The Evolutionary History of *Plasmodium vivax* as Inferred from Mitochondrial Genomes: Parasite Genetic Diversity in the Americas

Jesse E. Taylor,[†,1] M. Andreína Pacheco,[†,1] David J. Bacon,[†,2] Mohammad A. Beg,[†,3] Ricardo Luiz Machado,[†,4] Rick M. Fairhurst,[†,5] Socrates Herrera,[†,6] Jung-Yeon Kim,[†,7] Didier Menard,[†,8] Marinete Marins Póvoa,[†,9] Leopoldo Villegas,[†,10] Mulyanto,[†,11] Georges Snounou,[12,13] Liwang Cui,[14] Fadile Yildiz Zeyrek,[15] and Ananias A. Escalante*[,1,16]

[1]Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University

[2]Naval Research Laboratory, Washington, DC

[3]Department of Pathology and Microbiology, Aga Khan University, Karachi, Pakistan

[4]Centro de Investigação de Microrganismos, Departamento de Doenças Dermatológicas, Infecciosas e Parasitárias, Faculdade de Medicina de São José do Rio Preto, São José do Rio Preto, São Paulo, Brazil

[5]Laboratory of Malaria and Vector Research, NIAID, NIH, Rockville, Maryland

[6]Caucaseco Scientific Research Center/Immunology Institute, Universidad del Valle, Cali, Colombia

[7]Division of Malaria and Parasitic Diseases, Korea Centers for Disease Control and Prevention, Osong, Republic of Korea

[8]Institut Pasteur du Cambodge, Phnom Penh, Cambodia

[9]Evandro Chagas Institute, Belém, Pará, Brazil

[10]ICF International, International Health and Development Division, Calverton, Maryland

[11]Immunobiology Laboratory, Faculty of Medicine, University of Mataram, Mataram, Indonesia

[12]Université Pierre et Marie Curie - Paris VI, Paris, France

[13]Institut National de la Santé et de la Recherche Médicale, Paris, France

[14]Department of Entomology, Pennsylvania State University

[15]Department of Microbiology, Harran University School of Medicine, Sanliurfa, Turkey

[16]School of Life Sciences, Arizona State University

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: Ananias.Escalante@asu.edu.

Associate editor: Sarah Tishkoff

Article

Abstract

***Plasmodium vivax* is the most prevalent human malaria parasite in the Americas. Previous studies have contrasted the genetic diversity of parasite populations in the Americas with those in Asia and Oceania, concluding that New World populations exhibit low genetic diversity consistent with a recent introduction. Here we used an expanded sample of complete mitochondrial genome sequences to investigate the diversity of *P. vivax* in the Americas as well as in other continental populations. We show that the diversity of *P. vivax* in the Americas is comparable to that in Asia and Oceania, and we identify several divergent clades circulating in South America that may have resulted from independent introductions. In particular, we show that several haplotypes sampled in Venezuela and northeastern Brazil belong to a clade that diverged from the other *P. vivax* lineages at least 30,000 years ago, albeit not necessarily in the Americas. We propose that, unlike in Asia where human migration increases local genetic diversity, the combined effects of the geographical structure and the low incidence of *vivax* malaria in the Americas has resulted in patterns of low local but high regional genetic diversity. This could explain previous views that *P. vivax* in the Americas has low genetic diversity because these were based on studies carried out in limited areas. Further elucidation of the complex geographical pattern of *P. vivax* variation will be important both for diversity assessments of genes encoding candidate vaccine antigens and in the formulation of control and surveillance measures aimed at malaria elimination.**

Key words: **molecular clock, population structure, demographic history.**

Introduction

Despite remarkable progress made on its control, malaria remains one of the most important infectious diseases (WHO 2011). Nowadays, it is estimated that more than 2.6 billion people are at risk of infection and that they experience around 200 million episodes of malaria each year (WHO 2011; Cibulskis et al. 2011; Lynch et al. 2012; Murray et al. 2012). Of the five parasites that cause human malaria, *Plasmodium vivax* is responsible for most morbidity outside Africa (WHO 2011). This parasite has reemerged in many regions of the world where malaria was eliminated in the 1950–60s (Guerra et al. 2007). The so-called "benign tertian malaria" is by no means harmless; new evidence suggests that severe complications from *P. vivax* malaria may be more common than previously thought. Anemia, thrombocytopenia, and low neonatal birth weight at delivery are among those severe manifestations of disease (Kochar et al. 2005; Anstey et al. 2007; Alexandre et al. 2010; Douglas et al. 2012; Kute et al. 2012). Given the limitations of available tools for its research and control, there are serious gaps in basic information about this parasite that may delay malaria elimination (Mueller et al. 2009). One such knowledge gap is our limited understanding of the demographic history and structure of extant *P. vivax* populations.

Characterizing the evolutionary history of *P. vivax* is important for identifying adaptive genetic variation (Joy et al. 2008; Carlton et al. 2013), assessing the geographical distribution of variation in genes encoding candidate vaccine antigens, and designing population-based linkage studies that aim to identify genes associated with drug resistance and pathogenesis. In addition, such information is essential for implementing molecular surveillance approaches that aim to detect imported cases of malaria or to ascertain the efficacy of local malaria control measures (Cui et al. 2003; Arnott et al. 2012; Chan et al. 2012). Regardless of their clinical importance, the evolutionary history of the events leading to this parasite's contemporary broad geographical distribution remains unresolved.

There is a general agreement that *P. vivax* originated as a human parasite as a result of a host switch from a non-human primate, an event that likely took place somewhere in Asia (Escalante et al. 2005; Mu et al. 2005; Cornejo and Escalante 2006). Consistent with such an early origin, previous studies identified a relatively high genetic diversity in the populations of this parasite in Asia and Oceania (Jongwutiwes et al. 2005; Mu et al. 2005; Cornejo and Escalante 2006). In contrast, the origin and demographic history of *P. vivax* populations in the Americas have received limited attention (Li et al. 2001; Mu et al. 2005; Jongwutiwes et al. 2005; Cornejo and Escalante 2006; Grynberg et al. 2008; Joy et al. 2008; Culleton et al. 2011; Miao et al. 2012).

In this investigation, we used a suite of population genetic and phylogenetic analyses to characterize an enlarged sample of *P. vivax* mitochondrial genomes from the Americas in relation to global parasite diversity. We found that *P. vivax* populations in South America harbor levels of genetic diversity that are comparable to those observed in Asia and Oceania. Indeed, we identified a new and relatively divergent clade of *P. vivax* mitochondrial genomes found mainly in northeastern South America, as well as several other divergent lineages that may stem from independent introductions. These observations are consistent with recent comparative genomics investigations indicating that commonly used *P. vivax* lines obtained from the Americas exhibit great divergence (Chan et al. 2012; Neafsey et al. 2012).

Results

Global Mitochondrial Genome Diversity

Table 1 and supplementary table S1, Supplementary Material online, summarize the global, regional and national diversity of the 731 *P. vivax* mitochondrial genomes analyzed in this study. (A list of all the *P. vivax* sequences included in this study is provided in supplementary table S2, Supplementary Material online.) We identified 357 distinct haplotypes, including 216 haplotypes not found in earlier surveys of global mitochondrial diversity in this species, increasing the number of haplotypes found in South America from 9 to 105 (Mu et al. 2005; Culleton et al. 2011; Miao et al. 2012). As in previous studies, global haplotype diversity is very high in our sample ($H = 0.985$) and most (286) haplotypes were sampled just once (supplementary table S1, Supplementary Material online). Indeed, only 12 haplotypes have sample frequencies $>1\%$, including the mitochondrial haplotype of the genome reference strain Salvador I (Sal1 frequency = 1.9%), while the most common haplotypes, CA1 and Sal2, have sample frequencies of 7.4% and 6.3%. As expected, haplotype diversity is lower within regions but exceeds 0.9 in every region except Madagascar, Africa, and Central America. Furthermore, although the nucleotide diversity of our global sample ($\pi = 9.24 \times 10^{-4}$) is comparable to that found in previous studies, our estimate of nucleotide diversity within South America ($\pi = 6.18 \times 10^{-4}$) is more than three times larger than earlier estimates.

Phylogenetic Analyses and Haplotype Networks

The two consensus trees derived from the *BEAST skyline analyses (fig. 1) and the *BEAST analyses (supplementary fig. S1a, Supplementary Material online) have the same overall structure: each tree contains a large star-like cluster of *P. vivax* lineages distributed throughout the range of the parasite as well a second divergent clade composed primarily of lineages from eastern and southeastern Asia. A similar topology is evident in the global haplotype network (fig. 2), although here the divergent East Asian *P. vivax* lineages are connected to the star-like component by a group of haplotypes found mainly in Southeast Asia.

The roots of both consensus trees (fig. 1 and supplementary fig. S1a, Supplementary Material online) coincide with the most recent common ancestor of the star-shaped component and, in fact, 219 of the mitochondrial lineages branch directly from the root of each tree. Similarly, the two haplotypes with the highest outgroup probabilities in the median joining network (supplementary table S3, Supplementary Material online) are Sal2 ($P = 0.047$) and CA1 ($P = 0.039$),

Table 1. Regional Variation in *P. vivax* Mitochondrial Genomes.

Region	n	haps	priv	*H*	*S*	Π	*D* (*P*-value)	F_S (*P*-value)
East Asia	129	35	34	0.912	32	7.99×10^{-4}	−0.619 (0.308)	−14.546 (0.003)
South Asia	83	57	48	0.959	73	5.88×10^{-4}	−2.535 (0.000)	−26.218 (0.000)
Southeast Asia	154	81	74	0.968	84	8.42×10^{-4}	−2.109 (0.000)	−25.356 (0.000)
Melanesia	72	37	29	0.929	43	5.72×10^{-4}	−2.035 (0.002)	−26.264 (0.000)
Madagascar	66	34	30	0.808	51	4.25×10^{-4}	−2.564 (0.000)	−26.812 (0.000)
Africa	17	9	6	0.787	16	3.55×10^{-4}	−2.178 (0.004)	−3.768 (0.006)
Middle East	28	18	17	0.942	31	6.84×10^{-4}	−1.830 (0.009)	−9.435 (0.000)
Central America	10	3	0	0.511	2	0.95×10^{-4}	−0.691 (0.243)	−0.594 (0.113)
South America	172	105	103	0.947	106	6.18×10^{-4}	−2.530 (0.000)	−25.898 (0.000)
Global	731	357	—	0.985	311	9.24×10^{-4}	−2.576 (0.000)	−24.492 (0.000)

Note.—*n*, sample size; haps, number of haplotypes; priv, number of private haplotypes; *H*, haplotype diversity; *S*, number of polymorphic sites; π, nucleotide diversity; *D*, Tajima's *D* statistic; F_S, Fu's F_S statistic.

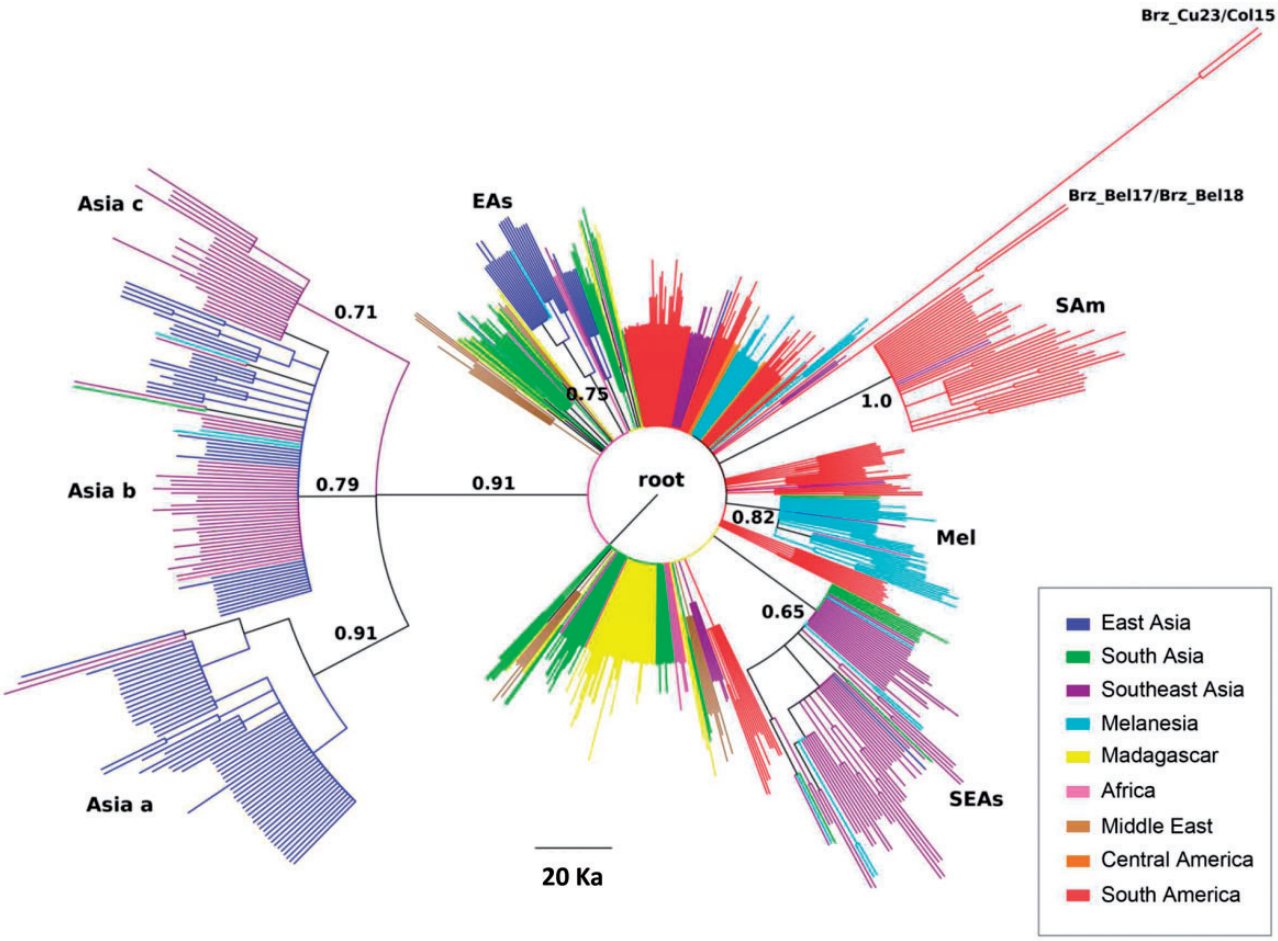

Fig. 1. Majority rule consensus tree obtained from the four Bayesian skyline analyses of the global *P. vivax* sample. The posterior probabilities of selected clades are listed next to the corresponding branches and the clades described in table 2 are labeled. Branches are colored if they subtend sequences sampled in the same region using the following key: blue = East Asia; green = South Asia; purple = Southeast Asia; light blue = Melanesia; yellow = Madagascar; pink = Africa; brown = Middle East; orange = Central America; red = South America.

which are both located at the center of the star-shaped component and are also the most common haplotypes in our sample. Although more phylogenetic structure is apparent in the maximum clade credibility (MCC) trees obtained from the *BEAST analyses (supplementary fig. S1b, Supplementary Material online) and the skyline analyses (supplementary fig. S1c, Supplementary Material online), this is largely an artifact of the method used to construct these trees, as they are required to be strictly bifurcating. Indeed, inspection of the two trees reveals that many of the clades have low posterior

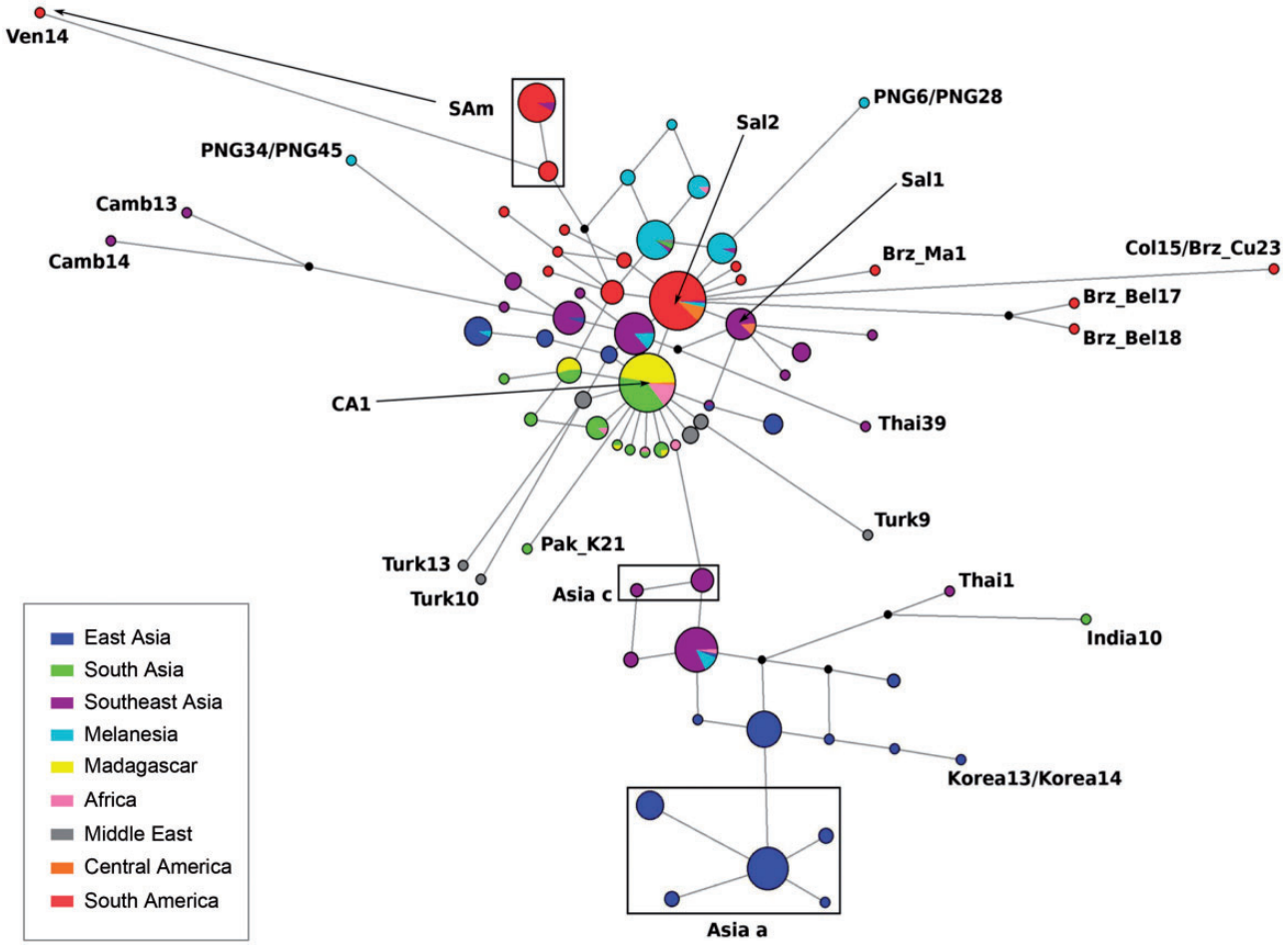

Fig. 2. Median joining network of global *P. vivax* mtDNA haplotypes sampled two or more times plus selected singleton haplotypes. Branch lengths are proportional to divergence, node sizes are proportional to total haplotype frequencies, and colors correspond to regions: blue = East Asia; green = South Asia; purple = Southeast Asia; light blue = Melanesia; yellow = Madagascar; pink = Africa; gray = Middle East; orange = Central America; red = South America. Because of star contraction, some nodes represent multiple haplotypes. Median nodes (inferred haplotypes at branch points) are shown in black. Black boxes enclose selected clades (SAm, Asia a, Asia b) and selected haplotypes are labeled using the nomenclature of supplementary table S2, Supplementary Material online.

probabilities and, in particular, the positions of the roots differ between the two trees and have only moderate support in each analysis.

Despite the overall lack of structure in the *P. vivax* trees, the skyline consensus tree contains several clades with estimated posterior probabilities (pp) exceeding 0.5. Those containing more than 15 sequences are described in table 2 and supplementary table S4, Supplementary Material online. The divergent cluster of Asian lineages (pp = 0.91) is itself split into three subclades. Two of these contain sequences that are largely restricted to either China and Korea (Asia a; pp = 0.91) or Indonesia, Thailand, and Cambodia (Asia c; pp = 0.71), but the clade labeled Asia b (pp = 0.79) includes lineages distributed throughout eastern and southeastern Asia as well as three sequences from Papua New Guinea, one from India, and one from Namibia. Four large clades are contained within the star-like component of the tree, including groups of sequences primarily from Korea and southern China (EAs; pp = 0.75), from Thailand, Cambodia, and Vietnam (SEAs; pp = 0.65), from Papua New Guinea and Vanuatu (Mel; pp = 0.82), and from Brazil and Venezuela (SAm; pp = 1.00). However, each of these clades also contains a small number of sequences that were sampled in countries far removed from the eponymous region. For example, the East Asian clade contains a sequence sampled in Vanuatu, the Southeast Asian clade includes six sequences sampled in

Table 2. Ages of the Major *P. vivax* mtDNA Clades.

Clade	*n*	haps	pp	TMRCA (ky)
Asia	160	61	0.91	121 (46–246)
Asia a	61	14	0.91	84 (27–198)
Asia b	83	39	0.79	88 (32–164)
Asia c	16	8	0.71	79 (26–155)
EAs	24	5	0.75	66 (24–124)
SEAs	76	38	0.65	98 (34–195)
Mel	42	19	0.82	66 (27–117)
SAm	43	28	1.00	76 (29–135)

Note.—*n*, number of sequences; haps, number of haplotypes; pp, posterior probability; TMRCA, time to most recent common ancestor in thousands of years (ky): cells show the mean and 95% highest probability density interval estimated by Bayesian skyline analysis of the complete *P. vivax* mtDNA data. Clades Asia a–c are subsets of clade Asia.

Papua New Guinea and the Solomon Islands, the Melanesian clade includes sequences sampled in Africa, India, and Iran, and the South American clade includes two identical sequences sampled in Thailand.

The 105 mtDNA haplotypes that we identified in the Americas belong to several phylogenetically distinctive groups (supplementary fig. S1d, Supplementary Material online). The largest of these is a paraphyletic collection of lineages that are present in both Central and South America and which belong to the star-like component of the consensus *P. vivax* tree. We refer to this as the Pan-American group as a way to emphasize the American lineages. However, this group also contains three of the four haplotypes that were found both within and outside of the Americas, including the two most common haplotypes, CA1 and Sal2, as well as the haplotype of the reference strain, Sal1. These three haplotypes are closely related, differing at only two sites, but have different distributions. While CA1 is represented by just one isolate from Central America and 53 isolates from India, Pakistan, Madagascar, and Sub-Saharan Africa, and Sal1 is represented by just two isolates from Central America and 12 from Myanmar, all but two of the 46 copies of Sal2 included in our sample were obtained from South and Central America, the two exceptions being from Papua New Guinea and Bangladesh.

The second largest group is a divergent South American clade which includes 41 mitochondrial genomes sampled in Brazil and Venezuela. This clade contains two-thirds (10/15) of our Venezuelan sequences, as well as 31 sequences from Brazil (supplementary table S4, Supplementary Material online) with the majority sampled from the northeastern localities of Belém and Macapá. Unexpectedly, this clade also contains two genomes that were reportedly sampled in Thailand (Mu et al. 2005). Assuming that the provenance of these sequences is correct, this could indicate that this clade was either introduced into South America from Southeast Asia or that it was introduced into Thailand from South America. Although we cannot conclusively exclude either hypothesis, we believe that the evidence more strongly favors the latter explanation. Indeed, the two Thai genomes are identical both to one another and to 12 additional genomes sampled in Venezuela and at several locations within Brazil. Furthermore, although this clade contains 27 haplotypes that were only sampled in South America, we did not find any additional haplotypes from outside the Americas that were closely related to the Thai sequences. Indeed, in both consensus trees (fig. 1 and supplementary fig. S1a, Supplementary Material online), the root of this clade is connected by a single long branch to the root of the entire *P. vivax* sample.

In addition to the Pan-American and the large structured South American groups, our sample contains two small groups from South America that are unusually divergent from the other haplotypes isolated in the Americas (supplementary fig. S1d, Supplementary Material online). One group consists of a pair of identical sequences (Brz_Cu23, Col15) that were sampled in the localities of Cuiabá in Brazil and Buenaventura in Colombia, which have eight fixed nucleotide differences relative to the other genomes in the global sample. These differences are found at sites located throughout the mitochondrial genome, in both protein and rRNA-coding regions, and, insofar as we can tell, cannot be attributed to either sequencing error or anomalous mutational processes. This lineage appears as a long branch in the Bayesian skyline analyses (fig. 1 and supplementary fig. S1c, Supplementary Material online) and is even positioned as an outgroup to all other *P. vivax* genomes in the maximum clade credibility tree derived from the Bayesian skyline analyses. On the other hand, this branch is much shorter in the trees obtained from the *BEAST analyses (supplementary fig. S1a and b, Supplementary Material online), although this could be due, in part, to the use of a relaxed molecular clock in these analyses, which will have a tendency to shorten unusually long branches. The other divergent group consists of two genomes (Brz_Bel17 and Brz_Bel18) that were sampled in the city of Belém in northeastern Brazil. These differ from one another at one nucleotide position, but also share four fixed nucleotide differences relative to all other sequences in our sample. Neither group appears to be closely related to any of the other clades present in the skyline and *BEAST consensus trees, and both groups branch directly off of the root of each tree.

We also inferred a median joining network for the haplotypes sampled in Central and South America (supplementary fig. S1d, Supplementary Material online). Like the global network, it consists of a large star-like cluster, with relatively little phylogenetic or geographical structure, as well as a more divergent cluster of haplotypes sampled in Venezuela and Brazil which coincides exactly with the South American clade identified in the phylogenetic analyses. Several haplotypes are linked to this network by relatively long branches, including the two haplotypes Col15 and Brz_Cu23 noted above, as well as two other Brazilian haplotypes, Brz_Bel17 and Brz_Bel18, sampled in the northeastern city of Belém, which differ at exactly one site. Also notable is an unusually divergent haplotype, Ven14, which belongs to the South American clade.

Substitution Rates and Coalescence Times

Two approaches were used to estimate the average genome-wide mtDNA substitution rate for *P. vivax*. In the first, we used the multispecies coalescent model implemented in *BEAST to analyze the 764 genomes from the 11 *Plasmodium* species included in this study (*P. vivax* and related species in non-human primates). The calibration was based on a proposed divergence time of 6–14.2 Ma between the clade comprised of *Plasmodium gonderi* and *Plasmodium* sp. from mandrills (Africa) and the clade containing the remaining nine *Plasmodium* species found in Southeast Asia non-human primates that together with *P. vivax* form a monophyletic group (Pacheco et al. 2012). This event has been used as calibration point in other studies (e.g., Mu et. al. 2005). The mean of the posterior distribution of the substitution rate, averaged over the entire tree, was 0.00383 substitutions per site per million years with a 95% highest probability density

interval (HPD) of 0.00189–0.00600. Although a relaxed molecular clock was used in this analysis, the fact that this model fits an independent substitution rate to each branch in the tree makes it impossible to extract a credibility interval for the substitution rate averaged only over lineages of *P. vivax*. Thus, we also used a simple birth–death model implemented in BEAST to analyze a smaller data set containing just one mitochondrial genome from each of the 11 *Plasmodium* species included in the previous analysis. With this approach, we found that the posterior mean of the substitution rate averaged over the entire tree was 0.00404 substitutions per site per million years (95% HPD: 0.00211–0.00601), while the posterior mean of the substitution rate for the single *P. vivax* lineage included in the alignment was 0.00378 (95% HPD: 0.00089–0.00729). Notice that the credibility interval for the lineage-specific substitution rate obtained in this second analysis is large because the estimate is based on a single branch. To account for these different results, we used a gamma distribution with shape parameter 10 and scale parameter 0.0004 as the prior distribution for the genome-wide average substitution in all subsequent phylogenetic analyses of alignments containing only *P. vivax* genomes. Since the mean of this distribution is 0.004 and the 95% HPD is 0.00192–0.00683, we believe that this is a good approximation to the posterior distribution of the substitution rates estimated in the *BEAST and the birth–death analyses. We also note that these rates are similar to the mitochondrial substitution rates estimated in previous studies of *P. vivax* (Mu et al. 2005: 0.0045–0.0056; Jongwutiwes et al. 2005: 0.0026–0.0036; Cornejo and Escalante 2006: 0.0032–0.0043 substitutions per site per million years).

The time to the most recent common ancestor (TMRCA) of the global *P. vivax* sample was estimated using four approaches. The *BEAST analysis of the multispecies data set produced the smallest estimate of the posterior mean of the TMRCA, which was 230 thousand of years ago (95% HPD: 78–446 ka), while the largest estimate was obtained from the skyline analyses and was 283 ka (95% HPD: 81–542 ka). Intermediate values were obtained from the two phylogeographical analyses (see below), these being 235 ka (95% HPD: 85–431 ka) for the 9 region subdivision and 259 ka (95% HPD: 97–480 ka) for the 18 region subdivision. As the models implemented in these analyses make different assumptions about the demography of *P. vivax*, it is unsurprising that they yield somewhat different estimates of the TMRCA. Nonetheless, the extensive overlap between the credibility intervals obtained from these different approaches suggests that the true value of the TMRCA of our *P. vivax* sample lies between 80 and 540 ka.

Skyline analyses calibrated by the gamma-distributed substitution rates (see above) were used to estimate the TMRCAs of both the regional samples of *P. vivax* listed in table 1 and the major *P. vivax* clades described in table 2. With the exception of the large Asian clade, the mean ages of all of the major clades lie between 66 and 98 ka. The large Asian clade is somewhat older, with a mean age of 121 ka, but this value is still much smaller than the ages estimated for the common ancestor of the global *P. vivax* sample. It is also worth noting that the TMRCA for the divergent South American clade from Brazil and Venezuela is 76 ka (95% HPD: 29–135 ka). In contrast, the TMRCAs of the regional samples of *P. vivax* sequences have a much greater spread (table 3). At one end, the mean age of the common ancestor of the 10 Central American sequences included in this study is only 23 ka (95% HPD: 3.7–64 ka), which is unsurprising given the limited amount of genetic variation found in this small sample. At the other extreme, the common ancestors of the samples of South Asian (India, Pakistan, Sri Lanka, and Bangladesh) and South American (Colombia, Venezuela, Brazil, and Peru) *P. vivax* sequences appear to be much older, with mean TMRCAs of 280 ka (95% HPD: 71–532 ka) and 309 ka (95% HPD: 72–605 ka), respectively.

Demographic History

Several lines of evidence are consistent with growth or expansion of *P. vivax* populations in most regions. Tajima's *D* is significantly negative in all regions except Central America and East Asia, while Fu's F_S is significantly negative everywhere except in Central America (table 1). Although these two statistics were originally introduced as tests of neutrality, significant negative values, such as those seen here, can also indicate population expansion or recent bottlenecks (Simonsen et al. 1995). Fewer significant results were obtained when the samples from individual countries were analyzed, perhaps in part because of the smaller sample sizes, but as with the regional analyses, the estimates of the two statistics were negative in most countries (supplementary table S1, Supplementary Material online). One notable exception is that Tajima's *D* is actually positive in Korea, albeit not significantly different from zero, while the value estimated for China is negative, but also not significant, which are both consistent with the nonsignificant result obtained for East Asia as a whole. More direct evidence of population growth is provided by the Bayesian skyline analyses of the global and regional *P. vivax* samples. Figure 3 shows the skyline plot obtained from the analysis of the entire *P. vivax* sample, which suggests that the global population size of the parasite grew slowly until about 60,000 years ago, after which it underwent a rapid exponential increase in size that began to taper off about 10,000 years before present.

Since regional differences in population dynamics would be obscured in the analysis of the global sample, we also carried out skyline analyses of each of the regional samples. The skyline plots for these analyses are shown in supplementary figure S2, Supplementary Material online, with the estimates of the relative ancestral population sizes at 0, 5, 10, 20 and 50 ka reported in table 3. Although the ancestral population size of *P. vivax* increased in every region except Central America, both the timing and the magnitude of the increases differ between the regions. For example, while the ancestral population dynamics of *P. vivax* in South Asia, Southeast Asia and South America resemble the S-shaped pattern exhibited by the global population, with diminishing growth rates over the past 20,000–50,000 years, the skyline plots for Melanesia, Madagascar, Africa, and the Middle East suggest that *P. vivax* populations have been slowly but steadily expanding in each

Table 3. Bayesian Skyline Analyses of Regional *P. vivax* Samples.

			Relative Population Size				
Region	cp	TMRCA (ky)	0 ka	5 ka	10 ka	20 ka	50 ka
East Asia	6	234 (71–458)	3.63	1.22	0.66	0.46	0.34
South Asia	4	280 (71–532)	32.31	28.46	24.62	17.27	5.73
Southeast Asia	7	199 (63–388)	40.47	33.64	24.78	12.18	2.06
Melanesia	3	130 (41–251)	2.83	2.13	1.64	1.11	0.56
Madagascar	3	79 (24–147)	37.25	33.88	30.51	23.85	12.94
Africa	2	85 (21–186)	1.78	1.58	1.40	1.09	0.67
Middle East	2	158 (51–313)	5.07	4.57	4.08	3.20	1.73
Central America	2	23 (3.7–64)	0.06	0.06	0.07	0.08	0.11
South America	8	309 (72–605)	52.67	45.78	39.64	30.20	13.30
Global	8	283 (81–542)	103.35	70.72	49.12	24.87	4.73

Note.—cp, number of change points in the piecewise linear skyline analysis; TMRCA, time to most recent common ancestor in thousands of years: cells show the mean and 95% highest probability density interval estimated by Bayesian skyline analysis of the global or regional data. The mean relative population sizes were estimated from the skyline analysis of each data set at the four times indicated and are expressed in units of populations size × generation time.

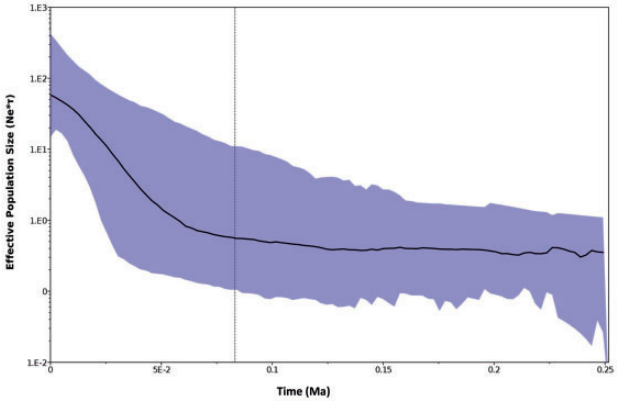

Fig. 3. Bayesian skyline plot of the entire *P. vivax* sample. The analysis was conducted assuming piecewise linear growth spread over eight epochs and using a gamma prior on the substitution rates (mean 0.004 substitutions per site per Ma; shape parameter = 10). The dark line depicts the median ancestral population size, while the colored blue region shows the 95% HPD for these estimates. Time is shown in units of million years ago (Ma) on the *x* axis, while the *y* axis shows the product of the ancestral effective population size N_e and the parasite generation time τ (also in Ma).

of these four regions. In contrast, the East Asian population appears to have been stable until about 10,000 years ago, at which point it began to increase rapidly without tapering off in the present. Only in Central America does the skyline analysis point to a small decline in the ancestral population size of the parasite, which is consistent with the results of the permutation tests reported in table 1. This could be an artifact of the small sample size of Central American *P. vivax* sequences available for this study, but an alternative explanation is that our sample consists of lineages that have only recently entered Central America from other regions with larger ancestral population sizes.

Phylogeography and Population Structure

Consistent with previous studies (Mu et al. 2005; Miao et al. 2012), we find evidence of substantial genetic differentiation among *P. vivax* mitochondrial genomes sampled in different regions. Only 15 haplotypes are shared between regions (supplementary table S5, Supplementary Material online) and most haplotypes (321/357) are restricted to single countries (supplementary table S1, Supplementary Material online). Furthermore, just four haplotypes were found in more than two regions and all four were sampled in Southern Asia (India or Bangladesh), while three were sampled in Papua New Guinea (supplementary table S2, Supplementary Material online). Indeed, although every region has at least one shared haplotype, two regions, Southern Asia and Melanesia, each have eight shared haplotypes, while East Asia and South America, despite having been sampled more extensively, have only one and two haplotypes shared with other regions, respectively (supplementary table S5, Supplementary Material online). Tests of differentiation based on F_{st} values, S_{nn} values, and mean numbers of pairwise nucleotide differences tell a similar story. The latter two test statistics are significantly greater than zero at the level of $P < 0.01$ (indicating genetic differentiation) in all pairwise comparisons except between Africa and Southern Asia and between South America and Central America, while the F_{st} test only fails to reach this level of significance in these two population pairs and in Africa and Madagascar (table 4 and supplementary table S6, Supplementary Material online). These results suggest that gene flow has been restricted (although not necessarily absent) between most regions on the time scale set by the mitochondrial genome-wide mutation rate. Similar results are obtained when differentiation tests are carried out between countries (plus the poorly sampled regions of Africa and Central America) (supplementary table S7, Supplementary Material online). For example, only 11 of the F_{st} tests fail to reach significance and these involve comparisons between India, Pakistan, Africa, and Madagascar, between Thailand and Vietnam, between Central America and Brazil, Colombia, or Peru, or between Colombia and Peru.

To obtain a more detailed picture of population structure within the Americas, we also carried out tests of differentiation using our *P. vivax* samples from Central America, Colombia, Venezuela, northeast Brazil, Amazonian Brazil,

Table 4. Differentiation of *P. vivax* mtDNA Between Regions.

	EAs	SAs	SEAs	Mel	Madg	Afr	ME	CAm	SAm
EAs		2.96	1.65	4.44	3.05	2.53	3.17	3.63	4.11
SAs	0.414		1.13	1.48	0.06	0.07	0.30	0.75	0.95
SEAs	0.255	0.204		1.70	1.40	1.02	1.34	1.03	1.28
Mel	0.514	0.304	0.277		1.71	1.60	1.75	1.03	0.92
Madg	0.436	0.019	0.249	0.369		0.07	0.30	0.90	1.14
Afr	0.360	0.010	0.170	0.336	0.025		0.29	0.74	1.05
ME	0.410	0.079	0.217	0.333	0.099	0.075		1.02	1.28
CAm	0.439	0.164	0.150	0.231	0.267	0.316	0.224		0.20
SAm	0.503	0.212	0.233	0.207	0.256	0.226	0.260	0.016	

Note.—Below diagonal: pairwise F_{st} values; above diagonal: corrected average pairwise nucleotide differences ($\pi XY - [\pi X + \pi Y]/2$). Values shown by underline are not significantly greater than zero ($P > 0.01$) by permutation test. Populations: EAs, East Asia; SAs, South Asia; SEAs, Southeast Asia; Mel, Melanesia; Madg, Madagascar; Afr, Africa; ME, Middle East; CAm, Central America; SAm, South America.

and northern and southern Peru (supplementary table S8, Supplementary Material online). These revealed that the populations of *P. vivax* found in Venezuela and in northeastern Brazil are differentiated from all other populations except each other, while the only other pairwise comparison with a significant F_{st} test is that between southern Peru and Amazonian Brazil. Furthermore, when differentiation is quantified either by F_{st} value or by the corrected average number of pairwise nucleotide differences, the Venezuelan population is consistently the most differentiated from all of the other American populations of *P. vivax* except that found in northeastern Brazil.

The results of the Bayesian phylogeographical analysis based on the subdivision into nine regions are summarized in table 5, which shows all pairs of regions for which the Bayes factor of the corresponding migration rate is >3, indicating strong support for gene flow between that pair of regions. The complete results of this analysis are given in supplementary table S9, Supplementary Material online, while the results of the analysis subdividing the range into individual countries are shown in supplementary table S10, Supplementary Material online. In the regional analysis, evidence of gene flow is found between ten pairs of regions, five of which include Southeast Asia and four of which include Southern Asia. In contrast, South America and East Asia appear to be much more isolated, with each of these being linked by gene flow only to Southeast Asia, while the Middle East is linked only to Melanesia. The isolation of South America is also suggested by the fact that the relative migration rate estimated between this region and Southeast Asia is the lowest (0.322) of the rates with Bayes factors >3. These relative rates can be converted to absolute rates by multiplying by the average pairwise migration rate which is also estimated by the Bayesian phylogeographical analysis. For this analysis, the mean of the posterior distribution of this parameter is 3.37 migrations per lineage per million years (95% HPD: 1.12–6.09), which is much less than the mean of the genome-wide mitochondrial substitution rate, which is 23.58 mutations per lineage per million years (95% HPD: 12.31–35.08). This is consistent with our observation that most mitochondrial haplotypes were only sampled in a single region.

Table 5. Regional Migration Rate Estimates from Bayesian Phylogeography.

Region1	Region 2	P_{inc}	BF	Rate
C America	SE Asia	0.519	3.389	0.695
C America	S Asia	0.512	3.296	0.843
S America	SE Asia	0.956	68.250	0.322
E Asia	SE Asia	1.000	Inf	0.787
SE Asia	S Asia	0.997	1043.923	0.760
SE Asia	Melanesia	1.000	Inf	1.319
S Asia	Madagascar	1.000	Inf	2.884
S Asia	Africa	0.991	345.880	1.779
Melanesia	Africa	0.815	13.838	0.785
Melanesia	Middle East	0.990	310.978	0.768

Note.—P_{inc}, inclusion probability; BF, Bayes factor; rate, mean relative migration rate. Larger inclusion probabilities and larger BFs indicate stronger support for migration between that pair of regions. Results are shown only for those pairs of regions with BF exceeding 3. See supplementary table S5, Supplementary Material online, for the complete results.

Repeating this analysis with the subdivision into 18 populations provides a more detailed picture of gene flow in *P. vivax*. In this case, gene flow is inferred between 22 pairs of populations, the majority of which belong to the same geographical region (supplementary table S10, Supplementary Material online). In particular, of the ten pairs of populations with Bayes factors >100, only two, namely Pakistan and Madagascar, and India and Thailand, are geographically distant from one another. Furthermore, Thailand appears to be a hub in the network defined by these gene flow estimates, being linked to seven other populations: Brazil, Indonesia, Cambodia, Vietnam, China, India, and Melanesia. In contrast, Colombia, Venezuela, Cambodia, Myanmar, Korea, and Madagascar are each linked to only one other population, although some of the pairwise relative migration rates are large. For example, the largest relative migration rate estimated in this analysis is between Colombia and Peru, which exchange migrants at a rate more than three times the average for all population pairs. On the other hand, the relative migration rate estimated between Myanmar and China is the lowest amongst those migration rates with Bayes factors >3 (here BF = 133) and is only 0.325 times as large as the overall migration rate. The posterior mean of the overall pairwise

migration rate estimated for this model is 8.87 migrations per lineage per million years (95% HPD: 3.31–15.01), which, as expected, is larger than the rate estimated using the coarser regional subdivision.

Lastly, the posterior distributions of the location of the most recent common ancestor (MRCA) of *P. vivax* for each analysis are shown in supplementary table S11, Supplementary Material online. When the analysis is conducted using the nine-region subdivision, the most likely location of the MRCA is in Southeast Asia (pp = 0.701), followed by South America (pp = 0.206) and then East Asia (pp = 0.051). In contrast, if the range is subdivided into 18 populations, the posterior distribution is more diffuse, with the greatest probability mass being assigned to Thailand (pp = 0.218), followed by Melanesia (pp = 0.184), Brazil (pp = 0.126), and China (pp = 0.077).

Discussion

Depending on the choice of marker, studies of genetic variation in populations of *P. vivax* in the Americas have reached different conclusions. While surveys of mitochondrial genomes have reported limited variation consistent with a recent introduction of this parasite to the New World (Cornejo and Escalante 2005; Mu et al. 2005; Culleton et al. 2011), many other studies have documented extensive microsatellite variation even within small geographical regions (Imwong et al. 2007; Joy et al. 2008; Karunaweera et al. 2008; Rezende et al. 2010; Van den Eede et al. 2010). In contrast, studies on nuclear genes encoding antigens have shown low local diversity (Grynberg et al. 2008; Chenet et al. 2012b) but strong geographic structure (Chenet et al. 2012b). Meanwhile, despite having small sample sizes, two recent studies of the nuclear genome of *P. vivax* have found that genomes sampled within the Americas are nearly as divergent from one another as they are from genomes sampled on other continents (Chan et al. 2012; Neafsey et al. 2012). Such discrepancies have also been noted in studies of global *P. vivax* diversity (De Brito and Ferreira 2011) and are likely caused by differences both in sampling and in the biological features of the markers themselves, such as mutation rate and selection.

Here, by analyzing a greatly expanded sample of mitochondrial genomes from multiple continents, we show that South American populations of *P. vivax* harbor much greater mitochondrial diversity than previously reported. Not only is the number of haplotypes identified in this study more than an order of magnitude greater than in earlier surveys, but the nucleotide diversity of our sample is also substantially higher (table 1). Of particular interest is the observation of a divergent South American clade that includes 41 mitochondrial genomes sampled primarily in northeastern Brazil and Venezuela. This clade may have originated 76 ka (29–135 ka, table 2) and its current geographic distribution probably explains why we found no evidence of genetic differentiation between northeastern Brazil and Venezuela. It is also responsible for the significant F_{st} values between each of these two localities and the other populations sampled in the Americas. Further sampling will be required to better define the distribution and characteristics of this clade, which appears to be most abundant along the Atlantic coast of South America.

On a global scale, our sample of *P. vivax* mitochondrial lineages can be divided into two divergent groups with very different geographical distributions (fig. 1 and supplementary fig. S1, Supplementary Material online). Approximately 20% of the mtDNA haplotypes belong to a moderately strongly supported clade (pp = 0.91) with a primarily east and southeast Asian distribution. This clade is estimated to be about half as old (121 ka [46–246]) as the entire *P. vivax* sample and exhibits pronounced phylogenetic and geographical substructure, including two large subclades that mainly contain sequences either from the Anhui and Guizhou provinces of central China (Asia a) or from southern China, Korea, and Indonesia (Asia b). Furthermore, although a few haplotypes belonging to this clade were sampled in Africa, India, and Papua New Guinea, we did not identify any such haplotypes in the Americas despite extensive sampling. Similar results were reported by Miao et al. (2012). It is worth noting that it is currently unclear whether this structure is mirrored in the nuclear genome. Although both of the neighbor joining trees inferred from genome-wide single-nucleotide polymorphism (SNP) data by Neafsey et al. (2012) and Chan et al. (2012) are star-shaped, this could either be an artifact of applying phylogenetic methods to recombining DNA sequences (Schierup and Hein 2000) or it could indicate that the small number of genomes do not include any strains belonging to this Asian clade.

Our estimates of the TMRCA of the global *P. vivax* mtDNA sample (283 ka [81–542], tables 2) are similar to those previously reported, which encompass the range 50–450 ka (Jongwutiwes et al. 2005; Mu et al. 2005; Cornejo and Escalante 2006; Miao et al. 2012). This suggests that our enlarged sample does not contain any new exceptionally divergent mitochondrial haplotypes, except possibly for the pair Col15/Brz_Cu23. In fact, when these two sequences are omitted from the global *P. vivax* sample, a repeat of the Bayesian skyline analysis yields a somewhat smaller estimate of the TMRCA (213 ka [71–403]), which is consistent with the tendency for this lineage to occur as an outgroup to the remaining sequences in some of the phylogenetic analyses (e.g., supplementary fig. S1c, Supplementary Material online). While few comparable estimates have been made for nuclear loci, those currently available are of a similar age. For example, using an average nuclear substitution rate of 2.2×10^{-9} substitutions per site per year at 4-fold degenerate sites, Neafsey et al. (2012) estimated the average coalescence time of six complete nuclear genomes included in their study to be 370 ka (the estimate reported in the original paper was incorrectly stated to be 768 ka; Neafsey D, personal communication). Likewise, using a substitution rate of $3.6–9.6 \times 10^{-9}$ substitutions per site per year, Gupta et al. (2012) estimated the TMRCA of a sample of 126 *P. vivax* isolates from India to be between 83 and 222 ka in a set of 12 noncoding regions spanning 5.6 kb on chromosome 13. Collectively, these results suggest that the parasites ancestral to the extant *P. vivax* populations existed between 50 and 550 ka before present.

Bayesian skyline analysis of our sample of mitochondrial genomes suggests that the growth of the global *P. vivax* population began to accelerate approximately 60 ka (fig. 3), a timeframe that is consistent with the demographic history of human populations (Gravel et al. 2011). Specifically, the divergence of African and Eurasian populations is estimated to have occurred 51 ky before present (95% HPD: 45–69 ka). Since most extant populations of *P. vivax* are found in contemporary or derived Eurasian populations, which are believed to have expanded more rapidly than African populations following their split (Gravel et al. 2011), it is plausible that such events had an effect on the demographic history of *P. vivax* (Jongwutiwes et al. 2005; Mu et al. 2005; Cornejo and Escalante 2006; Miao et al. 2012). The demographic histories of the regional samples of *P. vivax* are more complex. On the one hand, skyline analyses of several of the regional samples, including those from South Asia, Southeast Asia. and, surprisingly, South America, also point to rapid population growth in each of the corresponding ancestral populations beginning about 60–70 ka. Of course, these ancestral populations need not have existed in the regions where the extant populations are; thus, the similarity of their skyline plots may indicate that the parasite populations found in these three regions derive in large part from the same ancestral population existing tens of thousands of years ago. In contrast, the skyline plot for the East Asian samples reveals a population expansion that only began approximately 10 ka, around the time when rice and millet are both believed to have been domesticated in China (Lu et al. 2009; Molina et al. 2011). Since most of the East Asian haplotypes belong to a divergent clade that is distinct from all other *P. vivax* lineages for at least 121 ka (46–246), it is not surprising that the demographic history of this region is unlike that inferred elsewhere in Asia.

We were also able to use the Bayesian biogeographical method developed by Lemey et al. (2009) to investigate the geographical location of this ancestor. Depending on whether the range of the parasite is subdivided into regions or countries, this approach identifies either Southeast Asia (pp = 0.70) or Thailand (pp = 0.22), respectively, as the most likely location. Thus, the two analyses are broadly consistent. Furthermore, an ancestral presence of *P. vivax* or its immediate predecessor in Southeast Asia is consistent with the hypothesis that *P. vivax* is derived, by way of a host switch, from a lineage of *Plasmodium* infecting Southeast Asian macaques (Escalante et al. 2005; Cornejo and Escalante 2006). On the other hand, examination of the full posterior distributions of the two analyses paints a less coherent picture. In particular, while the regional analysis identifies South America (pp = 0.21) and East Asia (pp = 0.05) as the second and third most likely locations of the MRCA, the next two most likely locations in the country-level analysis are Papua New Guinea (pp = 0.18) and Brazil (pp = 0.13). Since Melanesia has a negligible posterior probability in the regional analysis (pp = 0.0002), this shows that the inferred location of the MRCA depends in part on how we choose to subdivide the range of the parasite.

The origins of the malaria parasites found in the Americas have been the focus of renewed interest (de Castro and Singer 2005; Gerszten et al. 2012; Yalcindag et al. 2012). However, while *Plasmodium falciparum* is believed to have been introduced to the Americas, at least in part, through the trans-Atlantic slave trade (Yalcindag et al. 2012), an African origin for the American populations of *P. vivax* is all but ruled out by the rarity of *P. vivax* in Sub-Saharan Africa, where the Duffy antigen null allele protects much of the population from this infection. Indeed, our investigation confirms that the relatively few *P. vivax* samples that were obtained from Africa are very closely related to parasites found in India and Pakistan, but share no recent common ancestry with the American haplotypes included in our sample (Krief et al. 2010; Culleton et al. 2011). Unfortunately, the lack of phylogenetic structure in the star-shaped component of the mitochondrial tree makes it difficult to identify a specific region that might be a source population for the American haplotypes. Similarly, although the Bayesian phylogeographical analysis provides strong support for migration between South America and Southeast Asia (table 5), the estimated migration rate between these two regions is lower than all of the other migration rates with Bayes factors >3. Furthermore, the support for this pairwise migration rate may largely be provided by the two Thai haplotypes which we believe were recently introduced to Thailand from South America. Indeed, this is reflected in the country-level phylogeographical analysis, where the only pairwise migration rate between a South American country and a region outside of the Americas with Bayes factor >3 is between Brazil and Thailand (supplementary table S10, Supplementary Material online). This analysis also finds support for migration between Central America and both India and Myanmar, but this is likely to be due to recent immigration of the common haplotypes CA1 and Sal1 into Central America. Thus, neither the phylogenetic nor the phylogeographical analyses point clearly to the origin of the American haplotypes.

In view of the lack of clear relationships between the American populations of *P. vivax* and those found outside of the New World, we believe that our data, which shows that the American populations harbor numerous private haplotypes, some of which may have diversified in situ, is best explained by two alternative but not mutually exclusive models for the introduction of *P. vivax* into the Americas. One model postulates a pre-Colombian introduction with humans, between 15 and 30 ka (Goebel et al. 2008), carrying a group of parasites originating in Beringia. Indeed, the capacity of some strains of *P. vivax* to cause chronic infections through the formation of long-lived dormant hypnozoites (a liver stage found in *P. vivax* but not in *P. falciparum*) might have facilitated such an introduction by allowing the parasite to persist during the period when the colonizing population occupied highly seasonal environments at high latitudes. Both acute and chronic strains of *P. vivax* are present in South America today, but it is unknown how evolutionarily labile these traits are and whether the common life history traits seen in different parts of the world have a common genetic basis. One argument against this model is

that if it were true, then we might expect American populations of *P. vivax* to share a common ancestry with the dominant haplotypes found in East Asia, which they clearly do not. For example, Amerindian genotypes of the bacterium *Helicobacter pylori* are more closely related to East Asian genotypes than to those found elsewhere in the world, which has been interpreted as evidence for a pre-Colombian introduction of this pathogen (Falush et al. 2001). On the other hand, our observation that the East Asian population of *P. vivax* underwent a rapid expansion around 10 ka makes it plausible that other genotypes were prevalent in this region during the period when the New World was first colonized.

The second model consistent with our data is based on post-European-contact introductions possibly involving multiple source *P. vivax* populations, including the now extinct European *P. vivax* population. Provided that the populations of *P. vivax* once found in Europe were both sufficiently diverse and sufficiently differentiated from the populations surviving in Asia and Melanesia, this model could explain the presence of numerous private haplotypes in the New World, especially if there were repeated introductions from multiple countries in Europe. A similar scenario is proposed by Culleton et al. (2011), who also suggest that both the European and Asian populations of *P. vivax* were derived from an ancient African population of parasites driven to extinction by the spread of Duffy negativity in the human host. They favor a European origin for *P. vivax* populations in the New World on the grounds that it would explain why a predominantly South American haplotype (what we call Sal2 and they call h2) occupies such a central position in the median joining network and has such a high outgroup probability. Notably, despite having a much larger sample size, the results of our network analyses largely agree with those presented in Culleton et al. (2011), i.e., our network topology is similar to theirs and we identify the same two haplotypes as having the highest outgroup probabilities. On the other hand, we believe that this inference must be greeted with some caution, as the identification of outgroup haplotypes using the method implemented in TCS (which relies on the frequency and the location of the haplotype in the network) is sensitive both to sampling and to modern day population dynamics, which may be of particular concern in a species with the potential to cause epidemics. Likewise, it should be noted that phylogenetic analysis of the multispecies dataset did not clearly identify the root of the mitochondrial tree of *P. vivax*. Unfortunately, the loss of the putative source populations makes it difficult to rigorously test this model, although the recovery of ancient DNA from a European sample of *P. vivax* could be used to this end. However, while ancient DNA has been recovered from *P. falciparum* samples (Taylor et al. 1997; Sallares and Gomzi 2001; Frías et al. 2013), we are unaware of any similar successes involving *P. vivax*.

Two scenarios that our data clearly rule out are 1) a single recent (post-European-contact) introduction of *P. vivax* into the Americas either accompanied or followed by a severe bottleneck (Grynberg et al. 2008) and 2) recurrent immigration of parasites into the Americas from other extant populations of *P. vivax*. Indeed, in view of the large number of distinct mitochondrial haplotypes that were sampled only in the Americas, as well as the presence of several divergent lineages, any model assuming introductions on a historical time scale or even during the late Pleistocene (Goebel et al. 2008) requires a diverse group of migrants from one or more source populations. For example, even if we conservatively assume that the genome-wide mitochondrial substitution rate (2.3×10^{-4} mutations per genome per year) is 10-fold higher than that estimated in the *BEAST analyses, the probability of identity by descent between two sequences will be 0.79 if their most recent common ancestor existed 500 years ago and 0.96 if that ancestor existed 100 years ago. This, along with the various phylogenetic and population genetic analyses described above, shows that the high levels of variation documented in South America are not consistent with either high rates of gene flow or a single recent founder event.

Regional geographical structure may explain why we have identified much higher levels of mitochondrial diversity in American *P. vivax* populations than have previous surveys. Not only were earlier studies based on smaller sample sizes, but perhaps more importantly, they were also limited to parasites sampled from fewer areas within South America. Here we have made a deliberate, if still insufficient, attempt to assemble a data set that covers a larger number of regions and better reflects the variation of *P. vivax* across the Americas. We believe that this is important because multiple studies, including this one, suggest that local *P. vivax* populations within countries are strongly structured (Imwong et al. 2007; Orjuela-Sánchez et al. 2010; Van den Eede et al. 2010; Chenet et al. 2012a). Thus, locally, there can be low levels of genetic variation even though regional genetic diversity is high. Such patterns have also been reported for genes encoding antigens (Chenet et al. 2012b). This model offers interesting perspectives in the context of malaria elimination. If such smaller geographic areas can be defined, these can be targeted by programs as "malaria elimination units" with limited risk of reintroduction. Also, low gene flow between areas should facilitate containment in the event of the emergence of drug resistance. If this hypothesis is correct, the implications are that malaria could be eliminated effectively in the Americas in many areas and contained in others. It will be a matter of defining the spatial connectivity and the factors leading to the observed gene flow at a time scale usable for elimination. For this purpose, microsatellites, with their much higher rates of evolution, may be more suitable markers at a local scale (Imwong et al. 2007; Van den Eede et al. 2010; Chenet et al. 2012a).

Despite the limitations of the mitochondrial genome, this study unveils patterns in the Americas that need to be considered in population genomics investigations. Although population genomics can uncover complex demographic processes, it is not immune to sampling bias. Thus, demographic studies based on mitochondrial genomes or nuclear SNPs can be used to design more cost-effective surveys of *P. vivax* genome-wide variation (Carlton et al. 2013). In particular, as the present study has documented high levels of mitochondrial variation in the Americas and the presence of multiple divergent lineages, we believe that population

genomic studies of global *P. vivax* diversity should include multiple isolates sampled from across this continent. Particular efforts should be made in sampling localities from northeastern Brazil and Venezuela since some of the isolates from this region are distinct, at least at the mitochondrial level. Such sampling should also consider that inbreeding can generate clonal expansions in the Americas that are stable in time and space (Chenet et al. 2012a; Echeverry et al. 2013); thus, temporal and spatial scales should be considered in population genomic study designs. Such data will help to resolve the still unclear picture of the origins of *P. vivax* in the Americas and will lead to a better understanding of the historical processes by which this parasite has colonized so much of the planet.

Material and Methods

Samples

As part of local surveillance, whole blood samples were collected in ethylenediaminetetraacetic acid from multiple countries in Latin America (Colombia, Peru, Venezuela, and Brazil), Madagascar, Turkey, Cambodia, and Pakistan between 2003 and 2007 (see supplementary table S2, Supplementary Material online). DNA was extracted using QIAamp® DNA Blood Mini Kit (Qiagen GmbH, Hilden, Germany). Approximately 5,800 bp of the parasite's 6KB mitochondrial genome (mtDNA) was amplified using the oligos Forward 5′ GAG GAT TCT CTC ACA CTT CAA TTC AAT TCG TAC TTC 3′ and Reverse 5′ CAG GAA AAT WAT AGA CCG AAC CTT GGA CTC 3′ with Takara LA Taq™ Polymerase (TaKaRa, Takara Mirus Bio). The PCR conditions were partial denaturation at 94 °C for 1 min and 30 cycles at 94 °C for 30 s and 68 °C for 7 min, and a final extension at 72 °C for 10 min. To detect mixed infections, samples were both cloned and sequenced directly. In all cases, at least two independent 5,800 bp PCR products were cloned using the pGEM®-T Easy Vector System (Promega, USA), and four clones were sequenced from each individual from the two independent PCR reactions. Both strands of the cloned inserts were sequenced using internal primers with at least 100 bp overlap. Mixed infections yield overlapping peaks in the sequence electropherogram when they are sequenced directly. In order to reduce further the risk of PCR errors, additional PCR amplifications and clones were sequenced until the haplotype was confirmed. All of the confirmed haplotypes were included in this study. The sequences reported in this investigation were deposited in Genbank under the accession numbers KC330370–KC330678.

Plasmodium mtDNA Genomes Data

In addition to the new 309 mtDNA genomes that were sequenced as part of this study, we retrieved all publicly available complete *P. vivax* mitochondrial genome sequences from GenBank, giving a total of 731 *P. vivax* mitochondrial genomes (Miao et al. 2012; Culleton et al. 2011). Two sequences were of unknown origin and were excluded from further analysis. Analyses requiring an outgroup were performed on an expanded data set that also included 33 mitochondrial genomes from the following ten *Plasmodium* species: *P. cynomolgi*, *P. inui*, *P. hylobati*, *P. simiovale*, *P. fieldi*, *P. coatneyi*, *P. knowlesi*, *P. fragile*, *P. gonderi*, and *Plasmodium* sp. from mandrills (Coatney et al. 1971; Mu. et al. 2005; Pacheco et al. 2012). This expanded set of sequences was aligned using ClustalX v2.0.12 and then edited by hand. Subsequent gap-stripping left 5,712 sites in the alignment containing all 11 *Plasmodium* species, whereas the *P. vivax* alignment (731 sequences) comprises 5,837 sites. A list of all the *P. vivax* sequences used in this study, along with their countries of origin and accession numbers, is provided in supplementary table S2, Supplementary Material online.

Diversity Estimates

Arlequin v3.5 (Excoffier and Lischer 2010) was used to estimate global, regional, and country levels of nucleotide and haplotype (gene) diversity in the *P. vivax* mtDNA genome and to conduct Tajima's and Fu's neutrality tests. These two tests compare different estimates of the scaled mutation rate Θ based on nucleotide and allelic diversity and can be used to detect violations of mutation–drift equilibrium caused by selection or changing population size. Statistical significance was assessed by using 1000 simulations of Kingman's coalescent and the infinite sites model to estimate the *P*-value of each test statistic. Regional analyses were based on the following assignments: East Asia (Korea, China), South Asia (India, Pakistan, Sri Lanka, and Bangladesh), Southeast Asia (Myanmar, Thailand, Cambodia, Vietnam, Philippines, Indonesia, Borneo, and East Timor), Melanesia (Papua New Guinea, Solomon Islands, and Vanuatu), Madagascar, Africa (Mauritania, Nigeria, São Tomé, Ethiopia, Tanzania, Rwanda, Angola, and Namibia), Middle East (Turkey and Iran), Central America (El Salvador, Honduras, Nicaragua, Panama, and Dominican Republic), and South America (Colombia, Venezuela, Brazil, and Peru). We also used Arlequin to estimate the fixation index (F_{ST}) and the mean number of nucleotide differences between each pair of regions and these were tested for significance using one-sided permutation tests with 1000 simulations. Because there is limited divergence between *P. vivax* mtDNA genomes, the genetic distances used to calculate the F_{ST} values were based on simple pairwise differences. Differentiation between regions was also assessed using Hudson's S_{nn} test (Hudson 2000), which compares the proportion of "nearest neighbors" (i.e., sequences more similar to one another than to any) that belong to the same locality with the proportions observed when sequences are randomly assigned to localities subject to the actual sample sizes. DnaSP v5.10.1 (Librado and Rozas 2009) was used to estimate the S_{nn} test statistics and their significance was assessed by permutation test with 1000 permutations.

Haplotype Networks

Network v4.6.1.0 (Fluxus Technologies 2011) was used to infer a median joining network for the complete set of *P. vivax* mtDNA haplotypes, with transversions weighted twice as much as transitions, with the epsilon parameter set equal to 0, and with up to three rounds of star contraction,

which collapses star-like structures in the network into single nodes. However, because the resulting networks were obscured by numerous intersecting edges, we repeated this analysis using a smaller data set containing the 70 haplotypes that were represented by at least two sequences in the global sample as well as an additional 14 singleton haplotypes that appeared unusually divergent in the complete network (India10, Pak_K21, Thai1, Thai39, Indo11, Camb13, Camb14, Turkey9, Turkey10, Turkey13, Ven4, Brz_Bel17, Brz_Bel18, Brz_Ma1; see supplementary table S2, Supplementary Material online, for haplotype definitions.) In this second analysis, the star contraction operation was applied three times, with contraction of star-like structures with radius equal to 1. We also inferred a median joining network for the full set of sequences sampled in the Americas, using the same settings employed in the preceding analysis. Outgroup probabilities for the complete set of haplotypes (including singletons) were calculated using TCS v1.21 (Clement et al. 2000), which uses an algorithm based on the frequencies and locations of the haplotypes within the network.

Phylogenetic Analyses

We used BEAST v1.7.1 to perform several Bayesian phylogenetic analyses of the *Plasmodium* mtDNA data (Drummond et al. 2012). Since the general time reversible model with gamma-distributed substitution rates and a proportion of invariant sites (GTR + Γ + I) had the highest Akaike information criterion score when the *P. vivax* mtDNA genomes were run through jModeltest (Posada 2008), this substitution model was used in all of the phylogenetic analyses described below. We began by using the multispecies coalescent model implemented in *BEAST (Heled and Drummond 2010) to analyze the complete set of 763 mitochondrial genomes from all 11 of the *Plasmodium* species included in this study. Speciation was modeled by a birth–death process and the genealogy of each species was assumed to follow the constant population size coalescent. In addition, rate variation between lineages was accommodated by a relaxed molecular clock with independent log normally distributed substitution rates. Phylogeographical considerations suggest that the *Plasmodium* clades infecting macaques diverged from those infecting other African non-human primate parasites between 6 and 14.2 million years ago (Pacheco et al. 2012), when *Macaca* branched from *Papio* (Mu et al. 2005; Pacheco et al. 2012). This information was used to calibrate the average genome-wide mitochondrial substitution rate by choosing the prior distribution of the TMRCA of the 763 *Plasmodium* lineages to be uniform on the interval (6, 14.2). Three independent chains were run for 100 million generations apiece and the first 20 million generations of each chain were discarded as burn-in. Convergence of these chains was assessed by examining both the trace files and the effective sample sizes reported by Tracer v1.5.

Because the within-species coalescent implemented in *BEAST assumes that each species has a constant population size, we conducted an additional set of analyses of our *P. vivax* sample by fitting a Bayesian skyline model with piecewise linear population growth spread over eight epochs to an alignment containing the 731 *P. vivax* mitochondrial genomes. Although we continued to use the GTR + Γ + I model, the limited mitochondrial polymorphism within this species led us to adopt a strict molecular clock, with a single mean substitution rate sampled from the gamma-distributed prior chosen as follows. Since we could not directly calibrate the average substitution rate in analyses of alignments containing only *P. vivax* sequences, we used *BEAST to fit a birth–death model of cladogenesis to an alignment containing one sequence from each of the 11 *Plasmodium* species. As in the *BEAST runs, we used a relaxed molecular clock to account for interspecific variation in substitution rates and we required the age of root of this tree to be between 6 and 14.2 Ma. Two independent chains were each run for 100 million steps and the method of moments was used to select a gamma distribution to approximate the posterior distribution of the *P. vivax* mean substitution rate obtained from these two runs. This gamma distribution was then used as a prior distribution for the mean substitution rate (measured in substitutions per Ma) in our skyline analyses of the *P. vivax* data.

The skyline analyses were run in two stages. We first ran two independent chains lasting 100 and 200 million generations, respectively, and then used the combined results of these two chains to estimate the posterior distribution of the genealogy of our *P. vivax* samples. Clades with strong support in this distribution were identified in two ways: first by using TreeAnnotator v1.7.1 to construct the MCC tree and second by using the SumTrees script in the DendroPy v3.9.0 library (Sukumaran and Holder 2010) to construct the majority rules consensus tree from the sampled genealogies. We then ran two additional chains lasting 100 million generations apiece to estimate the ages and posterior probabilities of the major clades identified in the first stage. The first 20 million generations of each run were discarded as burn-in and all four runs were combined to obtain improved estimates of the consensus tree and demography of the global *P. vivax* population.

To investigate the possibility that different regions may have experienced different demographic histories, we also conducted skyline analyses on alignments of *P. vivax* sequences sampled in the same geographical region. We followed the same procedure as in the analysis of the global data, with two exceptions. First, although we continued to use the piecewise linear skyline model, we reduced the number of skyline groups (i.e., change points) so that, where possible, there were approximately 20 coalescence events per group. Where the regional sample size was too small to adhere to this rule, we simply ran the analysis with two skyline groups, which is the minimum number allowed. The second modification was that we used the posterior distributions of the substitution rate parameters obtained from the global analyses to parametrize the corresponding prior distributions used in the region-specific analyses. This approach enabled us to continue using the parameter-rich GTR + Γ + I model of the substitution process despite having much smaller numbers of *P. vivax* sequences in some of the regional alignments.

Phylogeographical Analyses

The spatial dynamics of *P. vivax* mtDNA lineages were studied using the discrete phylogeographical model developed by Lemey et al. (2009) and implemented in *BEAST. In this model, the location of each lineage is treated as a discrete, neutrally evolving character and migration is assumed to follow a reversible Markov chain with a uniform stationary distribution on regions. To control the number of pairwise migration rates that need to be estimated, the method implemented by Lemey et al. (2009) uses a Bayesian variable selection strategy known as the stochastic search variable method (SSVM). This method assigns a prior distribution to the number of non-zero migration rates which favors sparse rate matrices with relatively few positive rates. Furthermore, SSVM not only estimates the magnitude of the migration rate between each pair of populations but also the posterior probability P_{inc} that the rate is nonzero. Notice that the closer the inclusion probability P_{inc} is to 1, the stronger the evidence of migration between that pair of populations. Alternatively, the inclusion probabilities can be used to estimate the Bayes factor for each migration rate using the formula $BF = (P_{inc}/1 - P_{inc})/(q/1 - q)$, where q is the prior probability of inclusion under a truncated Poisson distribution with parameter ln(2) conditioned to be at least as large as the number of populations minus one (Lemey et al. 2009). Larger Bayes factors indicate stronger evidence of migration and by convention variables with Bayes factors >3 are typically retained in a model. In particular, we note that low pairwise migration rates can still have large Bayes' factors if the sequence data provide strong evidence of low but non-zero rates of gene flow. We ran two analyses using two different subdivisions of the geographical range of *P. vivax*. In the first analysis, we subdivided the range into the nine regions described in the section of diversity estimates (East Asia, South Asia, Southeast Asia, Melanesia, Madagascar, Africa, the Middle East, Central America, and South America), while in the second analysis we subdivided the range into 18 regions mostly corresponding to individual countries (Colombia, Venezuela, Brazil, Peru, Indonesia, Cambodia, Vietnam, China, Thailand, Myanmar, Korea, Pakistan, India, and Madagascar) except for four trans-national regions (Central America, Melanesia, Sub-Saharan Africa, and the Middle East) that included sequences from countries with very small sample sizes. Both analyses were conducted using the Bayesian skyline model with piecewise linear population growth, but we assumed that there were three rather than eight change points in the slope of the population size to reduce the number of parameters that needed to be estimated. We also continued to use the GTR + Γ + I substitution model with a strict molecular clock and we used the results of the non-spatial analysis to elicit the substitution rate priors. Two independent Markov chains lasting 100 million generations were run for each analysis and a burn-in period consisting of the first 20 million generations was omitted from each chain. The combined results of each pair of runs were used to estimate the migration rates, their inclusion probabilities, and Bayes factors, as well as the posterior distribution of the location of the most recent common ancestor of the global *P. vivax* sample assuming a uniform prior on the set of locations used in each analysis.

Supplementary Material

Supplementary figures S1 and S2 and tables S1–S11 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank the DNA laboratory at the School of Life Sciences for their technical support. This work was supported by the US National Institutes of Health (R01 GM080586 to A.A.E., the Latin American Center for Malaria Research and Control U19AI089702, and the Intramural Research Program of the National Institute of Allergy and Infectious Diseases to R.M.F.). The content of this article is solely the responsibility of the authors and does not represent the views of the US National Institutes of Health.

References

Alexandre MA, Ferreira CO, Siqueira AM, Magalhaes BL, Mourao MP, Lacerda MV, Alecrim MG. 2010. Severe *Plasmodium vivax* malaria, Brazilian Amazon. *Emerg Infect Dis.* 16:1611–1614.

Anstey NM, Handojo T, Pain MC, Kenangalem E, Tjitra E, Price RN, Maguire GP. 2007. Lung injury in vivax malaria: pathophysiological evidence for pulmonary vascular sequestration and post treatment alveolar-capillary inflammation. *J Infect Dis.* 195:589–596.

Arnott A, Barry AE, Reeder JC. 2012. Understanding the population genetics of *Plasmodium vivax* is essential for malaria control and elimination. *Malar J.* 11:14.

Carlton JM, Das A, Escalante AA. 2013. Genomics, population genetics and evolutionary history of *Plasmodium vivax*. *Adv Parasitol.* 81: 203–222.

Chan ER, Menard D, David PH, et al. 11-coauthor). 2012. Whole genome sequencing of field isolates provides robust characterization of genetic diversity in *Plasmodium vivax*. *PLoS Negl Trop Dis.* 6(9):e1811.

Chenet SM, Schneider KA, Villegas L, Escalante AA. 2012a. Local population structure of *Plasmodium*: impact on malaria control and elimination. *Malar J.* 11:412.

Chenet SM, Tapia LL, Escalante AA, Durand S, Lucas C, Bacon DJ. 2012b. Genetic diversity and population structure of genes encoding vaccine candidate antigens of *Plasmodium vivax*. *Malar J.* 11:68.

Cibulskis RE, Aregawi M, Williams R, Otten M, Dye C. 2011. Worldwide incidence of malaria in 2009: estimates, time trends, and a critique of methods. *PLoS Med.* 8(12):e1001142.

Clement M, Posada D, Crandall K. 2000. TCS: a computer program to estimate gene genealogies. *Mol Ecol.* 9: 1657–1660.

Coatney RG, Collins WE, Warren M, Contacos PG. 1971. The primate malarias. (Washington) DC: U.S. Government Printing Office.

Cornejo OE, Escalante AA. 2006. The origin and age of *Plasmodium vivax*. *Trends Parasitol.* 22:558–563.

Cui L, Escalante AA, Imwong M, Snounou G. 2003. The genetic diversity of *Plasmodium vivax* populations. *Trend Parasitol.* 19:220–226.

Culleton R, Coban C, Zeyrek FY, et al. (13 co-authors). 2011. The origins of African *Plasmodium vivax*; insights from mitochondrial genome sequencing. *PLoS One* 6(12):e29137.

De Brito CFA, Ferreira MU. 2011. Molecular markers and genetic diversity of *Plasmodium vivax*. *Mem Inst Oswaldo Cruz* 106(Suppl I), 12–26.

de Castro MC, Singer BH. 2005. Was malaria present in the Amazon before European conquest? Available evidence and future research agenda. *J Archaeol Sci.* 32: 337–340.

Douglas NM, Anstey NM, Buffet PA, Poespoprodjo JR, Yeo TW, White NJ, Price RN. 2012. The anaemia of *Plasmodium vivax* malaria. *Malar J.* 11:135.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 29: 1969–1973.

Echeverry DF, Nair S, Osorio L, Menon S, Murillo C, Anderson TJ. 2013. Long term persistence of clonal malaria parasite *Plasmodium falciparum* lineages in the Colombian Pacific region. *BMC Genet.* 14:2.

Escalante AA, Cornejo OE, Freeland DE, Poe AC, Durrego E, Collins WE, Lal AA. 2005. A monkey's tale: the origin of *Plasmodium vivax as* a human malaria parasite. *Proc Natl Acad Sci U S A.* 102:1980–1985.

Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour.* 10: 564–567.

Falush D, With R, Linz B, et al. (15 co-authors). 2001. Traces of Human Migrations in *Helicobacter pylori* Populations. *Science* 299: 1582–1585.

Fluxus Technologies Ltd. 2011. Network v. 4.6.1.0. Available from: www.fluxus-engineering.com/network_terms.htm.

Frías L, Leles D, Araujo A. 2013. Studies on protozoa in ancient remains—a review. *Mem Inst Oswaldo Cruz.* 108(1):1–12.

Gerszten E, Allison MJ, Maguire B. 2012. Paleopathology in South American mummies: a review and new findings. *Pathobiology* 79: 247–256.

Goebel T, Waters MR, O'Rourke DH. 2008. The late Pleistocene dispersal of modern humans in the Americas. *Science* 319:1497–1502.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA 1000 Genomes Project, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A.* 108:11983–11988.

Grynberg P, Fontes CJ, Hughes AL, Braga EM. 2008. Polymorphism at the apical membrane antigen 1 locus reflects the world population history of *Plasmodium vivax*. *BMC Evol Biol.* 8:123.

Guerra CA, Hay SI, Lucioparedes LS, Gikandi PW, Tatem AJ, Noor AM, Snow RW. 2007. Assembling a global database of malaria parasite prevalence for the Malaria Atlas Project. *Malar J.* 6:17.

Gupta B, Srivastava N, Das A. 2012. Infering the evolutionary history of Indian Plasmodium vivax from population genetic analyses of multilocus nuclear DNA fragments. *Mol Ecol.* 21:1597–1616.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol Biol Evol.* 27:570–580.

Hudson RR. 2000. A new statistic for detecting genetic differentiation. *Genetics* 155(4):2011–2014.

Imwong M, Nair S, Pukrittayakamee S, et al. (14 co-authors). 2007. Contrasting genetic structure in *Plasmodium vivax* populations from Asia and South America. *Int J Parasitol.* 37:1013–1022.

Jongwutiwes S, Putaporntip C, Iwasaki T, Ferreira MU, Kanbara H, Hughes AL. 2005. Mitochondrial genome sequences support ancient population expansion in *Plasmodium vivax*. *Mol Biol Evol.* 22: 1733–1739.

Joy DA, Gonzalez-Ceron L, Carlton JM, Gueye A, Fay M, McCutchan TF, Su XZ. 2008. Local adaptation and vector-mediated population structure in *Plasmodium vivax* malaria. *Mol Biol Evol.* 25:1245–1252.

Karunaweera ND, Ferreira MU, Munasinghe A, Barnwell JW, Collins WE, King CL, Kawamoto F, Hartl DL, Wirth DF. 2008. Extensive microsatellite diversity in the human malaria parasite *Plasmodium vivax*. *Gene* 410:105–112.

Krief S, Escalante AA, Pacheco MA, et al. (17 co-authors). 2010. On the diversity of malaria parasites in African apes and the origin of *Plasmodium falciparum* from Bonobos. *PLoS Pathog.* 6:e1000765.

Kochar DK, Saxena V, Singh N, Kochar SK, Kumar SV, Das A. 2005. *Plasmodium vivax* malaria. *Emerg Infect Dis.* 11:132–134.

Kute VB, Trivedi HL, Vanikar AV, Shah PR, Gumber MR, Patel HV, Goswami JG, Kanodia KV. 2012. *Plasmodium vivax* malaria-associated acute kidney injury, India, 2010–2011. *Emerg Infect Dis.* 18: 842–845.

Li J, Collins WE, Wirtz RA, Rathore D, Lal A, McCutchan TF. 2001. Geographic subdivision of the range of the malaria parasite *Plasmodium vivax*. *Emerg Infect Dis.* 7:35–42.

Librado P, Rozas J. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25(11):1451–1452.

Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots. *PLoS Comput Biol.* 5(9):e1000520.

Lu H, Zhang J, Liu KB, et al. (10 co-authors). 2009. Earliest domestication of common millet (*Panicum miliaceum*) in East Asia extended to 10,000 years ago. *Proc Nat Acad Sci U S A.* 106(18):7367–7372.

Lynch M, Korenromp E, Eisele T, et al. (12 co-authors). 2012. New global estimates of malaria deaths. *Lancet* 380(9841):559.

Miao M, Yang Z, Patch H, Huang Y, Escalante AA, Cui L. 2012. *Plasmodium vivax* populations revisited: mitochondrial genomes of temperate strains in Asia suggest ancient population expansion. *BMC Evol Biol.* 12:22.

Molina J, Sikora M, Garud N, et al. (9 co-authors). 2011. Molecular evidence for a single evolutionary origin of domesticated rice. *Proc Nat Acad Sci U S A.* 108(20):8351–8356.

Mu J, Joy DA, Duan J, et al. (11 co-authors). 2005. Host switch leads to emergence of *Plasmodium vivax* malaria in humans. *Mol Biol Evol.* 22:1686–1693.

Mueller I, Galinski MR, Baird JK, Carlton JM, Kochar DK, Alonso PL, del Portillo HA. 2009. Key gaps in the knowledge of *Plasmodium vivax*, a neglected human malaria parasite. *Lancet Infect Dis.* 9:555–566.

Murray CJ, Rosenfeld LC, Lim SS, Andrews KG, Foreman KJ, Haring D, Fullman N, Naghavi M, Lozano R, Lopez AD. 2012. Global malaria mortality between 1980 and 2010: a systematic analysis. *Lancet* 379(9814):413–431.

Neafsey DE, Galinsky K, Jiang RH, et al. (18 co-authors). 2012. The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than Plasmodium falciparum. *Nat Genet.* 44(9):1046–1050.

Orjuela-Sánchez P, Karunaweera ND, da Silva-Nunes M, et al. (17 co-authors). 2010. Single-nucleotide polymorphism, linkage disequilibrium and geographic structure in the malaria parasite *Plasmodium vivax*: prospects for genome-wide association studies. *BMC Genet.* 11:65.

Pacheco MA, Reid MJ, Schillaci MA, Lowenberger CA, Galdikas BM, Jones-Engel L, Escalante AA. 2012. The origin of malarial parasites in orangutans. *PLoS One* 7:e34990.

Posada D. 2008. jModelTest: phylogenetic model averaging. *Mol Biol Evol.* 25:1253–1256.

Rezende AM, Tarazona-Santos E, Fontes CJ, Souza JM, Couto AD, Carvalho LH, Brito CF. 2010. Microsatellite loci: determining the genetic variability of *Plasmodium vivax*. *Trop Med Int Health.* 15(6):718–726.

Sallares R, Gomzi S. 2001. Biomolecular archaeology of malaria. *Anc Biomol.* 3:195–213.

Schierup MH, Hein J. 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics* 156(2):879–891.

Simonsen KL, Churchill GA, Aquadro CF. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* 141: 413–429.

Sukumaran J, Holder MT. 2010. A Python library for phylogenetic computing. *Bioinformatics* 26:1569–1571.

Taylor GM, Rutland P, Molleson T. 1997. A sensitive polymerase chain reaction method for the detection of *Plasmodium* species DNA in ancient human remains. *Anc Biomol.* 1:193–203.

Van den Eede P, Van der Auwera G, Delgado C, et al. (11 co-authors). 2010. Multilocus genotyping reveals high heterogeneity and strong local population structure of the *Plasmodium vivax* population in the Peruvian Amazon. *Malar J.* 9:151.

World Health Organization (WHO). 2011. World malaria report 2011. [cited 2012 July 24] Available from: http://www.who.int/malaria/world_malaria_report_2011/en.

Yalcindag E, Elguero E, Arnathau C, et al. (35 co-authors). 2012. Multiple independent introductions of *Plasmodium falciparum* in South America. *Proc Natl Acad Sci U S A.* 109:511–516.