

The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*

Daniel E Neafsey¹, Kevin Galinsky¹, Rays H Y Jiang¹, Lauren Young¹, Sean M Sykes¹, Sakina Saif¹, Sharvari Gujja¹, Jonathan M Goldberg¹, Sarah Young¹, Qiandong Zeng¹, Sinéad B Chapman¹, Aditya P Dash^{2,6}, Anupkumar R Anvikar², Patrick L Sutton³, Bruce W Birren¹, Ananias A Escalante⁴, John W Barnwell⁵ & Jane M Carlton³

We sequenced and annotated the genomes of four *P. vivax* strains collected from disparate geographic locations, tripling the number of genome sequences available for this understudied parasite and providing the first genome-wide perspective of global variability in this species. We observe approximately twice as much SNP diversity among these isolates as we do among a comparable collection of isolates of *P. falciparum*, a malaria-causing parasite that results in higher mortality. This indicates a distinct history of global colonization and/or a more stable demographic history for *P. vivax* relative to *P. falciparum*, which is thought to have undergone a recent population bottleneck. The SNP diversity, as well as additional microsatellite and gene family variability, suggests a capacity for greater functional variation in the global population of *P. vivax*. These findings warrant a deeper survey of variation in *P. vivax* to equip disease interventions targeting the distinctive biology of this neglected but major pathogen.

Half the world's population is estimated to be at risk for malaria caused by *P. vivax*¹, owing to this parasite's unique potential for lengthy remission and tolerance of cooler climates than those preferred by strictly tropical *Plasmodium* species. Although the *P. falciparum* parasite is responsible for the majority of contemporary malaria-related mortality, there is evidence that *P. vivax* may have been a more virulent parasite before the advent of modern medicine. As far north as England, death records indicate that *P. vivax* likely reduced the average lifespan from 58 to 33 years during the nineteenth century². More recently, studies have shown that *P. vivax* is capable of causing the severe malaria syndromes that have long been attributed only to *P. falciparum*³.

In spite of the past and present impact of *P. vivax* on human health, it remains chronically understudied relative to *P. falciparum*. The ability to continuously culture *P. falciparum* but not *P. vivax* in the laboratory, compounded with the differential mortality imposed by

the two species, has led to vast discrepancies in the state of knowledge regarding almost all aspects of the biology of these species. The recently renewed push for malaria eradication may remain incomplete unless this disparity in knowledge is addressed and the findings of new studies applied to disease control programs⁴.

Genetic diversity is important to characterize in order to understand the history of human association with a disease, to evaluate the direct effects of diversity on clinical disease and also because it may directly or indirectly reduce the efficacy of therapeutics such as drugs and vaccines. In contrast to *P. falciparum*, for which the genomes of many hundreds of isolates have now been sequenced or genotyped⁵, only the *P. vivax* genomic reference strain (Salvador I)⁶ and unassembled shotgun sequencing of a Peruvian isolate (IQ07)⁷ have been completed. We sequenced, assembled and annotated the genomes of four geographically disparate isolates of *P. vivax* to remedy the lack of genetic diversity data available for this species relative to *P. falciparum*. The designations and geographic origins of each *P. vivax* strain are given in **Table 1**, which also lists the names and origins of a concurrently sequenced comparator panel of *P. falciparum* isolates that came from similarly disparate geographic locales. Templates of both species were sequenced using the same next-generation sequencing platform (Illumina Genome Analyzer IIx and HiSeq 2000) and evaluated for diversity using the same bioinformatic tools. All four of the newly sequenced *P. vivax* reference strains (North Korean, India VII, Mauritania I and Brazil I) are clonal infections adapted for growth in monkeys, are publically available via the Malaria Research and Reference Reagent Resource Center (MR4; see URLs) and were sequenced from genomic DNA derived from leukocyte-depleted monkey blood (Online Methods). The *P. falciparum* isolates (from Honduras, India, Indochina (Thailand and Laos) and Senegal) were also clonal and were sequenced using template derived from *in vitro* cultures. We generated *de novo* assemblies for each of the four *P. vivax* isolates (**Table 2**) using the ALLPATHS LG assembly algorithm⁸. Assembly quality was substantially higher for *P. vivax*

¹Broad Institute, Cambridge, Massachusetts, USA. ²National Institute of Malaria Research, Indian Council of Medical Research, Dwarka, New Delhi, India.

³Department of Biology, Center for Genomics and Systems Biology, New York University, New York, New York, USA. ⁴Center for Evolutionary Medicine & Informatics, The Biodesign Institute, School of Life Sciences, Arizona State University, Tempe, Arizona, USA. ⁵Center for Global Health, Centers for Disease Control and Prevention, Division of Parasitic Diseases and Malaria, Atlanta, Georgia, USA. ⁶Present address: World Health House, New Delhi, India. Correspondence should be addressed to J.M.C. (jane.carlton@nyu.edu).

Received 24 January; accepted 9 July; published online 5 August 2012; doi:10.1038/ng.2373

Table 1 Strains and isolates of *P. vivax* and *P. falciparum* used in this study

Geographic origin	<i>P. vivax</i> strains and isolates	<i>P. falciparum</i> strains
Latin America	Salvador I (El Salvador) (ref. 6) Brazil I (ref. 23) IQ07 (Peru) (ref. 7)	HB3 (Honduras) (ref. 24)
South Asia (India)	India VII (ref. 25)	ML-14 87_239
East Asia	North Korean (ref. 26)	Dd2 (Indochina: Thailand and Laos) (ref. 27)
Africa	Mauritania I (ref. 28)	3D7 (ref. 29) Th231.08 (Senegal) (ref. 30)

Information in parentheses give more specific geographic origins, where known.

than for a *P. falciparum* assembly generated using the same approach (Supplementary Table 1), likely due to the more moderate AT nucleotide composition of the *P. vivax* genome (*P. vivax*: 57.7% AT versus *P. falciparum*: 80.6% AT). Synteny between the new *P. vivax* assemblies and the Salvador I reference assembly was found to be highly conserved ($\geq 97.9\%$; Supplementary Fig. 1).

We next used the sequencing data to evaluate genetic diversity within each species. The pairwise SNP rate for each sequenced isolate relative to the reference assembly for each species (*P. vivax*, Salvador I; *P. falciparum*, 3D7) is given in Figure 1a as a function of inferred SNP quality. Pairwise SNP rate relative to a reference assembly is expected to be a function of evolutionary or geographic distance, but, despite the varying geographic origin of isolates from both species, we found that the *P. vivax* SNP rates were uniformly higher than the *P. falciparum* SNP rates, regardless of the SNP quality threshold used. This finding suggests globally higher genetic diversity in *P. vivax* relative to *P. falciparum*, a genome-wide result confirming previous pilot surveys^{9,10}. We explored this finding by comparing mean SNP rates according to sequence class, using SNPs with a minimum PHRED-style quality score of 30 (estimated accuracy of at least 99.9%)¹¹. We found that *P. vivax* exhibited a significantly higher mean SNP rate than *P. falciparum* at intergenic and intronic fourfold-degenerate synonymous coding sites, within coding sequences overall and across all sequence classes (*t* test, $P = 0.0087$; Fig. 1b), suggesting that the difference in diversity is pervasive and occurs across the genome. To control for variation in the degree of functional constraint among genes within each genome, we next evaluated mean pairwise SNP diversity (π) in a collection of 3,401 genes for which we could confidently identify 1:1 orthologs in the two species using the reciprocal best BLAST hits (RBH) criterion. In this comparison, we observed approximately twice as much SNP diversity in *P. vivax* compared to *P. falciparum* (paired *t* test, $P = 2.2 \times 10^{-16}$; Fig. 1c), confirming the ubiquity and magnitude of the SNP diversity disparity between the species (Supplementary Fig. 2 and Supplementary Table 2).

To test whether a differing SNP mutation rate, rather than a different effective population size and/or demographic history, can account for the differences in SNP diversity observed, we compared the genome-wide diversity of microsatellites, a different class of mutation. Because microsatellite length variants are caused by replication

slippage rather than point substitution¹², we would expect their relative diversity profile to be different from that of SNPs under the null hypothesis that a different point mutation rate explains the disparity in SNP diversity. We applied a new method of evaluating microsatellite length variation from Illumina data and observed, as we had with the SNPs, significantly higher diversity in *P. vivax* than in *P. falciparum* (bootstrapping, 279 and 22,713 microsatellites with at least 8 repeats in *P. vivax* and *P. falciparum*, respectively; $P = 0.021$; Fig. 1d and Supplementary Fig. 3). Given the population genetic evidence that *P. falciparum* underwent multiple drug-induced selective sweeps and at least one significant bottleneck¹³, these results indicate that *P. vivax* may have a comparatively large effective population size due to an absence of such demographic events in recent history. Even in the face of common drug pressure, *P. vivax* may have disproportionate demographic stability due to its unique capacity for dormancy within infected hosts.

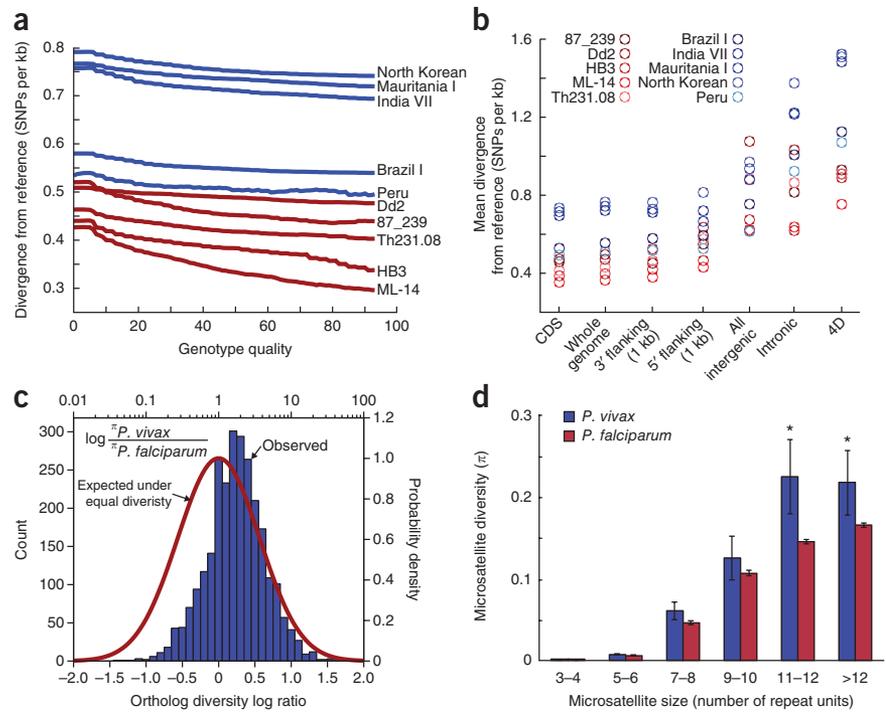
Our sample is smaller than is ideal for evaluating the time to the most recent common ancestor (TMRCA), but by comparing the deepest pairwise nucleotide divergence observed for each species at fourfold-degenerate synonymous sites (*P. vivax*: Mauritania versus Brazil I, 1.628×10^{-3} substitutions per site; *P. falciparum*: 3D7 versus Dd2, 9.59×10^{-4} substitutions per site) we can predict that the lower bound for TMRCA in *P. vivax* is approximately 70% greater than that for *P. falciparum*. The calculation of absolute TMRCA dates is dependent on the accuracy of the inferred mutation rate, and, therefore, results must be interpreted with caution. Nevertheless, if we assume a commonly accepted eukaryotic genome-wide mutation rate for fourfold-degenerate sites (2.2×10^{-9} substitutions per site per year¹⁴, similar to a *Plasmodium* rate estimated with less precision¹⁵), we can estimate the TMRCA as 768,000 and 452,000 years ago for *P. vivax* and *P. falciparum*, respectively. This *P. vivax* estimate is deeper than previous TMRCA estimates generated using a small number of loci in a larger population sample¹⁶, but reconciliation of the absolute estimates is difficult without knowledge of the true mutation rate. Assuming mutation rates are similar in both parasite lineages, however, the result stands that *P. vivax* has a much deeper TMRCA.

Other departures in the global population history of these two species are indicated by the topology and branch lengths of their respective phylogenies (Fig. 2). The relatively large degree of divergence between the IQ07 Peruvian isolate and the Brazil I and Salvador I strains of *P. vivax* suggests a distinct history in the New World relative to *P. falciparum*, which exhibits low diversity in the New World and is thought to have been introduced within the last 500 years via the African slave trade¹⁷. The high New World diversity of *P. vivax*, combined with the closer phylogenetic affinity of the three New World *P. vivax* isolates with the east Asian (North Korean) rather than the African (Mauritania I) or south Asian (India VII) strains, could suggest the precolonial arrival of *P. vivax* in the New World accompanying human dispersal from Asia by sea or, less likely, by the Bering land bridge during the last glacial maximum. Alternatively, this profile could be explained by recent but very large-scale (relative to *P. falciparum*) postcolonial introductions to the New World

Table 2 Assembly statistics of four *P. vivax* reference strains sequenced using Illumina technology

Strain	Assembly size (Mb)	Fold coverage	Contig N50 (kb)	Number of contigs	Scaffold N50 (kb)	Number of scaffolds	Coverage of Salvador I reference (%)
Brazil I	28.87	68.5	28.2	1,999	885.6	260	98.0
India VII	29.25	35.0	21.2	3,358	594.6	568	98.1
Mauritania I	28.43	91.1	39.4	1,510	945.1	205	97.9
North Korean	29.65	87.6	22.1	2,499	317.6	541	98.8

Figure 1 Disparity in SNP and microsatellite diversity between *P. vivax* and *P. falciparum*. (a) Quality score versus pairwise whole-genome SNP divergence rates compared to reference assemblies. Blue lines represent *P. vivax* isolates, and red lines represent *P. falciparum* isolates. (b) *P. falciparum* (red shades) versus *P. vivax* (blue shades) Q30 call rates for: coding sequences (CDS), whole-genome, 3' flanking sequence, 5' flanking sequence, all intergenic sequences, introns and fourfold-degenerate synonymous coding sites (4D). (c) Density distribution of *P. falciparum*-*P. vivax* diversity log ratios for genes with 1:1 orthologs compared to the null expected distribution centered on 1. (d) Histogram of microsatellite diversity in microsatellite loci with a repeat unit size of 2 bp. Error bars indicate standard errors. Asterisks indicate size bins for which *P. vivax* is significantly more diverse than *P. falciparum* (bootstrapping, $P < 0.05$).



via transoceanic trade from east Asia, the Pacific or Europe, the last of which presumably harbored a genetically distinct parasite clade before disease elimination¹⁸. Deeper genomic sampling of *P. vivax* populations will be required to explain these patterns in the diversity data.

We next explored the profile of variation in individual genes and gene families to evaluate the potential functional consequences of the extremely high genomic diversity we observe in *P. vivax*. Mean pairwise divergence among the sequenced *P. vivax* isolates is highest in gene families associated with red blood cell invasion and immune evasion (Fig. 3a). Functional enrichment analysis of diversity in individual genes also found nonsynonymous SNPs to be concentrated in invasion-related motility genes (Supplementary Table 3). This extremely high sequence diversity suggests that vaccines targeting polymorphic antigens may encounter an even greater hurdle in eliciting cross-protective immune responses than they do in *P. falciparum*, where strain-specific immunity has been recently observed to limit vaccine efficacy¹⁹.

Differences in the distribution of nonsynonymous SNPs among genes with orthologs in both species are potentially reflective of differences in disease biology in *P. vivax* and *P. falciparum* (Supplementary Table 4). Genes expressed in the pre-erythrocytic stages are the most enriched group among those with higher ratios of nonsynonymous-to-synonymous diversity (π_{NS}/π_S) in *P. vivax* relative to *P. falciparum* (Mann-Whitney U test; Z score = 2.2), whereas genes associated with

host-parasite interactions are the most enriched group with higher π_{NS}/π_S in *P. falciparum* relative to *P. vivax* (Mann-Whitney U test; Z score = 2.4) (Supplementary Table 5). Although neither of these enrichments was statistically significant after correction for multiple testing, this pattern bears further exploration when sequence data from more *P. vivax* genomes become available.

As expected, we observed enormous diversity in the *vir* gene family, the members of which are variably expressed and encode proteins that are exported to the host cell surface for the purpose of evading the host adaptive immune response²⁰. Of the 313 *vir* genes included in the Salvador I reference assembly, less than a third are also observed in the four new assemblies (Fig. 3b). Unexpectedly, we encountered 15 'ultra-conserved' *vir* genes that were present in all assemblies and showed very low SNP diversity; in particular, one locus (PVX_113230) was invariant and showed the highest similarity (70% protein sequence identity) of any *vir* to the homologous *kir* gene family in *Plasmodium knowlesi*, a zoonotic parasite (Fig. 3c). Unlike most *vir* genes, this locus also exhibits conserved synteny in more distantly related rodent malaria parasites. These attributes suggest that PVX_113230 is likely the founder of the *vir* family in the *P. vivax* lineage, and the lack of polymorphism suggests that the protein it encodes performs an ancestral role rather than host immune modulation. The molecular function of PVX_113230 could be related to erythrocyte invasion, as suggested by its distinct expression profile in erythrocytes relative to most *vir* genes²¹ (Supplementary Fig. 4).

This global census of genomic diversity in *P. vivax* has uncovered an unexpected degree of genetic polymorphism, much of which may

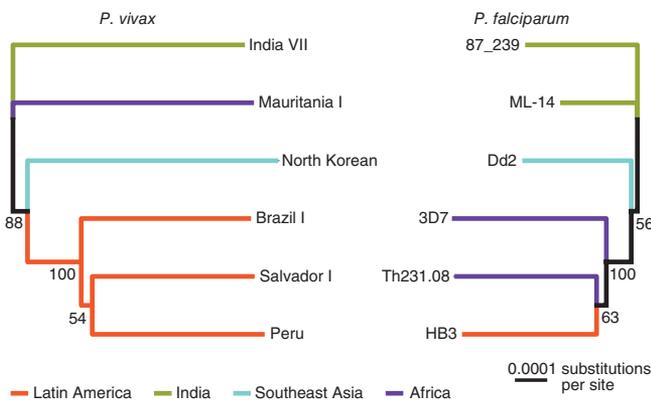
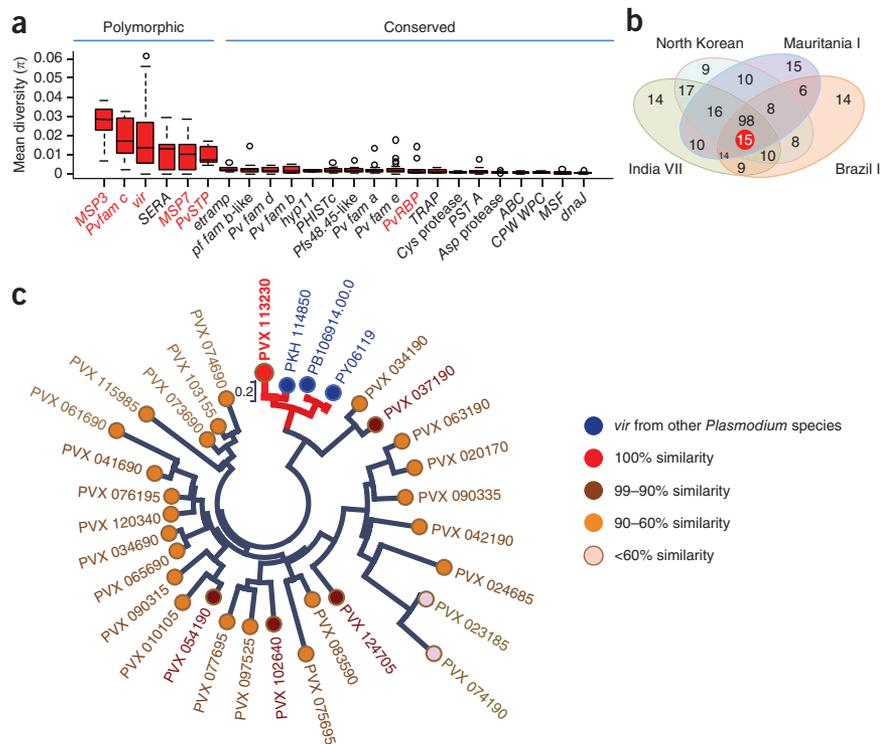


Figure 2 Neighbor-joining phylograms of *P. vivax* and *P. falciparum* constructed from presumably neutral SNPs occurring in fourfold-degenerate coding sites. Lineages are colored according to geographic origin. Branch lengths indicate considerable diversity in New World *P. vivax* strains and show no clear affiliation between New World and African *P. vivax* strains. Phylograms were constructed from 471,543 sites in *P. vivax* and 359,901 sites in *P. falciparum*. Numbers at nodes indicate percentages of bootstrap support.

Figure 3 Diversity of *P. vivax* gene families.

(a) Mean pairwise SNP diversity in *P. vivax* gene families. Gene families associated with merozoite invasion or immune response modulation (red text) show the highest diversity. Red bars on the box plots represent the 25th to 75th percentile range, and circles indicate outlier genes. (b) Limited overlapping *vir* repertoires of *P. vivax* isolates. *vir* genes exhibiting at least 70% sequence identity between isolates were included in the Venn diagram. A set of 15 ultra-conserved *vir* genes with more than 95% similarity in all comparisons are included in the central red circle. (c) A neighbor-joining phylogenetic tree of ultra-conserved *vir* genes and related paralogs from the *vir12* and *vir14* subfamilies. The most highly conserved *vir*, PVX_113230, has clear orthologs in other *Plasmodium* species.



translate into important functional variation. Our data stop short of suggesting the existence of distinct subspecies of *P. vivax*, similar to the subspecies *P. vivax vivax* and *P. vivax hibernans* proposed on the basis of relapse phenotype²². However, the extreme diversity we observe among these new reference strains suggests a more stable and older association of this parasite with humans than for *P. falciparum*, serving as a warning that the eradication of *P. vivax* could represent a qualitatively different task.

URLs. Malaria Research and Reference Reagent Resource Center (MR4): <http://www.mr4.org/>; TransposonPSI: <http://transposonpsi.sourceforge.net/>; *P. vivax* nuclear and apicoplast genome assemblies and gene calls and Sanger sequencing-based Dd2 assembly, http://www.broadinstitute.org/annotation/genome/plasmodium_vivax/MultiDownloads.html; PlasmoDB, <http://plasmodb.org/>.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Illumina sequencing reads have been submitted to the NCBI Short Read Archive (SRA; accessions SRP007883, SRP007923, SRP000493 and SRP000316). SNP calls have been submitted to dbSNP and can also be downloaded from the Broad Institute website (see URLs). *P. vivax* nuclear and apicoplast genome assemblies and gene calls may be downloaded from GenBank (accessions AFMK01000000, AFBK01000000, AFNI01000000, AFNJ01000000, JQ437257, JQ437258 and JQ437259) and the Broad Institute website (see URLs).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This work has been funded, in whole or in part, by federal funds from the US National Institute of Allergy and Infectious Diseases (NIAID), the US National Institutes of Health (NIH) and the US Department of Health and Human Services, under contracts HHSN266200400001C and HHSN2722009000018C. We gratefully acknowledge the Indian Council of Medical Research for financial support of the Malaria Parasite Bank at the National Institute of Malaria Research, New Delhi, and the NIMR Director for providing all facilities. We thank the NIAID/National Human Genome Research Institute (NHGRI) Eukaryotic Pathogens and Disease Vectors Working Group and the Broad Institute Genome Sequencing Platform for

significant contributions to this project. K.G. and L.Y. are supported by a Global Health Program grant from the Bill and Melinda Gates Foundation. A.A.E. is supported by a grant from the NIH (RO1GM084320), and J.M.C. and P.L.S. are supported by grant U19AI089676, an NIAID International Center of Excellence for Malaria Research. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

AUTHOR CONTRIBUTIONS

J.W.B. provided *P. vivax* strains, and A.R.A. and A.P.D. provided Indian *P. falciparum* material. S.M.S., S.S. and S.Y. performed genome assembly. S.G., J.M.G. and Q.Z. performed genome annotation. K.G., R.H.Y.J., L.Y. and D.E.N. performed analyses. S.B.C. performed project management. P.L.S. undertook experimental validation. D.E.N., A.A.E., B.W.B. and J.M.C. directed the analyses. D.E.N., A.A.E., J.W.B. and J.M.C. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/doi/10.1038/ng.2373>.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Guerra, C.A. *et al.* The international limits and population at risk of *Plasmodium vivax* transmission in 2009. *PLoS Negl. Trop. Dis.* **4**, e774 (2010).
- Dobson, M.J. Malaria in England: a geographical and historical perspective. *Parassitologia* **36**, 35–60 (1994).
- Price, R.N., Douglas, N.M. & Anstey, N.M. New developments in *Plasmodium vivax* malaria: severe disease and the rise of chloroquine resistance. *Curr. Opin. Infect. Dis.* **22**, 430–435 (2009).
- Carlton, J.M., Sina, B.J. & Adams, J.H. Why is *Plasmodium vivax* a neglected tropical disease? *PLoS Negl. Trop. Dis.* **5**, e1160 (2011).
- Winzler, E.A. Malaria research in the post-genomic era. *Nature* **455**, 751–756 (2008).
- Carlton, J.M. *et al.* Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* **455**, 757–763 (2008).
- Dharia, N.V. *et al.* Whole-genome sequencing and microarray analysis of *ex vivo Plasmodium vivax* reveal selective pressure on putative drug resistance genes. *Proc. Natl. Acad. Sci. USA* **107**, 20045–20050 (2010).
- Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **108**, 1513–1518 (2011).
- Feng, X. *et al.* Single-nucleotide polymorphisms and genome diversity in *Plasmodium vivax*. *Proc. Natl. Acad. Sci. USA* **100**, 8502–8507 (2003).

10. Mu, J. *et al.* Host switch leads to emergence of *Plasmodium vivax* malaria in humans. *Mol. Biol. Evol.* **22**, 1686–1693 (2005).
11. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
12. Gemayel, R., Vences, M.D., Legendre, M. & Verstrepen, K.J. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* **44**, 445–477 (2010).
13. Joy, D.A., Mu, J., Jiang, H. & Su, X. Genetic diversity and population history of *Plasmodium falciparum* and *Plasmodium vivax*. *Parassitologia* **48**, 561–566 (2006).
14. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci. USA* **99**, 803–808 (2002).
15. Paget-McNicol, S. & Saul, A. Mutation rates in the dihydrofolate reductase gene of *Plasmodium falciparum*. *Parasitology* **122**, 497–505 (2001).
16. Cornejo, O.E. & Escalante, A.A. The origin and age of *Plasmodium vivax*. *Trends Parasitol.* **22**, 558–563 (2006).
17. Conway, D.J. *et al.* Origin of *Plasmodium falciparum* malaria is traced by mitochondrial DNA. *Mol. Biochem. Parasitol.* **111**, 163–171 (2000).
18. Carter, R. Speculations on the origins of *Plasmodium vivax* malaria. *Trends Parasitol.* **19**, 214–219 (2003).
19. Thera, M.A. *et al.* A field trial to assess a blood-stage malaria vaccine. *N. Engl. J. Med.* **365**, 1004–1013 (2011).
20. Fernandez-Becerra, C. *et al.* *Plasmodium vivax* and the importance of the subtelomeric multigene *vir* superfamily. *Trends Parasitol.* **25**, 44–51 (2009).
21. Bozdech, Z. *et al.* The transcriptome of *Plasmodium vivax* reveals divergence and diversity of transcriptional regulation in malaria parasites. *Proc. Natl. Acad. Sci. USA* **105**, 16290–16295 (2008).
22. Garnham, P.C. *et al.* A strain of *Plasmodium vivax* characterized by prolonged incubation: morphological and biological characteristics. *Bull. World Health Organ.* **52**, 21–32 (1975).
23. Nayar, J.K. *et al.* Studies on a primaquine-tolerant strain of *Plasmodium vivax* from Brazil in Aotus and Saimiri monkeys. *J. Parasitol.* **83**, 739–745 (1997).
24. Bhasin, V.K. & Trager, W. Gametocyte-forming and non-gametocyte-forming clones of *Plasmodium falciparum*. *Am. J. Trop. Med. Hyg.* **33**, 534–537 (1984).
25. Sullivan, J.S. *et al.* Adaptation of a strain of *Plasmodium vivax* from India to New World monkeys, chimpanzees, and anopheline mosquitoes. *J. Parasitol.* **87**, 1398–1403 (2001).
26. Collins, W.E. *et al.* Studies on the North Korean strain of *Plasmodium vivax* in Aotus monkeys and different anophelines. *J. Parasitol.* **71**, 20–27 (1985).
27. Wellems, T.E. *et al.* Chromosome size variation occurs in cloned *Plasmodium falciparum* on *in vitro* cultivation. *Rev. Bras. Genet.* **11**, 813–825 (1988).
28. Collins, W.E. *et al.* Adaptation of a strain of *Plasmodium vivax* from Mauritania to New World monkeys and anopheline mosquitoes. *J. Parasitol.* **84**, 619–621 (1998).
29. Gardner, M.J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).
30. Melnikov, A. *et al.* Hybrid selection for sequencing pathogen genomes from clinical samples. *Genome Biol.* **12**, R73 (2011).

ONLINE METHODS

Parasite material. We chose four strains of *P. vivax* for whole-genome sequencing on the basis of geographic origin and phenotype to provide a resource of high-quality assembled and annotated sequences for the malaria research community. The North Korean strain has a long relapse phenotype that enables it to survive in the primate host through periods of drought and long winters when mosquito vectors for transmission are unavailable²⁶. The Brazil I strain is highly resistant to the anti-relapse drug primaquine²³. The India VII strain is the first *Plasmodium* species to be sequenced from India. The Mauritania I strain is a rare example of an African *P. vivax* strain that occurs among west Africans with Berber and Arab genetic backgrounds²⁸. Genomic DNA for *P. vivax* sequencing was obtained from the leukocyte-depleted blood of infected, splenectomized *Saimiri* monkeys as described previously⁶. Genome-wide fragment analysis of 15 microsatellites before sequencing confirmed their independent origins. DNA and frozen stabilates are available upon request at MR4 (see URLs).

Genome sequencing, assembly and annotation. Genomic DNA was used to construct two Illumina sequencing libraries for each *P. vivax* isolate, with library insert sizes of 180 bp and 3 kb. Each library was sequenced to a depth of at least 150-fold coverage (to account for contaminating monkey host DNA) using 76-bp paired-end Illumina reads on an Illumina Genome Analyzer Ix sequencer. After filtering out reads with sequence similarity to monkey or primate sequence, each genome was assembled using the ALLPATHS-LG algorithm⁸. Assembly quality was quantified using the N50 statistic for contigs and scaffolds, which describes the minimum contig or scaffold size such that the sum of the lengths of all contigs or scaffolds of equal or greater size accounts for at least half of the total assembly length. Synteny with the *P. vivax* Salvador I reference assembly was quantified for each *de novo* assembly by evaluating the concordance of approximately 60,000 randomly chosen pairs of 100-bp sequences, each with 100 kb of intervening sequence in the new assemblies. This method identifies breaks in synteny when the sequence pairs do not map to locations separated by 100 kb in the reference assembly or when only one member of a sequence pair is successfully mapped to the reference. Finally, for three of the four *P. vivax* isolates (North Korean, Mauritania I and Brazil I), we were able to recover a scaffold representing the complete apicoplast genome.

P. falciparum analyses were based on comparable Illumina read data generated from a single library for each isolate (insert size of 180 bp). Coverage for each of the isolates was as follows: Indian isolate 87_239, 42×; Indochina isolate Dd2, 196×; Honduran isolate HB3, 30×; Indian isolate ML-14, 41×; Senegalese isolate Th231.08, 59×.

The protein-coding genes in the *P. vivax* nuclear genomes were annotated using a combination of reference gene mapping, homology-based gene models (GeneWise)³¹, EST-based gene models and *ab initio* gene predictions. Ribosomal RNAs (rRNAs) were identified with RNAmmer³². The tRNA features were identified using tRNAScanSE³³. Other common RNA features were identified with RFAM³⁴.

The MUMmer algorithm³⁵ was used to align draft assemblies to the *P. vivax* Salvador I reference genome assembly. Neighboring syntenic alignment blocks were joined to form longer alignments, which were then used to map the gene coordinates from the reference genome to the draft assemblies. Homology-based gene models were created using tblastn to search against the draft genome assemblies with the UniRef90 protein database, a *Plasmodium* protein database created from the annotated proteins of *P. falciparum* 3D7 and *P. vivax* Salvador I. The resultant BLAST hits were used to create GeneWise gene models. Gene models were also built using 31,777 *P. vivax* ESTs available on GenBank, with an ORF cutoff length of 300 bp (EST ORFs). *Ab initio* gene models were predicted using self-training GeneMark-ES³⁶ and GeneId³⁷ with parameters trained on genes from the *P. vivax* Salvador I genome.

The final merged gene set for each of the four sequenced *P. vivax* strains was created using the following workflow. If a mapped reference gene in a given assembly had intact start and stop codons, no frameshift or in-frame stop and no exons in the contig gaps, then the mapped reference gene was used directly as the gene model. If a mapped gene had a frameshift or in-frame stop, then the corresponding GeneMarkES gene model was selected. If a GeneWise gene model had no overlap with gene models from the previous two sources but had

a non-generic gene product name or overlap with non-repeat PFAM domains, then the GeneWise feature was added as a gene model. Finally, if an EST ORF was at least 600 bp in length and exhibited no overlap with models identified from the previous three sources, then the EST ORF was added as a gene model. The initial gene set was compared against tRNA and rRNA features, and genes were filtered out where appropriate. Additional gene filtering was performed by removing genes with 30% or more coding sequence overlap with hits in TransposonPSI (see URLs) ($e < 1 \times 10^{-10}$).

SNP calling. Sequencing reads from each isolate were aligned using the Burrows-Wheeler Aligner (BWA)³⁸ to the reference assemblies of *P. falciparum* 3D7 (build v7.1) and *P. vivax* Salvador I (build v7.0), both downloaded from PlasmoDB (see URLs). SNPs were called using the Unified Genotyper¹¹ in the GATK package³⁹. SNPs with an estimated PHRED-style quality score of Q30 or greater were used for diversity analyses.

Microsatellite variants and validation. The mreps program⁴⁰ was used to identify microsatellites in the reference assemblies of *P. falciparum* and *P. vivax*. For the microsatellite search, the mreps parameters were set as minp = 1 and maxp = 9. Searching under these conditions identified 538,794 and 95,990 microsatellite loci in *P. falciparum* and *P. vivax*, respectively. For *P. falciparum*, the *allow-small* parameter was employed to allow identification of small microsatellites that the mreps algorithm might otherwise flag as biologically insignificant due to the high AT content (~80%) of that genome.

Illumina sequencing reads were mapped to the *P. falciparum* 3D7 or *P. vivax* reference assemblies using BWA³⁸ with the -q 5 and -l 32 options. Base-quality score recalibration and local realignment around microsatellites¹¹ were applied using GATK³⁹. Custom Python scripts were used to filter out reads that did not span the entirety of a microsatellite interval in the genome. A final local realignment around microsatellite sequences, and insertion and/or deletion (indel) calling with standard hard filtering parameters was applied using GATK; accepted indels were converted into microsatellite-length genotypes. Illumina-based microsatellite calls were validated by comparison with calls made by aligning the Sanger sequence-based Dd2 assembly (see URLs) to the 3D7 reference assembly to generate a truth set of indels for genotyped microsatellites. Comparison of the Illumina-based calls to the Sanger sequencing-based calls indicated that our Illumina-based calling method exhibits a specificity of 100% and a sensitivity of 97%, on the basis of 81,569 AT dinucleotide microsatellites callable in both data sets. *P* values for evaluating the significance of the difference in microsatellite diversity between species were generated by resampling π values for each species 10,000 times and noting the frequency with which a difference in mean π occurred that was equal to or greater than the observed difference in mean π . The overall comparison of microsatellite diversity across motif unit sizes was performed using only microsatellites with eight or more repeat units, given that, as previously observed for *P. vivax*⁴¹, these longer microsatellites were observed to be more polymorphic in both species (Supplementary Fig. 5).

Evolutionary analyses and TMRCA calculation. Mean pairwise diversity at all nucleotide sites (π), synonymous sites (π_s) and nonsynonymous sites (π_n) was calculated for each gene in the reference annotations for *P. falciparum* and *P. vivax* using the method described in Table 1 of ref. 42. *P. knowlesi* orthologs were employed to calculate interspecies d_N/d_S ratios using PAML v4.5 (ref. 43) with sequences from the Salvador I isolate of *P. vivax*. A matrix of pairwise nucleotide distances between isolates (Supplementary Table 6) was constructed for each species using SNPs identified in fourfold-degenerate synonymous coding sites of genes with orthologs in each species. To control for differences in the nucleotide substitution profile between species and enable direct comparison of branch lengths, pairwise distances were normalized by empirical nucleotide transition matrices (Supplementary Table 7) constructed for each species by rooting fourfold-degenerate polymorphisms using an outgroup species (*P. knowlesi* for *P. vivax* and *P. reichenowi* for *P. falciparum*). Phylograms were constructed from pairwise distance matrices using the Neighbor-Joining method. The lower bound of the TMRCA was estimated for each species as the deepest pairwise divergence. SNP functional enrichment analyses were carried out using Mann-Whitney *U* tests, and *Z* scores were interpreted using a Bonferroni correction for multiple testing.

31. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
32. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
33. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
34. Griffiths-Jones, S. Annotating non-coding RNAs with Rfam. *Curr. Protoc. Bioinformatics* **Chapter 12**, Unit 12.5 (2005).
35. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
36. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O. & Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).
37. Blanco, E. & Abril, J.F. Computational gene annotation in new genome assemblies using GeneID. *Methods Mol. Biol.* **537**, 243–261 (2009).
38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
39. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
40. Kolpakov, R., Bana, G. & Kucherov, G. mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.* **31**, 3672–3678 (2003).
41. Russell, B., Suwanarusk, R. & Lek-Uthai, U. *Plasmodium vivax* genetic diversity: microsatellite length matters. *Trends Parasitol.* **22**, 399–401 (2006).
42. Ina, Y. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* **40**, 190–226 (1995).
43. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).