HHS Public Access
Author manuscript
Nat Genet. Author manuscript; available in PMC 2013 March 01.

Published in final edited form as:
Nat Genet. 2012 September ; 44(9): 1046–1050. doi:10.1038/ng.2373.

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

Daniel E. Neafsey[1], **Kevin Galinsky**[1], **Rays H. Y. Jiang**[1], **Lauren Young**[1], **Sean M. Sykes**[1], **Sakina Saif**[1], **Sharvari Gujja**[1], **Jonathan M. Goldberg**[1], **Sarah Young**[1], **Qiandong Zeng**[1], **Sinéad B. Chapman**[1], **Aditya P. Dash**[2], **Anupkumar R. Anvikar**[2], **Patrick L. Sutton**[3], **Bruce W. Birren**[1], **Ananias A. Escalante**[4], **John W. Barnwell**[5], and **Jane M. Carlton**[3]

[1]Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA

[2]National Institute of Malaria Research, Indian Council of Medical Research, Sector 8, Dwarka, New Delhi 110 077, India

[3]Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY 10003, USA

[4]Center for Evolutionary Medicine & Informatics, The Biodesign Institute, School of Life Sciences, Arizona State University, PO Box 874501, Tempe AZ 85287-450, USA

[5]Divison of Parasitic Diseases and Malaria, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, GA 30329, USA

Abstract

We sequenced and annotated the genomes of four *Plasmodium vivax* strains collected from disparate geographical locations, tripling the number of genome sequences available for this

Users may view, print, copy, download and text and data- mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: http://www.nature.com/authors/editorial_policies/license.html#terms

URLs
MR4: http://www.mr4.org/
TransposonPSI: http://transposonpsi.sourceforge.net/

ACCESSION NUMBERS

Illumina sequencing reads have been submitted to the NCBI short read archive (accession numbers: SRP007883, SRP007923, SRP000493, SRP000316). SNP calls have been submitted to dbSNP (handle: BROAD-GENOMEBIO; SS-IDs: 522951448-525286517; 526103049-526165183; 526166468-526189295; 526189799-526214992; 526889101-526983923; 526985108-527023853; 527024052-527114619; 527114764-527224909; 527225105-527336921). *P. vivax* nuclear and apicoplast genome assemblies and gene calls may be downloaded from the Broad Institute website (http://www.broadinstitute.org/annotation/genome/plasmodium_vivax/MultiDownloads.html) or GenBank (accession numbers: AFMK01000000, AFBK01000000, AFNI01000000, AFNJ01000000, JQ437257, JQ437258, JQ437259).

AUTHOR CONTRIBUTIONS

JWB provided *P. vivax* strains and ARA and APD provided Indian *P. falciparum* material. SMS, SS, and SY performed genome assembly. SG, JMG, and QZ performed genome annotation. KG, RJ, LY, and DEN performed analyses. SBC performed project management. PLS undertook experimental validation. DEN, AAE, BWB, and JMC directed the analyses. DEN, AAE, JWB and JMC wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Editorial Summary (AOP and Month, same)

Jane Carlton and colleagues report the genome sequencing, *de novo* assembly and annotation of four *Plasmodium vivax* reference strains from diverse geographical locations. Their cross species comparisons show that *P. vivax* has greater genetic diversity than *P. falciparum*.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

understudied parasite and providing the first genome-wide perspective of global variability within this species. We observe approximately twice as much SNP diversity among these isolates as we do among a comparable collection of isolates of *Plasmodium falciparum*, a malaria parasite that causes higher mortality. This indicates a distinct history of global colonization and/or a more stable demographic history for *P. vivax* than *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 the capacity for greater functional variation within 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 of *P. vivax* malaria[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 prior to the advent of modern medicine. As far north as England, death records indicate that *P. vivax* likely reduced the average life span from 58 to 33 years during the 19th century[2]. More recently, studies have shown *P. vivax* to be capable of causing severe malaria syndromes long attributed only to *P. falciparum*[3].

In spite of the past and present significance of *P. vivax* to 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 our 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 applied to disease control programs[4].

Genetic diversity is important to characterize in order to understand the history of our association with a disease, to evaluate the direct impacts 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*, where the genomes of many hundreds of isolates have now been sequenced or genotyped[5], 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 indicated in Table 1, which also lists the names and origins of a concurrently-sequenced comparator panel of *P. falciparum* isolates hailing from similarly disparate geographic locales. Templates of both species were sequenced using the same next-generation sequencing platform (Illumina GAIIx/HiSeq) 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 (see URLs), and were sequenced from genomic DNA derived from leukocyte-depleted monkey blood (Online methods). The *P. falciparum* isolates (from

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Honduras, India, Indochina 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 algorithm[8]. Assembly quality was significantly higher for *P. vivax* than for a *P. falciparum* assembly generated using the same approach (Supplementary Table 1), likely due to the more moderate A/T nucleotide composition of the *P. vivax* genome (*P. vivax* 57.7% A/T *vs. P. falciparum* 80.6% A/T). Synteny between the new *P. vivax* assemblies and the Salvador I reference assembly was found to be highly conserved (≥ 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 indicated 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 the evolutionary or geographic distance, but despite the varying geographic origin of isolates from both species, we observed the *P. vivax* SNP rates to be uniformly higher than the *P. falciparum* SNP rates regardless of SNP quality threshold. 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 find *P. vivax* to exhibit a significantly higher mean SNP rate than *P. falciparum* at intergenic, and intronic, fourfold-degenerate (4D) synonymous coding sites, within coding sequence overall, and across all sequence classes (t-test, P = 0.0087; Fig. 1b), suggesting that the difference in diversity is pervasive and genome-wide. 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 between the species using a reciprocal best BLAST hits (RBH) criterion. For this comparison we observed approximately twice as much SNP diversity in *P. vivax* compared to *P. falciparum* (paired t-test, P = 2.2E-16; Fig. 1c), confirming the ubiquity and magnitude of SNP diversity disparity between the species (Supplementary Fig. 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 substitution[12], we would expect their relative diversity profile to be different from that of SNPs under the null hypothesis that a different point substitution mutation rate explains the SNP diversity disparity. We applied a novel 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 eight repeats in *P. vivax* and *P. falciparum*, respectively; P = 0.021, Fig. 1d; Supplementary Fig. 3). Given the population genetic evidence that *P. falciparum* underwent multiple drug-induced selective sweeps and at least one significant bottleneck[13], these results indicate that *P. vivax* may exhibit a comparatively large effective population size due to an absence of such demographic events in recent history. Even in the face of common

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

drug pressure, *P. vivax* may exhibit disproportional demographic stability due to its unique capacity for dormancy within infected hosts.

Our sample is smaller than 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 4D synonymous sites (*P. vivax*: Mauritania vs. Brazil I, 1.628E-3 substitutions/site; *P. falciparum*: 3D7 vs. Dd2, 9.59E-4 substitutions/site) we can predict that the lower bound for TMRCA in *P. vivax* is approximately 70% larger 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 4D sites (2.2 E-9 subs/site/yr[14]; similar to a *Plasmodium* rate estimated with less precision[15]), we can estimate the TMRCA as 768 KYA (thousands of years ago) and 452 KYA for *P. vivax* and *P. falciparum*, respectively. This *P. vivax* estimate is deeper than previous TMRCA estimates produced using a small number of loci in a larger population sample[16], 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* exhibits 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 phylograms (Figure 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 (NW) relative to *P. falciparum*, which exhibits low diversity in the NW and is thought to have been introduced within the last 500 years via the African slave trade[17]. The high NW diversity, combined with the significantly closer phylogenetic affinity of the three NW *P. vivax* isolates with the East Asian (North Korean) rather than the African (Mauritania I) or South Asian (India VII) strains, could suggest the pre-colonial arrival of *P. vivax* in the NW 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*) post-colonial introductions to the NW via trans-oceanic trade from East Asia, the Pacific, or Europe, the last of which presumably harbored a genetically distinct parasite clade prior to disease elimination[18]. 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 within individual genes and gene families to evaluate the potential functional consequences of the extraordinarily high genomic diversity we observe in *P. vivax*. As Figure 3a indicates, mean pairwise divergence among the sequenced *P. vivax* isolates is highest in gene families associated with red blood cell invasion and immune evasion. Functional enrichment analysis of diversity in individual genes also finds nonsynonymous SNPs to be concentrated in invasion-related motility genes (Supplementary Table 2). 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 efficacy[19].

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Differences in the distribution of nonsynymous SNPs among genes with orthologs in both species are potentially reflective of differences in disease biology between *P. vivax* and *P. falciparum* (Supplementary Table 3). Genes expressed in the pre-erythrocytic stages are the most enriched group among those exhibiting higher ratios of nonsynonymous to synonymous diversity (πNS/πS) 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 exhibiting higher πNS/πS in *P. falciparum* relative to *P. vivax* (Mann Whitney U test; Z score = 2.4; Supplementary Table 4). While neither of these enrichments is 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 evasion of the host adaptive immune response[20]. Of the 313 *vir* genes included in the Salvador I reference assembly, Figure 3b indicates that less than a third are also observed in the four new assemblies. Unexpectedly, we encountered 15 ‘ultra-conserved’ *vir* genes that were present in all assemblies and exhibited very low SNP diversity, in particular one locus (PVX_113230) that was invariant and exhibited 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 intra-erythrocytic expression profile relative to most *virs*[21] (Supplementary Fig. 4).

This global census of genomic diversity in *P. vivax* has uncovered an unexpected degree of genetic polymorphism, much of which may translate into important functional variation. Our data stop short of suggesting the existence of distinct sub-species of *P. vivax*, similar to the sub-species *P. vivax vivax* and *P. vivax hibernans* proposed on the basis of relapse phenotype[22]. 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*, and serves as a warning that *P. vivax* could present a qualitatively different eradication task.

ONLINE METHODS

Parasite material

We chose four strains of *P. vivax* for whole genome sequencing based upon geographical 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 unavailable[26]. The Brazil I strain is highly resistant to the anti-relapse drug primaquine[23]. The India VII strain is the first *Plasmodium* species to be sequenced from India. The Mauritania I strain is a rare example of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

an African *P. vivax* strain that occurs among West Africans with Berber and Arab genetic backgrounds[28]. Genomic DNA for *P. vivax* sequencing was obtained from the leukocyte-depleted blood of infected, splenectomized *Saimiri* monkeys as described previously[6]. Genome-wide fragment analysis of 15 microsatellites prior to 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 Illumina GAIIx sequencers. After filtering out reads with sequence similarity to monkey/primate sequence each genome was assembled using the ALLPATHS-LG algorithm[8]. 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 checking 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, 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 180 bp). Coverage for each of the isolates was as follows: Indian isolate 87_239, 42X; Indochina isolate Dd2, 196X; Honduran isolate HB3, 30X; Indian isolate ML-14, 41X; Senegalese isolate Th231.08, 59X.

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 RNAmmer[32]. The tRNA features were identified using tRNAScanSE[33]. Other common RNA features were identified with RFAM[34].

The MUMmer algorithm[35] 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* SaIvador 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

self-training GeneMark-ES[36] and Geneld[37] 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 frame-shift 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 frame-shift or in-frame stop, then the corresponding GeneMarkES gene model was selected. If a GeneWise gene model had no overlap to 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 checked against tRNA and rRNA features and filtered where appropriate. Additional gene filtering was performed by removing genes with 30% or more coding sequence overlap with TransposonPSI (see URLs) hits (e < 1E-10).

SNP calling

Sequencing reads from each isolate were aligned using BWA[38] to the references assemblies of *P. falciparum* 3D7 (build v7.1) and *P. vivax* Salvador I (build v7.0), both downloaded from PlasmoDB (http://plasmodb.org). SNPs were called using the Unified Genotyper[11] in the GATK package[39]. SNPs with an estimated PHRED-style quality score of Q30 or greater were used for diversity analyses.

Microsatellite variants and validation

The mreps program[40] was used to identify microsatellites in the references assemblies of *P. falciparum* and *P. vivax*. The following mreps parameters were used for the microsatellite search: (1) minp=1, (2) 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 BWA[38] with the “−q 5” and “ l 32” options. Base quality score recalibration and local realignment around microsatellites[11] was applied using GATK[39]. 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, followed by indel (insertion/deletion) calling with standard hard filtering parameters was applied to using GATK, and 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 (downloaded from http://www.broadinstitute.org/annotation/genome/plasmodium_falciparum_spp/Downloads.html) to the 3D7 reference assembly to generate a truth set of indels for genotyped microsatellites. Comparison of the Illumina-based calls to

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

the Sanger calls indicates that our Illumina calling method exhibits a specificity of 100% and a sensitivity of 97%, on the basis of 81,569 AT dinucleotide microsatellites callable in both datasets. P values for evaluating the significance of the difference in microsatellite diversity between species were generating 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 (πS), and nonsynonymous (πN) sites was calculated for each gene in the reference annotations for *P. falciparum* and *P. vivax* using the method described in Table 1 of Ina (1995)[42]. *P. knowlesi* orthologs were employed to calculate interspecific d_N/d_S ratios using PAML v4.5[43]. with sequences from the Salvador I isolate of *P. vivax*. A matrix of pairwise nucleotide distances between isolates (Supplementary Table 5) was constructed for each species using SNPs identified in fourfold-degenerate (4D) 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 6) constructed for each species by rooting 4D polymorphisms using an outgroup species (*P. knowlesi* for *P. vivax*, *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 a Mann Whitney U test and the Z scores were interpreted using a Bonferroni correction for multiple testing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work has been funded in whole or in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, 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. We thank the NIAID/NHGRI Eukaryotic Pathogens and Disease Vectors Working Group, and the Broad Institute Genome Sequencing Platform for its significant contributions to this project. KG and LY are supported by a Global Health Program grant from the Bill and Melinda Gates Foundation. AAE is supported by a grant from the US National Institutes of Health, GM084320, and JMC and PLS by U19AI089676, an NIH International Center of Excellence for Malaria Research.

References

1. Guerra CA, et al. The international limits and population at risk of Plasmodium vivax transmission in 2009. PLoS Negl Trop Dis. 2010; 4:e774. [PubMed: 20689816]
2. Dobson MJ. Malaria in England: a geographical and historical perspective. Parassitologia. 1994; 36:35–60. [PubMed: 7898959]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

3. Price RN, Douglas NM, Anstey NM. New developments in Plasmodium vivax malaria: severe disease and the rise of chloroquine resistance. Curr. Opin. Infect. Dis. 2009; 22:430–435. [PubMed: 19571748]

4. Carlton JM, Sina BJ, Adams JH. Why Is Plasmodium vivax a Neglected Tropical Disease? PLoS Negl Trop Dis. 2011; 5:e1160. [PubMed: 21738804]

5. Winzeler EA. Malaria research in the post-genomic era. Nature. 2008; 455:751–756. [PubMed: 18843360]

6. Carlton JM, et al. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. Nature. 2008; 455:757–763. [PubMed: 18843361]

7. Dharia NV, 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. U.S.A. 2010; 107:20045–20050. [PubMed: 21037109]

8. Gnerre S, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences. 2011; 108:1513–1518.

9. Feng X, et al. Single-nucleotide polymorphisms and genome diversity in Plasmodium vivax. Proc. Natl. Acad. Sci. U.S.A. 2003; 100:8502–8507. [PubMed: 12799466]

10. Mu J, et al. Host switch leads to emergence of Plasmodium vivax malaria in humans. Mol. Biol. Evol. 2005; 22:1686–1693. [PubMed: 15858201]

11. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 2011; 43:491–498. [PubMed: 21478889]

12. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu. Rev. Genet. 2010; 44:445–477. [PubMed: 20809801]

13. Joy DA, Mu J, Jiang H, Su X. Genetic diversity and population history of Plasmodium falciparum and Plasmodium vivax. Parassitologia. 2006; 48:561–566. [PubMed: 17688177]

14. Kumar S, Subramanian S. Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci U.S.A. 2002; 99:803–808. [PubMed: 11792858]

15. Paget-McNicol S, Saul A. Mutation rates in the dihydrofolate reductase gene of Plasmodium falciparum. Parasitology. 2001; 122:497–505. [PubMed: 11393822]

16. Cornejo OE, Escalante AA. The origin and age of Plasmodium vivax. Trends Parasitol. 2006; 22:558–563. [PubMed: 17035086]

17. Conway DJ, et al. Origin of Plasmodium falciparum malaria is traced by mitochondrial DNA. Mol. Biochem. Parasitol. 2000; 111:163–171. [PubMed: 11087926]

18. Carter R. Speculations on the origins of Plasmodium vivax malaria. Trends Parasitol. 2003; 19:214–219. [PubMed: 12763427]

19. Thera MA, et al. A field trial to assess a blood-stage malaria vaccine. N. Engl. J. Med. 2011; 365:1004–1013. [PubMed: 21916638]

20. Fernandez-Becerra C, et al. Plasmodium vivax and the importance of the subtelomeric multigene vir superfamily. Trends in Parasitology. 2009; 25:44–51. [PubMed: 19036639]

21. Bozdech Z, et al. The transcriptome of Plasmodium vivax reveals divergence and diversity of transcriptional regulation in malaria parasites. Proc. Natl. Acad. Sci. U.S.A. 2008; 105:16290–16295. [PubMed: 18852452]

22. Garnham PC, et al. A strain of Plasmodium vivax characterized by prolonged incubation: morphological and biological characteristics. Bull. World Health Organ. 1975; 52:21–32. [PubMed: 764993]

23. Nayar JK, et al. Studies on a primaquine-tolerant strain of Plasmodium vivax from Brazil in Aotus and Saimiri monkeys. J. Parasitol. 1997; 83:739–745. [PubMed: 9267419]

24. Bhasin VK, Trager W. Gametocyte-forming and non-gametocyte-forming clones of Plasmodium falciparum. Am. J. Trop. Med. Hyg. 1984; 33:534–537. [PubMed: 6383092]

25. Sullivan JS, et al. Adaptation of a strain of Plasmodium vivax from India to New World monkeys, chimpanzees, and anopheline mosquitoes. J. Parasitol. 2001; 87:1398–1403. [PubMed: 11780828]

26. Collins WE, et al. Studies on the North Korean strain of Plasmodium vivax in Aotus monkeys and different anophelines. J. Parasitol. 1985; 71:20–27. [PubMed: 3884764]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

27. Wellems TE, et al. Chromosome size variation occurs in cloned Plasmodium falciparum on in vitro cultivation. Rev. Bras. Genet. 1988; 11:813–825.

28. Collins WE, et al. Adaptation of a strain of Plasmodium vivax from Mauritania to New World monkeys and anopheline mosquitoes. J. Parasitol. 1998; 84:619–621. [PubMed: 9645868]

29. Gardner MJ, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature. 2002; 419:498–511. [PubMed: 12368864]

30. Melnikov A, et al. Hybrid selection for sequencing pathogen genomes from clinical samples. Genome Biol. 2011; 12:R73. [PubMed: 21835008]

31. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Research. 2004; 14:988–995. [PubMed: 15123596]

32. Lagesen K, et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007; 35:3100–3108. [PubMed: 17452365]

33. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997; 25:955–964. [PubMed: 9023104]

34. Griffiths-Jones S. Annotating non-coding RNAs with Rfam. Chapter 12. Curr Protoc Bioinformatics. 2005; (Unit 12.5)

35. Kurtz S, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5:R12. [PubMed: 14759262]

36. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 2008; 18:1979–1990. [PubMed: 18757608]

37. Blanco E, Abril JF. Computational gene annotation in new genome assemblies using GeneID. Methods Mol. Biol. 2009; 537:243–261. [PubMed: 19378148]

38. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

39. 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]

40. Kolpakov R, Bana G, Kucherov G. mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Research. 2003; 31:3672–3678. [PubMed: 12824391]

41. Russell B, Suwanarusk R, Lek-Uthai U. Plasmodium vivax genetic diversity: microsatellite length matters. Trends Parasitol. 2006; 22:399–401. [PubMed: 16837246]

42. Ina Y. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. 1995; 40:190–226. [PubMed: 7699723]

43. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007; 24:1586–1591. [PubMed: 17483113]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

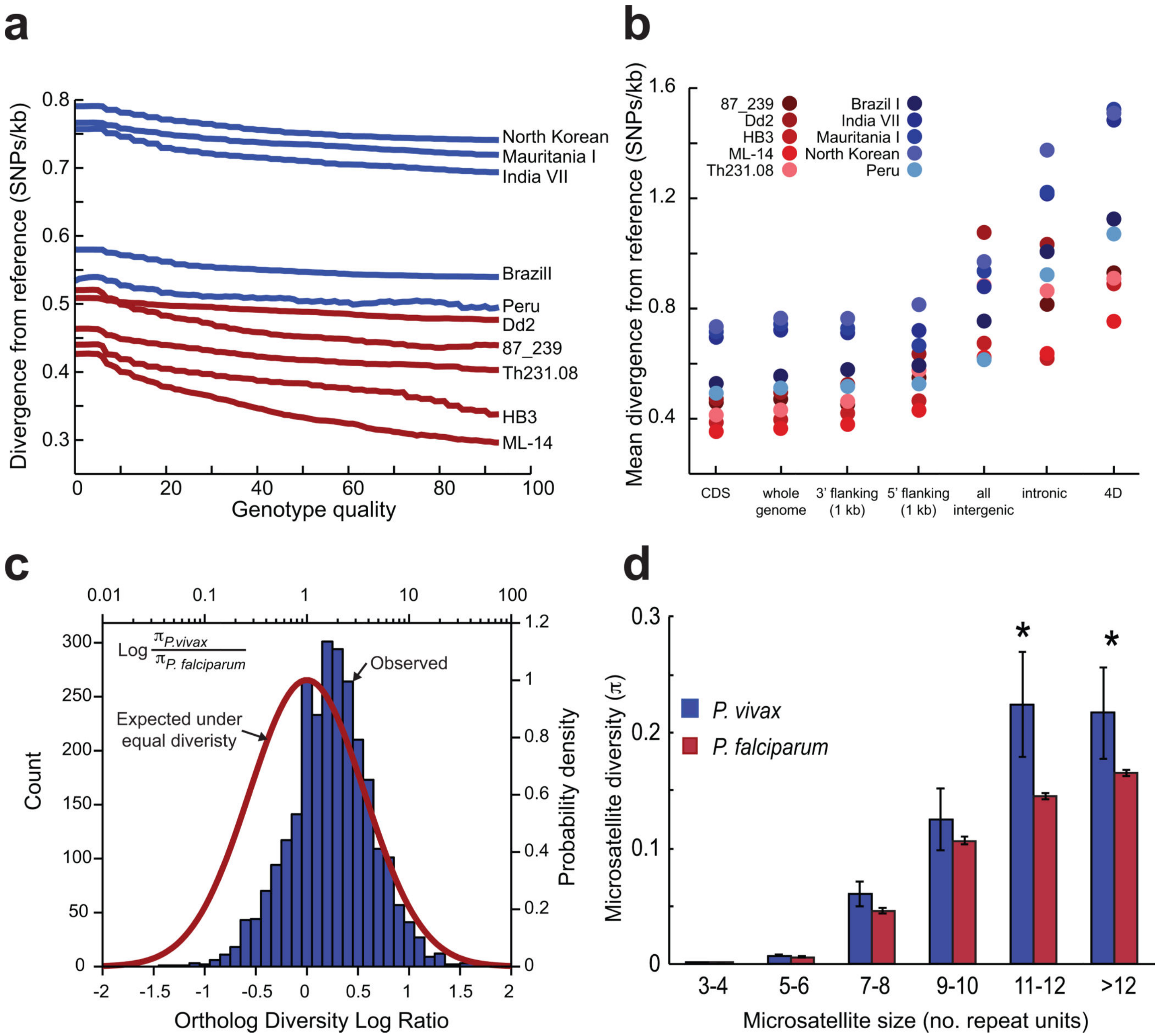

Figure 1.

Disparity in SNP and microsatellite diversity between *P. vivax* and *P. falciparum*. (**a**) Quality score vs. pairwise whole genome SNP rates against reference assemblies. Blue lines indicate *P. vivax* isolates and red lines indicate *P. falciparum* isolates. (**b**) *P. falciparum* vs. *P. vivax* Q30 call rates for: coding sequence (CDS), whole genome, 5' flanking sequence (1 kb), 3' flanking sequence (1 kb), all intergenic sequence, introns, and fourfold degenerate (4D) synonymous coding sites. **c**) Density distribution of *P. falciparum/P. vivax* diversity log ratios for genes with 1:1 orthologs, compared to null expected distribution centered on 1. (**d**) Histogram of microsatellite diversity in microsatellite loci with a repeat unit size of two 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).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

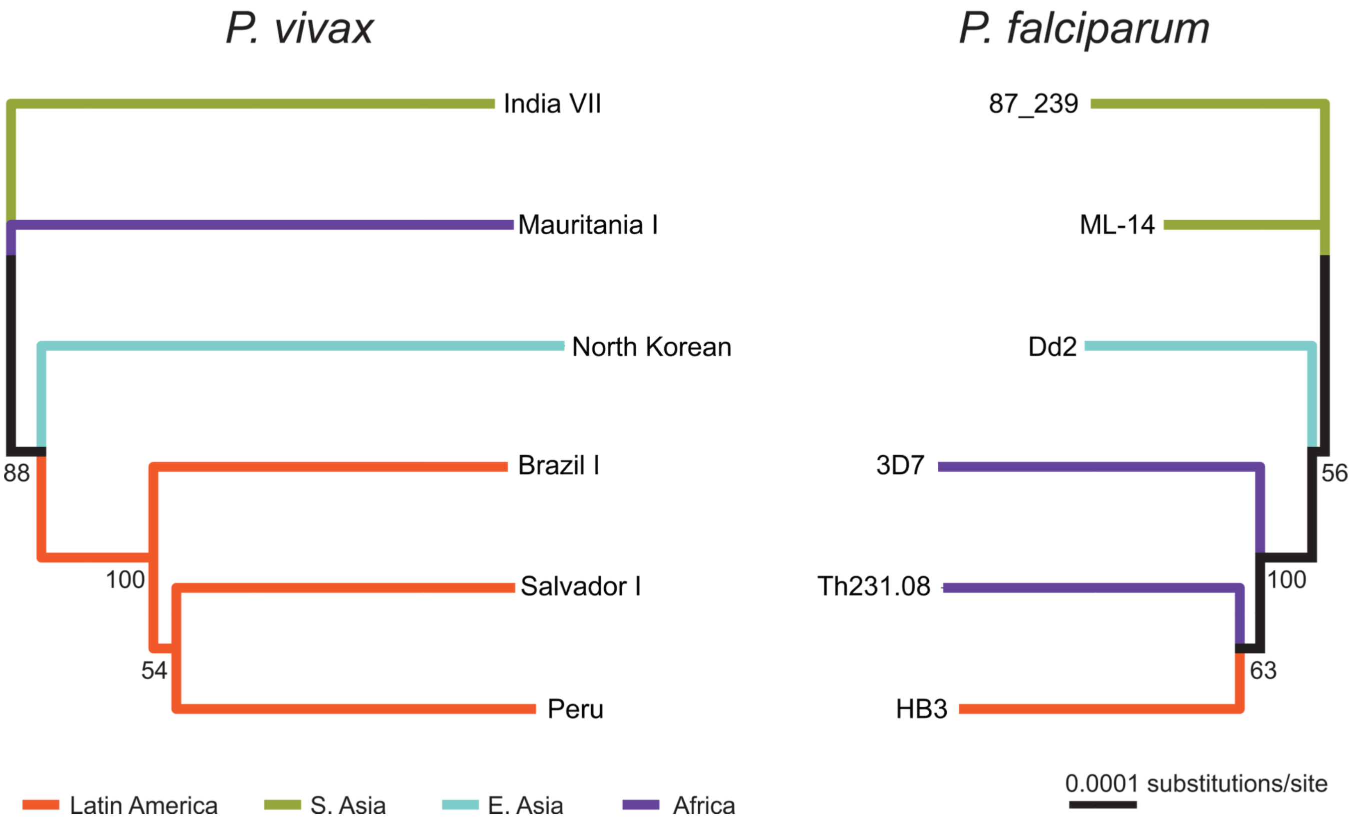

Figure 2.

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: Red = Central/South America, Purple = Africa, Green = India, Teal = Southeast Asia. Branch lengths indicate considerable diversity in New World *P. vivax* strains as well as 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 % bootstrap support.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

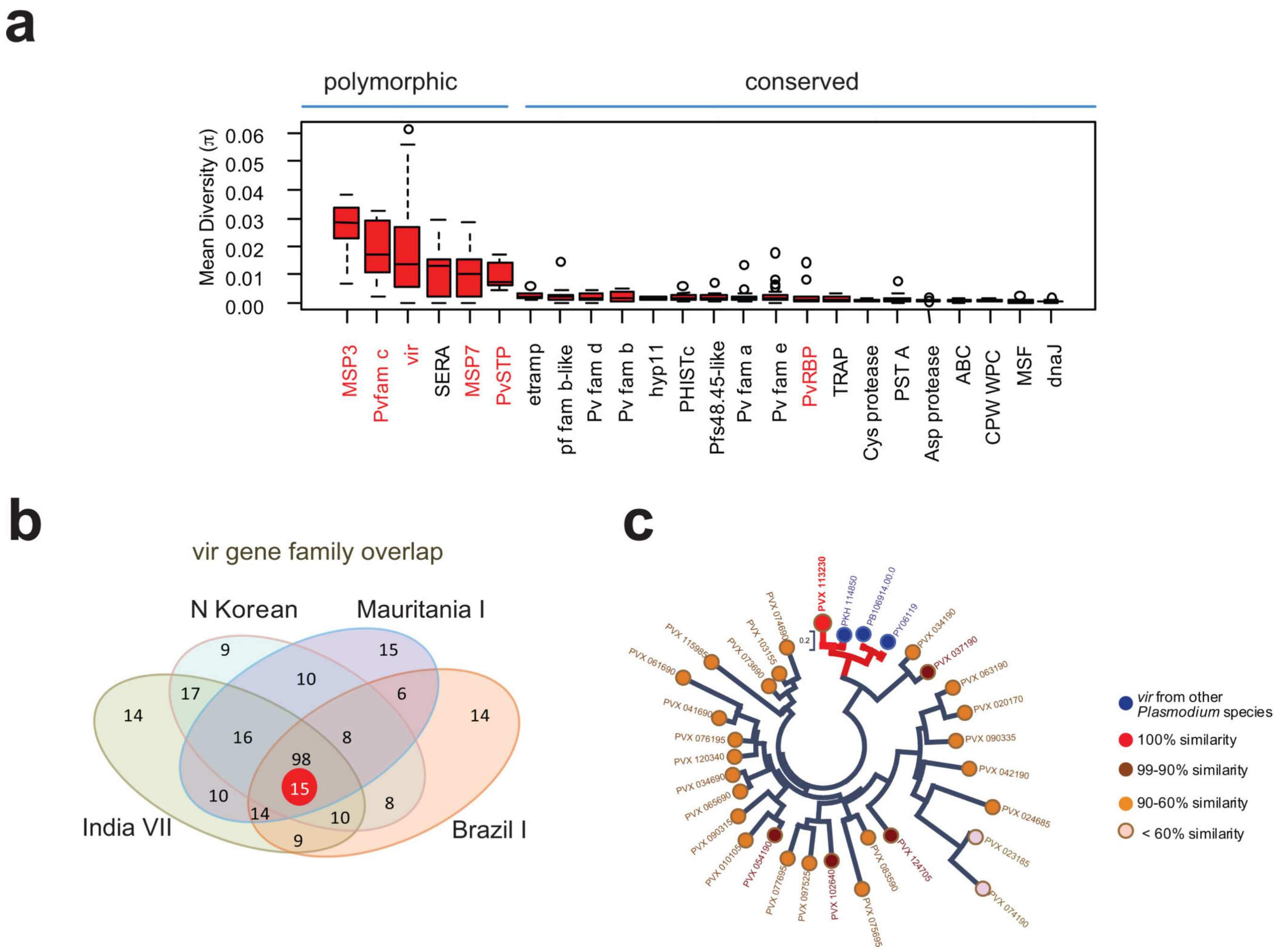

Figure 3.

Diversity of *P*. *vivax* gene families. (**a**) Mean pairwise SNP diversity (π) in *P*. *vivax* gene families. Gene families associated with merozoite invasion or immune response modulation (red text) exhibit highest diversity. Red bars on the box plots represent the 25–75th 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.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Table 1

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

Geographic Origin	*P. vivax* strains and isolates	*P. falciparum* strains
Latin America	Salvador I (El Salvador)[6] Brazil I[23] IQ07 (Peru)[7]	HB3 (Honduras)[24]
South Asia (India)	India VII[25]	ML-14 87_239
East Asia	North Korean[26]	Dd2 (Indochina [Thailand/Laos])[27]7/10/2012 3:22:00 AM
Africa	Mauritania I[28]	3D7[29] Th231.08 (Senegal)[30]

Parenthetic inclusions indicate more specific geographic origination, where known. Citations for the two reference sequences *P. vivax* Salvador I and *P. falciparum* 3D7 are their respective genome papers

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

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

Strain	Assembly size (Mb)	Fold coverage	Contig N50 (kb)	No. contigs	Scaffold N50 (kb)	No. 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