



HAL
open science

Population genomic evidence of *Plasmodium vivax* Southeast Asian origin

J. Daron, A. Boissiere, L. Boundenga, B. Ngoubangoye, S. Houze, C. Arnathau, C. Sidobre, J.-F. Trape, P. Durant, F. Renaud, et al.

► **To cite this version:**

J. Daron, A. Boissiere, L. Boundenga, B. Ngoubangoye, S. Houze, et al.. Population genomic evidence of *Plasmodium vivax* Southeast Asian origin. *Science Advances* , 2021, 7, pp.eabc3713. 10.1126/sciadv.abc3713 . hal-02915422

HAL Id: hal-02915422

<https://hal.science/hal-02915422>

Submitted on 24 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

20 Abstract

21 *Plasmodium vivax* is the most prevalent and widespread human malaria parasite,
22 with almost three billion people living at risk of infection. With the discovery of its closest
23 genetic relatives in African great apes (*Plasmodium vivax-like*), the origin of *P. vivax* has
24 been proposed to be located in the sub-Saharan African area. However, the limited number
25 of genetic markers and samples investigated questioned the robustness of this result. Here,
26 we examined the population genomic variation of 447 human *P. vivax* strains and 19 ape *P.*
27 *vivax-like* strains originating from 24 different countries across the world. We identified
28 2,005,455 high quality single-nucleotide polymorphism loci allowing us to conduct an
29 extensive characterization to date of *P. vivax* worldwide genetic variation. Phylogenetic
30 relationships between human and ape *Plasmodium* revealed that *P. vivax* is a sister clade of
31 *P. vivax-like*, not included within the radiation of *P. vivax-like*. By investigating a variety of
32 aspects of *P. vivax* variation, we identified several striking geographical patterns in summary
33 statistics as function of increasing geographic distance from Southeast Asia, suggesting that
34 *P. vivax* may derived from serial founder effects from a single origin located in Asia.

35 Introduction

36 *Plasmodium vivax* is the most prevalent human malarial parasite responsible for 70
37 to 80 million clinical cases each year (1). It is widespread along the world tropical belt where
38 almost three billion people live at risk of infection. It is the most widely distributed cause of
39 human malaria (2). The vast majority of *P. vivax* transmission is located in Central and South
40 East Asia, while populations present in the sub-Saharan African area are protected from
41 transmission due to the absence of the Duffy antigen (*i.e.*, Duffy negativity) at the surface of
42 their red blood cells (3, 4). Historically, *P. vivax* remained understudied compared to *P.*
43 *falciparum* because of its lower mortality rate. However, recent emergence of new
44 therapeutic resistances and the discovery of fatal cases due to *P. vivax* questioned the
45 benign status of *P. vivax* malaria (5). In addition, the identification of strains able to invade
46 Duffy-negative red blood cells have been raising concerns over potential spread of the
47 disease into regions that were previously assumed to be protected (6). Today, the
48 perception has changed and *P. vivax* is now recognized as a major threat for Public Health
49 (7). In order to carry effective strategies for malaria control and elimination, we need to
50 accumulate knowledge on the genetic structure of individual infections, as it provides
51 understanding on the local patterns of malaria transmission and the dynamics of genetic
52 recombination in natural populations of *P. vivax*.

53 Until recently, early works investigating patterns of genetic variation in worldwide *P.*
54 *vivax* populations have been limited to a reduced set of samples and genetic markers (based
55 on mitochondrial or autosomal markers) (8–10). Consequently, *P. vivax* genetic analyses
56 yielded an incomplete picture of the evolutionary history of this pathogen. Recently,
57 technological improvement has made possible to sequence the whole *P. vivax* genome *via*
58 an enrichment of the parasite DNA from clinical blood samples (11). This methodological
59 breakthrough marked the beginning of a new era in the field, with the released in a few years
60 of several projects characterizing the pathogen genetic variation at the whole genome scale
61 in several *P. vivax* populations (12–15). Those projects shed light on strong signals of recent
62 evolutionary selection partly due to known drug resistance genes. However, even though
63 they released hundreds of complete genomes, each of them restricted their investigation on
64 a sub-fraction of the worldwide *P. vivax* diversity, either located in Asia-Pacific or South
65 America. Consequently, a comprehensive picture of the worldwide genetic diversity and
66 structure of *P. vivax* populations is still missing and its evolutionary history and how it spread
67 over the world is still poorly understood. Moreover, despite growing evidence suggesting an
68 underlying widespread presence of *P. vivax* across all malaria-endemic regions of Africa
69 (16), this continent remains remotely covered with too few complete genomes sequenced

70 from this area. Thus, a key challenge is to provide a worldwide understanding of the genetic
71 variability of *P. vivax* at the genome level to bring insights on the past demographic history
72 and origin of this pathogen.

73 The origin of current *P. vivax* in humans has stimulated passionate and exciting
74 debates for years. Certain studies placed the origin of the human *P. vivax* in Southeast Asia
75 (“out of Asia” hypothesis) based on its phylogenetic position in a clade of parasites infecting
76 Asian monkeys (17–19). This scenario was also supported by genotyping data at 11
77 microsatellite markers collected across four continents, showing the highest microsatellite
78 diversity in Southeast Asia (20, 21). However, the Asian-origin has been challenged by an
79 “out of Africa” scenario, with the recent discovery of a closely related *Plasmodium* species
80 circulating in wild-living African great apes (chimpanzees and gorillas) (22). This new
81 lineage, hereafter referred to as *P. vivax-like*, was suggested to have given rise to *P. vivax* in
82 humans following the transfer of parasites from African apes (23). This new finding echoes a
83 70 years-old postulate speculating that the high prevalence of Duffy negativity among
84 human populations in sub-Saharan Africa is a consequence of a long interaction between
85 humans and its parasite, in favor of an African origin of *P. vivax* (24). Yet, despite over a
86 century of research, this monkey’s tale has not yet come to an end. Although now privileged,
87 the “out of Africa” hypothesis remains disputable as the exact series of events linking current
88 human *P. vivax* populations and African great ape *P. vivax-like* still remains unclear (25).

89 Here, to provide novel insight into the worldwide historical demography of *P. vivax*
90 populations, we analyze the genomic variation of 447 human *P. vivax* strains originating from
91 21 different countries across the world. By including 19 *P. vivax-like* strains sampled on
92 great apes, we explore the evolutionary history of *P. vivax* from its origins to contemporary
93 time and specifically assess whether the origin of the parasite is consistent with an “out-of-
94 Africa” or “out-of-Asia” hypothesis. To our knowledge, this data set provides the most
95 comprehensive characterization to date of the worldwide population genetic structure and
96 diversity of *P. vivax* and *P. vivax-like*, with the identification of two distinct non-recombining
97 *P. vivax-like* clades circulating in sympatry among great apes. Our result clearly demonstrate
98 that *P. vivax* is a sister clade to *P. vivax-like* and is not included within the radiation of *P.*
99 *vivax-like*, as previously suggested by Liu et al., (22). Finally, we discovered multiple lines of
100 evidence from summary statistics of *P. vivax* worldwide genetic variation increasing with
101 geographic distance from Southeast Asia, supporting the hypothesis of a serial founder
102 effect from a single origin located in South-East Asia.

103 Results and Discussion

104 Genomic data and diversity in *P. vivax* and *P. vivax-like*

105 Genome wide data from a total of 1,154 *P. vivax* isolates sampled from all around the
106 world were processed and analyzed (Figure 1A). This dataset included 20 newly sequenced
107 African isolates (Mauritania n=14, Ethiopia n=3 and Sudan n=3), as the worldwide sampling
108 was lacking *P. vivax* isolates coming from Africa (with only three genomes from Madagascar
109 (13) and 24 from Ethiopia (14)). *P. vivax* samples were obtained from 24 countries across
110 the globe (769 from Asia, 338 from America, and 47 from Africa) (Supplementary Table 1).
111 In order to trace the genetic ancestry of *P. vivax*, we also added a total of 27 genomes of *P.*
112 *vivax-like* isolated from African great apes. Among them, ten *P. vivax-like* genomes from
113 Gabon were newly sequenced in our laboratory and 17 genomes were obtained from public
114 databases (Gabon n=11, Cameroon n=5 and Ivory Coast n=1) (23, 25).

115 Due to the heterogeneity in DNA enrichment methods and sequencing technologies
116 used to access *Plasmodium* genome, our cohort exhibited a broad range of sequencing
117 depth coverage (Supplementary Figure 1). We selected genomes with a minimum average
118 sequencing depth of at least 5x to conduct reliable analyses, thus reducing sampling from
119 1,154 to 473 for the *P. vivax* isolates and from 27 to 20 for the *P. vivax-like* isolates. Next,
120 we discarded isolates and variants having a proportion of missing data above 50%. Finally,
121 we evaluated for each sample the within-sample parasite infection complexity by calculating
122 the reference allele frequency (RAF) distribution (Supplementary Figure 2). A RAF of 0% or
123 100% indicates the presence of either the reference or a single alternative allele for all
124 polymorphic sites, which suggests the presence of a single infection (26). Overall, nearly all
125 polymorphic variants exhibited either high or low RAF value suggesting the presence of a
126 single infection for all samples. Though, to avoid further possible bias due to multiple
127 infections with several strains, the SNP calling was designed to select within each sample
128 the variant with the highest frequency, resulting in having at each site, one single variant
129 called per sample. After these pre-processing steps, the final data set included a total of 466
130 genomes (447 of *P. vivax* and 19 of *P. vivax-like*) in which 2,005,455 high-quality SNPs were
131 identified.

132 Heterogeneous genetic structure along the genome of *P. vivax*

133 While population structuration leads to a global impact on the genomic variation in
134 natural populations, local heterogeneity can occur along the genome in regions under the
135 influence of non-random factors, including structural chromosomal features such as large
136 inversions, regions of heterochromatin, or due to selective forces impacting local genetic

137 diversity and recombination (27). These factors can lead to local variation in individual strain
138 genetic ancestry and relatedness along the genome confounding global patterns of genetic
139 diversity and population structure. We identified potential local variation in individual
140 relatedness along the genome using a local Principal Component Analysis (PCA) (28). As
141 shown in Figure 1B and 1C, contrasted patterns of individual relatedness were found along
142 the genome of *P. vivax*, defining two main genomic partitions: a small partition localized
143 mainly at the sub-telomeric ends of each chromosome (in orange in Figures 1B and 1C), and
144 a larger one at the of each chromosome encompassing 80% of the total SNPs set (hereafter
145 called the "core region", in purple in Figures 1B and 1C). Interestingly, the sub-telomeric
146 regions identified as having distinct ancestry from the rest of the genome also coincided with
147 hyper-variable genomic regions (in orange in Supplementary Figure 3). Those regions are
148 known to include hyper-variable repetitive regions causing high genotypic errors. Their exact
149 coordinate have only been reported on a former version of the *P. vivax* genome assembly
150 (12). The differences in genetic ancestry and strain relatedness provided by the two
151 partitions are evident when looking at the local PCA analyses (Supplementary Figure 4).
152 While the population structure recovered from the core genomic regions displayed
153 interpretable genetic patterns that were consistent with previous studies (12, 13), the genetic
154 picture obtained from the SNPs in the sub-telomeric hyper-variable regions was much harder
155 to interpret. We thus excluded these regions from the following analyses, and focused only
156 on the core genomic regions on the central chromosomal region, which represents 21 Mb of
157 the genome and includes 1,610,445 SNPs (~1 SNP every 13 bp).

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

159 In order to better understand the evolutionary origin of human *P. vivax*, we examined
160 its phylogenetic relationships with *P. vivax-like* infecting great apes, since chimpanzees and
161 gorillas harbor the closest relatives of human *P. vivax*. The neighbor joining (NJ) tree based
162 on the SNP data set composed of the 19 *P. vivax-like* and 447 *P. vivax* genomes revealed
163 three distinct clades (Figure 2A). The first bifurcation in the tree splits ape-infecting strains
164 from human-infecting strains and represented the strongest axis of genetic variation on the
165 PCA (Figure 2B, 29% of variance explained). This result clearly demonstrates that human *P.*
166 *vivax* lineage is a sister clade of the ape *P. vivax-like* strains and contradict previous results
167 from Liu et al., suggesting that *P. vivax* forms a monophyletic lineage within the ape parasite
168 radiation (22). Liu's phylogenetic topology, presented as the main argument in favor of an
169 African origin of the parasite, may result from a lack of phylogenetic signal, due to the small
170 number (11 single-genome amplification sequences) of genetic markers investigated. Our
171 result, consistent with previous studies based on a smaller sampling size of *P. vivax-like* (23,
172 25), re-opens the debate on the origin of *P. vivax* (see below). Interestingly, the second split

173 on the phylogenomic tree (Figure 2A) and on the PCA on the second PC axis (Figure 2B)
174 identified two distinct lineages among *P. vivax-like* strains (referred to PVL.grp1 and
175 PVL.grp2 on Figure 2), composed of five and 14 different strains respectively. Since *P.*
176 *vivax-like* infects both chimpanzees and gorillas, it would have been reasonable to expect
177 these two lineages to be associated with a specialization on each host. However, both
178 lineages were found on both host species.

179 With the identification of two genetically distinct *P. vivax-like* lineages, we tested the
180 extent of recombination between them, in order to assess whether gene flow still occurs
181 between them. We used the software fastGEAR (29) using the 19 *P. vivax-like* individual
182 genomes to identify recently exchanged genomic fragments (*i.e.* those shared between two
183 individuals) and their clade of origin. Through its clustering method, fastGEAR was able to
184 recover the two major *P. vivax-like* clades and identified among them 12,108 recently
185 imported fragments with a mean import length of 1,775 bp (Figure 2C). Interestingly, the
186 donor and recipient individuals of the recently imported fragments belong in all cases to the
187 same lineage, suggesting that no recent inter-lineage recombination occurred between the
188 two distinct *P. vivax-like* clades identified. This result was confirmed further by building a
189 reticulation network for the 19 *P. vivax-like* individuals (Supplementary Figure 5). The lack of
190 reticulation shared between individuals belonging to different lineages is evidence of the
191 absence of recombination between the two distinct *P. vivax-like* clades. Thus, our results
192 clearly demonstrate that although circulating in sympatry inside the same host populations of
193 great apes (*i.e.* within the La Lekedi Park in Gabon), the two lineages of *P. vivax-like* do not
194 recombine, which suggests that they may form distinct species. Intriguingly, a similar
195 subdivision was found in *Plasmodium praefalciparum*, the closest parasite of *P. falciparum*,
196 but the status of these two clades was never investigated as done here for *P. vivax-like* (30).

197 The genetic diversity (π) of both *P. vivax-like* lineages were found to be about nine
198 higher than the value observed in *P. vivax* (Figure 2D), in agreement with previous reports
199 (23, 25). Since it is unlikely that the mutation rates radically differ between *P. vivax-like* and
200 *P. vivax* (31), this difference in genetic diversity may be reflecting a higher historical effective
201 size for *P. vivax-like* than for *P. vivax*. The lower genetic diversity observed in *P. vivax* could
202 result from a bottleneck effect that occurred during the host shift, when the pathogen
203 colonized humans (23). A comparable contrasted genetic pattern exists between African
204 apes and humans (32), where humans exhibit a far lower genetic diversity than the African
205 apes due to the bottleneck effect that modern human populations underwent with the out-of-
206 Africa expansion from a small number of founders that replaced the archaic forms of humans
207 (e.g. Neanderthals). Is a similar history likely for *P. vivax* in humans? And if yes, does Africa
208 really represent the point of origin of human *P. vivax*, as recently suggested (22)? Within *P.*
209 *vivax-like*, although a higher genetic diversity was detected for PVL.grp1 compared to

210 PVL.grp2, this may reflect sampling differences as PVL.grp1 samples were collected across
211 three African countries, while samples from PVL.grp2 came from one single location, in the
212 Park of La Lékédi in Gabon.

213 Finally, we inferred historical fluctuations in effective population size (N_e) for the two
214 lineages of *P. vivax-like*. Using a multiple sequentially Markovian coalescent (MSMC)
215 approach, the analysis of genome variation indicated an ancient major expansion in genetic
216 diversity for both clades, follow by a recent dramatic decline (Figure 2E). Unexpectedly, a
217 much high increase in N_e was observed for the PVL.grp2 lineage compared to the PVL.grp1
218 lineage. Together, our analysis of historical N_e suggest a putative ancient speciation event
219 of the two lineage of *P. vivax-like*, consistent with our previous result demonstrating the
220 reproductively isolation of those two lineages.

221 Worldwide genetic structure of human *P. vivax*

222 To answer these questions, we investigated the genetic structure in our cohort of 447
223 *P. vivax* isolates collected from 21 countries around the world. Bi-allelic SNPs from the core
224 genome were analyzed using complementary approaches: a PCA that does not rely on any
225 model assumptions (33), a model-based individual ancestry analysis implemented in the
226 software ADMIXTURE (34), a neighbor joining phylogenetic tree, and by assessing the
227 amount genetic differentiation among pairs of populations using F_{ST} .

228 All analyses revealed consistent patterns splitting the genetic variation primarily by
229 continents (Figure 3). The first principal component of the genetic variation (EV1) split
230 Southeast Asian from Africa and American populations, while the second axis (EV2) split
231 African populations from the rest of the world (Figure 3A and Supplementary Figure 6).
232 These three major clusters were also identified on the neighbor-joining tree (Figure 3B), and
233 displayed clearly distinct genetic ancestry as estimated when simulating three ancestral
234 clusters (K) by the ADMIXTURE analysis (Supplementary Figure 7).

235 Within each cluster/continent, further subdivision was also evident. Both the PCA
236 (Supplementary Figure 6) and ADMIXTURE (Supplementary Figures 7 and 8) analyses
237 suggested that up to six distinct genetic pools were present. Three distinct ancestral
238 populations were identified in Asia, one in Africa and two in America (Figures 3C and D).

239 Interestingly, strains from India and Sri Lanka were genetically more closely related
240 to the African populations than to the other neighboring Asiatic groups (Figures 3A, 3B and
241 3D). This may results from the different trades that took place between the Indian
242 Subcontinent and Africa over the last two millennia, during which movements of populations
243 (and likely diseases) occurred either from India to East Africa or from Africa to India (35).
244 The most recent exchange has been the migration of the human Karana population from the
245 north-west India into Madagascar at the end of the 17th century (36).

246 Across the Americas, *P. vivax* strains were structured into two distinct ancestral
247 populations. This is consistent with recent findings suggesting two successive migratory
248 waves responsible for the introduction of *P. vivax* in America (37). The first wave has been
249 suggested to occur following a reverse Kon-Tiki route, with a long-range oceanic crossing
250 from the Western Pacific to the Americas. The second wave, has been recently attributed to
251 an introduction by the European colonization of the Americas during the 15th century, and
252 represents the major genetic contributors to the New World *P. vivax* lineages (38).

253 Lastly, individuals collected across the Southeast Asia/Pacific region were structured
254 in three distinct ancestral populations. Among each ancestral population located in
255 Southeast Asia, the genetic differentiation estimated with F_{ST} between countries was weak
256 (Supplementary Figure 9), consistent with relatively unrestricted gene flow between
257 countries. If a distinct ancestral population shared by populations located in Indonesia and
258 Papua New Guinea (PNG) is well explained by their insular isolation, the presence in
259 Southeast Asia continent of two ancestral populations may reflect the effect of the malaria-
260 free corridor previously established through central Thailand, consistent with previous
261 observations in *P. falciparum* (39). Population in Malaysia were found admix between two
262 ancestral populations, at a contact zone in which admixture proportions progressively
263 change from one cluster to the other.

264 Patterns of genetic diversity in agreement with an Asian origin of human *P. vivax*

265 The topology of the NJ tree suggested that populations split from each other
266 following a stepping stone colonization model along the world tropical belt. Such stepping
267 stone pattern is expected to lead to an isolation-by-distance (IBD) among populations. As
268 expected under a 2-dimensions stepping stone IBD model (40), we found a highly significant
269 relationship between the $F_{ST}/(1-F_{ST})$ and the logarithm of the geographic distance (Mantel
270 test $r^2 = 0.57$, $p < 10^{-4}$, Figure 3E), suggesting limited *P. vivax* dispersal across space. Based
271 on the NJ tree, the *P. vivax* clade containing Malaysia, Indonesia and PNG is located
272 nearest to the root of the tree, outward from which are branches that correspond,
273 sequentially to populations from Southeast Asia, Africa and America. Consequently, the
274 branching pattern largely supports an “out-of-Asia” model of *P. vivax* origin rather than the
275 previously suggested “out of Africa” model (22).

276 Under a scenario where the range expansion experienced by *P. vivax* results in a
277 series of founder events, it has been demonstrated that the newly founded populations
278 should experience striking geographical patterns in summary statistics describing *P. vivax*
279 genetic diversity (41). First, we investigated the nucleotide diversity (π) of our *P. vivax*
280 populations that includes at least five genomes per country, as it is expected to decrease
281 from the native origin to newly invaded areas (42). Interestingly, populations from Malaysia

282 and Indonesia displayed the highest genetic diversity, consistent with an Asian origin
283 hypothesis (Figure 4A). We then searched for the hypothetical origin and observed that the
284 decrease in genetic diversity with increasing geographic distance was maximal when
285 Southeast Asia was considered as the putative origin of *P. vivax*, consistent with an Asian
286 origin hypothesis (Supplementary Figure 10). Further evidence for this hypothesis was the
287 very strong correlation we observed considering the putative Asian origin located at the
288 border of Malaysia and Indonesia (spearman $r = -0.79$, p -value $< 8 \times 10^{-4}$, Figure 4B). By
289 using estimates of the number of cases per year obtained for each country (WHO 2010
290 report (43)), we observed that the correlation between the genetic diversity and distance
291 from Asia remains significant, regardless of the *P. vivax* number of cases (Generalized
292 Linear Models (GLM), p -value $< 1.98 \times 10^{-5}$).

293 Next, under a model of sequential founder effects during a range expansion, we
294 would expect linkage disequilibrium (LD) among loci to increase at each step of the
295 expansion, being lowest closed to native populations of origin and highest in populations that
296 are at the colonization front (44). We analyzed how LD decay within population varied as a
297 function of the geographic distance from the Asian putative origin identified based on the
298 genetic diversity (Figure 4C). A highly significant and strong positive correlation was
299 observed between LD at 250 bp within each population and the geographic distance to the
300 putative Asiatic origin ($\rho = 0.75$, p -value = 0.0015, Figure 4D). This result is thus also
301 consistent with a South Asian origin of *P. vivax* that would have spread to the rest of the
302 world.

303 Third, pattern of ancestral allele frequency (AAF) distributions can also inform on the
304 origin of *P. vivax* and its worldwide colonization routes (45). We polarized the AAF spectrum
305 of 94K SNPs in our panel using the *P. vivax-like* and *P. cynomolgi* and considered
306 populations with at least 20 genomes because sampling size can strongly impact the shape
307 of the AAF (Figure 4E). By comparing the shape of the AAF spectrum among populations,
308 we observed that populations located in Southeast Asia displayed more SNPs with high
309 AAFs (> 0.9) and fewer with low AAFs, while at the opposite, non-Asian population displayed
310 a progressive flattening of their AAFs distribution. Theoretical work and simulations
311 demonstrated that interplay between multiple demographic forces can have a major role in
312 the change of the AAF spectrum over time (41). Large size population would tend to
313 preserve the ancestral state of the variant loci, while small effective population size –that
314 experienced a severe bottleneck- would have more pronounced genetic drift, resulting in a
315 more rapid increase in derived allele frequencies. Our results based on the AAFs spectrum
316 show that population from Africa displayed a lower amount of ancestral allele than
317 population from Asia, consistent with a South Asian origin of *P. vivax*.

318 Finally, we analyzed the evolution of the Ne through time and their time to the most
319 recent common ancestor (TMRCA) for populations belonging to the different geographical
320 regions. In a serial founder-colonization model, the TMRCA is expected to decrease with
321 distance from the origin, as observed for humans (46). To test this hypothesis, we applied
322 the multiple sequentially Markovian coalescent (MSMC) approach to perform demographic
323 inference using five individuals from 14 *P. vivax* populations. As previously describe in Otto
324 et al. (31) we assumed a mutation rate (μ) per generation (g) of $\mu.g = 1.158 \times 10^{-9}$ and a
325 generation time of $g = 0.18$. As expected, our results show that the TMRCA values in the
326 populations from Southeast Asia were in most of the case older than the TMRCA values of
327 the populations from Africa and America (Figure 5). In addition, the MSMC curves revealed
328 that all populations displayed two distinct dynamics in their effective population sizes.
329 Populations from Southeast Asia experienced an increase in their effective population size in
330 a more distant past, followed by a steady decline. In contrast, populations from Africa and
331 America exhibited only a decrease in their effective population size, with a severe bottleneck
332 in a recent past.

333 Together, these results of *P. vivax* populations displaying a decrease in genetic
334 diversity and increase in LD with increasing geographic distance from Southeast Asia are
335 consistent with a model of serial founder events beginning from an origin located in
336 Southeast Asian, more precisely near Malaysia and Indonesia. The higher proportion of
337 ancestral alleles and older TMRCA found in the Southeast Asian populations are additional
338 evidence supporting this scenario. Therefore, these results are more consistent with a
339 scenario in which the ancestral populations that gave rise to the current human *P. vivax*
340 originated from Asia.

341 An Asian origin from a non-human primate raises the question about the fixation of
342 the Duffy negativity in sub-Saharan Africa as a result of an ancient presence of *P. vivax* on
343 that continent. If it is well admitted that the fixation of the FY*0 allele in sub-Saharan Africa
344 constitutes one of the fastest known selective sweep for any human gene (47). Given the
345 relatively mild clinical symptoms of modern *P. vivax*, it is unlikely that the evolutionary forces
346 that led to the allele fixation have been triggered by *P. vivax* alone (48). Especially in regard
347 of growing evidence showing the presence of endemic *P. vivax* circulating in sub-Saharan
348 Africa (including Duffy negative hosts), suggesting that the FY*0 allele may confer only
349 partial protection against *P. vivax* (16).

350 In conclusion, this study provides to our knowledge one of the most detailed view of
351 the worldwide distribution of the population genetic diversity and demographic history thus
352 far available. The present study focused intentionally on the global patterns of genetic
353 diversity in order to assess the origin of *P. vivax*, addressing especially the expectation
354 under the different hypotheses. We showed here that *P. vivax* is a sister group to *P. vivax*-

355 *like*, and not a sub-lineage. The genetic diversity in *P. vivax-like* is clearly richer than *P.*
356 *vivax*, consistent with a strong bottleneck in the lineage that gave rise to *P. vivax*. Our results
357 based on whole genome sequencing support of an out-of-Asia origin, rather than Africa, for
358 the world populations of *P. vivax*, with a clear signal of stepping stone colonization events
359 accompanied by serial founder effects. A specific effort should now be done to describe in
360 more details the demographic and selective history of the parasite in some particular regions
361 (e.g. South America, Africa, Europe) and estimate when and how these different regions
362 were colonized. The question of the host of origin still remain open.

363 Material and Methods

364 African *P. vivax* samples collection and ethical statements

365 Since only 27 human *P. vivax* genomes from the African continent were publicly
366 available, we augmented this dataset by sequencing 20 additional *P. vivax* genomes from
367 Mauritania (N=14), Ethiopia (N=3) and Sudan (N=3). *P. vivax* infections were diagnosed
368 using microscopy, PCR on the *Cytochrome b* gene and / or Rapid Detection Test (RDT).
369 Samples were collected from *P. vivax*-infected patients after obtaining informed consent and
370 following ethical approval at the local institutional review board of each country. The
371 informed consent procedure for the study consisted of a presentation of the aims of the
372 study to the community followed by invitation of adult individuals for enrolment. At the time of
373 sample collection, the purpose and design of the study was explained to each individual and
374 a study information sheet was provided before oral informed consent was collected. Oral
375 consent was provided since the study did not present any harm to subjects and did not
376 involve procedures for which written consent is required. The verbal consent process was
377 consistent with the ethical expectations for each country at the time of enrolment, approved
378 by each country and the ethics committees approved these procedures. Privacy and
379 confidentiality of the data collected were ensured by the anonymization of all samples before
380 the beginning of the study. For samples from Mauritania, the study received the approval
381 from the pediatric services of the National Hospital, the Cheikh Zayed Hospital and the
382 Direction régionale à l'Action Sanitaire de Nouakchott (DRAS)/Ministry of Health in
383 Mauritania. No ethics approval number obtained at this time. For samples from Sudan, no
384 specific consent was required, the human clinical, epidemiological and biological data were
385 collected in the CNRP database and analyzed in accordance with the common public health
386 mission of all French National Reference Centers
387 (<https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000810056&dateTexte=&categorieLien=id>). The study of the biological samples obtained in the medical care
388

389 context was considered as non-interventional research (article L1221-1.1 of the French
390 public health code) only requiring the non-opposition of the patient during sampling (article
391 L1211-2 of the French public health code). All data collected were anonymized before
392 analyses. For samples from Ethiopia, this study was approved by the Ethical Clearance
393 Committee of Haramaya University-College of Health and Medical Sciences, and from the
394 Harari and Oromia Regional State Health Bureau in Ethiopia.

395 *P. vivax-like* samples collection and ethical statements

396 *P. vivax-like* samples were obtained through a continuous survey of great ape
397 *Plasmodium* infections carried out in the Park of La Lékédi, in Gabon, in collaboration with
398 the Centre International de Recherches Médicales de Franceville (CIRMF). The park of La
399 Lékédi is a sanctuary for ape orphans in Gabon. During sanitary controls, blood samples
400 were collected and treated just after collection on the field using leukocyte depletion by CF11
401 cellulose column filtration (49), before being stored at -20°C at the CIRMF. The animals'
402 well-being was guaranteed by the veterinarians of the “Parc of La Lékédi” and the CIRMF,
403 who were responsible for the preceding sanitary procedures (including blood collection). All
404 animal work was conducted according to relevant national and international guidelines.
405 These investigations were approved by the Government of the Republic of Gabon and by
406 the Animal Life Administration of Libreville, Gabon (no. CITES 00956). It should be noted
407 that our study did not involve randomization or blinding.

408 *P. vivax-like* samples were also obtained from sylvatic anopheles mosquitoes
409 trapped with CDC light traps in the forest of the same park, during a longitudinal study (50).
410 Anopheles mosquitoes were morphologically identified by reference to standard keys (51),
411 stored in liquid nitrogen, returned at the CIRMF and kept at -80 °C until analysis.

412 Genomic DNA was extracted from each sample using DNeasy Blood and Tissue kit
413 (Qiagen, France) according to manufacturer’s recommendations. *P. vivax-like* samples were
414 identified by amplifying and sequencing either *Plasmodium* Cytochrome b (*Cytb*) or
415 Cytochrome oxidase 1 (*Cox1*) genes as described elsewhere (18, 52). This resulted in a
416 total of 10 *P. vivax-like* samples to sequence, including 3 from gorillas, 4 from chimpanzees
417 and 3 Anopheles mosquitoes.

418 African *P. vivax* and *P. vivax-like* genomes sequencing

419 To overcome host DNA contamination, selective whole genome amplification
420 (sWGA) was used to enrich submicroscopic DNA levels as already described elsewhere
421 (53). This technic preferentially amplifies *P. vivax* and *P. vivax-like* genomes from a set of
422 target DNAs. For each sample, the DNA amplification was carried out by the strand-
423 displacing phi29 DNA polymerase and two sets of *P. vivax*-specific primers that target short

424 (6 to 12 nucleotides) motifs commonly found in the *P. vivax* genome (set1920 and PvSet1)
425 (52, 53). For each set of primers separately, 30 ng of input DNA was added to a 50 µl
426 reaction mixture containing 3.5 µM of each sWGA primers, 30 units of phi29 DNA
427 polymerase enzyme (New England Biolabs), 1X phi29 buffer (New England Biolabs), 4 mM
428 of dNTPs (Invitrogen), 1% of Bovine Serum Albumine and sterile water. Amplifications were
429 carried out in a thermal cycler: a ramp down from 35°C to 30°C (10 min per degree), 16h at
430 30°C, 10 min at 65°C and hold at 4°C. For each sample, the two amplifications obtained
431 (one per set) were purified with AMPure XP beads (Beckman-Coulter) at a 1:1 ratio
432 according to the manufacturer's recommendations and pooled at equimolar concentrations.
433 Finally, each sWGA pooled library products were prepared using the Nextera XT DNA kit
434 (Illumina) according to the manufacturer's protocol. Samples were then pooled and clustered
435 on the HiSeq-2500 sequencer in Rapid Run mode with 2x250-bp paired end reads.

436 Genomic data from public and/or published data archives

437 To provide a worldwide comparative study of the *P. vivax* genetic diversity,
438 population structure and evolution, we performed a large literature search to identify
439 previously published genomic data set. We identified and collected fastq files from 1,134 *P.*
440 *vivax* isolates obtained from the following 12 different bioprojects: PRJEB10888 (14),
441 PRJEB2140 (12), PRJNA175266 (54), PRJNA240356 (13), PRJNA240452 , PRJNA240531,
442 PRJNA271480 (13), PRJNA284437 (55), PRJNA350554 (56), PRJNA432819 (57),
443 PRJNA432819 (58), PRJNA65119 (59). In addition, 17 *P. vivax-like* sequenced genomes
444 from Cameroon, Ivory Coast and Gabon were collected from two different bioprojects
445 (PRJNA474492 and PRJEB2579) (23, 25). Published genome were downloaded from the
446 National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (60)
447 and converted into fastq files using the NCBI SRA-Toolkit (*fastq-dump* -split-3). Details
448 about the genomic samples are provided in Supplementary Table 1.

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

450 Both newly generated and previously published sequencing reads were trimmed for
451 adapters and pre-processed to remove low quality reads (--quality-cutoff=30) using the
452 program *cutadapt* (61). Reads shorter than 50bp and containing 'N' were discarded (--
453 minimum-length=50 --max-n=0). Sequenced reads were aligned to the PVP01 (62) *P. vivax*
454 reference genome using *bwa-mem* (63). Here, a first round of filtration was applied to our
455 isolates by excluding isolates having an average genome coverage depth lower than 5x. We
456 used the *Genome Analysis Toolkit* (GATK, version 3.8.0) (64) to identify SNPs in each
457 isolate following GATK best practices. Duplicate reads were marked using the
458 *MarkDuplicates* tool from the Picard tools (65) with default options. Local realignment around

459 indels was performed using the *IndelRealigner* tool from *GATK*. Variant calling was carried
460 out using *HaplotypeCaller* module in *GATK*, on reads mapped with a “reads minimum
461 mapping quality” of 30 (-mmq 30) and a minimum base quality > 20 (--
462 min_base_quality_score 20). We filtered out low quality SNP according to *GATK* best-
463 practice recommendations and kept SNP call at site covered by at least 5 reads. Finally,
464 individual isolate VCF files were merge using the *GATK* module *GenotypeGVCFs*. Because
465 of our tolerant cutoff of genome depth of coverage set up at 5x, we apply a second round of
466 filtration to our data by excluding variants and isolates with a missing call rate >50%. To
467 evaluate the possibility of co-infections we follow the method developed by Chan et al. (26)
468 and examined the reference allele frequency distributions (RAF). Single infection would
469 display a strict RAF pattern as all variants will carry either the reference or a single alternate
470 allele (*i.e* RAF of 100% or 0%). For each of the samples, we calculated the proportion of
471 reads carrying the reference allele to display the RAF distribution.

472 Heterogenous patterns of relatedness along *Plasmodium* genomes.

473 In order to identify heterogenous patterns of relatedness along *Plasmodium*
474 genomes, we conducted a local-PCA analysis, as described in Li et al., 2018 (28). Briefly,
475 the *P. vivax* genome was divided into 1,439 contiguous and non-overlapping windows of 100
476 SNPs. On each window, we applied a principal component analysis (PCA) and stored the
477 individual isolate scores for the first two principal components. To measure the similarity
478 among windows, a Euclidian distance matrix is computed among windows based on the PC
479 scores. Finally, a multidimensional scaling (MDS) was used to visualize the relationships
480 among windows. We used a set of three coordinates to visualize the patterns of relatedness
481 shared among windows. Because the local PCA analysis was sensitive to missing data, we
482 selected the top 304 *P. vivax* genomes having a SNPs missing discovery rate lower than
483 25% to run this analysis.

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

485 Recombination among *P. vivax-like* strains was inferred using fastGEAR software (29). From
486 the SNPs present in the core genome we created a multi-fasta file used as input for
487 fastGEAR that we launched with the iteration number set to 15 (default). Recent
488 recombination events were detected with the Bayesian factor (BF) > 1 (default) and refers to
489 inter-lineage recombination for which the donor-recipient relation can be inferred.
490 Phylogenetic network tree was generated using SplitTree 4 software (66) using the
491 polymorphic sites present on the core genome.

492 Population structure analysis

493 Genetic relationships between populations and species were assessed and
494 visualized using a distance based NJ trees, PCA and by *ADMIXTURE* analyses. PCA and
495 *ADMIXTURE* analyses were performed after selecting only bi-allelic SNPs present in the
496 core *P. vivax* genome, excluding singleton from the dataset. The variants were linkage-
497 disequilibrium-pruned to obtain a set of unlinked variants using the option --indep-pairwise
498 50 5 1.5 in PLINK (67). PCA analysis was performed using PLINK and display in R.
499 Assessment of population structure and estimation of isolate individual ancestry to various
500 number (K) of ancestral populations was performed using *ADMIXTURE* (34). Each
501 *ADMIXTURE* analysis was repeated 100 times with different random seeds, with a K value
502 ranging from 2 to 20. The most likely number of ancestral populations (K) was determined
503 using the cross-validation error. We then used *CLUMPAK* (68) to analyze *ADMIXTURE*
504 outputs and compute ancestry proportions. NJ tree was estimated using the *ape* R package
505 (69) using the full SNP dataset present on the core *P. vivax* genome (without performing any
506 SNP LD-pruning) and plotted with *figtree* (70). The reliability of branching order was
507 assessed by performing 1,000 bootstrap replicates. In order to use *P. cynomolgi* (strain M)
508 as outgroup we used the tools *genome liftOver* (71) to translate our SNP coordinates from
509 the *P. vivax* assembly to the *P. cynomolgi* assembly. Genetic differentiation among
510 populations was estimated using the Weir and Cockerham's estimator of F_{ST} using *VCFtools*
511 (72). This metric accounts for differences in the sample size in each population.

512 Demographic history analysis

513 To investigate the demographic history of *P. vivax* and test distinct scenarios, we
514 compute several statistics that describing different features of genetic diversity within
515 populations. Each of these analyses was computed for each of the populations defined by
516 country of origin using the SNPs present on the core genome of the reference. To minimize
517 the impact of missing data on our analyses only individuals with less than 5% of missing
518 data were considered for the following analysis. The genome-wide nucleotide diversity (π)
519 was calculated by averaging the number of nucleotide differences between pairs of DNA
520 sequences at each SNP position using *VCFtools*. LD-decay was estimated by randomly
521 sampling five individuals from each population using the tools *PopLDdecay* (73), that
522 calculate the genotype correlation coefficient R^2 for pairs of SNPs at a maximum range of
523 distances of 10kb. Finally, we polarized the bi-allelic SNPs of each *P. vivax* population using
524 shared alleles identified between *P. vivax-like* individuals and *P. cynomolgi*. The final
525 unfolded SFS was computed for a total of 93,652 SNPs. Regression of nucleotide diversity
526 against geographic distance were calculate as described in Fontaine et al., (74). Briefly, we

527 divided the world map into a 400 x 500 pixel two-dimensional lattice, and considered each
528 pixel as a potential source for the geographic expansion of *P. vivax*. The geographic
529 distances between each population and the focal pixel were determined using the R
530 package *geosphere*, to then calculate the spearman coefficient of correlation between the
531 nucleotide diversity and the geographic distance. The pixel with the lowest negative
532 correlation coefficient is thus indicative of the origin.

533 Estimates of the effective population size

534 To infer historical changes in effective population size (N_e) and estimate the TMRCA
535 in the *P. vivax* and *P. vivax-like* populations, we ran the MSMC program (75) on the core
536 genome of all chromosomes. Homozygous SNPs present on each *Plasmodium*
537 chromosome were considered as single phased haplotype. We run MSMC for 20 iterations
538 with a fixed recombination rate. The MSMC method was applied by selecting randomly five
539 individuals from each population with a missing discovery rate lower than 5%. The error
540 around our estimates was estimated by bootstrapping 50 replicates by randomly resampling
541 from the segregating sites used.

542 Supplementary Materials

543 **Supplementary Table 1:** Genome statistics and metadata. Country of origin, GeneBank
544 accession number and several *in silico* results for the 1,154 human *P. vivax* and 27 *P. vivax-*
545 *like* strains analyzed in this study.

546

547 **Supplementary Figure 1:** Distribution of the mean sequencing depth of *P. vivax* and *P.*
548 *vivax-like* whole-genome sequence. Genomes with a sequencing depth $\geq 5x$ (dash line) were
549 selected for downstream analysis.

550

551 **Supplementary Figure 2:** Rare allele frequencies distribution. For each population of *P.*
552 *vivax* (PV) and *P. vivax-like* (PVL), the distribution of polymorphic sites (y-axis) is represented
553 as a function of the proportion of reads carrying the reference allele (x-axis). Almost all
554 polymorphic sites, are supported by reads carrying either the reference or a single alternate
555 allele, suggesting the presence of single infection.

556

557 **Supplementary Figure 3:** Genome scan of the SNP density (left y-axis) along the 14
558 chromosomes for *P. vivax* and *P. vivax-like* combined. The black line represents the number
559 of SNPs shared between the two species. The right y-axis represents the scores along the

560 first axis of the multidimensional scaling analysis (MDS1) for each genomic window along
561 the genome.

562

563 **Supplementary Figure 4:** Scatter plots showing the results of the two PCAs conducted
564 using the SNPs located either in the core genome or in the hypervariable sub-telomeric
565 regions.

566

567 **Supplementary Figure 5:** Reticulate network based on SplitTree4 (66), drawn from 19 *P.*
568 *vivax-like* genomes. The reticulate networks split the *P. vivax-like* genomes in two distinct
569 clades. On the tree, reticulation indicates likely occurrence of recombination. No reticulation
570 was observed among the two clades suggesting the absence of recombination between the
571 two clades.

572

573 **Supplementary Figure 6:** PCA analysis of the 447 of *P. vivax* using SNPs present in the
574 core genome. Scatter plot shows individual strain relationships along the principle
575 components 3 to 6. Dot colors indicate populations. The bar chart shows the percentage of
576 variance explained by each principal component axis.

577

578 **Supplementary Figure 7:** Individual ancestry proportion to each ancestral population tested
579 (from K=2 to K=10) estimated using the ADMIXTURE program for the 447 *P. vivax*
580 genomes. Individual strains are sorted by country with increasing longitude.

581

582 **Supplementary Figure 8:** Cross validation error rate estimated using the ADMIXTURE
583 program for each ancestral populations tested (K between 2 and 20) .We chose K=5 to
584 analyze the SNPs data, as the value that minimizes the error.

585

586 **Supplementary Figure 9:** Average allele frequency differentiation estimated using F_{ST}
587 values between pairs of populations.

588

589 **Supplementary Figure 10:** Genetic diversity of *P. vivax* regressed on geographic distance
590 across the world. The value at each pixel of the map corresponds to the Spearman
591 correlation coefficient (r) between the expected genetic diversity in each population and the
592 geographic distance between this population and the pixel. The black dots represent the
593 sampling sites used in the regression (where $n \geq 5$ individuals).

594

595 **Supplementary Figure 11:** Multiple sequentially Markovian coalescent (MSMC) estimates
596 of the effective population size (N_e) in 14 *P. vivax* populations. The y-axis shows the \log_{10}

597 of Ne. Gray lines represent the MSMC results obtained from 50 bootstrap resampling of the
598 segregating sites.

599

600 Reference

- 601 1. C. A. Guerra, R. E. Howes, A. P. Patil, P. W. Gething, T. P. Van Boeckel, W.
602 H. Temperley, C. W. Kabaria, A. J. Tatem, B. H. Manh, I. R. F. Elyazar, J. K. Baird,
603 R. W. Snow, S. I. Hay, The international limits and population at risk of *Plasmodium*
604 *vivax* transmission in 2009. *PLoS Negl. Trop. Dis.* **4**, e774 (2010).
- 605 2. R. E. Howes, K. E. Battle, K. N. Mendis, D. L. Smith, R. E. Cibulskis, J. K.
606 Baird, S. I. Hay, Global Epidemiology of *Plasmodium vivax*. *Am. J. Trop. Med. Hyg.*
607 **95**, 15–34 (2016).
- 608 3. R. Horuk, C. E. Chitnis, W. C. Darbonne, T. J. Colby, A. Rybicki, T. J. Hadley,
609 L. H. Miller, A receptor for the malarial parasite *Plasmodium vivax*: the erythrocyte
610 chemokine receptor. *Science*. **261**, 1182–1184 (1993).
- 611 4. A. Chaudhuri, V. Zbrzezna, J. Polyakova, A. O. Pogo, J. Hesselgesser, R.
612 Horuk, Expression of the Duffy antigen in K562 cells. Evidence that it is the human
613 erythrocyte chemokine receptor. *J. Biol. Chem.* **269**, 7835–7838 (1994).
- 614 5. H. Noedl, Y. Se, K. Schaecher, B. L. Smith, D. Socheat, M. M. Fukuda,
615 Artemisinin Resistance in Cambodia 1 (ARC1) Study Consortium, Evidence of
616 artemisinin-resistant malaria in western Cambodia. *N. Engl. J. Med.* **359**, 2619–2620
617 (2008).
- 618 6. D. Ménard, C. Barnadas, C. Bouchier, C. Henry-Halldin, L. R. Gray, A.
619 Ratsimbaoa, V. Thonier, J.-F. Carod, O. Domarle, Y. Colin, O. Bertrand, J. Picot, C.
620 L. King, B. T. Grimberg, O. Mercereau-Pujalon, P. A. Zimmerman, *Plasmodium*
621 *vivax* clinical malaria is commonly observed in Duffy-negative Malagasy people.
622 *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5967–5971 (2010).
- 623 7. J. K. Baird, Evidence and implications of mortality associated with acute
624 *Plasmodium vivax* malaria. *Clin. Microbiol. Rev.* **26**, 36–57 (2013).
- 625 8. S. Gunawardena, N. D. Karunaweera, M. U. Ferreira, M. Phone-Kyaw, R. J.
626 Pollack, M. Alifrangis, R. S. Rajakaruna, F. Konradsen, P. H. Amerasinghe, M. L.
627 Schousboe, G. N. L. Galappaththy, R. R. Abeyasinghe, D. L. Hartl, D. F. Wirth,
628 Geographic structure of *Plasmodium vivax*: microsatellite analysis of parasite
629 populations from Sri Lanka, Myanmar, and Ethiopia. *Am. J. Trop. Med. Hyg.* **82**,
630 235–242 (2010).
- 631 9. P. Orjuela-Sánchez, J. M. Sá, M. C. C. Brandi, P. T. Rodrigues, M. S. Bastos,
632 C. Amaratunga, S. Duong, R. M. Fairhurst, M. U. Ferreira, Higher microsatellite
633 diversity in *Plasmodium vivax* than in sympatric *Plasmodium falciparum* populations
634 in Pursat, Western Cambodia. *Exp. Parasitol.* **134**, 318–326 (2013).
- 635 10. C. Koepfli, L. Timinao, T. Antao, A. E. Barry, P. Siba, I. Mueller, I. Felger, A
636 Large *Plasmodium vivax* Reservoir and Little Population Structure in the South
637 Pacific. *PloS One.* **8**, e66041 (2013).

- 638 11. A. Melnikov, K. Galinsky, P. Rogov, T. Fennell, D. Van Tyne, C. Russ, R.
639 Daniels, K. G. Barnes, J. Bochicchio, D. Ndiaye, P. D. Sene, D. F. Wirth, C.
640 Nusbaum, S. K. Volkman, B. W. Birren, A. Gnirke, D. E. Neafsey, Hybrid selection
641 for sequencing pathogen genomes from clinical samples. *Genome Biol.* **12**, R73
642 (2011).
- 643 12. R. D. Pearson, R. Amato, S. Auburn, O. Miotto, J. Almagro-Garcia, C.
644 Amaratunga, S. Suon, S. Mao, R. Noviyanti, H. Trimarsanto, J. Marfurt, N. M.
645 Anstey, T. William, M. F. Boni, C. Dolecek, T. T. Hien, N. J. White, P. Michon, P.
646 Siba, L. Tavul, G. Harrison, A. Barry, I. Mueller, M. U. Ferreira, N. Karunaweera, M.
647 Randrianarivelojosa, Q. Gao, C. Hubbart, L. Hart, B. Jeffery, E. Drury, D. Mead, M.
648 Kekre, S. Campino, M. Manske, V. J. Cornelius, B. MacInnis, K. A. Rockett, A. Miles,
649 J. C. Rayner, R. M. Fairhurst, F. Nosten, R. N. Price, D. P. Kwiatkowski, Genomic
650 analysis of local variation and recent evolution in *Plasmodium vivax*. *Nat. Genet.* **48**,
651 959–964 (2016).
- 652 13. D. N. Hupalo, Z. Luo, A. Melnikov, P. L. Sutton, P. Rogov, A. Escalante, A. F.
653 Vallejo, S. Herrera, M. Arévalo-Herrera, Q. Fan, Y. Wang, L. Cui, C. M. Lucas, S.
654 Durand, J. F. Sanchez, G. C. Baldeviano, A. G. Lescano, M. Laman, C. Barnadas, A.
655 Barry, I. Mueller, J. W. Kazura, A. Eapen, D. Kanagaraj, N. Valecha, M. U. Ferreira,
656 W. Roobsoong, W. Nguitragool, J. Sattabonkot, D. Gamboa, M. Kosek, J. M. Vinetz,
657 L. González-Cerón, B. W. Birren, D. E. Neafsey, J. M. Carlton, Population genomics
658 studies identify signatures of global dispersal and drug resistance in *Plasmodium*
659 *vivax*. *Nat. Genet.* **48**, 953–958 (2016).
- 660 14. S. Auburn, S. Getachew, R. D. Pearson, R. Amato, O. Miotto, H. Trimarsanto,
661 S. J. Zhu, A. Rumaseb, J. Marfurt, R. Noviyanti, M. J. Grigg, B. Barber, T. William, S.
662 M. Goncalves, E. Drury, K. Sriprawat, N. M. Anstey, F. Nosten, B. Petros, A. Aseffa,
663 G. McVean, D. P. Kwiatkowski, R. N. Price, Genomic Analysis of *Plasmodium vivax*
664 in Southern Ethiopia Reveals Selective Pressures in Multiple Parasite Mechanisms.
665 *J. Infect. Dis.* **220**, 1738–1749 (2019).
- 666 15. S. Auburn, E. D. Benavente, O. Miotto, R. D. Pearson, R. Amato, M. J. Grigg,
667 B. E. Barber, T. William, I. Handayani, J. Marfurt, H. Trimarsanto, R. Noviyanti, K.
668 Sriprawat, F. Nosten, S. Campino, T. G. Clark, N. M. Anstey, D. P. Kwiatkowski, R.
669 N. Price, Genomic analysis of a pre-elimination Malaysian *Plasmodium vivax*
670 population reveals selective pressures and changing transmission dynamics. *Nat.*
671 *Commun.* **9**, 2585 (2018).
- 672 16. K. A. Twohig, D. A. Pfeffer, J. K. Baird, R. N. Price, P. A. Zimmerman, S. I.
673 Hay, P. W. Gething, K. E. Battle, R. E. Howes, Growing evidence of *Plasmodium*
674 *vivax* across malaria-endemic Africa. *PLoS Negl. Trop. Dis.* **13**, e0007140 (2019).
- 675 17. A. A. Escalante, O. E. Cornejo, D. E. Freeland, A. C. Poe, E. Durrego, W. E.
676 Collins, A. A. Lal, A monkey's tale: the origin of *Plasmodium vivax* as a human
677 malaria parasite. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 1980–1985 (2005).
- 678 18. J. M. Carlton, A. Das, A. A. Escalante, Genomics, population genetics and
679 evolutionary history of *Plasmodium vivax*. *Adv. Parasitol.* **81**, 203–222 (2013).
- 680 19. F. Prugnolle, V. Rougeron, P. Becquart, A. Berry, B. Makanga, N. Rahola, C.
681 Arnathau, B. Ngoubangoye, S. Menard, E. Willaume, F. J. Ayala, D. Fontenille, B.

- 682 Ollomo, P. Durand, C. Paupy, F. Renaud, Diversity, host switching and evolution of
683 *Plasmodium vivax* infecting African great apes. *Proc. Natl. Acad. Sci. U. S. A.* **110**,
684 8123–8128 (2013).
- 685 20. C. Koepfli, P. T. Rodrigues, T. Antao, P. Orjuela-Sánchez, P. Van den Eede,
686 D. Gamboa, N. van Hong, J. Bendezu, A. Erhart, C. Barnadas, A. Ratsimbaoa, D.
687 Menard, C. Severini, M. Menegon, B. Y. M. Nour, N. Karunaweera, I. Mueller, M. U.
688 Ferreira, I. Felger, *Plasmodium vivax* Diversity and Population Structure across Four
689 Continents. *PLoS Negl. Trop. Dis.* **9**, e0003872 (2015).
- 690 21. V. Rougeron, E. Elguero, C. Arnathau, B. Acuña Hidalgo, P. Durand, S.
691 Houze, A. Berry, S. Zakeri, R. Haque, M. Shafiul Alam, F. Nosten, C. Severini, T.
692 Gebru Woldearegai, B. Mordmüller, P. G. Kremsner, L. González-Cerón, G.
693 Fontecha, D. Gamboa, L. Musset, E. Legrand, O. Noya, T. Pumpaibool, P.
694 Harnyuttanakorn, K. M. Lekweiry, M. Mohamad Albsheer, M. Mahdi Abdel Hamid, A.
695 O. M. S. Boukary, J.-F. Trape, F. Renaud, F. Prugnolle, Human *Plasmodium vivax*
696 diversity, population structure and evolutionary origin. *PLoS Negl. Trop. Dis.* **14**
697 (2020), doi:10.1371/journal.pntd.0008072.
- 698 22. W. Liu, Y. Li, K. S. Shaw, G. H. Learn, L. J. Plenderleith, J. A. Malenke, S. A.
699 Sundararaman, M. A. Ramirez, P. A. Crystal, A. G. Smith, F. Bibollet-Ruche, A.
700 Ayoub, S. Locatelli, A. Esteban, F. Mouacha, E. Guichet, C. Butel, S. Ahuka-
701 Mundeke, B.-I. Inogwabini, J.-B. N. Ndjango, S. Speede, C. M. Sanz, D. B. Morgan,
702 M. K. Gonder, P. J. Kranzusch, P. D. Walsh, A. V. Georgiev, M. N. Muller, A. K. Piel,
703 F. A. Stewart, M. L. Wilson, A. E. Pusey, L. Cui, Z. Wang, A. Färnert, C. J.
704 Sutherland, D. Nolder, J. A. Hart, T. B. Hart, P. Bertolani, A. Gillis, M. LeBreton, B.
705 Tafon, J. Kiyang, C. F. Djoko, B. S. Schneider, N. D. Wolfe, E. Mpoudi-Ngole, E.
706 Delaporte, R. Carter, R. L. Culleton, G. M. Shaw, J. C. Rayner, M. Peeters, B. H.
707 Hahn, P. M. Sharp, African origin of the malaria parasite *Plasmodium vivax*. *Nat.*
708 *Commun.* **5**, 3346 (2014).
- 709 23. D. E. Loy, L. J. Plenderleith, S. A. Sundararaman, W. Liu, J. Gruszczyk, Y.-J.
710 Chen, S. Trimboli, G. H. Learn, O. A. MacLean, A. L. K. Morgan, Y. Li, A. N. Avitto,
711 J. Giles, S. Calvignac-Spencer, A. Sachse, F. H. Leendertz, S. Speede, A. Ayoub,
712 M. Peeters, J. C. Rayner, W.-H. Tham, P. M. Sharp, B. H. Hahn, Evolutionary history
713 of human *Plasmodium vivax* revealed by genome-wide analyses of related ape
714 parasites. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E8450–E8459 (2018).
- 715 24. J. Haldane, Disease and evolution. *Ric. Sci.* **19**, 68–76 (1949).
- 716 25. A. Gilabert, T. D. Otto, G. G. Rutledge, B. Franzon, B. Ollomo, C. Arnathau,
717 P. Durand, N. D. Moukodoum, A.-P. Okouga, B. Ngoubangoye, B. Makanga, L.
718 Boundenga, C. Paupy, F. Renaud, F. Prugnolle, V. Rougeron, *Plasmodium vivax*-like
719 genome sequences shed new insights into *Plasmodium vivax* biology and evolution.
720 *PLoS Biol.* **16**, e2006035 (2018).
- 721 26. E. R. Chan, J. W. Barnwell, P. A. Zimmerman, D. Serre, Comparative analysis
722 of field-isolate and monkey-adapted *Plasmodium vivax* genomes. *PLoS Negl. Trop.*
723 *Dis.* **9**, e0003566 (2015).
- 724 27. Anopheles gambiae 1000 Genomes Consortium, Data analysis group, Partner
725 working group, Sample collections—Angola:, Burkina Faso:, Cameroon:, Gabon:,

- 726 Guinea:, Guinea-Bissau:, Kenya:, Uganda:, Crosses:, Sequencing and data
727 production, Web application development, Project coordination, Genetic diversity of
728 the African malaria vector *Anopheles gambiae*. *Nature*. **552**, 96–100 (2017).
- 729 28. H. Li, P. Ralph, Local PCA Shows How the Effect of Population Structure
730 Differs Along the Genome. *Genetics*. **211**, 289–304 (2019).
- 731 29. R. Mostowy, N. J. Croucher, C. P. Andam, J. Corander, W. P. Hanage, P.
732 Marttinen, Efficient Inference of Recent and Ancestral Recombination within
733 Bacterial Populations. *Mol. Biol. Evol.* **34**, 1167–1182 (2017).
- 734 30. L. Boundenga, B. Ollomo, V. Rougeron, L. Y. Mouele, B. Mve-Ondo, L. M.
735 Delicat-Loembet, N. D. Moukodoum, A. P. Okouga, C. Arnathau, E. Elguero, P.
736 Durand, F. Liégeois, V. Boué, P. Motsch, G. Le Flohic, A. Ndoungouet, C. Paupy, C.
737 T. Ba, F. Renaud, F. Prugnolle, Diversity of malaria parasites in great apes in
738 Gabon. *Malar. J.* **14**, 111 (2015).
- 739 31. T. D. Otto, A. Gilabert, T. Crellen, U. Böhme, C. Arnathau, M. Sanders, S. O.
740 Oyola, A. P. Okouga, L. Boundenga, E. Willaume, B. Ngoubangoye, N. D.
741 Moukodoum, C. Paupy, P. Durand, V. Rougeron, B. Ollomo, F. Renaud, C. Newbold,
742 M. Berriman, F. Prugnolle, Genomes of all known members of a *Plasmodium*
743 subgenus reveal paths to virulent human malaria. *Nat. Microbiol.* **3**, 687–697 (2018).
- 744 32. J. Prado-Martinez, P. H. Sudmant, J. M. Kidd, H. Li, J. L. Kelley, B. Lorente-
745 Galdos, K. R. Veeramah, A. E. Woerner, T. D. O’Connor, G. Santpere, A. Cagan, C.
746 Theunert, F. Casals, H. Laayouni, K. Munch, A. Hobolth, A. E. Halager, M. Malig, J.
747 Hernandez-Rodriguez, I. Hernando-Herraez, K. Prüfer, M. Pybus, L. Johnstone, M.
748 Lachmann, C. Alkan, D. Twigg, N. Petit, C. Baker, F. Hormozdiari, M. Fernandez-
749 Callejo, M. Dabad, M. L. Wilson, L. Stevison, C. Camprubí, T. Carvalho, A. Ruiz-
750 Herrera, L. Vives, M. Mele, T. Abello, I. Kondova, R. E. Bontrop, A. Pusey, F.
751 Lankester, J. A. Kiyang, R. A. Bergl, E. Lonsdorf, S. Myers, M. Ventura, P. Gagneux,
752 D. Comas, H. Siegismund, J. Blanc, L. Agueda-Calpena, M. Gut, L. Fulton, S. A.
753 Tishkoff, J. C. Mullikin, R. K. Wilson, I. G. Gut, M. K. Gonder, O. A. Ryder, B. H.
754 Hahn, A. Navarro, J. M. Akey, J. Bertranpetit, D. Reich, T. Mailund, M. H. Schierup,
755 C. Hvilsom, A. M. Andrés, J. D. Wall, C. D. Bustamante, M. F. Hammer, E. E.
756 Eichler, T. Marques-Bonet, Great ape genetic diversity and population history.
757 *Nature*. **499**, 471–475 (2013).
- 758 33. G. McVean, A genealogical interpretation of principal components analysis.
759 *PLoS Genet.* **5**, e1000686 (2009).
- 760 34. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of
761 ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
- 762 35. A. Crowther, L. Lucas, R. Helm, M. Horton, C. Shipton, H. T. Wright, S.
763 Walshaw, M. Pawlowicz, C. Radimilahy, K. Douka, L. Picornell-Gelabert, D. Q.
764 Fuller, N. L. Boivin, Ancient crops provide first archaeological signature of the
765 westward Austronesian expansion. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 6635–6640
766 (2016).
- 767 36. L. M. Kootker, L. Mbeki, A. G. Morris, H. Kars, G. R. Davies, Dynamics of
768 Indian Ocean Slavery Revealed through Isotopic Data from the Colonial Era Cobern
769 Street Burial Site, Cape Town, South Africa (1750-1827). *PLoS One*. **11**, e0157750

- 770 (2016).
- 771 37. P. T. Rodrigues, H. O. Valdivia, T. C. de Oliveira, J. M. P. Alves, A. M. R. C.
772 Duarte, C. Cerutti-Junior, J. C. Buery, C. F. A. Brito, J. C. de Souza, Z. M. B. Hirano,
773 M. G. Bueno, J. L. Catão-Dias, R. S. Malafronte, S. Ladeia-Andrade, T. Mita, A. M.
774 Santamaria, J. E. Calzada, I. S. Tantular, F. Kawamoto, L. R. J. Raijmakers, I.
775 Mueller, M. A. Pacheco, A. A. Escalante, I. Felger, M. U. Ferreira, Human migration
776 and the spread of malaria parasites to the New World. *Sci. Rep.* **8**, 1993 (2018).
- 777 38. L. van Dorp, P. Gelabert, A. Rieux, M. de Manuel, T. de-Dios, S.
778 Gopalakrishnan, C. Carøe, M. Sandoval-Velasco, R. Fregel, I. Olalde, R. Escosa, C.
779 Aranda, S. Huijben, I. Mueller, T. Marquès-Bonet, F. Balloux, M. T. P. Gilbert, C.
780 Lalueza-Fox, Plasmodium vivax Malaria viewed through the lens of an eradicated
781 European strain. *Mol. Biol. Evol.* (2019), doi:10.1093/molbev/msz264.
- 782 39. O. Miotto, R. Amato, E. A. Ashley, B. MacInnis, J. Almagro-Garcia, C.
783 Amaratunga, P. Lim, D. Mead, S. O. Oyola, M. Dhorda, M. Imwong, C. Woodrow, M.
784 Manske, J. Stalker, E. Drury, S. Campino, L. Amenga-Etego, T.-N. N. Thanh, H. T.
785 Tran, P. Ringwald, D. Bethell, F. Nosten, A. P. Phy, S. Pukrittayakamee, K.
786 Chotivanich, C. M. Chuor, C. Nguon, S. Suon, S. Sreng, P. N. Newton, M. Mayxay,
787 M. Khanthavong, B. Hongvanthong, Y. Htut, K. T. Han, M. P. Kyaw, M. A. Faiz, C. I.
788 Fanello, M. Onyamboko, O. A. Mokuolu, C. G. Jacob, S. Takala-Harrison, C. V.
789 Plowe, N. P. Day, A. M. Dondorp, C. C. A. Spencer, G. McVean, R. M. Fairhurst, N.
790 J. White, D. P. Kwiatkowski, Genetic architecture of artemisinin-resistant
791 Plasmodium falciparum. *Nat. Genet.* **47**, 226–234 (2015).
- 792 40. F. Rousset, Genetic differentiation and estimation of gene flow from F-
793 statistics under isolation by distance. *Genetics*. **145**, 1219–1228 (1997).
- 794 41. M. DeGiorgio, M. Jakobsson, N. A. Rosenberg, Explaining worldwide patterns
795 of human genetic variation using a coalescent-based serial founder model of
796 migration outward from Africa. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 16057–16062
797 (2009).
- 798 42. F. Prugnolle, A. Manica, F. Balloux, Geography predicts neutral genetic
799 diversity of human populations. *Curr. Biol. CB.* **15**, R159-160 (2005).
- 800 43. WHO | World Malaria Report 2010. WHO, (available at
801 <https://www.who.int/malaria/publications/atoz/9789241564106/en/>).
- 802 44. M. Jakobsson, S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H.-C.
803 Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, R. Guerreiro, J. M. Bras, J. C.
804 Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton,
805 J. van de Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg,
806 A. B. Singleton, Genotype, haplotype and copy-number variation in worldwide
807 human populations. *Nature*. **451**, 998–1003 (2008).
- 808 45. J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S.
809 Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, R. M.
810 Myers, Worldwide human relationships inferred from genome-wide patterns of
811 variation. *Science*. **319**, 1100–1104 (2008).
- 812 46. W. Shi, Q. Ayub, M. Vermeulen, R. Shao, S. Zuniga, K. van der Gaag, P. de
813 Knijff, M. Kayser, Y. Xue, C. Tyler-Smith, A Worldwide Survey of Human Male

- 814 Demographic History Based on Y-SNP and Y-STR Data from the HGDP–CEPH
815 Populations. *Mol. Biol. Evol.* **27**, 385–393 (2010).
- 816 47. K. F. McManus, A. M. Taravella, B. M. Henn, C. D. Bustamante, M. Sikora, O.
817 E. Cornejo, Population genetic analysis of the DARC locus (Duffy) reveals
818 adaptation from standing variation associated with malaria resistance in humans.
819 *PLoS Genet.* **13**, e1006560 (2017).
- 820 48. B. Roche, V. Rougeron, L. Quintana-Murci, F. Renaud, J. L. Abbate, F.
821 Prugnolle, Might Interspecific Interactions between Pathogens Drive Host Evolution?
822 The Case of Plasmodium Species and Duffy-Negativity in Human Populations.
823 *Trends Parasitol.* **33**, 21–29 (2017).
- 824 49. M. Venkatesan, C. Amaratunga, S. Campino, S. Auburn, O. Koch, P. Lim, S.
825 Uk, D. Socheat, D. P. Kwiatkowski, R. M. Fairhurst, C. V. Plowe, Using CF11
826 cellulose columns to inexpensively and effectively remove human DNA from
827 Plasmodium falciparum-infected whole blood samples. *Malar. J.* **11**, 41 (2012).
- 828 50. B. Makanga, P. Yangari, N. Rahola, V. Rougeron, E. Elguero, L. Boundenga,
829 N. D. Moukodoum, A. P. Okouga, C. Arnathau, P. Durand, E. Willaume, D. Ayala, D.
830 Fontenille, F. J. Ayala, F. Renaud, B. Ollomo, F. Prugnolle, C. Paupy, Ape malaria
831 transmission and potential for ape-to-human transfers in Africa. *Proc. Natl. Acad.*
832 *Sci. U. S. A.* **113**, 5329–5334 (2016).
- 833 51. M. Gillies, M. Coetzee, A supplement to the Anophelinae of Africa South of
834 the Sahara. *Publ Afr Inst Med Res.* **55**, 1–143 (1987).
- 835 52. S. A. Sundararaman, W. Liu, B. F. Keele, G. H. Learn, K. Bittinger, F.
836 Mouacha, S. Ahuka-Mundeye, M. Manske, S. Sherrill-Mix, Y. Li, J. A. Malenke, E.
837 Delaporte, C. Laurent, E. Mpoudi Ngole, D. P. Kwiatkowski, G. M. Shaw, J. C.
838 Rayner, M. Peeters, P. M. Sharp, F. D. Bushman, B. H. Hahn, Plasmodium
839 falciparum-like parasites infecting wild apes in southern Cameroon do not represent
840 a recurrent source of human malaria. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 7020–7025
841 (2013).
- 842 53. A. N. Cowell, D. E. Loy, S. A. Sundararaman, H. Valdivia, K. Fisch, A. G.
843 Lescano, G. C. Baldeviano, S. Durand, V. Gerbasi, C. J. Sutherland, D. Nolder, J. M.
844 Vinetz, B. H. Hahn, E. A. Winzeler, Selective Whole-Genome Amplification Is a
845 Robust Method That Enables Scalable Whole-Genome Sequencing of Plasmodium
846 vivax from Unprocessed Clinical Samples. *mBio.* **8** (2017), doi:10.1128/mBio.02257-
847 16.
- 848 54. E. R. Chan, D. Menard, P. H. David, A. Ratsimbaoa, S. Kim, P. Chim, C. Do,
849 B. Witkowski, O. Mercereau-Puijalon, P. A. Zimmerman, D. Serre, Whole Genome
850 Sequencing of Field Isolates Provides Robust Characterization of Genetic Diversity
851 in Plasmodium vivax. *PLoS Negl. Trop. Dis.* **6** (2012),
852 doi:10.1371/journal.pntd.0001811.
- 853 55. S.-B. Chen, Y. Wang, K. Kassegne, B. Xu, H.-M. Shen, J.-H. Chen, Whole-
854 genome sequencing of a Plasmodium vivax clinical isolate exhibits geographical
855 characteristics and high genetic variation in China-Myanmar border area. *BMC*
856 *Genomics.* **18**, 131 (2017).
- 857 56. T. C. de Oliveira, P. T. Rodrigues, M. J. Menezes, R. M. Gonçalves-Lopes, M.

- 858 S. Bastos, N. F. Lima, S. Barbosa, A. L. Gerber, G. Loss de Moraes, L. Berná, J.
859 Phelan, C. Robello, A. T. R. de Vasconcelos, J. M. P. Alves, M. U. Ferreira,
860 Genome-wide diversity and differentiation in New World populations of the human
861 malaria parasite *Plasmodium vivax*. *PLoS Negl. Trop. Dis.* **11**, e0005824 (2017).
862 57. J. Popovici, L. R. Friedrich, S. Kim, S. Bin, V. Run, D. Lek, M. V. Cannon, D.
863 Menard, D. Serre, Genomic Analyses Reveal the Common Occurrence and
864 Complexity of *Plasmodium vivax* Relapses in Cambodia. *mBio.* **9** (2018),
865 doi:10.1128/mBio.01888-17.
866 58. C. Delgado-Ratto, D. Gamboa, V. E. Soto-Calle, P. Van den Eede, E. Torres,
867 L. Sánchez-Martínez, J. Contreras-Mancilla, A. Rosanas-Urgell, H. Rodriguez
868 Ferrucci, A. Llanos-Cuentas, A. Erhart, J.-P. Van geertruyden, U. D'Alessandro,
869 Population Genetics of *Plasmodium vivax* in the Peruvian Amazon. *PLoS Negl. Trop.*
870 *Dis.* **10** (2016), doi:10.1371/journal.pntd.0004376.
871 59. D. E. Neafsey, K. Galinsky, R. H. Y. Jiang, L. Young, S. M. Sykes, S. Saif, S.
872 Gujja, J. M. Goldberg, S. Young, Q. Zeng, S. B. Chapman, A. P. Dash, A. R.
873 Anvikar, P. L. Sutton, B. W. Birren, A. A. Escalante, J. W. Barnwell, J. M. Carlton,
874 The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than
875 *Plasmodium falciparum*. *Nat. Genet.* **44**, 1046–1050 (2012).
876 60. R. Leinonen, H. Sugawara, M. Shumway, The Sequence Read Archive.
877 *Nucleic Acids Res.* **39**, D19–D21 (2011).
878 61. M. Martin, Cutadapt removes adapter sequences from high-throughput
879 sequencing reads. *EMBnet.journal.* **17**, 10–12 (2011).
880 62. S. Auburn, U. Böhme, S. Steinbiss, H. Trimarsanto, J. Hostetler, M. Sanders,
881 Q. Gao, F. Nosten, C. I. Newbold, M. Berriman, R. N. Price, T. D. Otto, A new
882 *Plasmodium vivax* reference sequence with improved assembly of the subtelomeres
883 reveals an abundance of pir genes. *Wellcome Open Res.* **1**, 4 (2016).
884 63. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-
885 Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).
886 64. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytzky,
887 K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis
888 Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing
889 data. *Genome Res.* **20**, 1297–1303 (2010).
890 65. *Picard toolkit* (Broad Institute, 2019; <http://broadinstitute.github.io/picard/>).
891 66. D. H. Huson, D. Bryant, Application of phylogenetic networks in evolutionary
892 studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
893 67. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender,
894 J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: a tool set for
895 whole-genome association and population-based linkage analyses. *Am. J. Hum.*
896 *Genet.* **81**, 559–575 (2007).
897 68. N. M. Kopelman, J. Mayzel, M. Jakobsson, N. A. Rosenberg, I. Mayrose,
898 Clumpak: a program for identifying clustering modes and packaging population
899 structure inferences across K. *Mol. Ecol. Resour.* **15**, 1179–1191 (2015).
900 69. E. Paradis, K. Schliep, ape 5.0: an environment for modern phylogenetics and
901 evolutionary analyses in R. *Bioinformatics.* **35**, 526–528 (2019).

- 902 70. A. Rambaut, FigTree, a graphical viewer of phylogenetic trees (2007).
903 71. H. Zhao, Z. Sun, J. Wang, H. Huang, J.-P. Kocher, L. Wang, CrossMap: a
904 versatile tool for coordinate conversion between genome assemblies. *Bioinforma.*
905 *Oxf. Engl.* **30**, 1006–1007 (2014).
906 72. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R.
907 E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, The
908 variant call format and VCFtools. *Bioinformatics.* **27**, 2156–2158 (2011).
909 73. C. Zhang, S.-S. Dong, J.-Y. Xu, W.-M. He, T.-L. Yang, PopLDdecay: a fast
910 and effective tool for linkage disequilibrium decay analysis based on variant call
911 format files. *Bioinforma. Oxf. Engl.* **35**, 1786–1788 (2019).
912 74. M. C. Fontaine, F. Austerlitz, T. Giraud, F. Labbé, D. Papura, S. Richard-
913 Cervera, F. Delmotte, Genetic signature of a range expansion and leap-frog event
914 after the recent invasion of Europe by the grapevine downy mildew pathogen
915 *Plasmopara viticola*. *Mol. Ecol.* **22**, 2771–2786 (2013).
916 75. S. Schiffels, K. Wang, MSMC and MSMC2: The Multiple Sequentially
917 Markovian Coalescent. *Methods Mol. Biol. Clifton NJ.* **2090**, 147–166 (2020).

918 Acknowledgements

919 **Acknowledgements:** The authors acknowledge the IRD itrop HPC (South Green Platform)
920 at IRD Montpellier for providing HPC resources that have contributed to the research results
921 reported within this paper. URL: <https://bioinfo.ird.fr/>- <http://www.southgreen.fr>. **Funding:**
922 ANR T-ERC EVAD, PEPS ECOMOB MOV 2019, CNRS, plateforme sequencage MGX
923 Montpellier. JD was supported by the Fondation pour la recherche Médicale (FRM,
924 ARF20170938823) as well as by the Marie-Curie EU Horizon 2020 Marie-Sklodowska-Curie
925 research and innovation program grant METHYVIREVOL (contract number 800489). **Author**
926 **contributions:** Conception: V.R. and F.P.; funding acquisition: V.R. and J.D.; biological data
927 acquisition and management: V.R., A.B., B.N., B.L., S.H., C.A., C.S., J-F.T., P.D.; sequence
928 data acquisition: A.B. and V.R.; method development and data analysis: J.D. and M.C.F.;
929 interpretation of the results: J.D., M.C.F., F.P., and V.R.; drafting of the manuscript: J.D. and
930 V.R.; reviewing and editing of the manuscript: J.D., F.R., M.C.F., F.P., and V.R. **Competing**
931 **interests:** The authors declare that they have no competing interests. **Data and materials**
932 **availability:** The Illumina sequence reads generated on the new *P. vivax* samples from
933 Mauritania, Sudan and Ethiopia have been deposited in the European Nucleotide Archive
934 (ENA) under the accession codes listed in Supplementary table 1. The authors declare that
935 all other data supporting the findings of this study are available within the article and its
936 Supplementary Information files, or are available from the authors upon request.
937

938 Figures

939 **Figure 1: Geographical origin of *Plasmodium* isolates and patterns of genomic** 940 **variation**

941 A. Country of origin of 447 *P. vivax* and 19 *P. vivax-like*. Note that the isolates are coming
942 from various locations within each country. The chimpanzee pictogram represents African *P.*
943 *vivax-like* isolates.

944 B. Local variation along the genome in *P. vivax* individual genetic relatedness along the
945 genome visualized with a multidimensional scaling (MDS) based on the local PCA approach
946 (28). Each point represents a non-overlapping genomic window of 100 SNPs. Based on the
947 variation of the MDS-1 coordinate, each window has been classified into two groups, the
948 sub-telomeric hyper-variable regions (orange) or the core region (purple).

949 C. Chromosome 5 genome scan of the SNPs count (left y-axis) that has been identified for
950 both *Plasmodium vivax* and *Plasmodium vivax-like*. The black line represents the number of
951 SNPs shared between the two species. The right y-axis represents the MDS-1 coordinates
952 against the middle point of each window. The color represents windows classified within the
953 sub-telomeric hyper-variable regions (orange) or the core region (purple).

954

955 **Figure 2: *P. vivax-like* strains are structured into two distinct clades forming a sister** 956 **monophyletic lineage to the human *P. vivax*.**

957 A. Neighbor-joining phylogenetic tree illustrating the relatedness between *P. vivax* (PV, grey
958 triangle) and *P. vivax-like* strains. Based on the phylogeny, two distinct groups of *P. vivax-*
959 *like* (PVL) were identified: PVL.grp1 (in orange) and PVL.grp2 (in brown). The animal
960 pictograms on each leaf indicate the primate host, gorilla or chimpanzee, colored by country
961 of origin (Gabon, Cameroon, Ivory Coast). The mosquito pictogram represents samples
962 collected on anopheles' mosquitoes for which the primate host was unknown. The
963 phylogenetic tree was rooted using one outgroup strain of *P. cynomolgi* (strain M). Black
964 dots indicate the bootstrap values, ranging from 0.9 to 1.

965 B. Top two principal components of PCA based on the genome-wide SNPs data present on
966 the core genome of 447 of *P. vivax* and 19 *P. vivax-like* isolates.

967 C. Genome-wide visualization of recent recombination events between the two *P. vivax-like*
968 groups. The horizontal white rectangles represent the genomic position of each *P. vivax-like*
969 individual and the vertical colored lines represent a genomic segment that has recently
970 recombined. The colors indicate donor lineage of each segment.

971 D. Boxplots showing the differences in nucleotide genetic diversity (π) using variants present
972 on the core genome between *P. vivax-like* groups 1 and 2 and *P. vivax*.

973 E. Multiple sequentially Markovian coalescent estimates of the effective population size (N_e)
974 in the *P. vivax-like* groups 1 (in orange) and 2 (in brown) populations. The y-axis shows the
975 \log_{10} of N_e . Light pink and light brown lines represents 50 bootstrap resampling replicates of
976 randomly sampled segregating sites along the core genome.

977

978 **Figure 3: *P. vivax* genetic structure on the core genome is mostly due to genetic**
979 **isolation of natural populations**

980 A. Principal Component Analysis on the SNPs data from the core genome for the 447 *P.*
981 *vivax* isolates.

982 B. Neighbour-joining phylogenetic tree constructed using SNPs data present on the core
983 genome. The star indicates the location of the outgroup.

984 C. Isolation by distance among population. Pairwise estimates of F_{ST} ($1 - F_{ST}$) are plotted
985 against the corresponding geographical distances between countries (as \log_{10} value). The
986 spearman correlation coefficient and the p-value estimated using a Mantel test with 1000
987 permutations are shown.

988 D. Individual strain genetic ancestry proportion, which is depicted as a vertical bar, estimated
989 using the program ADMIXTURE (34) for each of the $K=6$ inferred ancestral populations.

990 E. These same ancestral proportions summed over all individuals for each population plotted
991 as pie charts on the world map.

992

993 **Figure 4: Southeast Asian origin of *P. vivax* is supported by patterns of nucleotide**
994 **diversity, linkage disequilibrium, and ancestral allele frequency spectrum**

995 A. Boxplot of nucleotide diversity (π) for the different populations of *P. vivax*. π values were
996 calculated at each bi-allelic site among all individuals present in each population.

997 B. Decay of the nucleotide diversity as a function of the \log_{10} distance from the Indonesia-
998 Malaysia border.

999 C. LD-decay of each population of *P. vivax*. LD was measured by R^2 as a function of
1000 physical distance in base pairs (bp). Dotted line represents the physical distance threshold at
1001 which the R^2 values were taken to draw the following regression with the geographic
1002 distance.

1003 D. LD at 250bp measured by normalized R^2 as a function of the \log_{10} distance from
1004 Malaysia.

1005 E. Histograms of the Ancestral Allele Frequency (AAF) in populations with a minimum
1006 sample size of 20 individuals are shown. Refer to Figure 1 for the color code that is
1007 consistent across the figure.

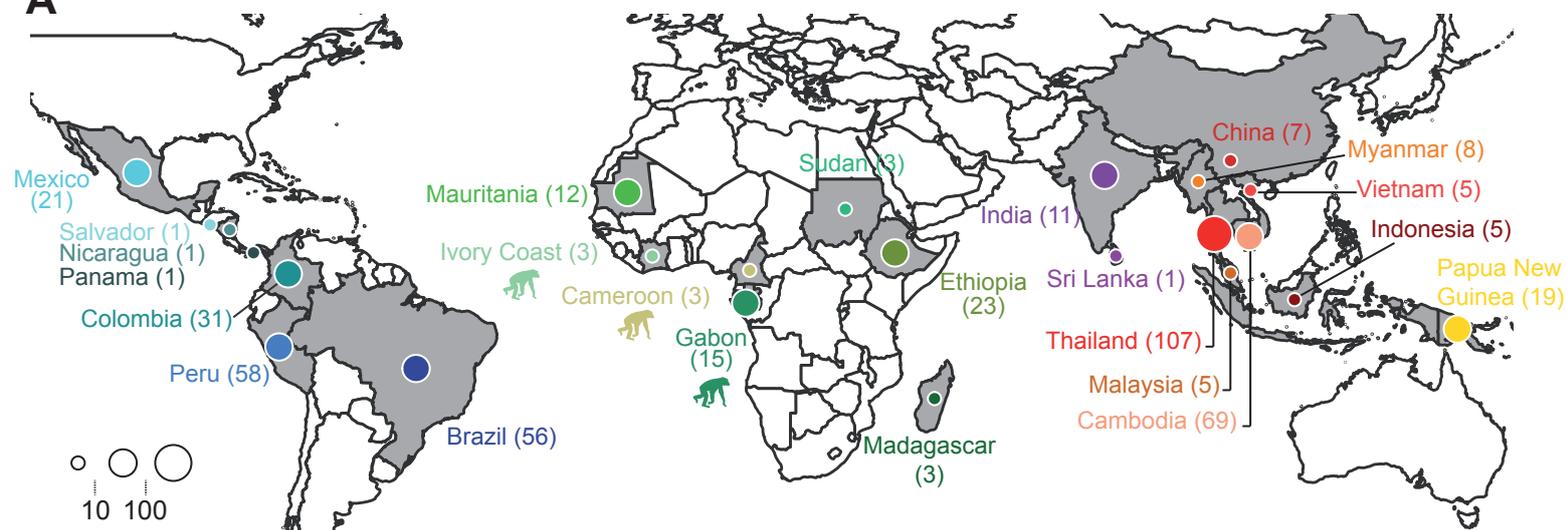
1008

1009 **Figure 5: Coalescent-based inference of demographic history in each *P. vivax***
1010 **population.**

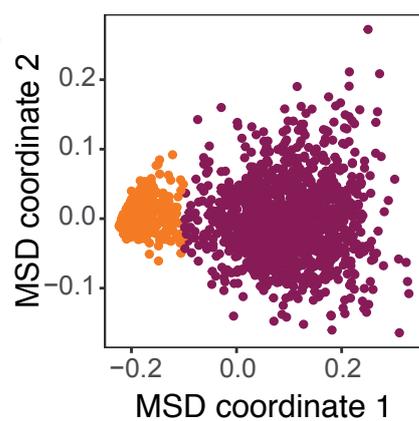
1011 (A) Effective population size variation back to the time to the most recent common ancestor
1012 (TMRCA) inferred using MSMC and (B) inferred TMRCA in each population. Analyses were
1013 conducted considering 5 individuals in each population, assuming a mutation rate per
1014 generate ($\mu.g = 1.158 \times 10^{-9}$) and a generation time (g) of 0.18.

Figure 1

A



B



C

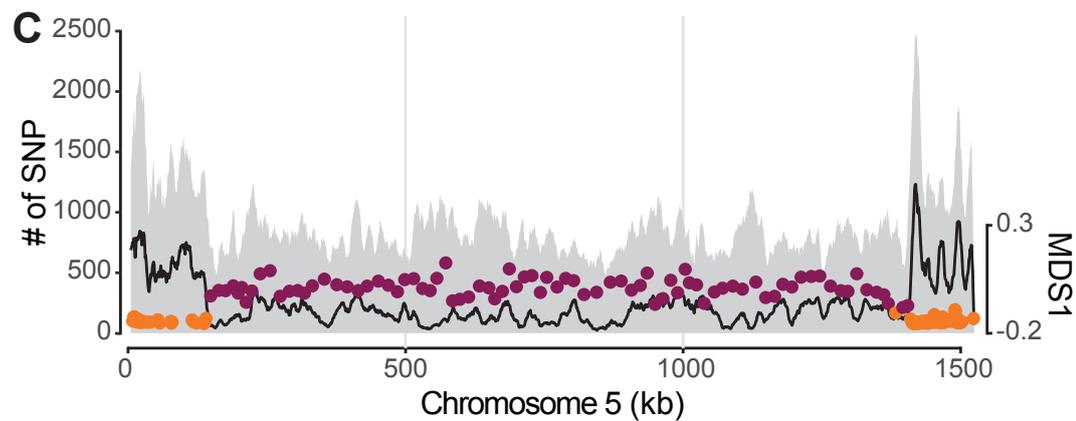


Figure 2

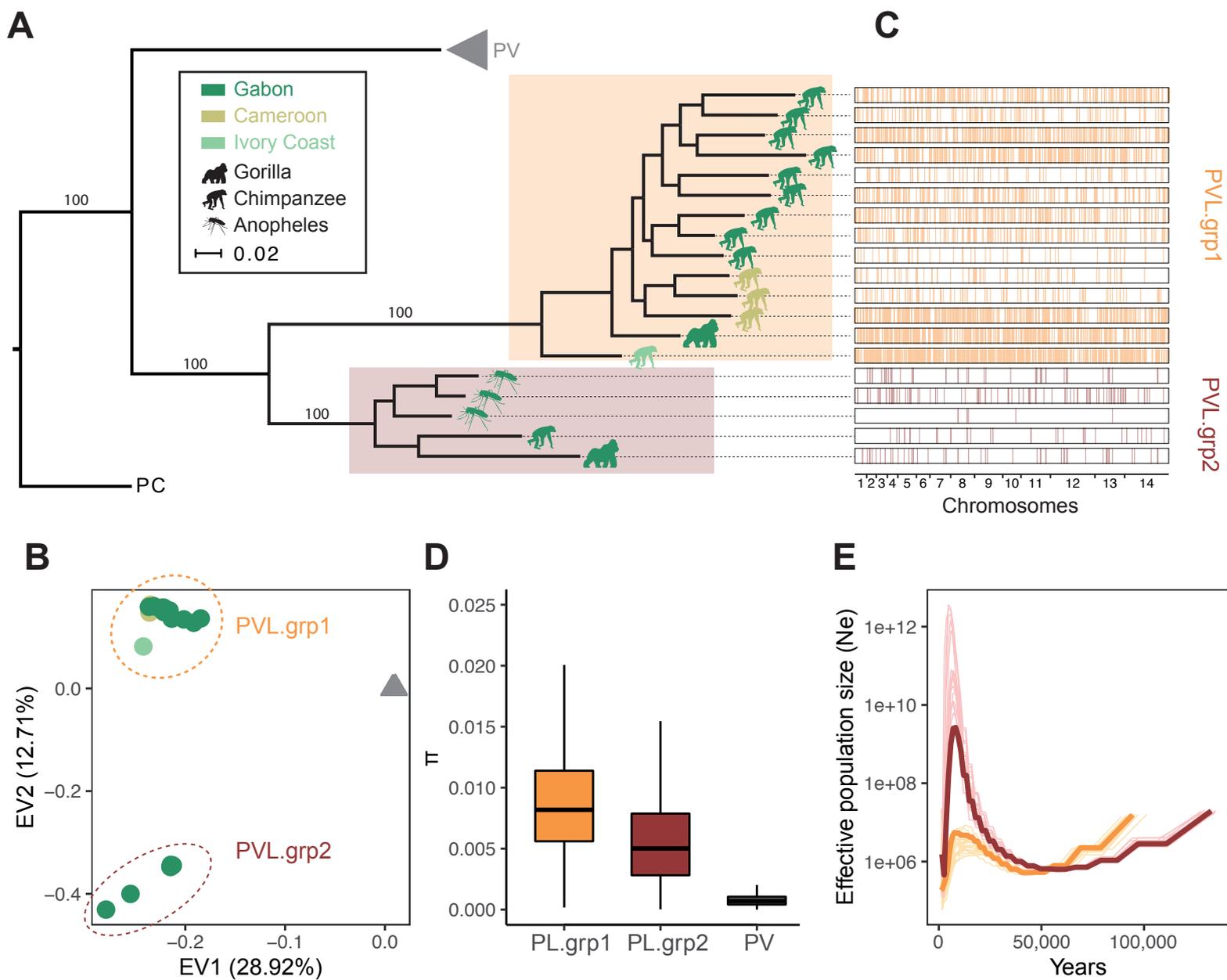


Figure 3

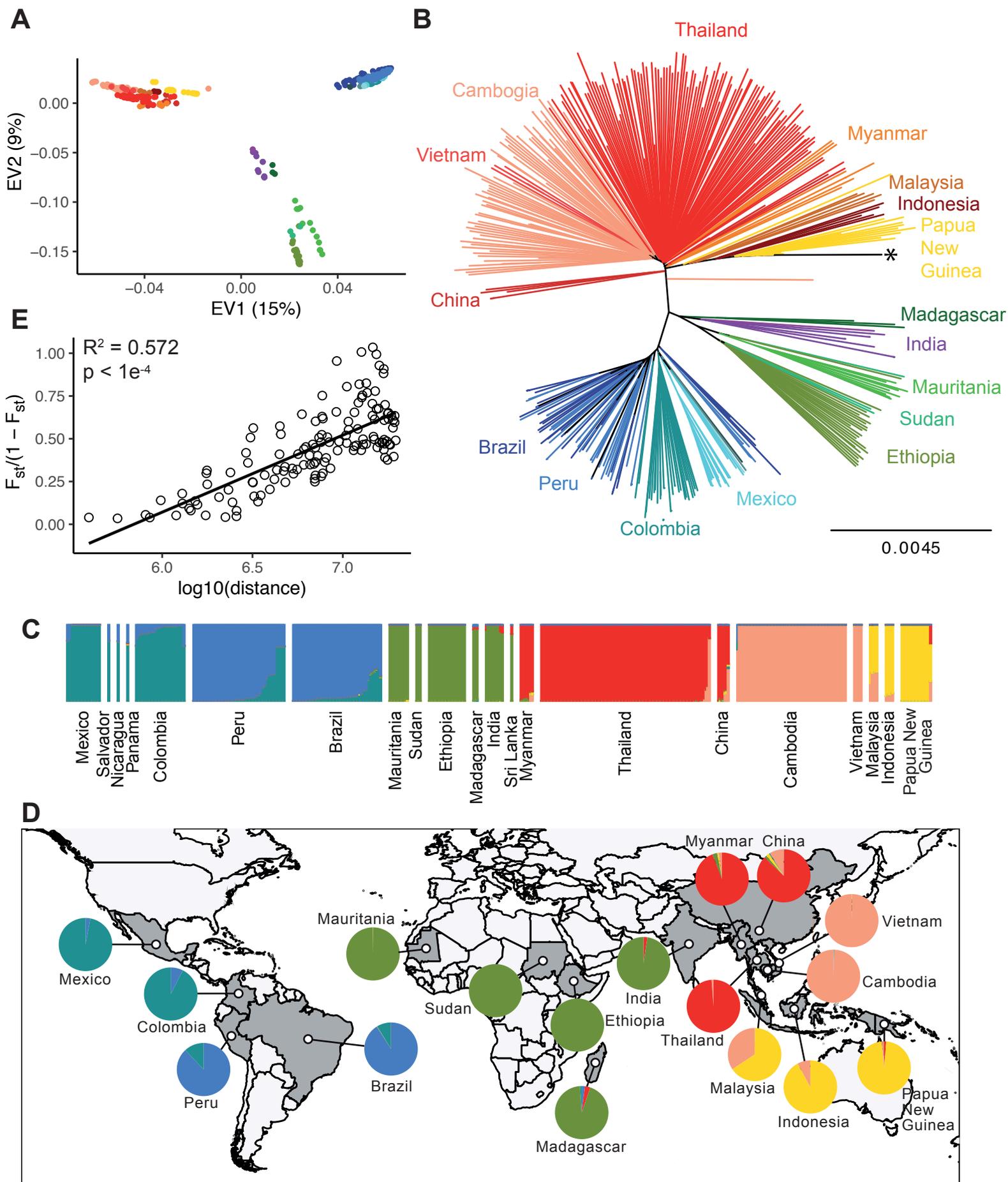
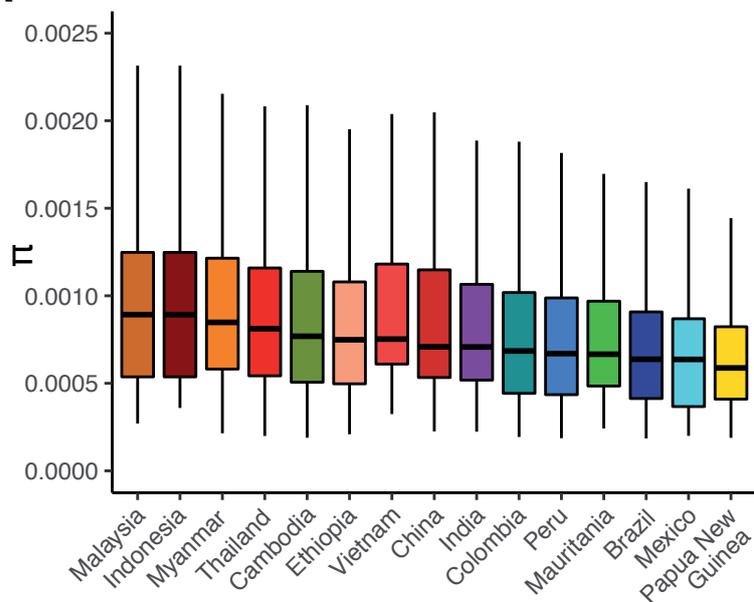
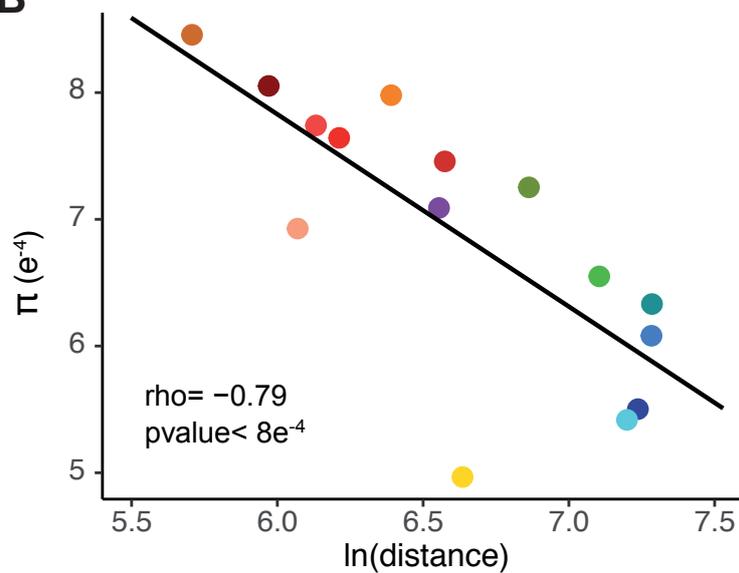


Figure 4

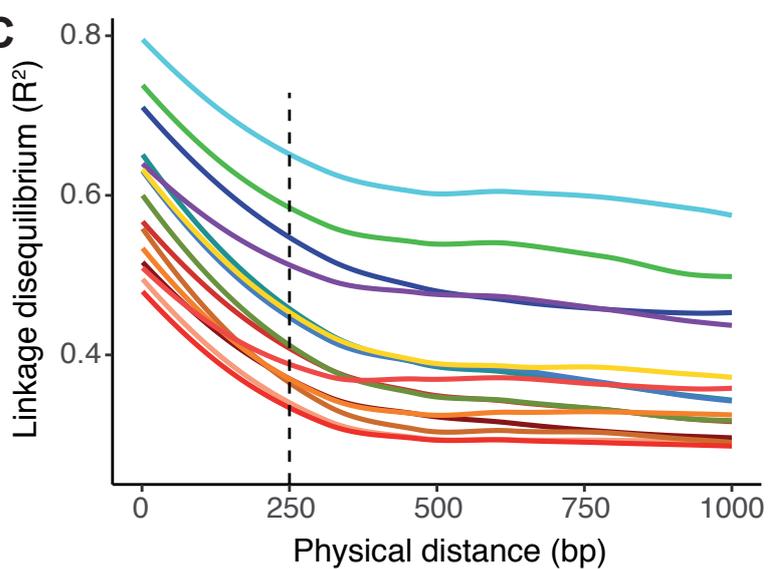
A



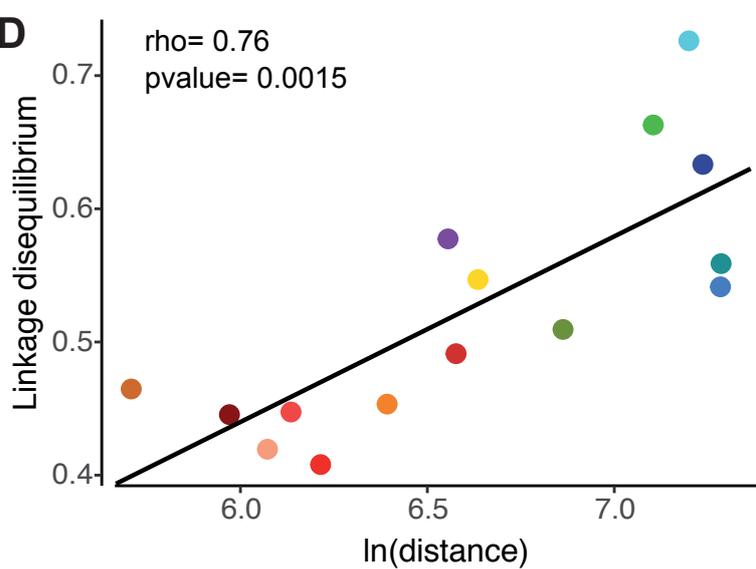
B



C



D



E

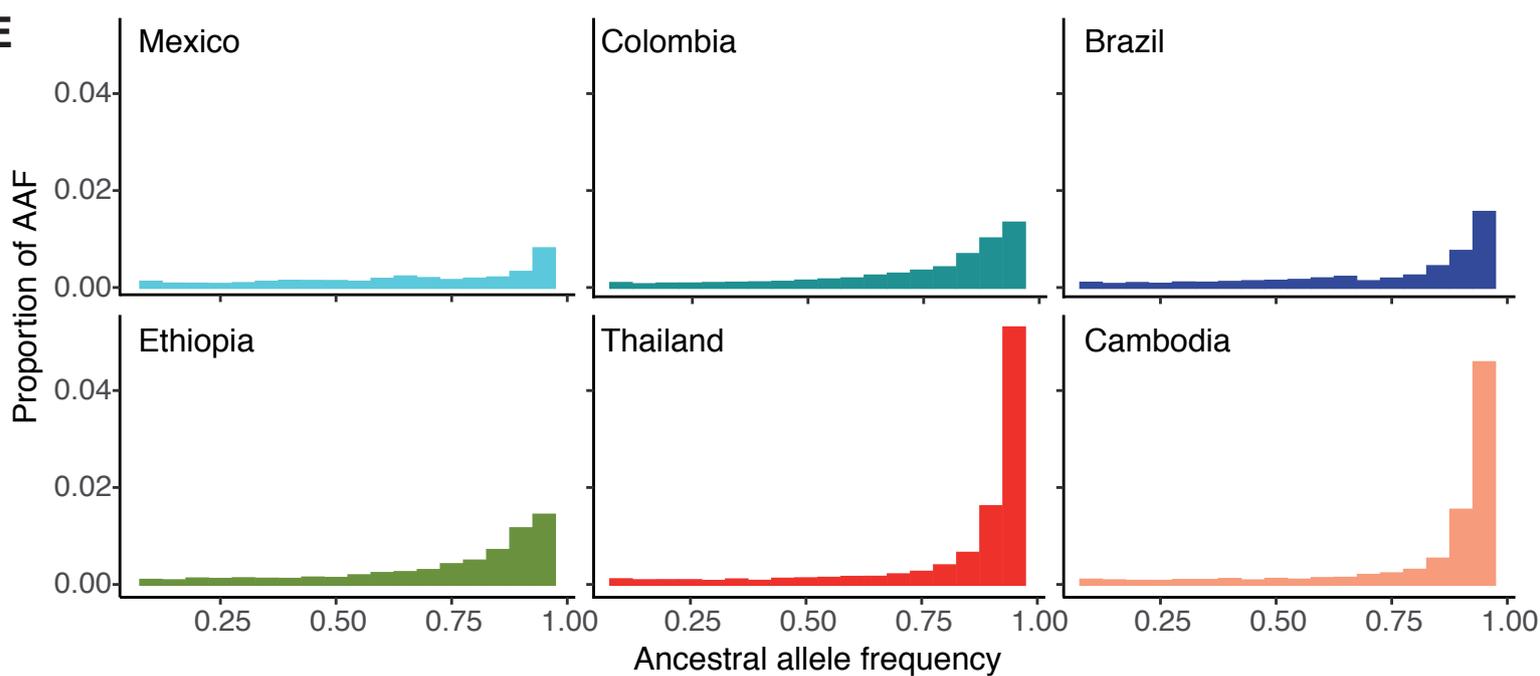


Figure 5

