HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2017 March 13.

Published in final edited form as:

Nat Genet. 2016 August ; 48(8): 953–958. doi:10.1038/ng.3588.

Population genomics studies identify signatures of global dispersal and drug resistance in *Plasmodium vivax*

Daniel N Hupalo[1], **Zunping Luo**[1], **Alexandre Melnikov**[2], **Patrick L Sutton**[1], **Peter Rogov**[2], **Ananias Escalante**[3], **Andrés F Vallejo**[4], **Sócrates Herrera**[4], **Myriam Arévalo-Herrera**[4,5], **Qi Fan**[6], **Ying Wang**[7], **Liwang Cui**[8], **Carmen M Lucas**[9], **Salomon Durand**[9], **Juan F Sanchez**[9], **G Christian Baldeviano**[9], **Andres G Lescano**[9], **Moses Laman**[10], **Celine Barnadas**[11,12], **Alyssa Barry**[13,14], **Ivo Mueller**[13,14,15], **James W Kazura**[16], **Alex Eapen**[17], **Deena Kanagaraj**[17], **Neena Valecha**[18], **Marcelo U Ferreira**[19], **Wanlapa Roobsoong**[20], **Wang Nguitragool**[21], **Jetsumon Sattabonkot**[20], **Dionicia Gamboa**[22,23], **Margaret Kosek**[24], **Joseph M Vinetz**[22,23,25], **Lilia González-Cerón**[26], **Bruce W Birren**[2], **Daniel E Neafsey**[2], and **Jane M Carlton**[1]

[1]Center for Genomics and Systems Biology, Department of Biology, New York University, New York, New York, USA [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA [3]Institute for Genomics and Evolutionary Medicine, Department of Biology, Temple University, Philadelphia, Pennsylvania, USA [4]Caucaseco Scientific Research Center, Cali, Colombia [5]Faculty of Health, Universidad del Valle, Cali, Colombia [6]Dalian Institute of Biotechnology, Dalian, Liaoning, China [7]Third Military Medical University, Shapingba, Chongqing, China [8]Department of Entomology, Pennsylvania State University, University Park, Pennsylvania, USA [9]US Naval Medical Research Unit No. 6, Callao, Peru [10]Papua New Guinea Institute of Medical Research, Madang, Papua, New Guinea [11]Vector Borne Diseases Unit, Papua New Guinea Institute of Medical Research, Goroka, Papua New Guinea [12]Division of Infection and Immunity, Walter & Eliza Hall Institute of Medical Research, Parkville, Australia [13]Division of Population Health and Immunity, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia [14]Department of Medical Biology, University of Melbourne, Carlton, Victoria, Australia [15]Institute of Global Health (ISGLOBAL), Barcelona, Spain [16]Center for Global Health and Diseases, Case Western Reserve University, Cleveland, Ohio, USA [17]National Institute of Malaria Research Field Unit, Indian Council of Medical Research, National Institute of Epidemiology Campus, Chennai, Tamil Nadu, India [18]National Institute of Malaria Research, Indian Council of Medical Research, New Delhi, India [19]Department of Parasitology, Institute of Biomedical Sciences, University of São

Correspondence should be addressed to J.M.C. (jane.carlton@nyu.edu) or D.E.N. (neafsey@broadinstitute.org).

Urls. Malaria Research and Reagent Resource Repository (MR4), http://www.malaria.mr4.org/; PlasmoDB, http://www.plasmodb.org/; Broad *Plasmodium* 100 Genomes project, https://olive.broadinstitute.org/projects/plasmodium_100_genomes.

Accession codes. Illumina sequencing reads are available through the NCBI Sequence Read Archive with BioProject accession numbers PRJNA240356–PRJNA240533.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

Author Contributions: J.M.C., I.M. and D.E.N. conceived and conducted the study. A.M., P.L.S., P.R., A.F.V., Q.F., Y.W., C.M.L., S.D., J.F.S., M.L., C.B., D.K., W.R., W.N. and M.K. undertook field and/or wet-lab work and sequencing of the samples. D.N.H., Z.L., J.M.C. and D.E.N. analyzed data. D.N.H., J.M.C., Z.L. and D.E.N. wrote the manuscript, and A.E., S.H., M.A.-H., L.C., G.C.B., A.G.L., A.B., I.M., J.W.K., A.E., N.V., M.U.F., J.S., D.G., J.M.V., L.G.-C. and B.W.B. revised the manuscript and made comments.

Competing Financial Interests: The authors declare no competing financial interests.

Paulo, São Paulo, Brazil [20]Mahidol Vivax Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand [21]Department of Molecular Tropical Medicine and Genetics, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand [22]Instituto de Medicine Tropical Alexander von Humboldt, Universidad Peruana Cayetano Heredia, Lima, Peru [23]Departamento de Ciencias Celulares y Moleculares, Universidad Peruana Cayetano Heredia, Lima, Peru [24]Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA [25]Division of Infectious Diseases, Department of Medicine, University of California San Diego, La Jolla, California, USA [26]Regional Centre for Research in Public Health, National Institute for Public Health, Tapachula, Chiapas, México

Abstract

Plasmodium vivax is a major public health burden, responsible for the majority of malaria infections outside Africa. We explored the impact of demographic history and selective pressures on the *P. vivax* genome by sequencing 182 clinical isolates sampled from 11 countries across the globe, using hybrid selection to overcome human DNA contamination. We confirmed previous reports of high genomic diversity in *P. vivax* relative to the more virulent *Plasmodium falciparum* species; regional populations of *P. vivax* exhibited greater diversity than the global *P. falciparum* population, indicating a large and/or stable population. Signals of natural selection suggest that *P. vivax* is evolving in response to antimalarial drugs and is adapting to regional differences in the human host and the mosquito vector. These findings underline the variable epidemiology of this parasite species and highlight the breadth of approaches that may be required to eliminate *P. vivax* globally.

P. vivax causes an estimated 15.8 million clinical malaria cases each year[1] but remains understudied in comparison to *P. falciparum* because of its lower mortality and lack of a continuous *in vitro* culture system. Small-scale surveys of samples collected from human patients ('isolates') in the last century and adapted for propagation in splenectomized monkeys have indicated higher genetic diversity for *P. vivax* relative to *P. falciparum* and discordant phylogeographical histories for the two species[2], likely caused, in part, by the unique ability of *P. vivax* to persist in a dormant state in the liver for months or years. This capacity for dormancy has enabled *P. vivax* to historically cause infections in Europe and North America, where the temperate climate disfavors year-round transmission by mosquitoes[3].

We used hybrid selection[4] to enrich parasite DNA in blood samples heavily contaminated with human DNA obtained from patients infected with *P. vivax*, collected as part of the International Centers of Excellence for Malaria Research program[5]. We sequenced 182 isolates from 11 countries across the globe (Fig. 1) and obtained high sequencing depth and quality for the majority of samples. Hybrid selection using synthetic baits (Online Methods and Supplementary Fig. 1) increased the fraction of parasite DNA in all but four samples, with the greatest enrichment in samples obtained starting with <1% parasite DNA (Fig. 2). We supplemented these new data with published sequences from patient isolates and monkey-adapted laboratory lines for a final data set of 195 *P. vivax* genomes

(Supplementary Table 1). We aligned reads to the Salvador I reference genome assembly[6] and 'called' SNP variants as described in Supplementary Figure 2. Because *P. vivax* is haploid in the human host, we used rates of heterozygous variant calls to classify 122 of the newly sequenced isolates as single-genotype infections and inferred the remaining 41 samples with sufficient coverage as multigenotype ('complex') infections (Online Methods and Supplementary Fig. 3). We restricted our population genetic analyses of diversity and divergence to single-genotype isolates.

We performed a principal-component analysis (PCA; Fig. 3a and Supplementary Fig. 4) on the SNP variants to explore the global population structure of *P. vivax*. Similar to a recent microsatellite study[7], the major axis of differentiation divides the New World from Old World samples. Given recent evidence for an African origin of *P. vivax*[8], it has been proposed that New World *P. vivax* populations had been founded during the colonial era by a now-extirpated European parasite population[9]. The second and third principal components define a distinct Papua New Guinea (PNG) cluster and a cluster of isolates from Mexico, respectively. We estimated the fraction of the genome exhibiting identity by descent among pairs of samples from geographical regions harboring at least 12 single-genotype isolates (Online Methods) and predicted that isolates sampled from Mexico have recent common ancestry (Supplementary Fig. 5), explaining their unusual clustering. Malaria in Mexico has steadily declined over the last decade to <10,000 cases[1], and the circulation of genetically similar *P. vivax* may improve prospects for disease control. In contrast to the *P. vivax* population in Mexico, the PNG population is very diverse from other *P. vivax* populations. Our data support a set of SNPs that differentiate *P. vivax* populations from those in the Americas, Asia and PNG, with applications for extending the ability to identify the source of imported infections[10] via SNP barcoding[11] or other genotyping methodologies.

Maximum-likelihood phylogenetic analysis yielded a tree concordant with the PCA (Fig. 3b), highlighting the distinctiveness of the PNG population (alternative reference genome maximum-likelihood trees with bootstrap values are available in Supplementary Fig. 6). We also observed a cluster of African (Mauritania and Madagascar) isolates that grouped closely with Indian isolates midway between the clades from Asia and the Americas. Previous studies of *P. vivax* mitochondrial diversity have identified shared or similar haplotypes among small numbers of sequenced African and Indian/Pakistani samples[9,12], suggesting that *P. vivax* was potentially reintroduced to Africa from South Asia by colonial seafarers after its virtual extirpation in Africa owing to the spread of the infection-deterring Duffy-negative mutation in human populations[13]. Given that the nuclear genome undergoes sexual recombination, the present data sets potentially offer a more sensitive history of gene flow between populations than mitochondrial data. Admixture analysis of the nuclear SNP data (Fig. 3c and Supplementary Fig. 7) suggests that the isolates from India and Africa exhibit a highly heterogeneous ancestry relative to isolates from other populations, and we inferred that they may derive from both New and Old World lineages. This finding adds support to the hypothesis that contemporary African and South Asian *P. vivax* populations are genetically similar and suggests that South Asian *P. vivax* populations may have genetically mingled with European ('New World') lineages during the colonial era or could reflect ancient connections between human populations in the eastern Mediterranean, the Middle East and the Indian subcontinent inferred from admixture signals in human genetic data[14].

The transatlantic slave trade that occurred between the sixteenth and nineteenth centuries continues to shape the contemporary epidemiology of malaria in South America. Genotyping the human Duffy mutation identified no homozygous samples in our collection (Supplementary Table 1) but identified ten samples from Colombian individuals that were heterozygous, five of whom self-identified as having African ancestry. In this setting of non-African parasites infecting people of partial African descent, we observed the dominance of disease susceptibility conferred by the wild-type Duffy allele. Further, we found no evidence of increased copy number of the *P. vivax* Duffy-binding protein gene hypothesized to overcome the barrier to infection presented by Duffy negativity in Madagascar[15] (Supplementary Table 1). South American *P. vivax* epidemiology reflects recent contact between humans and parasites introduced to this region from disparate geographical locations.

Our previous studies of *P. vivax* have highlighted its high diversity in comparison to *P. falciparum*, suggesting a larger and/or more stable effective population size, but were limited by sample size[2]. In this expanded data set, we found that even the least genetically diverse *P. vivax* subpopulation in our study was more diverse than a sampling of diverse *P. falciparum* isolates (Fig. 4). All of our sampled *P. vivax* populations were significantly more diverse than *P. falciparum* ($P < 0.001$, Wilcoxon signed-rank test), with Old World populations showing nearly twice as much nucleotide diversity. The population samples from Myanmar and Thailand were as diverse as the PNG samples, suggesting that the high divergence of the PNG population is due to protracted genetic isolation and not an elevated mutation rate. The most diverse genes are antigenic, such as members of the merozoite surface protein 3 (*MSP3*)[16] and *MSP7* multigene families and the serine-repeat antigen family (*SERA*)[17], genes important for immune evasion (Supplementary Fig. 8).

To explore the genomic profile of divergence and selection among *P. vivax* populations, we calculated the fixation index (F_{ST}) across each chromosome for pairs of well-sampled populations separated by differing geographical distances. The divergence profile we observed in a comparison of Old World versus New World single-genotype samples (Fig. 5a) was not generally similar to the profile that we observed in a comparison of samples from PNG and Thailand. Indeed, there were some notable differences suggesting population-specific selection pressures (Fig. 5b and Supplementary Table 2). Among the genes exhibiting similar signals of divergence for both the proximal and distal geographical comparisons were two loci associated with antifolate drug resistance in *P. falciparum*: *DHPS* (dihydropteroate synthase) and *DHFR-TS* (dihydrofolate reductase–thymidylate synthase). *DHPS* and *DHFR-TS* genes also reside in regions of heightened linkage disequilibrium (Fig. 5c,d), suggesting that they have been subjected to selective sweeps. These results indicate that sulfadoxine, pyrimethamine and other antifolate drugs constitute the most important recent evolutionary pressure on *P. vivax* parasite populations. Application of a McDonald–Kreitman test to *P. vivax* genes with orthologs in the sister taxon *Plasmodium cynomolgi*[18] of Southeast Asian macaques identified strong signals of selection on other genes, including several associated with red blood cell invasion, for example, *MAEBL* (merozoite-adhesive erythrocytic binding protein; Supplementary Fig. 9 and Supplementary Table 3). Other notable genes identified as potentially adapting to regional differences in the human host or

the mosquito vector, or due to the influence of other environmental variables, are indicated in Table 1 and Supplementary Figures 10–15.

We also observed a divergence signal on chromosome 12 enhanced in the New World versus Old World comparison relative to the PNG versus Thailand comparison. The divergence signal was centered on a surface-expressed gamete gene, *Pvs47* (Fig. 5e), orthologous to a locus in *P. falciparum* (*Pfs47*) that is a strong determinant of successful evasion of the mosquito JNK immune response[19] and that was likely subjected to selection as *P. falciparum* adapted to New World vectors during the colonial era[20]. New World anopheline vectors belong to the subgenus *Nyssorhynchus*, which diverged from the lineage leading to Old World anopheline mosquitoes approximately 100 million years ago[21], and likely presented comparable evolutionary challenges to both *P. falciparum* and *P. vivax* after human-mediated introduction of malaria parasites to the New World in the last 500 years. Concordant with the hypothesis that *Pvs47* was subject to strong selection upon the arrival of *P. vivax* in the New World, we observed considerably reduced haplotype diversity in Mexican, Colombian, Brazilian and Peruvian *P. vivax* populations in comparison with Old World populations (Fig. 5f), a pattern that was not exhibited by *DHFR-TS* or *DHPS* (Supplementary Fig. 16) but which has been observed previously in *P. falciparum* for *Pfs47* (ref. 22). Despite reduced haplotype diversity and strong divergence between populations, *Pvs47* did not exhibit heightened linkage disequilibrium (Fig. 5e and Supplementary Fig. 14), suggesting that it was subject to selection less recently than the *DHFR-TS* and *DHPS* drug resistance loci.

Our global survey of genomic diversity in *P. vivax* provides insights into the population structure, demographic history and selective pressures acting upon this species, contextualizing and extending previous, smaller-scale observations[2]. Just as malaria has influenced the human genome[23], this *P. vivax* global genomic data set reflects ongoing evolutionary interactions with the parasite's human and mosquito hosts, interactions that are likely to intensify as disease elimination efforts escalate.

Online Methods

Sample collection, hybrid selection and DNA sequencing

We collected venous blood samples from *P. vivax*-infected patients after obtaining informed consent and following ethical approval at the local institutional review board of each participating site. Details of the sites are available in the Supplementary Note. The monkey-adapted *P. vivax strains* Panama (MRA-343), Vietnam ONG (MRA-341) and Nicaragua I (MRA-340) were obtained from the Malaria Research and Reference Reagents Resource (see URLs). We extracted genomic DNA from each sample and used a modified nested PCR method targeting the multicopy small-subunit ribosomal gene of *P. vivax* for species confirmation[33]. We constructed 200-bp-insert Illumina libraries for each isolate and performed hybrid selection on each sequencing library using genome-wide baits to enrich parasite DNA, with synthetic oligonucleotide baits to test maximum possible enrichment, as described[4]. We sequenced the hybrid-selected libraries with 101-bp paired-end reads in indexed batches of 12 on Illumina HiSeq 2000 or 2500 lanes and evaluated parasite DNA enrichment by comparing the fraction of HiSeq reads mapping to the *P. vivax* reference

genome assembly to the mapping fraction observed in low sequencing coverage generated on the MiSeq platform.

Human Duffy antigen/chemokine receptor and *P. vivax* Duffy-binding protein genotyping

We determined the genotype of the human Duffy antigen/chemokine receptor (DARC) of *P. vivax*-infected patients by using Burrows-Wheeler aligner (BWA)[34] to align fastq files of 177 sequenced isolates against the DARC nucleotide sequence (human genome Build 38) and visualizing read alignments. These were classified as either Duffy homozygous positive, homozygous negative or heterozygous according to the presence of a point mutation in the promoter region (a T>C mutation at position −33 in the GATA box). The *in silico* results were validated using a modified wet-lab assay[35] (Supplementary Table 4 and Supplementary Note). We mapped all paired-end reads for each isolate to the *P. vivax* Salvador I reference genome and used the indel detection software Pindel[36] to detect *P. vivax* Duffy-binding protein (*DBP*) gene duplications. We validated these results by PCR amplification of flanking regions of *DBP*[37].

Sequence processing and alignment

We aligned reads from all isolates to the Salvador I reference genome assembly using BWA[34] with default parameters and processed the resulting alignments using SAMtools[38] and Picard v1.66. We obtained sequence data for 12 previously published *P. vivax* isolates from the Sequence Read Archive, converted them into fastq format and aligned them to Salvador I. We used REALPHY[39] software to assess whether our choice of reference genome assembly biased variant calls used for downstream analyses, using two alternative *P. vivax* reference genomes, North Korean and Mauritania I (ref. 2) (Supplementary Fig. 6).

SNP discovery and quality filtering

We used the Genome Analysis Toolkit (GATK) UnifiedGenotyper[40] to identify SNPs in each isolate according to GATK best practices[41] as part of a SNP identification pipeline (Supplementary Fig. 2). We used two parameters to exclude low-quality sites: GQ ≥40 and QUAL ≥30. We filtered SNP variants in single and complex infections according to separate criteria and removed all heterozygous calls from the single-genotype data set (Supplementary Fig. 2). After merging variant calls from all isolates, we removed calls from sites exhibiting a low call rate across the sample set and sites occurring in members of the highly repetitive *P. vivax* variant interspersed repeat (*vir*) multigene family.

Single-genotype versus multigenotype infection classification

To distinguish legitimate heterozygous variants in infections containing multiple genotypes from spurious heterozygous calls, we identified a trusted set of variants exhibiting high-quality homozygous calls for two or more segregating alleles in individual samples. This curated homozygous reference variant set was then used to measure the prevalence of heterozygous SNPs for each isolate, such that only in mixed infections should multiple alternative genotypes be observed. We classified samples as containing more than one genotype if they exhibited greater than 1,236 heterozygous variants (Supplementary Fig. 3).

Population structure

We used a set of filtered variant calls without heterozygous polymorphisms to visualize relationships between *P. vivax* populations using PCA implemented in the SNPRelate R library[42]. We used a combination of two and three eigenvectors to represent the structure of the polymorphism in the population. A more conservative PCA, limited to high-quality (at least 1 million callable sites) single-genotype isolates, was used to confirm the topology (Supplementary Fig. 4). We constructed maximum-likelihood trees using the filtered variant call set limited to homozygous polymorphisms in high-quality *P. vivax* isolates using RAxML[43] and performed admixture analysis with the Admixture software package[44] using the full set of homozygous filtered variant SNPs.

Genetic diversity

We calculated genome-wide nucleotide diversity (π) and Tajima's D from high-quality, single-genotype isolates (Supplementary Table 1) using R and Python custom scripts to tabulate all classes of segregating sites using filtered homozygous variant calls. In addition, we calculated π for 574 one-to-one high-quality orthologs in *P. falciparum* and *P. vivax.* We calculated F_{ST} using the VCFTools package[45]. We identified identical-by-descent genomic regions in pairwise comparisons of isolates using a hidden Markov model analysis of SNP variants[46].

Tests of neutrality/selection

We used data for 81 high-coverage, single-genotype infection isolates to search for signals of natural selection using a McDonald-Kreitman test and the *P. cynolmolgi* B strain reference genome as an outgroup[47]. We created coding-sequence alignments using *P. vivax* population sequence and the single-copy orthologs between species for 2,254 genes using the program MUSCLE[48]. We performed a McDonald-Kreitman test on the observed polymorphic and divergent synonymous and nonsynonymous mutations in each alignment and evaluated significance using a two-tailed χ^2 test to obtain a P value. A q-value error correction was applied to the set of 2,254 gene tests to correct for multiple sampling[49]. We also calculated α for each gene, a summary statistic that estimates the proportion of substitutions fixed by natural selection[50], using a custom script.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge J. Bochicchio and S. Chapman for project management, A. Gnirke for technical support, and members of the Broad Institute Genomics Platform and NYU's Genomics Core for data generation. We thank F. Santillan and P. Michon for technical assistance and MR4 for providing us with malaria parasites deposited by W.E. Collins. The following grants supported this work: National Institute of Allergy and Infectious Diseases (NIAID)/ National Institutes of Health (NIH) International Centers of Excellence for Malaria Research U19AI089676 to J.M.C.; U19AI089681, K24AI068903 and D43TW007120 to J.M.V.; U19AI089672 to L.C.; São Paulo Research Foundation 2009/52729-9 to M.U.F.; National Council for Science and Technology Mexico 29005-M SALUD-2004-119 and National Institute of Public Health Mexico project 476191 to L.G.-C.; Victorian State Government Operational Infrastructure Support and Australian Government National Health and Medical Research Council Independent Medical Research Institutes Infrastructure Support Scheme (NHMRC IRIISS) to A.B. and I.M.; 5U19AI089702 to S.H. and M.A.-H.; Armed Forces Health Surveillance Center, Global Emerging Infections

Surveillance and Response System and US NIH grant D43TW007393 to A.G.L.; NIH U19AI089686 to J.W.K.; and Bill and Melinda Gates Foundation grant to J.S. Sequencing and analysis work at the Broad Institute was supported by federal funds from the NIAID, NIH, US Department of Health and Human Services, under contract HHSN272200900018C. M.U.F. is supported by a senior research scholarship from the Conselho Nacional de Desenvolvimento Científico e Tecnológico of Brazil, I.M. is supported by NHMRC senior research fellowship 1043345 and D.N.H. is supported by NIH training grant T32AI007180. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official policy or position of the US Department of the Navy, the US Department of Defense, the US government or the National Institutes of Health.

References

1. World Health Organization. World Malaria Report. 2014
2. Neafsey DE, et al. The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. Nat Genet. 2012; 44:1046–1050. [PubMed: 22863733]
3. Carter R. Speculations on the origins of *Plasmodium vivax* malaria. Trends Parasitol. 2003; 19:214–219. [PubMed: 12763427]
4. Melnikov A, et al. Hybrid selection for sequencing pathogen genomes from clinical samples. Genome Biol. 2011; 12:R73. [PubMed: 21835008]
5. Carlton JM, et al. Population genetics, evolutionary genomics, and genome-wide studies of malaria: a view across the International Centers of Excellence for Malaria Research. Am J Trop Med Hyg. 2015; 93(suppl):87–98.
6. Carlton JM, et al. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature. 2008; 455:757–763. [PubMed: 18843361]
7. Koepfi C, et al. *Plasmodium vivax* diversity and population structure across four continents. PLoS Negl Trop Dis. 2015; 9:e0003872. [PubMed: 26125189]
8. Liu W, et al. African origin of the malaria parasite *Plasmodium vivax*. Nat Commun. 2014; 5:3346. [PubMed: 24557500]
9. Culleton R, et al. The origins of African *Plasmodium vivax*; insights from mitochondrial genome sequencing. PLoS One. 2011; 6:e29137. [PubMed: 22195007]
10. Rodrigues PT, et al. Using mitochondrial genome sequences to track the origin of imported *Plasmodium vivax* infections diagnosed in the United States. Am J Trop Med Hyg. 2014; 90:1102–1108. [PubMed: 24639297]
11. Baniecki ML, et al. Development of a single nucleotide polymorphism barcode to genotype *Plasmodium vivax* infections. PLoS Negl Trop Dis. 2015; 9:e0003539. [PubMed: 25781890]
12. Taylor JE, et al. The evolutionary history of *Plasmodium vivax* as inferred from mitochondrial genomes: parasite genetic diversity in the Americas. Mol Biol Evol. 2013; 30:2050–2064. [PubMed: 23733143]
13. Miller LH, Mason SJ, Clyde DF, McGinniss MH. The resistance factor to *Plasmodium vivax* in blacks. The Duffy-blood-group genotype, FyFy. N Engl J Med. 1976; 295:302–304. [PubMed: 778616]
14. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009; 461:489–494. [PubMed: 19779445]
15. Ménard D, et al. *Plasmodium vivax* clinical malaria is commonly observed in Duffy-negative Malagasy people. Proc Natl Acad Sci USA. 2010; 107:5967–5971. [PubMed: 20231434]
16. Rice BL, et al. The origin and diversification of the merozoite surface protein 3 (msp3) multi-gene family in *Plasmodium vivax* and related parasites. Mol Phylogenet Evol. 2014; 78:172–184. [PubMed: 24862221]
17. Arisue N, Hirai M, Arai M, Matsuoka H, Horii T. Phylogeny and evolution of the SERA multigene family in the genus *Plasmodium*. J Mol Evol. 2007; 65:82–91. [PubMed: 17609844]
18. Tachibana S, et al. *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. Nat Genet. 2012; 44:1051–1055. [PubMed: 22863735]
19. Molina-Cruz A, et al. The human malaria parasite Pfs47 gene mediates evasion of the mosquito immune system. Science. 2013; 340:984–987. [PubMed: 23661646]

20. Molina-Cruz A, Barillas-Mury C. The remarkable journey of adaptation of the *Plasmodium falciparum* malaria parasite to New World anopheline mosquitoes. Mem Inst Oswaldo Cruz. 2014; 109:662–667. [PubMed: 25185006]
21. Moreno M, et al. Complete mtDNA genomes of *Anopheles darlingi* and an approach to anopheline divergence time. Malar J. 2010; 9:127. [PubMed: 20470395]
22. Anthony TG, Polley SD, Vogler AP, Conway DJ. Evidence of non-neutral polymorphism in *Plasmodium falciparum* gamete surface protein genes Pfs47 and Pfs48/45. Mol Biochem Parasitol. 2007; 156:117–123. [PubMed: 17826852]
23. Kwiatkowski DP. How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet. 2005; 77:171–192. [PubMed: 16001361]
24. Gething PW, et al. A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. PLoS Negl Trop Dis. 2012; 6:e1814. [PubMed: 22970336]
25. Lacroix C, Ménard R. TRAP-like protein of *Plasmodium* sporozoites: linking gliding motility to host-cell traversal. Trends Parasitol. 2008; 24:431–434. [PubMed: 18760672]
26. Thompson J, et al. *Plasmodium* cysteine repeat modular proteins 1-4: complex proteins with roles throughout the malaria parasite life cycle. Cell Microbiol. 2007; 9:1466–1480. [PubMed: 17253978]
27. Chuquiyauri R, et al. Genome-scale protein microarray comparison of human antibody responses in *Plasmodium vivax* relapse and reinfection. Am J Trop Med Hyg. 2015; 93:801–809. [PubMed: 26149860]
28. Pacheco MA, et al. Evidence of purifying selection on merozoite surface protein 8 (MSP8) and 10 (MSP10) in *Plasmodium* spp. Infect Genet Evol. 2012; 12:978–986. [PubMed: 22414917]
29. Mbengue A, et al. A molecular mechanism of artemisinin resistance in *Plasmodium falciparum* malaria. Nature. 2015; 520:683–687. [PubMed: 25874676]
30. Schousboe ML, et al. Multiple origins of mutations in the mdr1 gene—a putative marker of chloroquine resistance in *P. vivax*. PLoS Negl Trop Dis. 2015; 9:e0004196. [PubMed: 26539821]
31. Sidhu AB, Verdier-Pinard D, Fidock DA. Chloroquine resistance in *Plasmodium falciparum* malaria parasites conferred by pfcrt mutations. Science. 2002; 298:210–213. [PubMed: 12364805]
32. Ariey F, et al. A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. Nature. 2014; 505:50–55. [PubMed: 24352242]
33. Rubio JM, et al. Semi-nested, multiplex polymerase chain reaction for detection of human malaria parasites and evidence of *Plasmodium vivax* infection in Equatoria Guinea. Am J Trop Med Hyg. 1999; 60:183–187. [PubMed: 10072133]
34. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
35. Tournamille C, Colin Y, Cartron JP, Le Van Kim C. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat Genet. 1995; 10:224–228. [PubMed: 7663520]
36. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009; 25:2865–2871. [PubMed: 19561018]
37. Menard D, et al. Whole genome sequencing of field isolates reveals a common duplication of the Duffy binding protein gene in Malagasy *Plasmodium vivax* strains. PLoS Negl Trop Dis. 2013; 7:e2489. [PubMed: 24278487]
38. Li H, et al. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]
39. Bertels F, Silander OK, Pachkov M, Rainey PB, van Nimwegen E. Automated reconstruction of whole-genome phylogenies from short-sequence reads. Mol Biol Evol. 2014; 31:1077–1088. [PubMed: 24600054]
40. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–1303. [PubMed: 20644199]
41. Van der Auwera GA, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013; 43:11.10.1–11.10.33. [PubMed: 25431634]

42. Zheng X, et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012; 28:3326–3328. [PubMed: 23060615]

43. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30:1312–1313. [PubMed: 24451623]

44. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19:1655–1664. [PubMed: 19648217]

45. Danecek P, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27:2156–2158. [PubMed: 21653522]

46. Daniels RF, et al. Modeling malaria genomics reveals transmission decline and rebound in Senegal. Proc Natl Acad Sci USA. 2015; 112:7067–7072. [PubMed: 25941365]

47. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature. 1991; 351:652–654. [PubMed: 1904993]

48. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. [PubMed: 15034147]

49. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA. 2003; 100:9440–9445. [PubMed: 12883005]

50. Smith NG, Eyre-Walker A. Adaptive protein evolution in *Drosophila*. Nature. 2002; 415:1022–1024. [PubMed: 11875568]

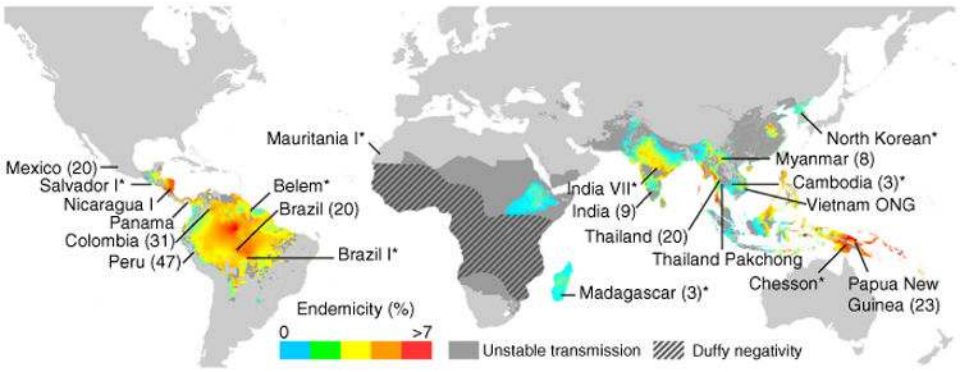

Figure 1.

Country of origin for 195 *P. vivax* patient isolates indicated on a map of *P. vivax* malaria endemicity. Asterisks mark the five published monkey-adapted reference genomes Brazil I, Salvador I, Mauritania I, North Korean and India VII and the published Belem and Chesson strains. The four additional monkey-adapted strains, Vietnam ONG, Nicaragua I, Panama and Thailand Pakchong, sequenced as part of this study are also indicated. Numbers of clinical isolates are given in parentheses. Lines do not represent accurate within-country locations; locations are listed in supplementary table 1. The map was adapted from the Malaria MAP Project[24]

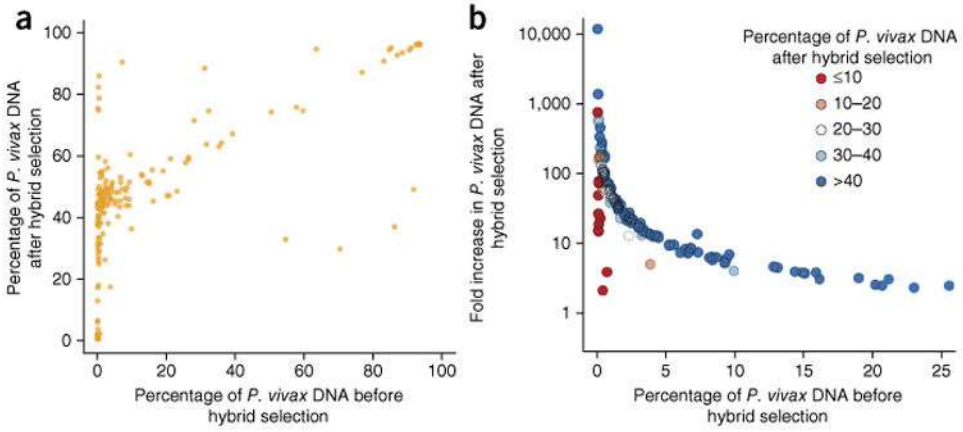

Figure 2.

Enrichment of *P. vivax* DNA extracted from clinical samples using the hybrid selection method. (**a**) Enrichment of parasite DNA measured as the percentage of total sequence reads alignable to the *P. vivax* reference genome before versus after hybrid selection. Enrichment was variable but nearly universal. (**b**) Fold enrichment of parasite DNA, as compared to the percentage of alignable reads before hybrid selection. Hybrid selection resulted in the greatest fold enrichment for samples containing the lowest amount of parasite DNA before hybrid selection.

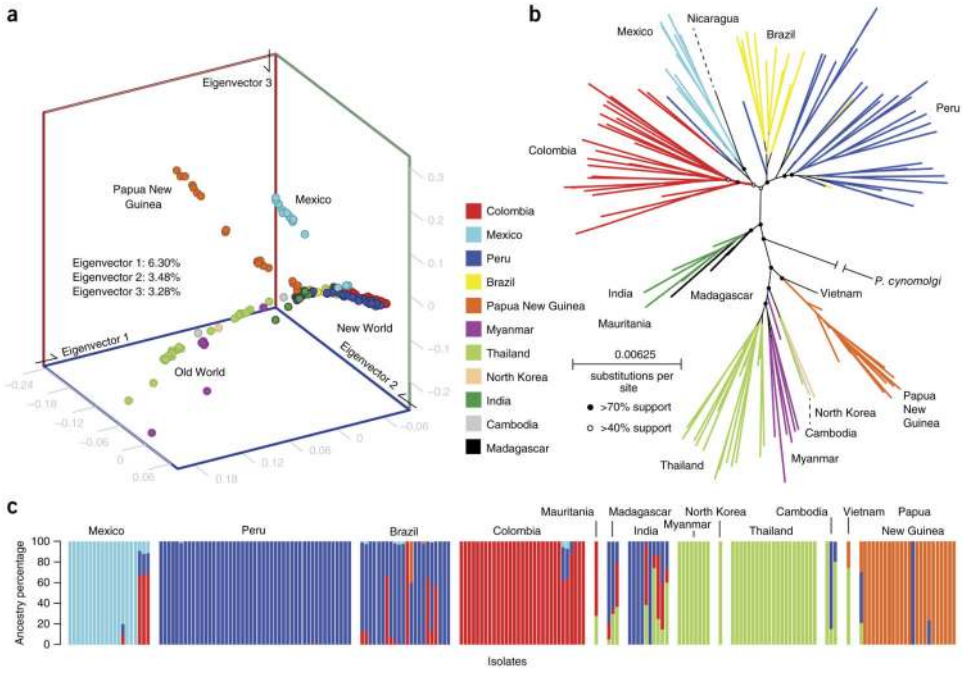

Figure 3.

Global population structure of *P. vivax*. (**a**) Three-eigenvector PCA using variation data from all *P. vivax* isolates. Each isolate is colored according to its origin, and the percentage contribution of each eigenvector is indicated. (**b**) A maximum-likelihood phylogenetic tree with 50 bootstraps computed through alignment to the Salvador 1 reference genome. Colors and labels correspond to those in **a**, and the sister taxon to *P. vivax*, *P. cynomolgi*, was included as an outgroup. Select internal nodes are highlighted with open and closed circles, representing >40% and >70% bootstrap support, respectively. (**c**) Admixture analysis of the full set of variation data from 195 isolates using 100 bootstraps. Ancestry for each isolate was assessed according to an optimized cluster value of $K = 5$ (see supplementary Fig. 7 for admixture plots under six different K cluster values). Colors correspond to the subpopulations seen in **a** and **b** and were assigned to the five most deeply sampled subpopulations of *P. vivax*.

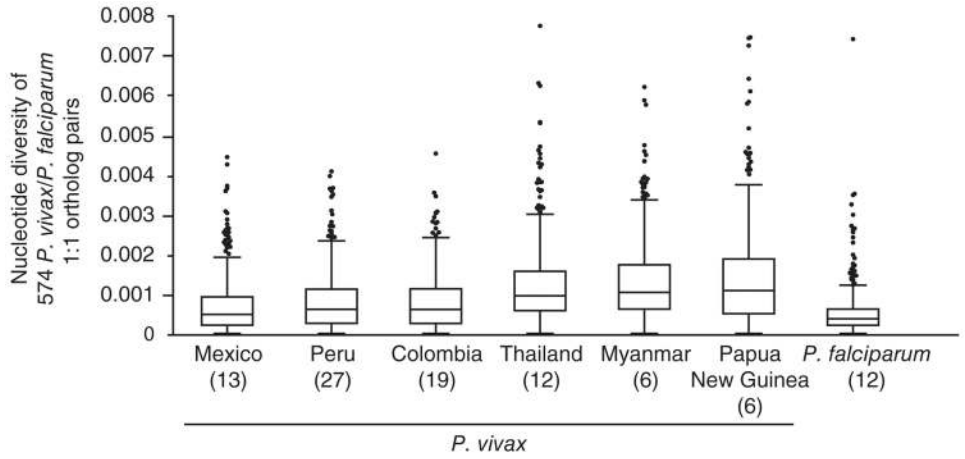

Figure 4.

Diversity in orthologs across six *P. vivax* subpopulations and *P. falciparum*. Box-and-whisker plots of nucleotide diversity (π) calculated from 574 one-to-one orthologs in single-infection, high-quality isolates from Mexico, Colombia, Peru, Thailand, Myanmar and Papua New Guinea subpopulations are shown. The center and ends of the boxes represent the 25th, 50th and 75th percentiles of π values for each subpopulation, and the whiskers on each plot represent the range of π within 1.5 times the interquartile range of the lower and upper quartiles. Data points beyond the range of the whiskers are plotted as black dots. For *P. falciparum*, π was calculated from orthologs in a population of 12 African isolates previously sequenced at the Broad Institute (see URLs). *P. falciparum* was significantly less diverse than all plotted *P. vivax* populations ($P < 0.001$) using both a paired Student's *t* test and a Wilcoxon signed-rank test.

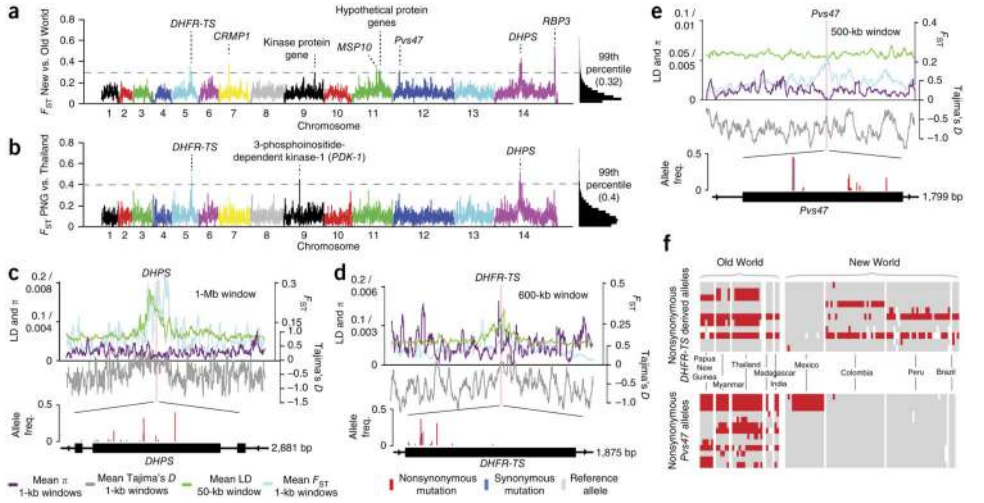

Figure 5.

Signals of selection in *P. vivax* subpopulations. (**a**) Genome-wide view of the divergence (F_{ST}) between Old World and New World populations calculated from 1-kb windows. Plotted is a moving average across each of the 14 chromosomes, with a histogram of F_{ST} values on the right. The 99th percentile of the distribution of 1-kb windows forms a cutoff for outliers, which are labeled with the nearest gene annotation to the peak F_{ST}. (**b**) F_{ST} comparison as in **a** between the Thailand and Papua New Guinea populations. (**c–e**) Detailed plots for *DHPS* (**c**), *DHFR-TS* (**d**) and *Pvs47* (**e**), overlaying F_{ST} values from **a** and **b**, the collapsed estimate of mean linkage disequilibrium (LD) calculated across a 50-kb window for each nucleotide, a moving average of π in 1-kb windows for isolates from Southeast Asia and Tajima's *D* in 1-kb windows. The location of each gene is shown (pink bar) above a schematic of the gene. Homozygous variants within coding sequences are indicated, including nonsynonymous changes, synonymous changes and deletions. Polymorphism columns are scaled by their allele frequency in the global population. (**f**) Haplotype map for the *P. vivax* genes *DHFR-TS* and *Pvs47*. Red blocks denote nonsynonymous changes, gray blocks indicate the presence of the reference allele and white blocks show regions of missing data. Nonsynonymous sites in *DHFR-TS* are limited to derived alleles.

Table 1

Notable *P. vivax* genes with their associated population genetic statistics

Gene name (identifier)	MK test *z* score	New World/Old World F_{ST} *z* score	Diversity (π)	Function
Mosquito transmission				
TRAP-like protein (*TLP*; PVX_113965)	−0.723	6.82	0.001	Important for cell traversal of invasive sporozoite stage[25]
Cysteine-repeat modular protein-1 (*CRMP1*; PVX_099005)	−0.164	7.27	0.0016	Expressed in both vertebrate and mosquito hosts for host tissue targeting and invasion[26]
Red blood cell invasion				
Merozoite-adhesive erythrocytic binding protein (*MAEBL*; PVX_092975)	−3.958	4.04	0.0011	Erythrocyte-binding-like protein involved in host receptor recognition in *P. falciparum*
Merozoite surface protein 10 (*MSP10*; PVX_114145)	1.178	6.10	0.0019	Single-copy gene with high human antibody reactivity[27]; possible vaccine candidate[28]
Drug resistance				
Phosphoinositide-dependent protein kinase 1 (*PDK1*; PVX_091715)	0.0518	−0.109	0.0009	Associated with artemisinin resistance pathway in *P. falciparum*[29]; exhibited strong divergence between PNG and Old World *P. vivax* in this study
Phosphatidylinositol 3 kinase (*PI3K*; PVX_080480)	0.899	0.537	0.0009	Associated with artemisinin resistance pathway in *P. falciparum*[29]; showed high polymorphism in our global set of *P. vivax* isolates
Multidrug resistance 1 gene (*MDR1*; PVX_080100)	−3.150	3.41	0.001	Implicated in *P. vivax* drug resistance[30]; exhibited high divergence in this study
Chloroquine resistance transporter (*CRT*; PVX_087980)	−0.101	1.61	0.0009	Ortholog in *P. falciparum* implicated in chloroquine resistance[31]; exhibited high polymorphism in some *P. vivax* Thailand samples in this study
Kelch12 propeller domain gene (*K12*; PVX_083080)	−1.486	1.65	<0.0001	Ortholog in *P. falciparum* implicated in artemisinin resistance[32]; highly conserved in our global set of *P. vivax* isolates

Some of the genes listed were identified as potentially adapting to regional differences in the human host or mosquito vector or due to the influence of other environmental variables because of a significant McDonald–Kreitman (MK) test, F_{ST} score or both. McDonald–Kreitman test results and the New World/Old World F_{ST} values for each gene are shown as *z* scores, and putative functions are given for the genes in *P. falciparum* or *P. vivax* where known. The *P. vivax* kelch12 propeller domain gene, the ortholog of an important drug resistance locus in *P. falciparum*, did not appear to be under selective pressure in our global set of *P. vivax* isolates and was also included.