GENETICS

Population genomic evidence of *Plasmodium vivax* Southeast Asian origin

Josquin Daron[1]*, Anne Boissière[1,2], Larson Boundenga[3], Barthelemy Ngoubangoye[3], Sandrine Houze[4], Celine Arnathau[1,2], Christine Sidobre[1], Jean-François Trape[1], Patrick Durand[1,2], François Renaud[1,2], Michael C. Fontaine[1,2,5†], Franck Prugnolle[1,2†], Virginie Rougeron[1,2*†]

***Plasmodium vivax* is the most common and widespread human malaria parasite. It was recently proposed that *P. vivax* originates from sub-Saharan Africa based on the circulation of its closest genetic relatives (*P. vivax-like*) among African great apes. However, the limited number of genetic markers and samples investigated questions the robustness of this hypothesis. Here, we extensively characterized the genomic variations of 447 human *P. vivax* strains and 19 ape *P. vivax-like* strains collected worldwide. Phylogenetic relationships between human and ape *Plasmodium* strains revealed that *P. vivax* is a sister clade of *P. vivax-like*, not included within the radiation of *P. vivax-like*. By investigating various aspects of *P. vivax* genetic variation, we identified several notable geographical patterns in summary statistics in function of the increasing geographic distance from Southeast Asia, suggesting that *P. vivax* may have derived from a single area in Asia through serial founder effects.**

INTRODUCTION

Plasmodium vivax is the most common human malarial parasite responsible for 12 million to 22 million clinical cases per year (*1*). It is widespread in the tropical belt where almost 3 billion people are at risk of infection. It is the most frequent cause of human malaria (*2*), particularly in Central and Southeast Asia. Conversely, populations living in sub-Saharan Africa are protected from *P. vivax* transmission due to the absence of the Duffy antigen (i.e., Duffy negativity) at the surface of their red blood cells (*3*, *4*). Historically, *P. vivax* remained understudied compared with *Plasmodium falciparum* because of the lower mortality rate of malaria caused by *P. vivax*. However, the recent emergence of new therapeutic resistance and the discovery of fatal cases due to *P. vivax* have questioned the benign status of *P. vivax* malaria (*5*). Moreover, the identification of *P. vivax* strains that can invade Duffy-negative red blood cells has been raising concerns over the potential spread of malaria in regions that are assumed to be protected (*6*). Today, *P. vivax* is considered a major public health threat (*7*). To put in place effective strategies for malaria control and elimination, we need to accumulate knowledge on the genetic structure of isolates from infected individuals, because this helps to understand the local patterns of malaria transmission and the dynamics of genetic recombination in natural *P. vivax* populations.

Early work on the patterns of genetic variation in *P. vivax* populations worldwide was limited to a small set of samples and genetic markers (mainly mitochondrial and autosomal markers) (*8*–*16*). Consequently, these genetic analyses yielded an incomplete picture of *P. vivax* evolutionary history. Recent technological breakthroughs allowed sequencing the whole *P. vivax* genome after parasite DNA enrichment from clinical blood samples (*17*, *18*). In few years, several projects characterized *P. vivax* genetic variation at the whole-genome scale (*19*–*22*) and highlighted strong signals of recent evolutionary selection, partly due to known drug resistance genes. These studies released hundreds of complete genome sequences, but they restricted their investigation to a very small fraction of the worldwide *P. vivax* diversity. Consequently, a comprehensive picture of the worldwide genetic diversity and structure of *P. vivax* populations is still missing, and its evolutionary history and how it spread in the world are still poorly understood. Moreover, despite growing evidence suggesting an underlying widespread presence of *P. vivax* in all African malaria-endemic regions (*23*), too few complete genome sequences have been released for this area. Thus, a key challenge is to provide a worldwide understanding of *P. vivax* genetic variability genome wide with the ultimate aim of better understanding its past demographic history and origin.

The origin of the current *P. vivax* in humans has stimulated passionate and exciting debates for years. Some studies placed the origin of human *P. vivax* in Southeast Asia ("out-of-Asia" hypothesis) based on its phylogenetic position in a clade of malaria parasites that infect Asian monkeys (*10*, *13*, *24*–*28*). This scenario is also supported by genotyping data at 11 microsatellite markers collected in four continents, showing the highest microsatellite diversity in Southeast Asia (*29*). However, the Asian origin has been challenged by an "out-of-Africa" scenario after the recent discovery of a closely related *Plasmodium* species that circulates in wild-living African great apes (chimpanzees and gorillas) (*28*, *30*–*34*). It has been suggested that this new lineage, hereafter referred to as *P. vivax-like*, gave rise to *P. vivax* in humans following a parasite transfer from African apes (*33*, *35*). This new finding brought to the limelight a 70-year-old hypothesis according to which the high prevalence of Duffy negativity in sub-Saharan African human populations is the consequence of the long interaction between humans and malaria parasites, thus supporting an African origin of *P. vivax* (*36*). Yet, despite a century of research, the debate has not been settled yet. Although privileged, the out-of-Africa hypothesis remains debatable because the exact series of events linking the current human *P. vivax* populations and African great ape *P. vivax-like* remains unclear (*37*).

[1]Laboratoire MIVEGEC (Université de Montpellier-CNRS-IRD), 34394 Montpellier, France. [2]Centre of Research in Ecology and Evolution of Diseases (CREES), Montpellier, France. [3]Centre Interdisciplinaire de Recherches Médicales de Franceville, Franceville, Gabon. [4]Service de Parasitologie-mycologie CNR du Paludisme, AP-HP Hôpital Bichat, 46 rue H. Huchard, 75877 Paris Cedex 18, France. [5]Groningen Institute for Evolutionary Life Sciences (GELIFES), University of Groningen, PO Box 11103 CC, Groningen, Netherlands.
*Corresponding author. Email: josquin.daron@gmail.com (J.D.); virginie.rougeron@ird.fr, virginie.rougeron@gmail.com (V.R.)
†These authors co-supervised this work.

Here, to provide novel insights into the worldwide historical demography of *P. vivax* populations, we analyzed the genomic variations of 447 human *P. vivax* strains from 21 different countries. We also included 19 *P. vivax-like* strains from great apes to explore *P. vivax* evolutionary history from its origins to contemporary time and to specifically assess whether the parasite origin is consistent with the out-of-Africa or out-of-Asia hypothesis. To our knowledge, this dataset provides the most comprehensive characterization of *P. vivax* and *P. vivax-like* population genetic structure and diversity worldwide, with the identification of two distinct non-recombining *P. vivax-like* clades that circulate in sympatry among great apes. Our results demonstrate that *P. vivax* is a sister clade of *P. vivax-like* and is not included within the radiation of *P. vivax-like*, as previously suggested (*33*, *35*). Last, several lines of evidence from summary statistics on geographic trends in *P. vivax* worldwide genetic variation support the hypothesis of a serial founder effect from a single origin in Southeast Asia.

RESULTS AND DISCUSSION

P. vivax and *P. vivax-like* genomic data and diversity

We processed and analyzed genome-wide data from 1154 *P. vivax* isolates sampled from all around the world. This dataset included 20 newly sequenced African isolates (Mauritania, $n = 14$; Ethiopia, $n = 3$; and Sudan, $n = 3$) because the worldwide sampling lacked *P. vivax* isolates from Africa [only 3 isolates from Madagascar (*20*) and 24 from Ethiopia (*21*)]. *P. vivax* samples were from 21 countries (769 from Asia, 338 from America, and 47 from Africa) (table S1). To trace *P. vivax* genetic ancestry, we also included 27 genome sequences of *P. vivax-like* isolated from African great apes ($n = 10$ newly sequenced in our laboratory and $n = 17$ from public databases: Gabon, $n = 11$; Cameroon, $n = 5$; and Ivory Coast, $n = 1$) (*35*, *37*).

Because of the heterogeneity of DNA enrichment methods and sequencing technologies used to obtain the sequences in these datasets, our genome-wide data exhibited a broad range of sequencing depth coverage (fig. S1A). Therefore, to obtain reliable results, we retained only full-genome sequences with a minimum average sequencing depth of at least 5×. This reduced the *P. vivax* isolates from 1154 to 473 and the *P. vivax-like* isolates from 27 to 20. Then, we discarded isolates and variants with a fraction of missing call rate of >50% (fig. S1B). The final dataset included 466 full-genome sequences (447 for *P. vivax* and 19 for *P. vivax-like* isolates; Fig. 1A) in which we identified 2,005,455 high-quality single-nucleotide polymorphisms (SNPs).

The *P. vivax* genome has a heterogeneous genetic structure

Population structuration and demographic fluctuations have a global impact on genomic variations in natural populations. Conversely, local heterogeneity can be observed in genomic regions due to non-random factors, including structural chromosomal features, such as

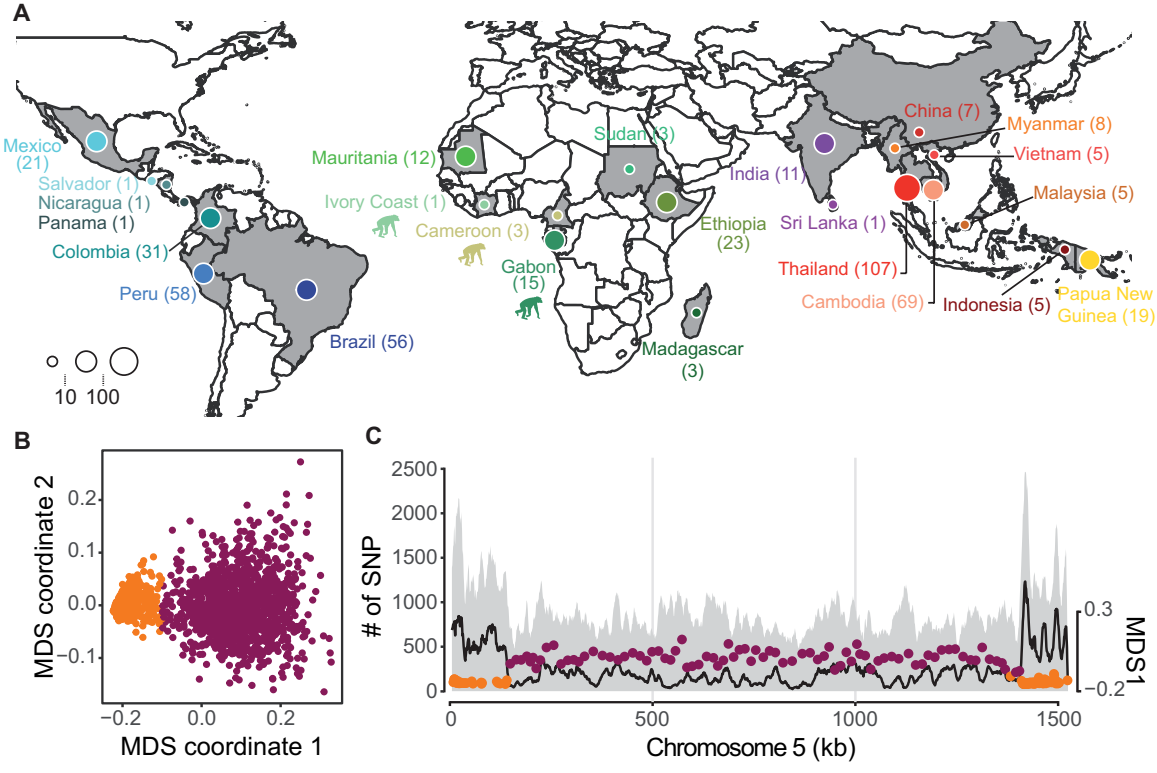

Fig. 1. Geographical origin of *Plasmodium* isolates and patterns of genomic variation. (**A**) Country of origin of the 447 *P. vivax* and 19 *P. vivax-like* isolates used in this study. Within each country, isolates were collected at different locations. The chimpanzee pictogram represents African *P. vivax-like* isolates. (**B**) Local variation of genetic relatedness along the genome of *P. vivax* isolates visualized by multidimensional scaling (MDS) based on the local PCA approach (*39*). Each point represents a non-overlapping genomic window of 100 SNPs. On the basis of the variation of the MDS-1 coordinate, each window was classified in two groups: subtelomeric hypervariable region (orange) and core region (purple). (**C**) Genome scan of the SNPs on chromosome 5 (left *y* axis) identified for both *P. vivax* and *P. vivax-like*. The gray area in the background shows the total number of SNPs identified (left *y* axis), and the black line represents the number of SNPs shared by the two species. The right *y* axis represents the MDS-1 coordinates against the middle point of each window. Windows were classified in the subtelomeric hypervariable region (orange) or the core region (purple).

chromosomal inversions and heterochromatin, or due to selective forces that affect the local genetic diversity and recombination (*38*). These factors can lead to local genome variations in the genetic ancestry and in the relatedness of individual strains, confounding global patterns of genetic diversity and population structure. We identified potential local variations in individual relatedness along the genome using a local principal components analysis (PCA) (*39*). We detected contrasted patterns of individual relatedness along the *P. vivax* genome that defined two main genomic partitions: a small partition localized mainly at the subtelomeric ends of each chromosome (in orange in Fig. 1, B and C) and a larger partition at the center of each chromosome that encompassed 80% of the total SNP set (hereafter called the "core region," in purple in Fig. 1, B and C). The subtelomeric regions that we identified as having distinct ancestry from the rest of the genome also coincided with hypervariable genomic regions (in orange in Fig. 1C and fig. S2). These regions are known to include hypervariable repetitive regions that often cause genotyping errors. Their exact coordinates have only been reported on a former version of the *P. vivax* genome assembly (*19*). The differences in genetic ancestry and strain relatedness provided by the two partitions were evident when looking at the local PCA (fig. S3). While the population structure recovered from the core regions displayed interpretable genetic patterns that were consistent with previous studies (*19*, *20*), the genetic picture obtained from the SNPs in the subtelomeric hypervariable regions was much harder to interpret. Therefore, we excluded these regions from the subsequent analyses that focused only on the core regions on the central chromosomal region (table S2). This region represented ~21 Mb of the genome and included 1,610,445 high-quality SNPs (~1 SNP every 13 bp).

Last, we evaluated in each sample the within-sample parasite infection complexity by calculating the within-sample F_{ws} metric (*40*). Monoclonal infections occurred in approximately 71% of samples, with the highest proportion of mixed infections observed in Thai and Cambodian isolates (fig. S4). The proportion of monoclonal infections was higher in the newly sequenced *P. vivax* individuals from Mauritania (92%, 11 of 12) but was not significantly different from the percentage found in Ethiopian isolates (61%, 14 of 23; Pearson's χ^2 test, $P = 0.13$), although it was skewed toward higher F_{ws} scores, suggesting less frequent superinfection and/or cotransmission in this area. To avoid further possible bias due to multiple infections by several strains, we designed an SNP calling procedure to select within each sample the variant with the highest frequency. Therefore, for each site, one single variant was called per sample. In addition, we estimated the genome fraction that exhibited identity by descent (IBD) among pairs of samples from each population and selected only samples with low level of relatedness (IBD of >70%) for the downstream population genetic analyses (fig. S5). Our results provided insights into the parasite recent epidemiology history because populations with high levels of relatedness were considered as coming from regions with lower malaria transmission.

Evolutionary relationships between ape *P. vivax-like* and human *P. vivax*

First, we analyzed the phylogenetic relationships between *P. vivax* and *P. vivax-like* infecting great apes, the closest relatives of human *P. vivax*, by building a maximum likelihood (ML) genetic tree based on the SNP data for the 19 *P. vivax-like* and 447 *P. vivax* full-genome sequences. This analysis revealed three distinct clades (Fig. 2A). The first bifurcation in the tree splits ape-infecting strains from human-infecting strains and represented the strongest axis of genetic variation on the PCA (Fig. 2B; explaining 28.92% of the total genetic variance). Therefore, the whole-genome ML phylogenetic tree showed that the human *P. vivax* strains formed a sister clade of ape *P. vivax-like* strains, unlike previous results based on single-gene phylogenies (mitochondrial and autosomal markers) (*33*, *35*). Liu *et al.* (*33*) and Loy *et al.* (*35*) suggested that *P. vivax* diversity was nested within the *P. vivax-like* radiation (*33*, *35*). Incongruences between whole-genome (*37*) and single-gene (*33*, *35*) tree topologies are common (*41*, *42*) and may be caused, for instance, by incomplete lineage sorting (ILS) (*43*), introgression, or selection at a specific locus. ILS denotes the persistence of ancestral polymorphisms across multiple successive speciation events, followed by stochastic allele fixation in each descendant lineage. This process leads to phylogenetic incongruences among individual loci (*44*, *45*). Disagreement between gene tree topologies may also be linked to the usage of mitochondrial markers that poorly represent the population history (*46*). Loy *et al.* (*35*) proposed that the shift in tree topologies—from sister clade to a nested relationship—was due to the incorporation of gorilla-infecting *P. vivax-like* samples and/or the inclusion of a new outgroup, *Plasmodium carteri*, a very rare parasite found in chimpanzees (*35*). However, the presence of two gorilla-infecting *P. vivax-like* lineages in our phylogeny partly contradicts this assertion. We could not build a whole-genome phylogeny using *P. carteri* sequences due to the unavailability of whole-genome sequencing data on this rare and elusive parasite. Nevertheless, our finding that *P. vivax* is a sister clade of *P. vivax-like* instead of a lineage included within the radiation of *P. vivax-like* challenges the view that *P. vivax* is included inside *P. vivax-like* diversity.

The second split in the phylogenomic tree (Fig. 2A) and on the PCA along the second PC axis (Fig. 2B; explaining 12.71% of the genetic variation) identified two distinct lineages among *P. vivax-like* strains (PVL.grp1 and PVL.grp2 in Fig. 2), composed of 5 and 14 strains, respectively. As *P. vivax-like* infects chimpanzees and gorillas, these two lineages might have been associated with host specialization. However, we found both lineages in both host species. Then, we tested the extent of recombination between these two genetically distinct *P. vivax-like* lineages to assess whether gene flow still occurs between them. We used the fastGEAR software (*47*) and the 19 *P. vivax-like* full-genome sequences to identify recently exchanged genomic fragments (i.e., fragments shared between two individuals) and their clade of origin. The fastGEAR clustering method recovered the two major *P. vivax-like* clades and identified 12,108 recently imported fragments, with a mean imported fragment length of 1775 bp (base pairs) (Fig. 2C). All the donor and recipient individuals of the recently imported fragments belonged to the same lineage, suggesting that no recent interlineage recombination occurred between the *P. vivax-like* clades. We confirmed this result by building a reticulation network for the 19 *P. vivax-like* individuals (fig. S6). The lack of reticulation between individuals belonging to different lineages confirmed the absence of recombination between the two *P. vivax-like* clades. Thus, our results demonstrate that although circulating in sympatry within the same host populations of great apes (i.e., the La Lékédi Park in Gabon), the two *P. vivax-like* lineages do not recombine, which suggests that they may represent distinct species. A similar subdivision was found in *Plasmodium praefalciparum*, the closest parasite to *P. falciparum*, but the status of these two clades was never investigated, as done here for *P. vivax-like* (*48*).

The nucleotide diversity (π) values of the two *P. vivax-like* lineages were approximately nine times higher than the value for

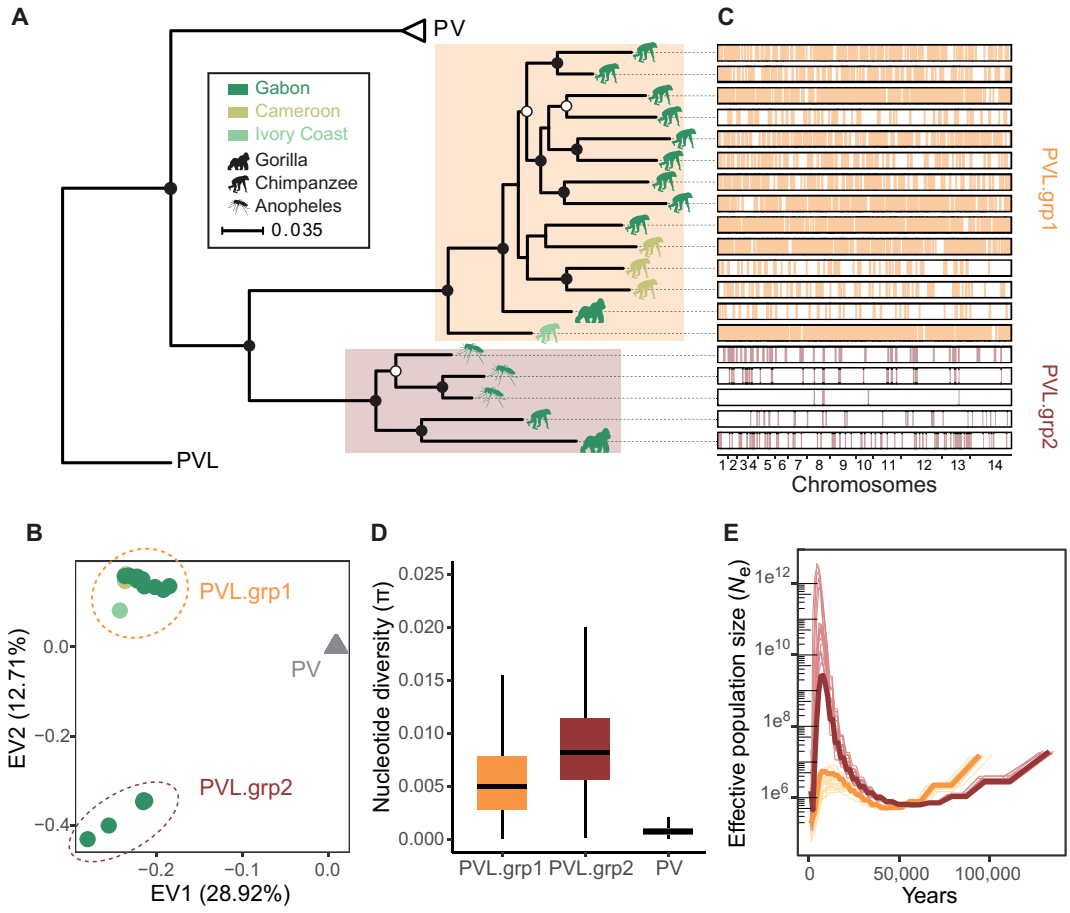

Fig. 2. *P. vivax-like* strains are structured in two distinct clades that form a sister monophyletic lineage to the human *P. vivax*. (**A**) ML phylogenetic tree illustrating the relationships between *P. vivax* (PV) and *P. vivax-like* (PVL) strains, rooted using *P. cynomolgi* (PC, M strain). Two distinct *P. vivax-like* groups were identified: PVL.grp1 (in orange) and PVL.grp2 (in brown). The animal pictograms on each leaf indicate the primate host, gorilla, chimpanzee, or mosquito (unknown primate host), colored according to the country of origin (Gabon, Cameroon, and Ivory Coast). Open and closed circles indicate nodes with >80 and >90% bootstrap support, respectively. (**B**) PCA displaying the two first eigen vectors (EVs) and the proportion of genetic variance they explain. (**C**) Genome-wide visualization of recent recombination events between *P. vivax-like* individuals. The horizontal black rectangles represent the 19 *P. vivax-like* genome sequences, and the vertical colored lines represent the recombining genomic segments on the recipient individual. Colors indicate the lineage membership of the donor individual. (**D**) Differences in nucleotide diversity (π) between *P. vivax-like* lineages and *P. vivax* strains. (**E**) Multiple sequentially Markovian coalescent (MSMC) estimates of the effective population size (N_e) in the two *P. vivax-like* groups (PVL.grp1 in orange and PVL.grp2 in brown). Lines in lighter color represent 50 bootstrap resampling replicates of randomly sampled segregating sites.

P. vivax (Fig. 2D), in agreement with previous reports (*35*, *37*). As it is unlikely that the mutation rates radically differ between *P. vivax-like* and *P. vivax* (*49*), this difference may reflect a higher effective size for *P. vivax-like* than for *P. vivax*. The lower nucleotide diversity observed in *P. vivax* could result from a bottleneck effect that occurred during the host shift, when the pathogen colonized humans (*35*). Similar differences were observed between African apes and humans (*50*). Genetic diversity is much lower in humans than in African apes due to the bottleneck effect linked to the modern human populations' expansion out of Africa from a small number of founders who replaced the archaic forms of humans (e.g., Neanderthals). This could suggest a similar history for *P. vivax* in humans, and Africa as the point of origin of human *P. vivax*, as recently proposed (*33*). However, within the *P. vivax-like* samples, genetic diversity was higher in the PVL.grp1 than PVL.grp2 lineage. This may reflect sampling differences because PVL.grp1 isolates were collected in three African countries, whereas PVL.grp2 isolates came from a single location in the Park of La Lékédi in Gabon. The identification of two new *Plasmodium* lineages in African great apes, in such a limited geographic range, highlights the putative underestimated current and ancestral diversity of this genus in the African continent.

Last, we inferred historical fluctuations in the effective population size (N_e) for the two *P. vivax-like* lineages using a multiple sequentially Markovian coalescent (MSMC) approach. The analysis of genome variations indicated an ancient major expansion in genetic diversity for both clades, followed by a recent important decline (Fig. 2E). Unexpectedly, N_e increase was much higher for the PVL.grp2 lineage than for the PVL.grp1 lineage. The diverging trends in N_e values observed in each lineage (about 50 thousand years ago) suggest a putative ancient split of the two *P. vivax-like* lineages. This is consistent with the lack of recombination between these lineages and further supports their reproductive isolation.

Genetic structure of human *P. vivax* isolates collected worldwide

To gain insight into the genetic relationships among *P. vivax* populations around the world, we investigated the genetic structure of

the 447 *P. vivax* isolates collected from 21 countries. We analyzed biallelic SNPs from the core region of the genome of unrelated samples (IBD of ≤0.7) using complementary population genomic approaches: a PCA that does not rely on any model assumptions (*51*), a model-based individual ancestry analysis implemented in the ADMIXTURE software (*52*), an ML phylogenetic tree of individual strains, and an analysis of the amount of genetic differentiation among population pairs using the F_{ST} index.

All analyses revealed consistent patterns that split the genetic variation primarily by continent (Fig. 3). The first principal component (PC) of the genetic variation (EV1) split Southeast Asian from African and American *P. vivax* populations, while the second axis (EV2) split African populations from the rest of the world (Fig. 3A and fig. S7). These three major geographical clusters were also identified on the ML tree (Fig. 3B) and displayed clearly distinct genetic ancestries, as estimated by simulating three ancestral clusters (K = 3) with the ADMIXTURE analysis (fig. S8). Within each cluster/continent, we observed additional subdivisions. The ADMIXTURE analysis (figs. S8 and S9) detected six main distinct genetic pools (three distinct ancestral populations in Asia, one in Africa, and two in America) (Fig. 3, C and D).

Strains from India and Sri Lanka were genetically closer to African *P. vivax* populations than to other neighboring Asiatic groups (Fig. 3, A, B, and D). Such genetic exchange may result from the different trades that took place between the Indian subcontinent and Africa over the last two millennia. Moreover, movements of populations, and of diseases, occurred from India to East Africa and vice versa (*8*). The most recent exchange has been the migration of

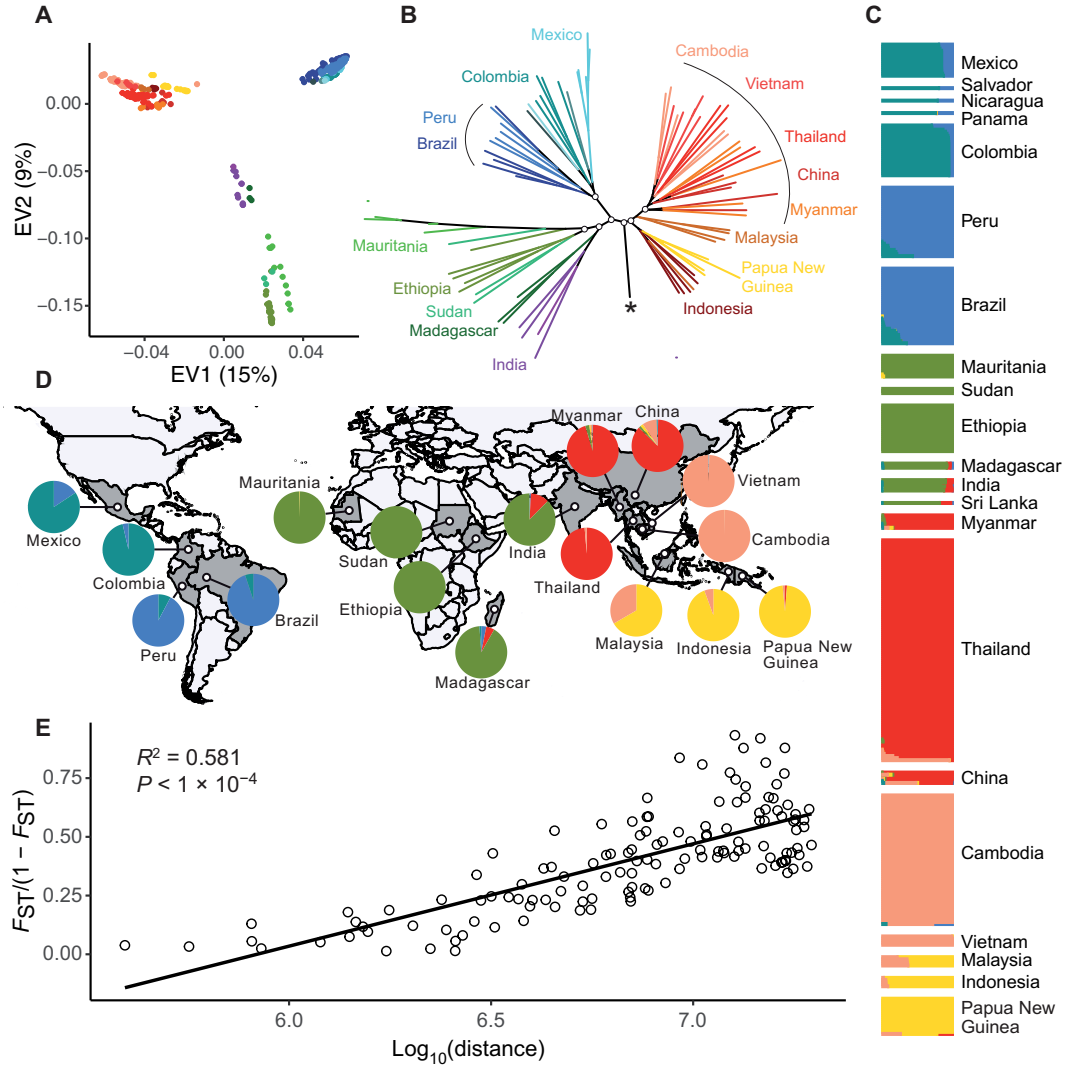

Fig. 3. The structure of *P. vivax* core region of the genome is mostly due to genetic isolation of natural populations. (**A**) PCA using the SNPs from the core region of the genome of the 447 *P. vivax* isolates showing the top two EVs and the accounted proportion of genetic variance. (**B**) ML phylogenetic tree built using SNP data from the core region of the genome. The star indicates the root of the tree, leading to the *P. vivax-like* outgroup. Specific internal nodes are highlighted with open circles and represent >80% bootstrap support. (**C**) The genetic ancestry proportion for individual strains, depicted as a vertical bar, was estimated using ADMIXTURE for each of the K = 6 inferred ancestral populations. (**D**) These same ancestral proportions for each population are displayed as pie charts on the world map. (**E**) Isolation by distance among populations. Pairwise estimates of $F_{ST}/1 - F_{ST}$ were plotted against the corresponding geographical distances between countries. The Spearman correlation coefficient (R^2) and the P value estimated using a Mantel test with 1000 permutations are shown.

the human Karana population from the northwest of India to Madagascar at the end of the 17th century (*53*).

Across the Americas, *P. vivax* strains were structured in two distinct ancestral populations (Fig. 3, B to D). This is consistent with two putative evolutionary scenarios of the parasite in America. The first scenario hypothesizes that both ancestral populations originated from the isolation between Amazonian and non-Amazonian populations, after a single introduction during the European colonization in the 15th century (*54*). In the second scenario, two distinct waves of introduction occurred from the same or from different sources (*15*). Disentangling these two hypotheses will require more samples and statistical population genetic modeling (*55*).

P. vivax individuals collected in the Southeast Asia/Pacific region were structured in three distinct ancestral populations. In each ancestral population, the genetic differentiation between countries, estimated with the F_{ST} index, was weak (fig. S10), consistent with a relatively unrestricted gene flow between countries. Among these three ancestral populations, one was shared by *P. vivax* populations located in Indonesia and Papua New Guinea and was explained by their insular isolation. The other two ancestral populations may reflect the effect of the malaria-free corridor previously established through central Thailand, consistent with observations in *P. falciparum* (*56*). The Malaysian population was an admixture between two ancestral populations, at a contact zone in which admixture proportions progressively changed from one cluster to the other.

Last, at the global scale, we observed a very significant relationship between the genetic differentiation among populations [estimated using $F_{ST}/(1 - F_{ST})$] and the logarithm of the geographic distance (Mantel test $r^2 = 0.56$, $P < 10^{-4}$; Fig. 3E). This is consistent with an isolation by distance in a two-dimensional habitat and suggests limited *P. vivax* dispersal across space (*57*).

Genetic evidences in favor of an Asian origin of human *P. vivax*

If *P. vivax* originated in Africa (*33*, *35*), the phylogeny of *Plasmodium* individuals should have displayed a topology where the *P. vivax-like* root was located within or near the monophyletic clade of African *P. vivax* individuals. However, our ML tree based on whole-genome SNP data did not display such topology (Fig. 3B). On the contrary, the branching of *P. vivax-like* lineages splits *P. vivax* populations from Asia on one side and from India, Africa, and America on the other side. The most common ancestor to all *P. vivax* populations was closely related to populations from Indonesia and Papua New Guinea (Fig. 3B and fig. S11). Although *P. vivax-like* is the closest outgroup to all *P. vivax* populations, the exact branching position of *P. vivax-like* into *P. vivax* populations could be affected by various caveats: (i) the relatively large genetic divergence between *P. vivax* and *P. vivax-like* (net average genetic divergence $d_a = 0.43 \pm 0.051$) compared with the divergence among *P. vivax* populations (average $d_a = 0.015 \pm 0.0025$) could make *P. vivax-like* branching position uncertain; (ii) at the recent evolutionary scale, ILS and gene flow among populations could also influence the tree topology of *P. vivax* populations, making it more reticulated than a simple model of bifurcating populations.

Therefore, we estimated the historical relationships among *P. vivax* populations by modeling shared and population-specific genetic drifts among populations along a reticulated evolutionary tree using TreeMix (*58*). This allowed us to model how the sampled populations were related to their common ancestor through a graph of ancestral populations that included split and migration edges. The ML population tree topology rooted with *P. vivax-like* without migrations (fig. S12) displayed a similar topology as the individual-based ML phylogenetic tree (Fig. 3B). The addition of sequential migration events increased significantly the model likelihood with an optimum found with four migration edges (fig. S13). In this model, the TreeMix graph suggested that gene flow or admixture events might have occurred between neighboring territories (i.e., between Myanmar and India) (Fig. 4, A and B). Unlike the model without migration events, the four migration edges changed slightly the branching of the tree root and placed the most recent common ancestor to all *P. vivax* populations within populations from Southeast Asia. Despite the low node support (≤50% of the bootstrap replicates) on the branching of the ancestral group leading to China, Thailand, Vietnam, and Cambodia, these results are still consistent with the

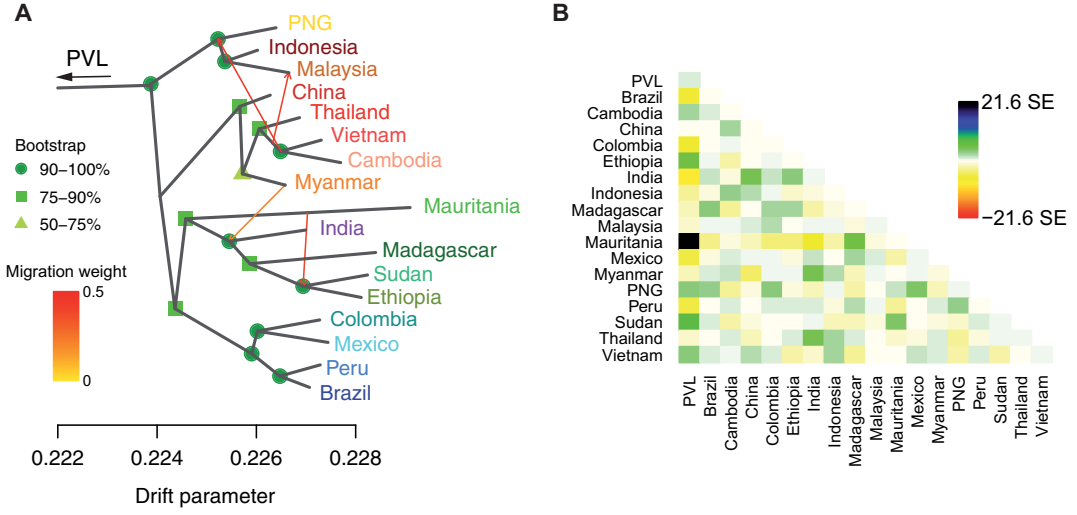

Fig. 4. Relationships and gene flow between *P. vivax* populations. (**A**) TreeMix ML tree of *P. vivax* populations with four migration edges (arrows) rooted with *P. vivax-like* (PVL), including bootstrap node support. (**B**) TreeMix residual matrix from the tree in (A). PNG, Papua New Guinea.

ML trees showing that the outgroup branch was closer to Southeast Asian populations than to African or American populations. This suggests that *P. vivax* Asian populations have kept a higher proportion of ancestral alleles compared with the African and American populations, which appeared more derived. However, similar to the ML phylogenetic analysis, the large genetic distances of the *P. vivax-like* outgroup call for caution in the placement of the root in *P. vivax* population tree built by TreeMix.

To further explore *P. vivax* origin, we analyzed the distribution of genetic variability among populations in a spatial context. In a model of population range expansion and progressive colonization of new areas from a unique origin with serial founder events, several genetic trends are expected as a function of the increasing geographical distance from the origin (*59*–*62*): (i) The population genetic diversity should decline; (ii) the linkage disequilibrium (LD) should increase; (iii) the ancestral allele frequency (AAF) spectrum should flatten, indicating that derived alleles tend to be more frequent in populations at a greater distance; and (iv) the time to the most recent common ancestor (TMRCA) estimates within each population should decrease with the distance from the origin.

First, analysis of the geographical distribution of nucleotide diversity (π) showed that most *P. vivax* populations from Southeast Asia exhibited the highest genetic diversities, while African populations and American populations displayed intermediate and low values, respectively (figs. S14 and S15). Moreover, we observed the highest decrease in genetic diversity with increasing geographic distance when we considered Southeast Asia as the putative origin of *P. vivax* (Fig. 5, A and B). However, this geographic trend in genetic diversity could also be related to variations in the local population size or to very recent population dynamics (e.g., malaria eradication treatments, such as artemisinin). As the reported cases of *P. vivax* infections are higher in Asia than in the rest of the world [World Health Organization 2010 report (*63*)], we used the estimated number of cases as a covariate in our model. Nevertheless, the correlation between genetic diversity and geographic distance from Asia remained highly significant after controlling for the number of infections [generalized linear models (GLM), $P < 2.37 \times 10^{-5}$].

Next, analysis of LD variations among *P. vivax* populations showed a very strong positive correlation between LD at 100 bp within each population and the geographic distance to the previously identified putative Asian origin (Fig. 5C and fig. S15). This pattern remained significant when controlling for variations in the number of cases per country (GLM, $P = 0.0396$).

The patterns of AAF distributions can also inform on *P. vivax* origin and its worldwide colonization routes (*60*). Using *P. vivax-like* and *Plasmodium cynomolgi* as outgroups, we could polarize the allele frequency spectrum and identify the ancestral and derived allelic states for ~94,000 SNPs. Figure 5D shows the corresponding AAF spectrum when considering an increasing sample size (5, 7, 10, and 20 genomes) in each population to compare our heterogeneously sized populations. Regardless of the number of full-genome sequences considered, populations located in Southeast Asia (red-orange shades) always displayed more SNPs with high AAFs and fewer SNPs with low AAFs. Conversely, in non-Asian populations, AAF distributions progressively flattened. This is consistent with the phylogenetic and population tree analyses done with TreeMix (Figs. 3B and 4A).

Last, we used the MSMC approach and five individuals from 14 *P. vivax* populations to obtain estimates of the historical population size changes and of the TMRCA. The MSMC curves revealed that *P. vivax* populations displayed two effective population size dynamics (Fig. 6A). The effective population sizes (N_e) of Southeast Asian populations were relatively stable, and some experienced a steady increase in a more distant past, followed by a quick decline. Conversely, in African and American populations, N_e have mainly steadily decreased since the TMRCA. The MSMC analysis showed that most populations from Southeast Asia exhibited much older TMRCA values than the African and American populations (Fig. 6B and fig. S16).

Together, these results indicate that *P. vivax* populations display a decrease in genetic diversity and an increase in LD with the increasing geographic distance from Southeast Asia, and are consistent with a historical model of colonization with serial founder events from an origin located in Southeast Asia. The higher proportion of ancestral alleles and older TMRCA found in the Southeast Asian populations are additional evidences supporting this scenario. These four different analyses constitute only two independent lines of evidence because our genetic trends based on π, LD, and TMRCA are interrelated and rely on genetic diversity. On the basis of *P. vivax* and *P. vivax-like* phylogenetic position within a radiation of *Plasmodium* parasites infecting Asian monkeys, we hypothesize that these two parasite lineages resulted from two independent waves of colonization from Asia. The first wave gave rise to *P. vivax-like* in African apes (with some unknown host species bridging Africa to Asia), and the second wave gave rise to the current *P. vivax* populations.

However, our conclusions need to be considered with caution for the following reasons. Although population genetic theories offer testable hypotheses to investigate *P. vivax* origin, the recent epidemiological history of each population might have influenced the local patterns of genetic variation. For instance, malaria eradication campaigns, such as the programs carried out in the 1950s and 1960s to stop malaria transmission (*64*), could have generated strong bottlenecks that affected local populations, thus preventing us from accurately inferring the evolutionary history of the affected populations based on their summary statistics. Populations from Mexico, Mauritania, and Papua New Guinea may be examples of populations in which the high level of recent common ancestry is associated with low genetic diversity and high LD. To take this issue into account, we corrected all our genetic trends for the current demographical variations between populations. As all correlations remained significant, it is unlikely that the local epidemiology might have affected the full global patterns at the world scale (i.e., local epidemiology is different for each *P. vivax* population). Furthermore, we used the powerful MCMC2 method to trace the evolutionary history of the entire population back to the most recent common ancestor, without being negatively influenced by very recent demographic changes.

An additional limitation of our study was the heterogeneity in sample sizes and sample collection schemes that prevented us to compute summary statistics with the same resolution for all populations. Concerning sample sizes, we calculated nucleotide diversity and LD for populations with at least five individuals, as a trade-off between including the highest possible number of populations and minimizing the bias in nucleotide diversity and LD estimation. Consequently, as presented in our rarefaction analysis, small-size populations exhibited high variation around their expectation. We think that this issue is not a major limitation because the large number of used genetic markers should offset the small sample sizes (*65*, *66*). Nevertheless, increasing the size of the small populations and collecting new populations, especially those lying between the

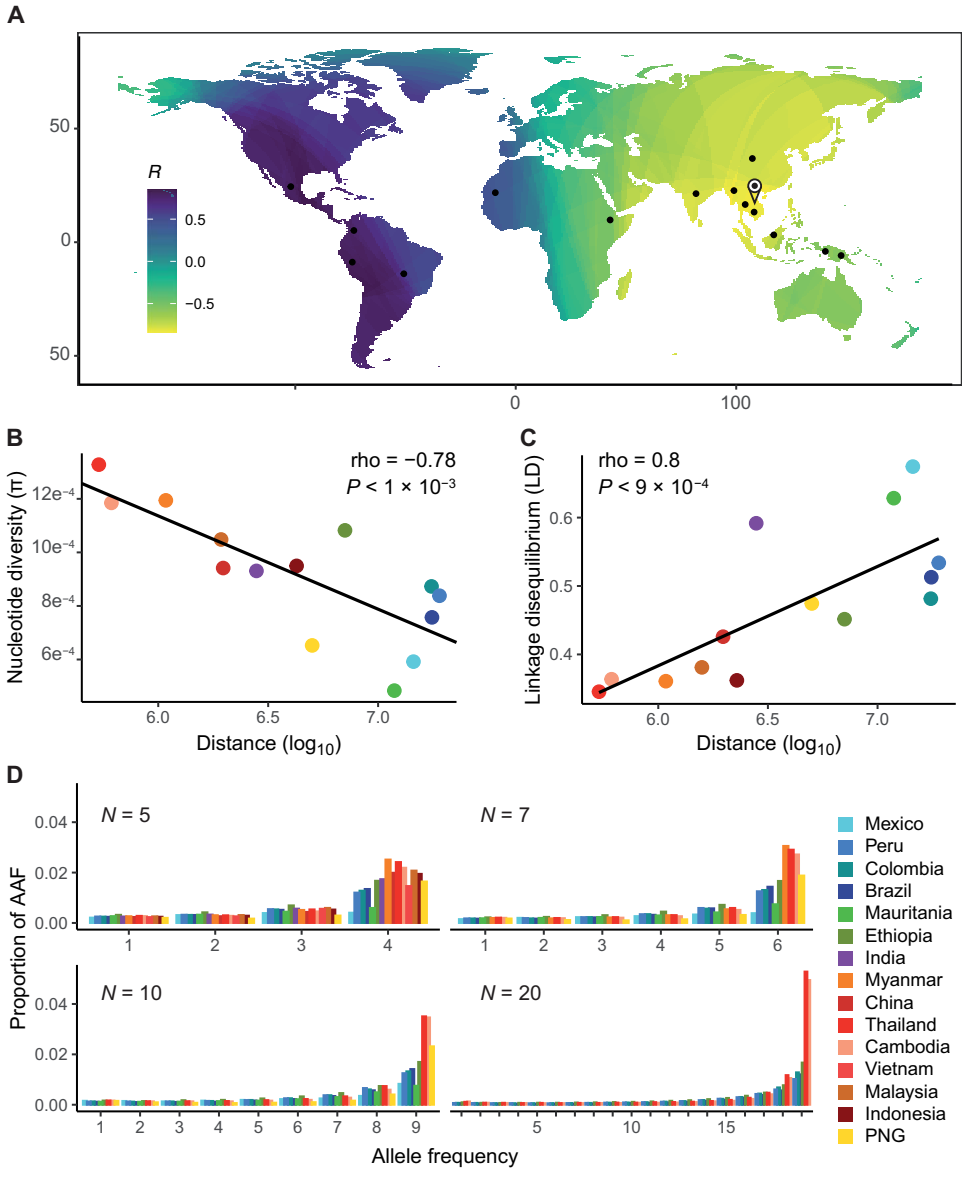

Fig. 5. The Southeast Asian origin of *P. vivax* is supported by patterns of nucleotide diversity, LD, and AAF. (**A**) Genetic diversity of *P. vivax* regressed on the great circle distance across the world. The value at each pixel of the map corresponds to the Spearman correlation coefficient (*R*) between the expected genetic diversity in each population and the geographic distance between this population and the pixel. Black dots represent the sampling sites used in the regression analysis (where $n \geq 5$ individuals). The landmark on the map represents the most negative correlation coefficient indicative of the most probable source of the range expansion. (**B**) Correlation between nucleotide diversity (π) and geographic distance, measured as the distance between the milestone indicated in (A) and each population. (**C**) Correlation between LD at 100 bp (measured as the normalized R^2) and the geographic distance. (**D**) Histograms of AAF proportions in *P. vivax* populations with an increasing sample size ($N = 5$, 7, 10, and 20).

supposed source populations in Asia and in Africa (e.g., populations from the Middle East), would be particularly useful to better characterize the patterns of *P. vivax* genetic variation worldwide and to refine the question of the origin of *P. vivax* range expansion. In addition, collecting samples using the same method, at the same time and during the same period, and in similar geographic areas would be the best sampling approach to study the evolutionary origin and history of a pathogen. Unfortunately, such ideal sampling is currently unavailable and extremely complex to put in place. Therefore, this study presents results obtained using up-to-date population genomic tools that can efficiently identify the origin of organisms (e.g., *P. falciparum*) (*61*). Last, to determine the impact of closely related isolates on nucleotide diversity, LD, and TMRCA estimations, we analyzed our dataset using a stringent IBD cutoff of 0.25 (instead of 0.7). This new analysis gave similar patterns as before, with a stronger coefficient of regression (figs. S17 and S18).

Loy *et al.* (*35*) proposed an alternative hypothesis according to which *P. vivax* spread from Africa to Asia, was then eliminated from Africa due to the evolution of Duffy negativity, and was reintroduced in Africa by more recent migration waves from Asia.

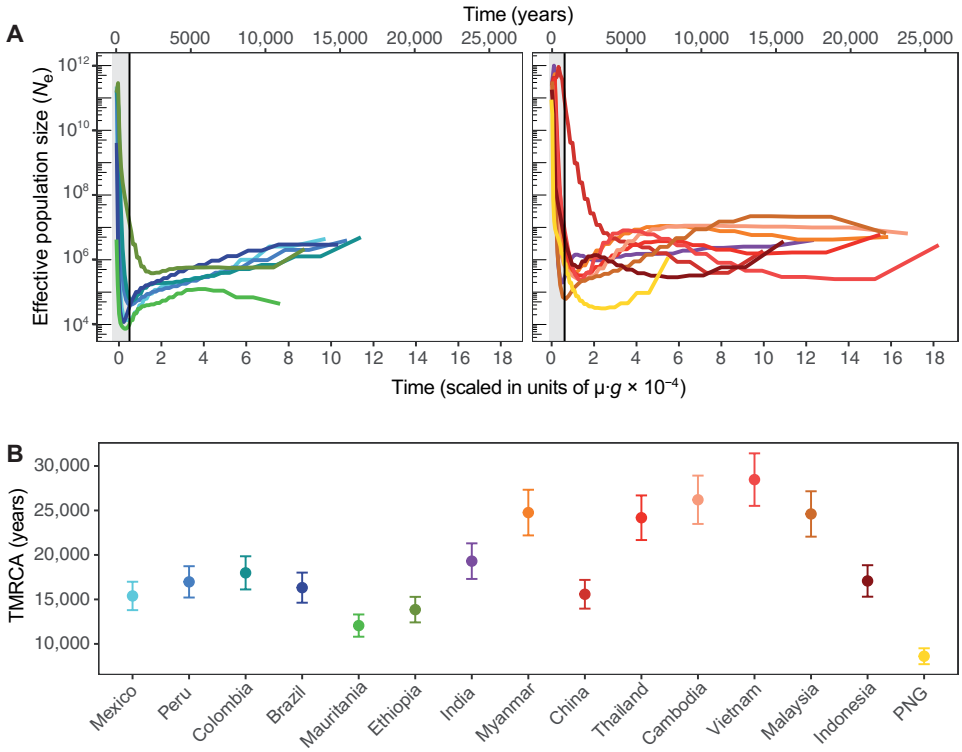

Fig. 6. Coalescent-based inference of the demographic history in each *P. vivax* population. (**A**) Effective population size (N_e) variations back to the TMRCA inferred using MSMC and (**B**) inferred TMRCA in each population. The analysis was performed using five individuals from each population, assuming a mutation rate per generation (μ·*g*) of 1.158×10^{-9} and a generation time (*g*) of 0.18. The gray area represents the first 1000 generations, at which low resolution of the recent past effective population size is observed, likely due to the presence of sites with high probability of being called inaccurately.

Although this scenario provides a potential explanation of Duffy negativity appearance in Africa, unfortunately it does not fit with the genetic variability patterns of *P. vivax* populations. According to the scenario by Loy *et al.* (*35*), populations from the Middle East and India (located on the migratory route between Africa and Southeast Asia) should display traces of admixture from both migration waves and higher genetic diversity. Our current results on the genetic variation of populations from India (Figs. 3C and 4B) and from the Middle East (*29*) do not support this scenario. Nevertheless, this hypothesis remains possible if all ancient populations disappeared from the world, except the Southeast Asian populations. However, testing this scenario can be very difficult in the absence of ancient *P. vivax* strains to study.

This study provides one of the most detailed views of the worldwide distribution of *P. vivax* population genetic diversity and demographic history. Our work investigated not only the phylogenetic relationship between *P. vivax* infecting human and apes but also different features of *P. vivax* genetic variation worldwide to test different hypotheses of an African or Asian origin. We showed that *P. vivax* is a sister group and not a sublineage of *P. vivax-like*. Genetic diversity in *P. vivax-like* is richer than in *P. vivax*, consistent with a strong bottleneck in this lineage that gave rise to *P. vivax*. Our results based on whole-genome sequencing data support an out-of-Asia origin, rather than an African origin, for the world populations of *P. vivax*, with a signal of stepping-stone colonization events accompanied by serial founder effects. Efforts should now be focused on precisely describing *P. vivax* demographic and selective history in specific regions (e.g., South America, Africa, and Europe) and on estimating when and how these different regions were colonized. The question of the host of origin still remains open.

MATERIALS AND METHODS

African *P. vivax* sample collection and ethical statements

Twenty *P. vivax* isolates from Mauritania ($N = 14$), Ethiopia ($N = 3$), and Sudan ($N = 3$) were sequenced. *P. vivax* infection was detected by microscopy analysis, polymerase chain reaction amplification of the *cytochrome b* gene, and/or rapid diagnostic test. Samples were collected from *P. vivax*–infected patients after informed consent and ethical approval by the local institutional review board of each country. The informed consent procedure for the study consisted in the presentation of the study aims to the community followed by inviting adults for enrollment. At the time of sample collection, the study purpose and design were explained to each individual and a study information sheet was provided before collection of the oral informed consent. Only oral consent was required because the study did not present any harm for the subjects and did not involve procedures that needed a written consent. The oral consent process was consistent with the ethical expectations for each country at the time of enrollment and was approved by each country, and these procedures were approved by local ethics committees. The privacy and confidentiality of the collected data were ensured by the sample anonymization before the study start. For samples from Mauritania, the study was approved by the pediatric services of the National

Hospital, the Cheikh Zayed Hospital, and the Direction régionale à l'Action Sanitaire de Nouakchott (DRAS)/Ministry of Health in Mauritania. No ethics approval number was given. For samples from Sudan, no specific consent was required; the human clinical, epidemiological, and biological data were collected in the Centre National de Référénce du Paludisme (CNRP) database and analyzed in accordance with the common public health mission of all French National Reference Centers (www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000810056&dateTexte=&categorieLien=id). The study of biological samples obtained in the framework of medical care was considered as noninterventional research (article L1221-1.1 of the French public health code) and only required the patient's non-opposition during sampling (article L1211-2 of the French public health code). All data collected were anonymized before analyses. For samples from Ethiopia, this study was approved by the Ethical Clearance Committee of the Haramaya University—College of Health and Medical Sciences, and from the Harari and Oromia Regional State Health Bureau in Ethiopia.

P. vivax-like sample collection and ethical statements

P. vivax-like samples were obtained during a continuous survey of great ape *Plasmodium* infections carried out in the Park of La Lékédi, in Gabon, in collaboration with the Centre International de Recherches Médicales de Franceville (CIRMF). The La Lékédi Park is a sanctuary for ape orphans in Gabon. During sanitary controls, blood samples were collected and leukocyte depletion was carried by CF11 cellulose column filtration (*67*) in the field, before storage at −20°C at the CIRMF. The animal well-being was guaranteed by the veterinarians of the "Park of La Lékédi" and the CIRMF who were responsible for all sanitary procedures (including blood collection). All animal work was performed according to the relevant national and international guidelines. These investigations were approved by the government of the Republic of Gabon and by the Animal Life Administration of Libreville, Gabon (no. CITES 00956). Our study did not involve randomization or blinding.

P. vivax-like samples were also obtained from sylvatic anopheles mosquitoes collected with Centers for Disease Control (CDC) light traps in the forest of the Park of La Lékédi during a longitudinal study (*34*). Anopheles mosquitoes were morphologically identified using standard keys (*68*), stored in liquid nitrogen, and then kept at −80°C at the CIRMF until analysis.

Genomic DNA was extracted from each sample using the DNeasy Blood and Tissue Kit (Qiagen, France) according to the manufacturer's recommendations. *P. vivax-like* samples were identified by amplifying and sequencing the *Plasmodium cytochrome b* (*Cytb*) or *cytochrome oxidase 1* (*Cox1*) gene, as described elsewhere (*24*, *69*). In total, 10 *P. vivax-like* samples were selected: 3 from gorillas, 4 from chimpanzees, and 3 from anopheles mosquitoes.

African *P. vivax* and *P. vivax-like* genome sequencing

To avoid host DNA contamination, selective whole-genome amplification (sWGA) was used to enrich submicroscopic DNA levels, as already described elsewhere (*70*). This technique preferentially amplifies *P. vivax* and *P. vivax-like* genomes from a set of target DNAs. For each sample, DNA amplification was carried out using the strand-displacing phi29 DNA polymerase and two sets of *P. vivax*–specific primers that target short (6 to 12 nucleotides) motifs commonly found in the *P. vivax* genome (set1920 and PvSet1) (*69*, *70*). For each set of primers, 30 ng of input DNA was added to a 50-µl reaction mixture containing 3.5 µM of each sWGA primer, 30 U of phi29 DNA polymerase enzyme (New England Biolabs), 1× phi29 buffer (New England Biolabs), 4 mM deoxynucleoside triphosphates (Invitrogen), 1% bovine serum albumin, and sterile water. DNA amplifications were carried out in a thermal cycler with the following program: a ramp down from 35° to 30°C (10 min per degree), 16 hours at 30°C, 10 min at 65°C, and hold at 4°C. For each sample, the products of the two amplifications (one per primer set) were purified with AMPure XP beads (Beckman Coulter) at a 1:1 ratio according to the manufacturer's recommendations and pooled at equimolar concentrations. Last, each sWGA library was prepared using the two pooled amplification products and a Nextera XT DNA kit (Illumina) according to the manufacturer's protocol. Samples were then pooled and clustered on a HiSeq 2500 sequencer in Rapid Run mode with 2 × 250-bp paired end reads.

Genomic data from public and/or published data archives

For the worldwide comparative study of *P. vivax* genetic diversity, population structure, and evolution, a large literature search was carried out to identify previously published genomic datasets. This allowed the identification and extraction of the fastq files of 1134 *P. vivax* isolates from the following 12 projects: PRJEB10888 (*21*), PRJEB2140 (*19*), PRJNA175266 (*71*), PRJNA240356 (*20*), PRJNA240452, PRJNA240531, PRJNA271480 (*20*), PRJNA284437 (*72*), PRJNA350554 (*73*), PRJNA420510 (*74*), PRJNA432819 (*75*), and PRJNA65119 (*76*). In addition, 17 *P. vivax-like* sequenced genomes from Cameroon, Ivory Coast, and Gabon were from two other projects (PRJNA474492 and PRJEB2579) (*35*, *37*). Published genomes were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (*77*) and converted into fastq files using the NCBI SRA Toolkit (fastq-dump –split-3). Details about the genomic samples are provided in table S1.

P. vivax and *P. vivax-like* read mapping and variant calling

Newly generated and previously published sequencing reads were trimmed to remove adapters and preprocessed to eliminate low-quality reads (--quality-cutoff=30) using the program cutadapt (*78*). Reads shorter than 50 bp and containing "N" were discarded (--minimum-length=50 --max-n=0). Sequenced reads were aligned to the PVP01 (*79*) *P. vivax* reference genome using bwa-mem (*80*). Here, a first filter was applied to exclude isolates with an average genome coverage depth lower than 5×. The Genome Analysis Toolkit (GATK, version 3.8.0) (*81*) was used to identify SNPs in each isolate following the GATK best practices. Duplicate reads were marked using the MarkDuplicates tool from the Picard tools (*82*) with default options. Local realignment around indels was performed using the IndelRealigner tool from GATK. Variant calling was carried out using the HaplotypeCaller module in GATK on reads mapped with a "reads minimum mapping quality" of 30 (-mmq 30) and a minimum base quality of >20 (--min_base_quality_score 20). Low-quality SNPs were eliminated according to GATK best-practice recommendations, and SNPs at sites covered by at least five reads were kept. Last, the VCF (variant call format) files for the different isolates were merged using the GATK module GenotypeGVCFs. Because of our tolerant genome coverage depth cutoff (5×), a second round of data filtering was carried out to exclude variants and isolates with a missing call rate higher than 50%. The within-host infection complexity was assessed by calculating the F_{WS} (*40*) with the R package moimix (http://bahlolab.github.io/moimix/). An F_{WS} threshold of >0.95 was used as a proxy of monoclonal infection.

Heterogenous patterns of relatedness among *Plasmodium* genomes

Highly related samples and clones can generate spurious population structure signals, bias estimators of population genetic variation, and violate the assumptions of the model-based population genetic approaches used in this study (e.g., ADMIXTURE, TreeMix, and MSMC2) (*83*). Therefore, the relatedness between haploid genotype pairs was measured by estimating the pairwise IBD between strains within populations using the hmmIBD program, with the default parameters for recombination rate, and genotyping error rate, and using the allele frequencies estimated by the program (*84*). Isolate pairs that shared >70% of IBD were considered highly related. Only one strain of each related pair was kept for the downstream analyses.

To identify heterogenous patterns of relatedness among *Plasmodium* genomes, a local PCA was performed, as described by Li and Ralph (*39*). Briefly, the *P. vivax* genome was divided in 1439 contiguous and nonoverlapping windows of 100 SNPs. On each window, the PCA was applied and the individual isolate scores for the first two PC were stored. To measure the similarity among windows, a Euclidian distance matrix was computed among windows based on the PC scores. Last, multidimensional scaling (MDS) was used to visualize the relationships among windows. A set of three coordinates was used to visualize the patterns of relatedness shared by windows. Because the local PCA was sensitive to missing data, the top 304 *P. vivax* genomes with an SNP missing discovery rate lower than 25% were selected for this analysis.

Analysis of genetic recombination among *P. vivax-like* genomes

Recombination among *P. vivax-like* strains was inferred using the fastGEAR software (*47*). From the SNPs present in the core region of the genomes, a multi-fasta file was created and used as input for fastGEAR that was launched with the iteration number set to 15 (default). Recent recombination events were detected with the Bayesian factor (BF) > 1 (default) and referred to as interlineage recombination for which the donor-recipient relation can be inferred. A phylogenetic network was generated using the SplitsTree 4 software (*85*) and the polymorphic sites present in the core region of the genome.

Phylogenetic analysis

ML phylogenetic trees were inferred with IQ-TREE (*86*) using the general time reversible (GTR) model with ascertainment bias correction. The amount of missing data present in the genome alignment was minimized by selecting 85 *P. vivax* individuals with the lowest missing call rate (up to 5 individuals per population). For *P. vivax-like*, all 19 isolates were used in the phylogenetic analysis presented in Fig. 2A, but only the 12 isolates with a missing call rate lower than 5% were used for the phylogeny presented in Fig. 3B. Consequently, the results presented in Fig. 2A were inferred on the basis of 28,000 SNPs, and the results in Fig. 3B were inferred on the basis of 87,000 SNPs. Phylogenies were plotted with FigTree v.1.4.3 (*87*), and the node reliability was assessed by performing 100 bootstrap replicates. *P. cynomolgi* [strain M (*88*)] was used as outgroup, and *P. cynomolgi* genomic coordinates were converted to *P. vivax* coordinates using the liftOver genome tool (*89*).

Population structure analysis and population evolutionary history

PCA and ADMIXTURE analyses were performed after selecting only biallelic SNPs present in the *P. vivax* core region of the genome and excluding singletons from the dataset. The variants were LD-pruned to obtain a set of unlinked variants using the option --indep-pairwise 50 5 0.7 in PLINK (*90*). PCA was performed using PLINK and plotted in R. The population structure and the estimation of individual genetic ancestry to various numbers (K) of ancestral populations were performed using ADMIXTURE (*52*). Each ADMIXTURE analysis was repeated 100 times using a subset of 100,000 SNPs randomly selected throughout the core region of the genome. Each run was launched with a random seed, with a *K* value ranging from 2 to 20. The most likely number of ancestral populations (*K*) was determined using the cross-validation error rate (*52*). Then, CLUMPAK (*91*) was used to analyze ADMIXTURE outputs and compute ancestry proportions.

Genetic differentiation among populations was estimated with the Weir and Cockerham's estimator of the F_{ST} index using VCFtools (*92*). This metric accounts for sample size differences among populations.

Demographic history analysis and range expansion

TreeMix was used to investigate the historical population relationships by estimating an ML population tree, the amount of genetic drift in each population, and the number of migration events that best fitted the data (*58*). The number of migration events (*m*) that best fitted the data was calculated by running TreeMix 15 times for each *m* value, with *m* ranging from 1 to 12. LD among SNPs was taken into account by grouping SNPs by blocks of 500 (-k 500). The optimal *m* value ($m = 4$) was estimated using the OptM R package (https://cran.r-project.org/web/packages/OptM/index.html). Then, a consensus ML tree including bootstrap node support was obtained by running TreeMix 100 times ($m = 4$ -k 500) and post-processing using the BITE R package (*93*).

To determine which model (out-of-Asia or out-of-Africa) of population range expansion best described the demographic history of *P. vivax*, summary statistics were computed to capture different features of genetic variation within populations. Each analysis was computed for each of the populations defined by their country of origin using the SNPs present in the reference core region of the genome. To minimize the impact of missing calls in our analyses, only individuals with <5% of missing data were included in the following analysis. The genome-wide nucleotide diversity (π) was calculated per site using VCFtools and divided by the number of nucleotides in the core region of the genome (20,844,131 bp) to derive an approximation of the genome-wide average nucleotide diversity (π) for each population, as described in (*21*, *22*). To take into account the sample size heterogeneity effect on π estimation, the rarefaction approach was used with five genomes resampled in each population to estimate and compare values among populations. This number provided a trade-off between minimizing the π value estimation variance due to sample size heterogeneity and keeping as many populations as possible in the analysis (fig. S14).

The nucleotide diversity regression in function of the geographic distance was calculated by dividing the world map into a 400 pixel × 500 pixel two-dimensional lattice and by considering each pixel as a potential source of *P. vivax* geographic range expansion. The geographic distances between populations and the focal pixel were determined using the R package geosphere to calculate the Spearman correlation coefficient between nucleotide diversity and geographic distance. The pixel with the lowest negative correlation coefficient was considered as indicative of the origin of the expansion range (*94*, *95*).

LD decay was estimated by randomly sampling five individuals from each population using the PopLDdecay tool (*96*) that calculates

the genotype correlation coefficient R^2 for pairs of SNPs at a maximum distance range of 10 kb. Regression between LD at 100 bp and the geographic distance to the potential source was calculated using the R package geosphere. Last, the biallelic SNPs of each *P. vivax* population were polarized using the shared alleles identified between *P. vivax-like* individuals and *P. cynomolgi* (custom scripts). The final unfolded site frequency spectrum (SFS) was computed for a total of 93,652 SNPs.

Estimates of historical variation in effective population size

The MSMC2 program (*97*) and the core region of the genome of all chromosomes were used to infer historical changes in N_e and to estimate the TMRCA in the *P. vivax* and *P. vivax-like* populations. Using the mutation rate (μ) per generation (g) previously estimated for *P. falciparum* (*49*), the estimates were scaled to years and N_e, assuming $\mu \cdot g = 1.158 \times 10^{-9}$ and a generation time of $g = 0.18$. MSMC2 was run for 20 iterations with a fixed recombination rate. The MSMC2 method was applied by randomly selecting five individuals with a missing call rate lower than 5% from each population. The error around the parameter estimates was estimated using 50 bootstrapped replicates by randomly resampling the segregating sites used in the analysis.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/18/eabc3713/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. K. E. Battle, T. C. D. Lucas, M. Nguyen, R. E. Howes, A. K. Nandi, K. A. Twohig, D. A. Pfeffer, E. Cameron, P. C. Rao, D. Casey, H. S. Gibson, J. A. Rozier, U. Dalrymple, S. H. Keddie, E. L. Collins, J. R. Harris, C. A. Guerra, M. P. Thorn, D. Bisanzio, N. Fullman, C. K. Huynh, X. Kulikoff, M. J. Kutz, A. D. Lopez, A. H. Mokdad, M. Naghavi, G. Nguyen, K. A. Shackelford, T. Vos, H. Wang, S. S. Lim, C. J. L. Murray, R. N. Price, J. K. Baird, D. L. Smith, S. Bhatt, D. J. Weiss, S. I. Hay, P. W. Gething, Mapping the global endemicity and clinical burden of *Plasmodium vivax*, 2000–17: A spatial and temporal modelling study. *Lancet* **394**, 332–343 (2019).
2. R. E. Howes, K. E. Battle, K. N. Mendis, D. L. Smith, R. E. Cibulskis, J. K. Baird, S. I. Hay, Global epidemiology of *Plasmodium vivax*. *Am. J. Trop. Med. Hyg.* **95**, 15–34 (2016).
3. R. Horuk, C. E. Chitnis, W. C. Darbonne, T. J. Colby, A. Rybicki, T. J. Hadley, L. H. Miller, A receptor for the malarial parasite *Plasmodium vivax*: The erythrocyte chemokine receptor. *Science* **261**, 1182–1184 (1993).
4. A. Chaudhuri, V. Zbrzezna, J. Polyakova, A. O. Pogo, J. Hesselgesser, R. Horuk, Expression of the Duffy antigen in K562 cells. Evidence that it is the human erythrocyte chemokine receptor. *J. Biol. Chem.* **269**, 7835–7838 (1994).
5. R. N. Price, E. Tjitra, C. A. Guerra, S. Yeung, N. J. White, N. M. Anstey, *Vivax Malaria: Neglected and Not Benign* (American Society of Tropical Medicine and Hygiene, 2007).
6. D. Ménard, C. Barnadas, C. Bouchier, C. Henry-Halldin, L. R. Gray, A. Ratsimbasoa, V. Thonier, J.-F. Carod, O. Domarle, Y. Colin, O. Bertrand, J. Picot, C. L. King, B. T. Grimberg, O. Mercereau-Puijalon, P. A. Zimmerman, *Plasmodium vivax* clinical malaria is commonly observed in Duffy-negative Malagasy people. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 5967–5971 (2010).
7. J. K. Baird, Evidence and implications of mortality associated with acute *Plasmodium vivax* malaria. *Clin. Microbiol. Rev.* **26**, 36–57 (2013).
8. R. Culleton, C. Coban, F. Y. Zeyrek, P. Cravo, A. Kaneko, M. Randrianarivelojosia, V. Andrianaranjaka, S. Kano, A. Farnert, A. P. Arez, P. M. Sharp, R. Carter, K. Tanabe, The origins of African *Plasmodium vivax*; insights from mitochondrial genome sequencing. *PLOS ONE* **6**, e29137 (2011).
9. S. Gunawardena, N. D. Karunaweera, M. U. Ferreira, M. Phone-Kyaw, R. J. Pollack, M. Alifrangis, R. S. Rajakaruna, F. Konradsen, P. H. Amerasinghe, M. L. Schousboe, G. N. L. Galappaththy, R. R. Abeyasinghe, D. L. Hartl, D. F. Wirth, Geographic structure of *Plasmodium vivax*: Microsatellite analysis of parasite populations from Sri Lanka, Myanmar, and Ethiopia. *Am. J. Trop. Med. Hyg.* **82**, 235–242 (2010).
10. S. Jongwutiwes, C. Putaporntip, T. Iwasaki, M. U. Ferreira, H. Kanbara, A. L. Hughes, Mitochondrial genome sequences support ancient population expansion in *Plasmodium vivax*. *Mol. Biol. Evol.* **22**, 1733–1739 (2005).
11. M. C. Leclerc, P. Durand, C. Gauthier, S. Patot, N. Billotte, M. Menegon, C. Severini, F. J. Ayala, F. Renaud, From The Cover: Meager genetic variability of the human malaria agent *Plasmodium vivax*. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 14455–14460 (2004).
12. M. Miao, Z. Yang, H. Patch, Y. Huang, A. A. Escalante, L. Cui, *Plasmodium vivax* populations revisited: Mitochondrial genomes of temperate strains in Asia suggest ancient population expansion. *BMC Evol. Biol.* **12**, 22 (2012).
13. J. Mu, D. A. Joy, J. Duan, Y. Huang, J. Carlton, J. Walker, J. Barnwell, P. Beerli, M. A. Charleston, O. G. Pybus, X. Su, Host switch leads to emergence of *Plasmodium vivax* malaria in humans. *Mol. Biol. Evol.* **22**, 1686–1693 (2005).
14. P. Orjuela-Sánchez, J. M. Sá, M. C. C. Brandi, P. T. Rodrigues, M. S. Bastos, C. Amaratunga, S. Duong, R. M. Fairhurst, M. U. Ferreira, Higher microsatellite diversity in *Plasmodium vivax* than in sympatric *Plasmodium falciparum* populations in Pursat, Western Cambodia. *Exp. Parasitol.* **134**, 318–326 (2013).
15. P. T. Rodrigues, H. O. Valdivia, T. C. de Oliveira, J. M. P. Alves, A. M. R. C. Duarte, C. Cerutti-Junior, J. C. Buery, C. F. A. Brito, J. C. de Souza Jr., Z. M. B. Hirano, M. G. Bueno, J. L. Catão-Dias, R. S. Malafronte, S. Ladeia-Andrade, T. Mita, A. M. Santamaria, J. E. Calzada, I. S. Tantular, F. Kawamoto, L. R. J. Raijmakers, I. Mueller, M. A. Pacheco, A. A. Escalante, I. Felger, M. U. Ferreira, Human migration and the spread of malaria parasites to the New World. *Sci. Rep.* **8**, 1993 (2018).
16. J. E. Taylor, M. A. Pacheco, D. J. Bacon, M. A. Beg, R. L. Machado, R. M. Fairhurst, S. Herrera, J.-Y. Kim, D. Menard, M. M. Póvoa, L. Villegas, M. Mulyanto, G. Snounou, L. Cui, F. Y. Zeyrek, A. A. Escalante, The evolutionary history of *Plasmodium vivax* as inferred from mitochondrial genomes: Parasite genetic diversity in the Americas. *Mol. Biol. Evol.* **30**, 2050–2064 (2013).
17. S. Auburn, J. Marfurt, G. Maslen, S. Campino, V. Ruano Rubio, M. Manske, B. MacHunter, E. Kenangalem, R. Noviyanti, L. Trianty, B. Sebayang, G. Wirjanata, K. Sriprawat, D. Alcock, B. MacInnis, O. Miotto, T. G. Clark, B. Russell, N. M. Anstey, F. Nosten, D. P. Kwiatkowski, R. N. Price, Effective preparation of *Plasmodium vivax* field isolates for high-throughput whole genome sequencing. *PLOS ONE* **8**, e53160 (2013).
18. A. Melnikov, K. Galinsky, P. Rogov, T. Fennell, D. Van Tyne, C. Russ, R. Daniels, K. G. Barnes, J. Bochicchio, D. Ndiaye, P. D. Sene, D. F. Wirth, C. Nusbaum, S. K. Volkman, B. W. Birren, A. Gnirke, D. E. Neafsey, Hybrid selection for sequencing pathogen genomes from clinical samples. *Genome Biol.* **12**, R73 (2011).
19. R. D. Pearson, R. Amato, S. Auburn, O. Miotto, J. Almagro-Garcia, C. Amaratunga, S. Suon, S. Mao, R. Noviyanti, H. Trimarsanto, J. Marfurt, N. M. Anstey, T. William, M. F. Boni, C. Dolecek, T. T. Hien, N. J. White, P. Michon, P. Siba, L. Tavul, G. Harrison, A. Barry, I. Mueller, M. U. Ferreira, N. Karunaweera, M. Randrianarivelojosia, Q. Gao, C. Hubbart, L. Hart, B. Jeffery, E. Drury, D. Mead, M. Kekre, S. Campino, M. Manske, V. J. Cornelius, B. MacInnis, K. A. Rockett, A. Miles, J. C. Rayner, R. M. Fairhurst, F. Nosten, R. N. Price, D. P. Kwiatkowski, Genomic analysis of local variation and recent evolution in *Plasmodium vivax*. *Nat. Genet.* **48**, 959–964 (2016).
20. D. N. Hupalo, Z. Luo, A. Melnikov, P. L. Sutton, P. Rogov, A. Escalante, A. F. Vallejo, S. Herrera, M. Arévalo-Herrera, Q. Fan, Y. Wang, L. Cui, C. M. Lucas, S. Durand, J. F. Sanchez, G. C. Baldeviano, A. G. Lescano, M. Laman, C. Barnadas, A. Barry, I. Mueller, J. W. Kazura, A. Eapen, D. Kanagaraj, N. Valecha, M. U. Ferreira, W. Roobsoong, W. Nguitragool, J. Sattabonkot, D. Gamboa, M. Kosek, J. M. Vinetz, L. González-Cerón, B. W. Birren, D. E. Neafsey, J. M. Carlton, Population genomics studies identify signatures of global dispersal and drug resistance in *Plasmodium vivax*. *Nat. Genet.* **48**, 953–958 (2016).
21. S. Auburn, S. Getachew, R. D. Pearson, R. Amato, O. Miotto, H. Trimarsanto, S. J. Zhu, A. Rumaseb, J. Marfurt, R. Noviyanti, M. J. Grigg, B. Barber, T. William, S. M. Goncalves, E. Drury, K. Sriprawat, N. M. Anstey, F. Nosten, B. Petros, A. Aseffa, G. McVean, D. P. Kwiatkowski, R. N. Price, Genomic analysis of *Plasmodium vivax* in Southern Ethiopia reveals selective pressures in multiple parasite mechanisms. *J. Infect. Dis.* **220**, 1738–1749 (2019).
22. S. Auburn, E. D. Benavente, O. Miotto, R. D. Pearson, R. Amato, M. J. Grigg, B. E. Barber, T. William, I. Handayuni, J. Marfurt, H. Trimarsanto, R. Noviyanti, K. Sriprawat, F. Nosten, S. Campino, T. G. Clark, N. M. Anstey, D. P. Kwiatkowski, R. N. Price, Genomic analysis of a pre-elimination Malaysian *Plasmodium vivax* population reveals selective pressures and changing transmission dynamics. *Nat. Commun.* **9**, 2585 (2018).
23. K. A. Twohig, D. A. Pfeffer, J. K. Baird, R. N. Price, P. A. Zimmerman, S. I. Hay, P. W. Gething, K. E. Battle, R. E. Howes, Growing evidence of *Plasmodium vivax* across malaria-endemic Africa. *PLOS Negl. Trop. Dis.* **13**, e0007140 (2019).
24. J. M. Carlton, A. Das, A. A. Escalante, Genomics, population genetics and evolutionary history of *Plasmodium vivax*. *Adv. Parasitol.* **81**, 203–222 (2013).
25. O. E. Cornejo, A. A. Escalante, The origin and age of *Plasmodium vivax*. *Trends Parasitol.* **22**, 558–563 (2006).
26. A. A. Escalante, D. E. Freeland, W. E. Collins, A. A. Lal, The evolution of primate malaria parasites based on the gene encoding cytochrome B from the linear mitochondrial genome. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 8124–8129 (1998).

27. A. A. Escalante, O. E. Cornejo, D. E. Freeland, A. C. Poe, E. Durrego, W. E. Collins, A. A. Lal, A monkey's tale: The origin of *Plasmodium vivax* as a human malaria parasite. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1980–1985 (2005).
28. F. Prugnolle, V. Rougeron, P. Becquart, A. Berry, B. Makanga, N. Rahola, C. Arnathau, B. Ngoubangoye, S. Menard, E. Willaume, F. J. Ayala, D. Fontenille, B. Ollomo, P. Durand, C. Paupy, F. Renaud, Diversity, host switching and evolution of *Plasmodium vivax* infecting African great apes. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 8123–8128 (2013).
29. V. Rougeron, E. Elguero, C. Arnathau, B. A. Hidalgo, P. Durand, S. Houze, A. Berry, S. Zakeri, R. Haque, M. S. Alam, F. Nosten, C. Severini, T. G. Woldearegai, B. Mordmüller, P. G. Kremsner, L. González-Cerón, G. Fontecha, D. Gamboa, L. Musset, E. Legrand, O. Noya, T. Pumpaibool, P. Harnyuttanakorn, K. M. Lekweiry, M. M. Albsheer, M. M. A. Hamid, A. O. M. S. Boukary, J.-F. Trape, F. Renaud, F. Prugnolle, Human *Plasmodium vivax* diversity, population structure and evolutionary origin. *PLOS Negl. Trop. Dis.* **14**, e0008072 (2020).
30. S. Krief, A. A. Escalante, M. A. Pacheco, L. Mugisha, C. André, M. Halbwax, A. Fischer, J.-M. Krief, J. M. Kasenene, M. Crandfield, O. E. Cornejo, J.-M. Chavatte, C. Lin, F. Letourneur, A. C. Grüner, T. F. McCutchan, L. Rénia, G. Snounou, On the diversity of malaria parasites in African apes and the origin of *Plasmodium falciparum* from Bonobos. *PLOS Pathog.* **6**, e1000765 (2010).
31. W. Liu, Y. Li, G. H. Learn, R. S. Rudicell, J. D. Robertson, B. F. Keele, J.-B. N. Ndjango, C. M. Sanz, D. B. Morgan, S. Locatelli, M. K. Gonder, P. J. Kranzusch, P. D. Walsh, E. Delaporte, E. Mpoudi-Ngole, A. V. Georgiev, M. N. Muller, G. M. Shaw, M. Peeters, P. M. Sharp, J. C. Rayner, B. H. Hahn, Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. *Nature* **467**, 420–425 (2010).
32. M. Kaiser, A. Löwa, M. Ulrich, H. Ellerbrok, A. S. Goffe, A. Blasse, Z. Zommers, E. Couacy-Hymann, F. Babweteera, K. Zuberbühler, S. Metzger, S. Geidel, C. Boesch, T. R. Gillespie, F. H. Leendertz, Wild chimpanzees infected with 5 *Plasmodium* species. *Emerg. Infect. Dis.* **16**, 1956–1959 (2010).
33. W. Liu, Y. Li, K. S. Shaw, G. H. Learn, L. J. Plenderleith, J. A. Malenke, S. A. Sundararaman, M. A. Ramirez, P. A. Crystal, A. G. Smith, F. Bibollet-Ruche, A. Ayouba, S. Locatelli, A. Esteban, F. Mouacha, E. Guichet, C. Butel, S. Ahuka-Mundeke, B.-I. Inogwabini, J.-B. N. Ndjango, S. Speede, C. M. Sanz, D. B. Morgan, M. K. Gonder, P. J. Kranzusch, P. D. Walsh, A. V. Georgiev, M. N. Muller, A. K. Piel, F. A. Stewart, M. L. Wilson, A. E. Pusey, L. Cui, Z. Wang, A. Färnert, C. J. Sutherland, D. Nolder, J. A. Hart, T. B. Hart, P. Bertolani, A. Gillis, M. LeBreton, B. Tafon, J. Kiyang, C. F. Djoko, B. S. Schneider, N. D. Wolfe, E. Mpoudi-Ngole, E. Delaporte, R. Carter, R. L. Culleton, G. M. Shaw, J. C. Rayner, M. Peeters, B. H. Hahn, P. M. Sharp, African origin of the malaria parasite *Plasmodium vivax*. *Nat. Commun.* **5**, 3346 (2014).
34. B. Makanga, P. Yangari, N. Rahola, V. Rougeron, E. Elguero, L. Boundenga, N. D. Moukodoum, A. P. Okouga, C. Arnathau, P. Durand, E. Willaume, D. Ayala, D. Fontenille, F. J. Ayala, F. Renaud, B. Ollomo, F. Prugnolle, C. Paupy, Ape malaria transmission and potential for ape-to-human transfers in Africa. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 5329–5334 (2016).
35. D. E. Loy, L. J. Plenderleith, S. A. Sundararaman, W. Liu, J. Gruszczyk, Y.-J. Chen, S. Trimboli, G. H. Learn, O. A. MacLean, A. L. K. Morgan, Y. Li, A. N. Avitto, J. Giles, S. Calvignac-Spencer, A. Sachse, F. H. Leendertz, S. Speede, A. Ayouba, M. Peeters, J. C. Rayner, W.-H. Tham, P. M. Sharp, B. H. Hahn, Evolutionary history of human *Plasmodium vivax* revealed by genome-wide analyses of related ape parasites. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E8450–E8459 (2018).
36. J. Haldane, Disease and evolution. *Ric. Sci.* **19**, 68–76 (1949).
37. A. Gilabert, T. D. Otto, G. G. Rutledge, B. Franzon, B. Ollomo, C. Arnathau, P. Durand, N. D. Moukodoum, A.-P. Okouga, B. Ngoubangoye, B. Makanga, L. Boundenga, C. Paupy, F. Renaud, F. Prugnolle, V. Rougeron, *Plasmodium vivax*-like genome sequences shed new insights into *S* biology and evolution. *PLOS Biol.* **16**, e2006035 (2018).
38. The Anopheles gambiae 1000 Genomes Consortium, Genetic diversity of the African malaria vector *Anopheles gambiae*. *Nature* **552**, 96–100 (2017).
39. H. Li, P. Ralph, Local PCA shows how the effect of population structure differs along the genome. *Genetics* **211**, 289–304 (2019).
40. M. Manske, O. Miotto, S. Campino, S. Auburn, J. Almagro-Garcia, G. Maslen, J. O'Brien, A. Djimde, O. Doumbo, I. Zongo, J.-B. Ouedraogo, P. Michon, I. Mueller, P. Siba, A. Nzila, S. Borrmann, S. M. Kiara, K. Marsh, H. Jiang, X.-Z. Su, C. Amaratunga, R. Fairhurst, D. Socheat, F. Nosten, M. Imwong, N. J. White, M. Sanders, E. Anastasi, D. Alcock, E. Drury, S. Oyola, M. A. Quail, D. J. Turner, V. R. Rubio, D. Jyothi, L. Amenga-Etego, C. Hubbart, A. Jeffreys, K. Rowlands, C. Sutherland, C. Roper, V. Mangano, D. Modiano, J. C. Tan, M. T. Ferdig, A. Amambua-Ngwa, D. J. Conway, S. Takala-Harrison, C. V. Plowe, J. C. Rayner, K. A. Rockett, T. G. Clark, C. I. Newbold, M. Berriman, B. MacInnis, D. P. Kwiatkowski, Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature* **487**, 375–379 (2012).
41. J. H. Degnan, N. A. Rosenberg, Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24**, 332–340 (2009).
42. J. H. Degnan, N. A. Rosenberg, Discordance of species trees with their most likely gene trees. *PLOS Genet.* **2**, e68 (2006).
43. J. C. Avise, T. J. Robinson, Hemiplasy: A new term in the lexicon of phylogenetics. *Syst. Biol.* **57**, 503–507 (2008).
44. N. A. Rosenberg, Discordance of species trees with their most likely gene trees: A unifying principle. *Mol. Biol. Evol.* **30**, 2709–2713 (2013).
45. A. Suh, L. Smeds, H. Ellegren, The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLOS Biol.* **13**, e1002224 (2015).
46. N. Galtier, B. Nabholz, S. Glémin, G. D. D. Hurst, Mitochondrial DNA as a marker of molecular diversity: A reappraisal. *Mol. Ecol.* **18**, 4541–4550 (2009).
47. R. Mostowy, N. J. Croucher, C. P. Andam, J. Corander, W. P. Hanage, P. Marttinen, Efficient inference of recent and ancestral recombination within bacterial populations. *Mol. Biol. Evol.* **34**, 1167–1182 (2017).
48. L. Boundenga, B. Ollomo, V. Rougeron, L. Y. Mouele, B. Mve-Ondo, L. M. Delicat-Loembet, N. D. Moukodoum, A. P. Okouga, C. Arnathau, E. Elguero, P. Durand, F. Liégeois, V. Boué, P. Motsch, G. Le Flohic, A. Ndoungouet, C. Paupy, C. T. Ba, F. Renaud, F. Prugnolle, Diversity of malaria parasites in great apes in Gabon. *Malar. J.* **14**, 111 (2015).
49. T. D. Otto, A. Gilabert, T. Crellen, U. Böhme, C. Arnathau, M. Sanders, S. O. Oyola, A. P. Okouga, L. Boundenga, E. Willaume, B. Ngoubangoye, N. D. Moukodoum, C. Paupy, P. Durand, V. Rougeron, B. Ollomo, F. Renaud, C. Newbold, M. Berriman, F. Prugnolle, Genomes of all known members of a *Plasmodium* subgenus reveal paths to virulent human malaria. *Nat. Microbiol.* **3**, 687–697 (2018).
50. J. Prado-Martinez, P. H. Sudmant, J. M. Kidd, H. Li, J. L. Kelley, B. Lorente-Galdos, K. R. Veeramah, A. E. Woerner, T. D. O'Connor, G. Santpere, A. Cagan, C. Theunert, F. Casals, H. Laayouni, K. Munch, A. Hobolth, A. E. Halager, M. Malig, J. Hernandez-Rodriguez, I. Hernando-Herraez, K. Prüfer, M. Pybus, L. Johnstone, M. Lachmann, C. Alkan, D. Twigg, N. Petit, C. Baker, F. Hormozdiari, M. Fernandez-Callejo, M. Dabad, M. L. Wilson, L. Stevison, C. Camprubí, T. Carvalho, A. Ruiz-Herrera, L. Vives, M. Mele, T. Abello, I. Kondova, R. E. Bontrop, A. Pusey, F. Lankester, J. A. Kiyang, R. A. Bergl, E. Lonsdorf, S. Myers, M. Ventura, P. Gagneux, D. Comas, H. Siegismund, J. Blanc, L. Agueda-Calpena, M. Gut, L. Fulton, S. A. Tishkoff, J. C. Mullikin, R. K. Wilson, I. G. Gut, M. K. Gonder, O. A. Ryder, B. H. Hahn, A. Navarro, J. M. Akey, J. Bertranpetit, D. Reich, T. Mailund, M. H. Schierup, C. Hvilsom, A. M. Andrés, J. D. Wall, C. D. Bustamante, M. F. Hammer, E. E. Eichler, T. Marques-Bonet, Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013).
51. G. McVean, A genealogical interpretation of principal components analysis. *PLOS Genet.* **5**, e1000686 (2009).
52. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
53. L. M. Kootker, L. Mbeki, A. G. Morris, H. Kars, G. R. Davies, Dynamics of Indian Ocean slavery revealed through isotopic data from the colonial era Cobern Street Burial Site, Cape Town, South Africa (1750-1827). *PLOS ONE* **11**, e0157750 (2016).
54. L. van Dorp, P. Gelabert, A. Rieux, M. de Manuel, T. de Dios, S. Gopalakrishnan, C. Carøe, M. Sandoval-Velasco, R. Fregel, I. Olalde, R. Escosa, C. Aranda, S. Huijben, I. Mueller, T. Marquès-Bonet, F. Balloux, M. T. P. Gilbert, C. Lalueza-Fox, *Plasmodium vivax* Malaria viewed through the lens of an eradicated European strain. *Mol. Biol. Evol.* **37**, 773–785 (2020).
55. A. Estoup, T. Guillemaud, Reconstructing routes of invasion using genetic data: Why, how and so what? *Mol. Ecol.* **19**, 4113–4130 (2010).
56. O. Miotto, R. Amato, E. A. Ashley, B. MacInnis, J. Almagro-Garcia, C. Amaratunga, P. Lim, D. Mead, S. O. Oyola, M. Dhorda, M. Imwong, C. Woodrow, M. Manske, J. Stalker, E. Drury, S. Campino, L. Amenga-Etego, T.-N. N. Thanh, H. T. Tran, P. Ringwald, D. Bethell, F. Nosten, A. P. Phyo, S. Pukrittayakamee, K. Chotivanich, C. M. Chuor, C. Nguon, S. Suon, S. Sreng, P. N. Newton, M. Mayxay, M. Khanthavong, B. Hongvanthong, Y. Htut, K. T. Han, M. P. Kyaw, M. A. Faiz, C. I. Fanello, M. Onyamboko, O. A. Mokuolu, C. G. Jacob, S. Takala-Harrison, C. V. Plowe, N. P. Day, A. M. Dondorp, C. C. A. Spencer, G. McVean, R. M. Fairhurst, N. J. White, D. P. Kwiatkowski, Genetic architecture of artemisinin-resistant *Plasmodium falciparum*. *Nat. Genet.* **47**, 226–234 (2015).
57. F. Rousset, Genetic differentiation and estimation of gene flow from F-Statistics under isolation by distance. *Genetics* **145**, 1219–1228 (1997).
58. J. Pickrell, J. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLOS Genet.* **8**, e1002967 (2012).
59. M. DeGiorgio, M. Jakobsson, N. A. Rosenberg, Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 16057–16062 (2009).
60. J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, R. M. Myers, Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
61. K. Tanabe, T. Mita, T. Jombart, A. Eriksson, S. Horibe, N. Palacpac, L. Ranford-Cartwright, H. Sawai, N. Sakihama, H. Ohmae, M. Nakamura, M. U. Ferreira, A. A. Escalante, F. Prugnolle, A. Björkman, A. Färnert, A. Kaneko, T. Horii, A. Manica, H. Kishino, F. Balloux, *Plasmodium falciparum* accompanied the human expansion out of Africa. *Curr. Biol.* **20**, 1283–1289 (2010).

62. L. Excoffier, M. Foll, R. J. Petit, Genetic consequences of range expansions. *Annu. Rev. Ecol. Evol. Syst.* **40**, 481–501 (2009).
63. World Health Oranization, *World Malaria Report 2010* (World Health Organization, 2010); www.who.int/malaria/publications/atoz/9789241564106/en/.
64. J. A. Nájera, M. González-Silva, P. L. Alonso, Some lessons for the future from the global malaria eradication programme (1955–1969). *PLOS Med.* **8**, e1000412 (2011).
65. E.-M. Willing, C. Dreyer, C. van Oosterhout, Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. *PLOS ONE* **7**, e42649 (2012).
66. A. G. Nazareno, J. B. Bemmels, C. W. Dick, L. G. Lohmann, Minimum sample sizes for population genomics: An empirical study from an Amazonian plant species. *Mol. Ecol. Resour.* **17**, 1136–1147 (2017).
67. M. Venkatesan, C. Amaratunga, S. Campino, S. Auburn, O. Koch, P. Lim, S. Uk, D. Socheat, D. P. Kwiatkowski, R. M. Fairhurst, C. V. Plowe, Using CF11 cellulose columns to inexpensively and effectively remove human DNA from *Plasmodium falciparum*-infected whole blood samples. *Malar. J.* **11**, 41 (2012).
68. M. Gillies, M. Coetzee, A supplement to the Anophelinae of Africa South of the Sahara (Afrotropical Region). *Publ. Afr. Inst. Med. Res.* **55**, 1–143 (1987).
69. S. A. Sundararaman, W. Liu, B. F. Keele, G. H. Learn, K. Bittinger, F. Mouacha, S. Ahuka-Mundeke, M. Manske, S. Sherrill-Mix, Y. Li, J. A. Malenke, E. Delaporte, C. Laurent, E. Mpoudi Ngole, D. P. Kwiatkowski, G. M. Shaw, J. C. Rayner, M. Peeters, P. M. Sharp, F. D. Bushman, B. H. Hahn, *Plasmodium falciparum*-like parasites infecting wild apes in southern Cameroon do not represent a recurrent source of human malaria. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 7020–7025 (2013).
70. A. N. Cowell, D. E. Loy, S. A. Sundararaman, H. Valdivia, K. Fisch, A. G. Lescano, G. C. Baldeviano, S. Durand, V. Gerbasi, C. J. Sutherland, D. Nolder, J. M. Vinetz, B. H. Hahn, E. A. Winzeler, Selective whole-genome amplification is a robust method that enables scalable whole-genome sequencing of *Plasmodium vivax* from unprocessed clinical samples. *MBio* **8**, e02257-16 (2017).
71. E. R. Chan, D. Menard, P. H. David, A. Ratsimbasoa, S. Kim, P. Chim, C. Do, B. Witkowski, O. Mercereau-Puijalon, P. A. Zimmerman, D. Serre, Whole genome sequencing of field isolates provides robust characterization of genetic diversity in *Plasmodium* vivax. *PLOS Negl. Trop. Dis.* **6**, e1811 (2012).
72. S.-B. Chen, Y. Wang, K. Kassegne, B. Xu, H.-M. Shen, J.-H. Chen, Whole-genome sequencing of a *Plasmodium vivax* clinical isolate exhibits geographical characteristics and high genetic variation in China-Myanmar border area. *BMC Genomics* **18**, 131 (2017).
73. T. C. de Oliveira, P. T. Rodrigues, M. J. Menezes, R. M. Gonçalves-Lopes, M. S. Bastos, N. F. Lima, S. Barbosa, A. L. Gerber, G. Loss de Morais, L. Berná, J. Phelan, C. Robello, A. T. R. de Vasconcelos, J. M. P. Alves, M. U. Ferreira, Genome-wide diversity and differentiation in New World populations of the human malaria parasite *Plasmodium vivax*. *PLOS Negl. Trop. Dis.* **11**, e0005824 (2017).
74. J. Popovici, L. R. Friedrich, S. Kim, S. Bin, V. Run, D. Lek, M. V. Cannon, D. Menard, D. Serre, Genomic analyses reveal the common occurrence and complexity of *Plasmodium vivax* relapses in Cambodia. *MBio* **9**, e01888-17 (2018).
75. C. Delgado-Ratto, D. Gamboa, V. E. Soto-Calle, P. Van den Eede, E. Torres, L. Sánchez-Martínez, J. Contreras-Mancilla, A. Rosanas-Urgell, H. Rodriguez Ferrucci, A. Llanos-Cuentas, A. Erhart, J.-P. Van Geertruyden, U. D'Alessandro, Population genetics of *Plasmodium vivax* in the Peruvian Amazon. *PLOS Negl. Trop. Dis.* **10**, e0004376 (2016).
76. D. E. Neafsey, K. Galinsky, R. H. Y. Jiang, L. Young, S. M. Sykes, S. Saif, S. Gujja, J. M. Goldberg, S. Young, Q. Zeng, S. B. Chapman, A. P. Dash, A. R. Anvikar, P. L. Sutton, B. W. Birren, A. A. Escalante, J. W. Barnwell, J. M. Carlton, The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. *Nat. Genet.* **44**, 1046–1050 (2012).
77. R. Leinonen, H. Sugawara, M. Shumway; International Nucleotide Sequence Database Collaboration, The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
78. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J.* **17**, 10–12 (2011).
79. S. Auburn, U. Böhme, S. Steinbiss, H. Trimarsanto, J. Hostetler, M. Sanders, Q. Gao, F. Nosten, C. I. Newbold, M. Berriman, R. N. Price, T. D. Otto, A new *Plasmodium vivax* reference sequence with improved assembly of the subtelomeres reveals an abundance of pir genes. *Wellcome Open Res.* **1**, 4 (2016).
80. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
81. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
82. *Picard Toolkit* (Broad Institute, 2019); http://broadinstitute.github.io/picard/.
83. J. Wang, Effects of sampling close relatives on some elementary population genetics analyses. *Mol. Ecol. Resour.* **18**, 41–54 (2018).
84. S. F. Schaffner, A. R. Taylor, W. Wong, D. F. Wirth, D. E. Neafsey, hmmIBD: Software to infer pairwise identity by descent between haploid genotypes. *Malar. J.* **17**, 196 (2018).
85. D. H. Huson, D. Bryant, Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
86. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
87. A. Rambaut, FigTree, a graphical viewer of phylogenetic trees (2007).
88. S.-I. Tachibana, S. A. Sullivan, S. Kawai, S. Nakamura, H. R. Kim, N. Goto, N. Arisue, N. M. Q. Palacpac, H. Honma, M. Yagi, T. Tougan, Y. Katakai, O. Kaneko, T. Mita, K. Kita, Y. Yasutomi, P. L. Sutton, R. Shakhbatyan, T. Horii, T. Yasunaga, J. W. Barnwell, A. A. Escalante, J. M. Carlton, K. Tanabe, *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. *Nat. Genet.* **44**, 1051–1055 (2012).
89. H. Zhao, Z. Sun, J. Wang, H. Huang, J.-P. Kocher, L. Wang, CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
90. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
91. N. M. Kopelman, J. Mayzel, M. Jakobsson, N. A. Rosenberg, I. Mayrose, Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* **15**, 1179–1191 (2015).
92. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin; 1000 Genomes Project Analysis Group, The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
93. M. Milanesi, S. Capomaccio, E. Vajana, L. Bomba, J. F. Garcia, P. Ajmone-Marsan, L. Colli, BITE: An R package for biodiversity analyses. bioRxiv 181610 [**Preprint**]. 29 August 2017. https://doi.org/10.1101/181610.
94. O. François, M. G. B. Blum, M. Jakobsson, N. A. Rosenberg, Demographic history of european populations of Arabidopsis thaliana. *PLOS Genet.* **4**, e1000075 (2008).
95. M. C. Fontaine, F. Austerlitz, T. Giraud, F. Labbé, D. Papura, S. Richard-Cervera, F. Delmotte, Genetic signature of a range expansion and leap-frog event after the recent invasion of Europe by the grapevine downy mildew pathogen Plasmopara viticola. *Mol. Ecol.* **22**, 2771–2786 (2013).
96. C. Zhang, S.-S. Dong, J.-Y. Xu, W.-M. He, T.-L. Yang, PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).
97. S. Schiffels, K. Wang, MSMC and MSMC2: The multiple sequentially markovian coalescent. *Methods Mol. Biol.* **2090**, 147–166 (2020).

Acknowledgments: We acknowledge the IRD itrop HPC (South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: https://bioinfo.ird.fr and www.southgreen.fr. We are grateful to E. Andermarcher for proofreading this article. **Funding:** This study was supported by ANR T-ERC EVAD, ANR JCJC GENAD, PEPS ECOMOB MOV 2019, CNRS, and MGX Montpellier sequencing facility. J.D. was supported by the Fondation pour la recherche Médicale (FRM, ARF20170938823) as well as by the Marie-Curie EU Horizon 2020 Marie-Skłodowska-Curie research and innovation program grant METHYVIREVOL (contract number 800489).
Author contributions: Conception: V.R. and F.P.; funding acquisition: V.R. and J.D.; biological data acquisition and management: V.R., A.B., B.N., L.B., S.H., C.A., C.S., J.-F.T., and P.D.; sequence data acquisition: A.B. and V.R.; method development and data analysis: J.D. and M.C.F.; interpretation of the results: J.D., M.C.F., F.P., and V.R.; drafting of the manuscript: J.D. and V.R.; reviewing and editing of the manuscript: J.D., F.R., M.C.F., F.P., and V.R. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The Illumina sequence reads generated on the new *P. vivax* and *P. vivax*-like samples have been deposited in the National Center for Biotechnology Information (NCBI) under the bioproject name PRJNA720520. The Illumina sequence reads generated on the new *P. vivax* samples from Mauritania, Sudan, and Ethiopia have been deposited in the European Nucleotide Archive (ENA) under the accession codes listed in table S1. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 28 April 2020
Accepted 10 March 2021
Published 28 April 2021
10.1126/sciadv.abc3713

Citation: J. Daron, A. Boissière, L. Boundenga, B. Ngoubangoye, S. Houze, C. Arnathau, C. Sidobre, J.-F. Trape, P. Durand, F. Renaud, M. C. Fontaine, F. Prugnolle, V. Rougeron, Population genomic evidence of *Plasmodium vivax* Southeast Asian origin. *Sci. Adv.* **7**, eabc3713 (2021).