 2943 1-1 orthologous, an alignment was generated with MUSCLE[50] and the alignment was finally cleaned with Gblocks (parameters: -t=p -b5=h -p=n -b4=2)[57].

To build the phylogenetic tree, the software RAxML v.8.2.8[58] was used on the concatenated alignments of 1000 random picked orthologous. The PROTGAMALG substitution model was then used, as proposed in Rutledge et al[14], 100 bootstraps were run confirming the tree.

To date the speciation, the methods from Silva et al.[27] was applied. The dAA was obtained through a pairwise comparison using paML v.4.7 [59]. An R script from the authors of the method[27] allowed the estimation of alpha with the error bound for each pair, based on a Total Least Squares regression. Results are reported in Figure 2. As a fix point, we used the relative distance of the split of the speciation of the two *P. ovale* genomes and the *P. malariae-like* and *P. malariae*[14]. The split of *P. reichenowi* and *P. falciparum* was also dated based on the *P. malariae* and *P. malariae-like* split estimation[14]. However, this will need to be confirmed with other methods, because the GC content could bias this estimation. Indeed, the models used in our study assume a strict molecular clock, which would not apply to all *Plasmodium* species, specifically for *P. falciparum* because of its extreme GC content in comparison to other *Plasmodium* species.

Phylogenetic tree of *P. vivax* and *P. vivax-like* strains. We constructed for Figure 3 a maximum-likelihood tree using the filtered variant call set of SNPs limited to the higher allelic frequency genotypes identified within each sample using RAxML and PhyML (using general-time reversible GTR models)[58,60]. Trees were visualized using Geneious software[46]. All approaches showed the same final phylogenetic tree described in the results section.

Genome wide nucleotide diversity. For the *P. vivax* and *P. vivax-like* populations, we calculated the genome-wide nucleotide diversity (π)[32] using VCFTools[61]. The nucleotide diversity was compared between *P. vivax* and *P. vivax-like* species based on the Wilcoxon-Mann-Whitney non-parametric test.

Detection of genes under selection. In order to identify genomic regions involved in the parasite adapting to the human host, meaning regions under positive selection, we performed branch site tests. To search for genes that have been subjected to positive selection in the *P. vivax* lineage alone, after the divergence from *P. vivax-like*, we used the updated Branch-site test of positive selection[62] implemented in the package PAML v4.4c[59]. This test detects sites that have undergone positive selection in a specific branch of the phylogenetic tree (foreground branch). All coding sequences in the core genome were used for the test (4,056 gene sets of orthologous genes). A set of 4056 orthologous groups between *P. vivax*, *P. vivax-like*, *P. knowlesi* and *P. cynomolgi* was used for this test. dN/dS ratio estimates per branch and gene were obtained using Codeml (PAML v4.4c) with a *free-ratio* model of evolution[59].

Data availability. All sequences are being submitted to the European Nucleotide Archive. The accession numbers of the raw reads and assembly data will be found in Supplementary Table 2. As the assemblies are private, they will be available on request.

References.

1. World Health Organization. *World Health Organization Report*. (2014).

2. Carlton, J. M. *et al.* Comparative genomics of the neglected human malaria parasite Plasmodium vivax. *Nature* **455,** 757–763 (2008).

3. Carlton, J. M., Das, A. & Escalante, A. A. Genomics, population genetics and evolutionary history of Plasmodium vivax. *Adv. Parasitol.* **81,** 203–222 (2013).

4. Price, R. N. *et al.* Clinical and pharmacological determinants of the therapeutic response to dihydroartemisinin-piperaquine for drug-resistant malaria. *Antimicrob. Agents Chemother.* **51,** 4090–4097 (2007).

5. Galinski, M. R., Meyer, E. V. S. & Barnwell, J. W. Plasmodium vivax. in *Advances in Parasitology* **81,** 1–26 (Elsevier, 2013).

6. Guerra, C. A. *et al.* The International Limits and Population at Risk of Plasmodium vivax Transmission in 2009. *PLoS Negl. Trop. Dis.* **4,** e774 (2010).

7. Mu, J. *et al.* Host switch leads to emergence of Plasmodium vivax malaria in humans. *Mol. Biol. Evol.* **22,** 1686–1693 (2005).

8. Tachibana, S.-I. *et al.* Plasmodium cynomolgi genome sequences provide insight into Plasmodium vivax and the monkey malaria clade. *Nat. Genet.* **44,** 1051–1055 (2012).

9. Prugnolle, F. *et al.* Diversity, host switching and evolution of Plasmodium vivax infecting African great apes. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 8123–8128 (2013).

10. Liu, W. *et al.* African origin of the malaria parasite Plasmodium vivax. *Nat. Commun.* **5,** (2014).

11. Makanga, B. *et al.* Ape malaria transmission and potential for ape-to-human transfers in Africa. *Proc. Natl. Acad. Sci.* **113,** 5329–5334 (2016).

12. Auburn, S. *et al.* An Effective Method to Purify Plasmodium falciparum DNA Directly from Clinical Blood Samples for Whole Genome High-Throughput Sequencing. *PLoS ONE* **6,** (2011).

13. Otto, T. D. *et al.* Genomes of an entire Plasmodium subgenus reveal paths to virulent human malaria. *bioRxiv* 095679 (2016). doi:10.1101/095679

14. Rutledge, G. G. *et al.* Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. *Nature* **542,** 101–104 (2017).

15. 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 (2016).

16. Steinbiss, S. *et al.* Companion: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res.* **44,** W29–W34 (2016).

17. Pain, A. *et al.* The genome of the simian and human malaria parasite Plasmodium knowlesi. *Nature* **455,** 799–803 (2008).

18. Su, X. Z. *et al.* The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. *Cell* **82,** 89–100 (1995).

19. Otto, T. D. *et al.* A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biol.* **12,** 86 (2014).

20. Desai, S. A. Ion and nutrient uptake by malaria parasite-infected erythrocytes. *Cell. Microbiol.* **14,** 1003–1009 (2012).

21. Gupta, A., Thiruvengadam, G. & Desai, S. A. The conserved clag multigene family of malaria parasites: essential roles in host-pathogen interaction. *Drug Resist. Updat. Rev. Comment. Antimicrob. Anticancer Chemother.* **18,** 47–54 (2015).

22. Iyer, J., Grüner, A. C., Rénia, L., Snounou, G. & Preiser, P. R. Invasion of host cells by malaria parasites: a tale of two protein families. *Mol. Microbiol.* **65,** 231–249 (2007).

23. Gunalan, K. *et al.* Role of *Plasmodium vivax* Duffy-binding protein 1 in invasion of Duffy-null Africans. *Proc. Natl. Acad. Sci.* **113,** 6271–6276 (2016).

24. Langhi, D. M. & Bordin, J. O. Duffy blood group and malaria. *Hematol. Amst. Neth.* **11,** 389–398 (2006).

25. Miller, L. H., Mason, S. J., Clyde, D. F. & McGinniss, M. H. The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. *N. Engl. J. Med.* **295,** 302–304 (1976).

26. Délicat-Loembet, L. *et al.* No evidence for ape Plasmodium infections in humans in Gabon. *PloS One* **10,** e0126933 (2015).

27. Silva, J. C., Egan, A., Arze, C., Spouge, J. L. & Harris, D. G. A New Method for Estimating Species Age Supports the Coexistence of Malaria Parasites and Their Mammalian Hosts. *Mol. Biol. Evol.* **32,** 1354–1364 (2015).

28. Hupalo, D. N. *et al.* Population genomics studies identify signatures of global dispersal and drug resistance in Plasmodium vivax. *Nat. Genet.* **48,** 953–958 (2016).

29. Degnan, J. H. & Rosenberg, N. A. Gene tree discordance, phylogenetic inference

and the multispecies coalescent. *Trends Ecol. Evol.* **24,** 332–340 (2009).

30. Choleva, L. *et al.* Distinguishing between Incomplete Lineage Sorting and Genomic Introgressions: Complete Fixation of Allospecific Mitochondrial DNA in a Sexually Reproducing Fish (Cobitis; Teleostei), despite Clonal Reproduction of Hybrids. *PLoS ONE* **9,** (2014).

31. Maddison, W. P. & Knowles, L. L. Inferring Phylogeny Despite Incomplete Lineage Sorting. *Syst. Biol.* **55,** 21–30 (2006).

32. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* **76,** 5269–5273 (1979).

33. Gillies, M. T. & Coetzee, M. A Supplement to the Anophelinae of Africa South of the Sahara(Afrotropical Region). (1987). Available at: http://ihi.eprints.org/1274/. (Accessed: 7th April 2017)

34. Venkatesan, M. *et al.* Using CF11 cellulose columns to inexpensively and effectively remove human DNA from Plasmodium falciparum-infected whole blood samples. *Malar. J.* **11,** 41 (2012).

35. Chan, E. R. *et al.* Whole Genome Sequencing of Field Isolates Provides Robust Characterization of Genetic Diversity in Plasmodium vivax. *PLoS Negl. Trop. Dis.* **6,** e1811 (2012).

36. Oyola, S. O. *et al.* Optimized whole-genome amplification strategy for extremely AT-biased template. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* **21,** 661–671 (2014).

37. Otto, T. D. *et al.* Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat. Commun.* **5,** 4754 (2014).

38. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinforma. Oxf. Engl.* **29,** 2669–2677 (2013).

39. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinforma. Oxf. Engl.* **27,** 578–579 (2011).

40. Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* **13 Suppl 14,** S8 (2012).

41. Tsai, I. J., Otto, T. D. & Berriman, M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* **11,** R41 (2010).

42. Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25,** 1968–1969 (2009).

43. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10,** 563–569 (2013).

44. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496 (2011).

45. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25,** 2078–2079 (2009).

46. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinforma. Oxf. Engl.* **28,** 1647–1649 (2012).

47. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419,** 498–511 (2002).

48. Chen, F., Mackey, A. J., Stoeckert, C. J. & Roos, D. S. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **34,** D363–D368 (2006).

49. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).

50. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5,** 113 (2004).

51. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **13,** 1028–1040 (2006).

52. Statistics of Local Complexity in Amino Acid Sequences and Sequence Databases. *ResearchGate* Available at: https://www.researchgate.net/publication/220667755_Statistics_of_Local_Complexity_in_Amino_Acid_Sequences_and_Sequence_Databases. (Accessed: 7th April 2017)

53. Castresana, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol. Biol. Evol.* **17,** 540–552 (2000).

54. Pasini, E. M. *et al.* An improved Plasmodium cynomolgi genome assembly reveals an unexpected methyltransferase gene expansion. *Wellcome Open Res.* **2,** 42 (2017).

55. Boehme, U. *et al.* Complete avian malaria parasite genomes reveal host-specific parasite evolution in birds and mammals. *bioRxiv* 086504 (2016). doi:10.1101/086504

56. Wootton, J. C. & Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17,** 149–163 (1993).

57. Talavera, G., Castresana, J., Kjer, K., Page, R. & Sullivan, J. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Syst. Biol.* **56,** 564–577 (2007).

58. Stamatakis, A. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* btu033 (2014). doi:10.1093/bioinformatics/btu033

59. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).

60. Guindon, S., Lethiec, F., Duroux, P. & Gascuel, O. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res.* **33,** W557–W559 (2005).

61. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

62. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22,** 2472–2479 (2005).

Acknowledgements. Authors thank ANR ORIGIN JCJC 2012, LMI ZOFAC, CNRS-INEE, CIRMF, IRD, Sanger Institute for financial support and Société d'Exploitation du Parc de la Lékédi, Bakoumba, GABON. GGR is supported by the Medical Research Council (MR/J004111/1) and the Wellcome Trust (098051).

Author contributions. DTO, FR, FP and VR designed the study. CA, , PD, BO, NDM, APO, BN, BM, LB, CP, FP and VR collected and assessed samples. CA performed the WGA. TDO managed the sequencing. AG, BF and TDO did assembly and annotation. TDO, VR and FP performed the evolutionary analyses on core genomes. TDO and GGR performed the dating analyses. AG, TDO, FR and VR wrote the manuscript. All authors read and approved the paper.

Competing financial interest. None.

Materials & Correspondance. Rougeron Virginie; Laboratoire MIVEGEC (UM-CNRS-IRD), 34394 Montpellier, France; rougeron.virginie@gmail.com / virginie.rougeron@ird.fr;

Tables

Table 1. Genome features of the *P. vivax-like* Pvl01 (Illumina HiSeq sequenced) and Pvl06 (PacBio sequenced) strains, *P. vivax* reference strains SalI and PvP01[15], *P. cynomolgi* B and M isolates[8,54] and *P. knowlesi* H strain[17]. *Unassigned contigs indicated in parentheses.

	***P. vivax-like* (Pvl01)**	***P. vivax-like* (Pvl06)**	***P. vivax* (PvP01)**	***P. vivax* (SalI)**	***P. cynomolgi* (B strain)**	***P. cynomolgi* (M strain)**	***P. knowlesi* (H strain)**
Assembly size (Mb)	27.5	18.8	29	26.8	26.2	30.6	24.1
Scaffolds	14 (226)*	14 (370)*	14	14	14	14	14
Overall G + C content (%)	44.9	45.8	39.8	42.3	40.4	37.3	37.5
Number of genes	5,532	4,953	6,642	6,690	5,722	6,632	5,188
Gene density (genes/Mb)	201.2	264.5	229	249.6	218.4	216.7	215.3
# 1:1 orthologues with PvP01	4,570	2,925	-	-	4,870	-	4,804

Table 2. Multigene families copy number description of *P. vivax-like* (Pvl01 and Pvl06), *P. vivax* strains SalI and PvP01[15], *P. cynomolgi* B and M strains[8,54] and *P. knowlesi* H strain[17]. *For *P. vivax-like* Pvl01 and Pvl06, a non-exhaustive list of family genes is represented since only partial genomes were obtained.

	P. vivax-like* (Pvl01)	***P. vivax-like* (Pvl06)***	***P. vivax* (PvP01)**	***P. vivax* (SalI)**	***P. cynomolgi* (B strain)**	***P. cynomolgi* (M strain)**	***P. knowlesi* (H strain)**
Vir/Pir (Subtelomeric)	148	14	1212	303	265	1373	64
Msp3 (Central)	9	0	12	11	12	14	3
Msp7 (Central)	11	12	13	13	13	11	5
DBP (Subtelomeric)	1	0	2	1	2	2	3
RBP (Subtelomeric)	9	3	10	9	8	6	2
Pv-fam-a (trag) (Subtelomeric)	33	10	40	34	36	39	26
Pv-fam-e (rad) (Subtelomeric)	38	15	40	34	27	27	16
PST-A (Subtelomeric and central)	6	3	10	11	9	8	7
ETRAMP (Subtelomeric)	7	4	10	9	9	9	9
CLAG (RhopH-1) (Subtelomeric)	4	2	3	3	2	2	2
PvSTP1 (Subtelomeric)	4	0	9	10	3	51	0
PHIST (Pf-fam-b)	20	12	84	64	48	54	15

(Subtelomeric)							
SERA (Central)	13	7	13	13	13	13	8

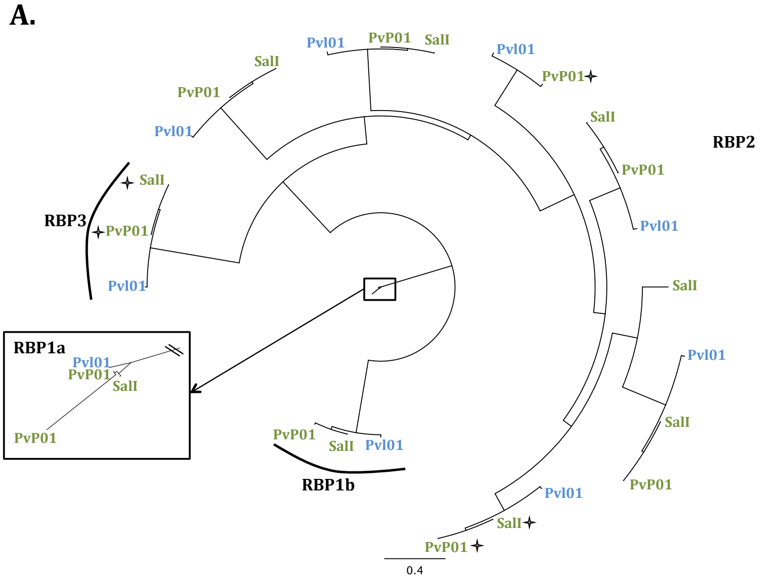
Figures legends.

Figure 1. Reticulocyte binding proteins in *P. vivax-like* and *P. vivax*. A. Phylogenetic tree of all full-length RBPs in *P. vivax-like* Pvl01 (in blue) and *P. vivax* SalI and PvP01 strains (in green). The different subclasses of RBPs are indicated as RBP1a, RBP1b, RBP2 and RPB3. The scale bar represents the genetic distance. The stars indicate pseudogenes. B. Table representing the number of variants (including the ones that are pseudogenised) observed in each RBP subclass in *P. vivax-like* (Pvl01), *P. vivax* (SalI and PvP01), *P. cynomolgi* (B, Berok and Cambodian strains) and *P. knowlesi* (H strain). Pseudogenes detected among each subclass of RBP are indicated within each subclass between brackets.

Figure 2. Divergence dating between *P. vivax* and *P. vivax-like*. Maximum likelihood phylogenetic tree of 12 *Plasmodium* species including *P. vivax* and *P. vivax-like*. The analysis was based on an alignment of 2943 1-1 orthologous of 12 *Plasmodium* reference genomes. The relative split between *P. vivax* and *P. vivax-like* is estimated at around three times shorter than the split between *P. ovale wallikeri* and *P. ovale curtisi* and 1.5 times longer than *Plasmodium malariae* and *Plasmodium malariae-like*. The 'X' indicates relative split values, based on the fixed point as the relative distance of the split of the speciation of the two *P. ovale* genomes and the *P. malariae-like* and *P. malariae*[14].

Figure 3. Maximum likelihood phylogenetic tree with 1000 bootstraps computed through alignment to the *P. cynomolgi* B strain genome, based on 100,616 SNPs shared by 11 *P. vivax-like* and 19 *P. vivax* samples. Bootstrap values superior to 70% are indicated. The host in which the *Plasmodium* parasite was detected is indicated by the pictograms (human, chimpanzee and *Anopheles*). This phylogeny showed the presence of a significantly distinct clade (high bootstrap values associated to each clade) composed of *P. vivax-like* strains on one side (light blue) and human *P. vivax* isolates on the other side (light green).

Figure 1.



B.

	P. vivax-like (Pvl01)	*P. vivax* (PvP01)	*P. vivax* (SalI)	*P. cynomolgi* (B, Berok, Cambodian strains)	*P. knowlesi* (H strain)
RBP1	2	3	2	1-2	1 (1)
RBP2	6	6 (2)	6 (1)	4-6	1
RBP3	1	1 (1)	1 (1)	1	1

Figure 2.

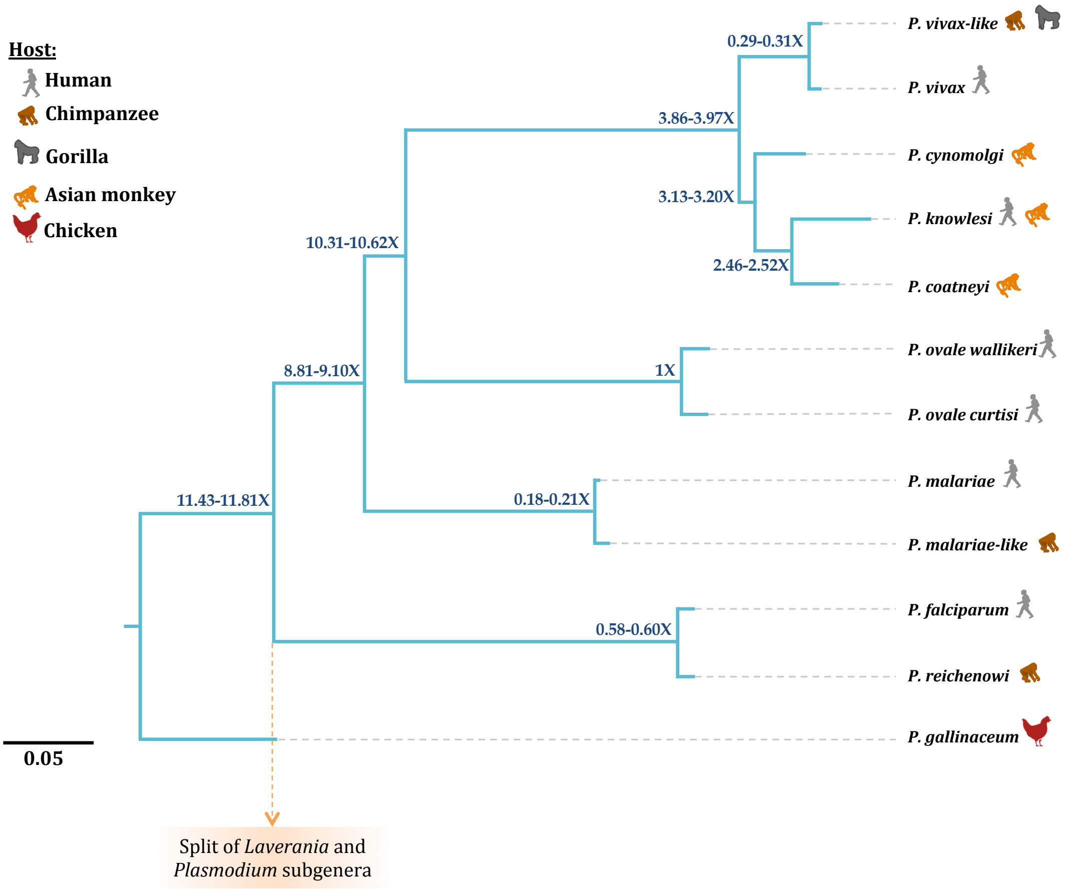
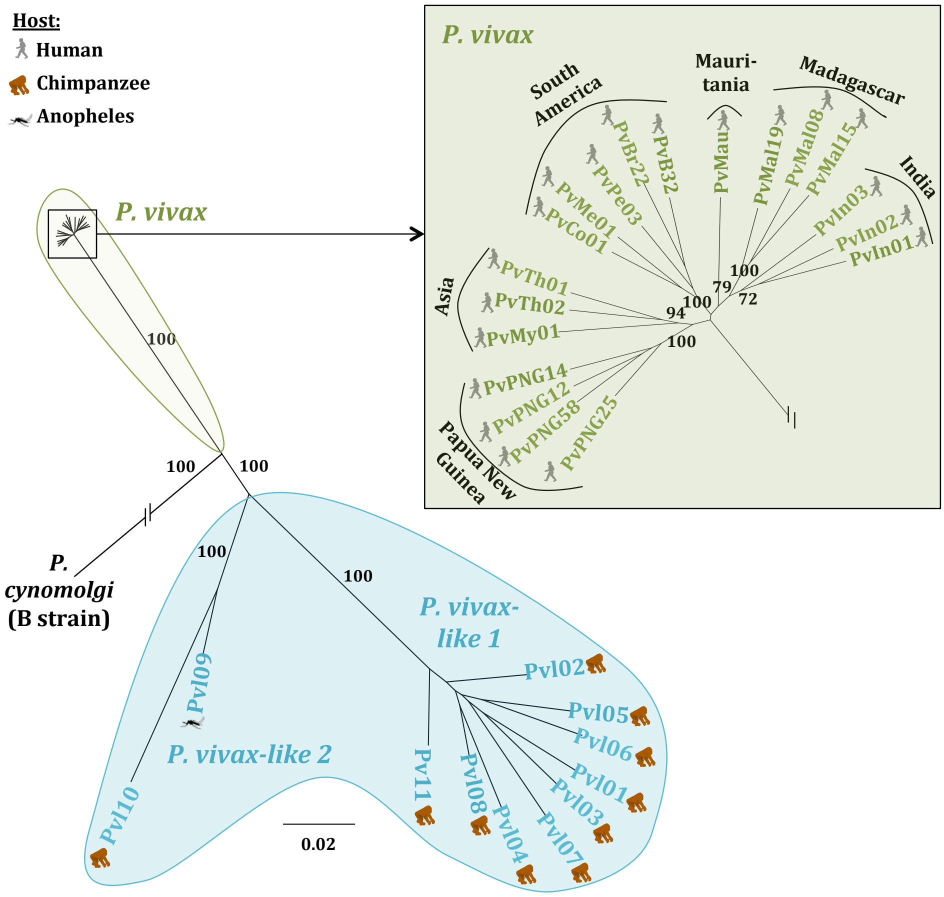

Host:
Human
Chimpanzee
Gorilla
Asian monkey
Chicken
0.29-0.31X
3.86-3.97X
3.13-3.20X
2.46-2.52X
10.31-10.62X
8.81-9.10X
1X
11.43-11.81X
0.18-0.21X
0.58-0.60X
P. vivax-like
P. vivax
P. cynomolgi
P. knowlesi
P. coatneyi
P. ovale wallikeri
P. ovale curtisi
P. malariae
P. malariae-like
P. falciparum
P. reichenowi
P. gallinaceum
0.05
Split of *Laverania* and *Plasmodium* subgenera

Figure 3.


Host:
Human
Chimpanzee
Anopheles
P. vivax
100
100
100
P. cynomolgi (B strain)
100
100
P. vivax-like 1
P. vivax-like 2
Pvl02
Pvl05
Pvl06
Pvl01
Pvl03
Pvl07
Pvl04
Pvl08
Pvl11
Pvl09
Pvl10
0.02
P. vivax
South America
Mauritania
Madagascar
India
Asia
Papua New Guinea
PvBr22
PvB32
PvPe03
PvMe01
PvCo01
PvMau
PvMal19
PvMal08
PvMal15
PvIn03
PvIn02
PvIn01
PvTh01
PvTh02
PvMy01
PvPNG14
PvPNG12
PvPNG58
PvPNG25
100
79
100
72
94
100