PLOS | SUBMISSION

Whole genome sequencing of field isolates reveals extensive genetic diversity in *Plasmodium vivax* from Colombia

David J. Winter[1*], M. Andreína Pacheco[1,2], Andres F. Vallejo[3], Rachel S.Schwartz[1], Myriam Arevalo-Herrera[3,4], Socrates Herrera[3], Reed A. Cartwright[1,5‡], Ananias A. Escalante[1,2*‡]

1 The Biodesign Institute, Arizona State University, Tempe, AZ, USA
2 Institute for Genomics and Evolutionary Medicine (iGEM), Temple University, Philadelphia, PA, USA
3 Caucaseco Scientific Research Center, Cali, Colombia
4 Faculty of Health, Universidad del Valle, Cali, Colombia
5 The School of Life Sciences, Arizona State University, Tempe, AZ, USA
‡These authors contributed equally to this work.

* Corresponding Authors.
DJW: djwinter@asu.edu
AAE: Ananias.Escalante@temple.edu

Abstract

Plasmodium vivax is the most prevalent malarial species in South America and exerts a substantial burden on the populations is affects. Its control and eventual elimination are a global health priority. Genomic research contributes to this objective by improving our understanding of the biology of *P. vivax* and through the development of new genetic markers that can be used to monitor efforts to reduce malaria transmission.

Here we analyze whole genome data from eight field samples from a region in Cordóba, Colombia where malaria is endemic. We find considerable genetic diversity within this population, a result that contrasts with earlier studies suggesting that *P. vivax* had limited diversity in the Americas. We also identify a selective sweep around a substitution known to confer resistance to sulphadoxine-pyrimethamine (SP). This is the first observation of a selective sweep for SP resistance in this parasite. These results indicate that *P. vivax* has been exposed to SP pressure even when the drug is not in use as a first line treatment for patients afflicted by this parasite. We identify multiple non-synonymous substitutions in three other genes known to be involved with drug resistance in *Plasmodium* species. Finally, we found extensive microsatellite polymorphisms. Using this information we developed 18 microsatellite loci that are polymorphic and easy to score and can thus be used in epidemiological investigations in South America.

Author Summary

Although *P. vivax* is not as deadly as the more widely studied *P. falciparum*, it remains a pressing global health problem. Here we report the results of a whole genome study of *P. vivax* from Cordóba, Colombia, in South America. This parasite is the most prevalent in this region. We show that the parasite population is genetically diverse, contrary the expectations from earlier studies from the Americas. We also find molecular evidence that resistance to an anti-malarial drug has arisen recently in this region. This selective sweep indicates that the parasite has been exposed to a drug that is not used as first line treatment for this malaria parasite. In addition to extensive SNP and microsatellite polymorphism, we report 18 new genetic loci that might be helpful for fine-scale studies of this species in the Americas.

Introduction

Despite significant advancements toward malaria control and elimination, about 40% of the world's population remains at risk of infection by one of the four protozoan species that commonly cause the disease [1]. Among the human malarias, *P. vivax* is the parasite with most morbidity outside Africa [1]. *P. vivax* differs from the more widely studied *P. falciparum* in aspects of its life cycle, disease severity, geographic distribution, ecology, and evolutionary history [2–7], raising concerns that gaps in our knowledge about its basic biology may compromise its control [8].

Genomic approaches provide important tools to study hard-to-culture parasites such as *P. vivax*. For example, genome-wide scans performed on samples from a natural parasite population can identify regions of the genome subject to strong selection. Studies using this approach in *P. falciparum* have contributed to the study of drug resistance and adaptation to the host immune system in that species. [9–12].However, this approach has not yet been widely applied in *P. vivax*.

Previous population genetic studies on genes encoding antigens have found that *P. vivax* populations in many regions of the Americas are less diverse than those from Asia or Oceania [8, 13–16]. Similarly, a recent whole genome study found limited genetic diversity in a population from the Amazon basin of Peru [17]. It remains unclear whether these results are reflective of populations in the New World generally, the geographical sampling of those particular studies, or the loci sampled in earlier studies. Two recent studies have suggested that *P. vivax* populations in the Americas may harbor more genetic diversity than previously thought. First, a genome-wide comparison revealed substantial genetic divergence among three parasite lineages isolated in the Americas and maintained in non-human primates [18]. Second, recent population studies on the mitochondrial genome have shown high levels of divergence and limited gene flow among populations in the region [19]. This pattern indicates that *P. vivax* populations in the Americas likely have a complex history, with divergent populations harboring differing levels of genetic diversity.

Genomic studies can also support efforts to control and eliminate malaria from a given region by identifying genetic markers that will be informative for fine-scale population genetic studies in that region. Molecular epidemiological investigations rely on multilocus genotyping of SNPs or microsatellites [20] to investigate patterns of population structure and gene flow. Although the high mutation rates of microsatellite loci makes them ideal markers for such studies, only a few loci are currently in use [20]. These existing loci were developed using data from a small number of populations. As a result, some of these loci fail to amplify in samples from other localities [21]. Whole genome studies with multiple samples offer the opportunity to identify new loci, and focus on those that are known to be polymorphic in a given region and have the sort of simple repeat motifs that lead to reliable scoring of genotypes. In addition to identifying patterns of transmission within a region, these markers can be used to distinguish local cases (the result of remaining malaria transmission) from those that are introduced from another region. Ascertaining the source of the parasites detected in a given case is critical in order to evaluate the success of interventions during an elimination program.

Here we take a whole-genome approach to characterize the genetic variation of field isolates with single-lineage *P. vivax* infections from Northern Colombia (specifically, Tierralta, Department of Córdoba), an area in South America with seasonal transmission [22]. In addition to assessing the genetic diversity within this population, we examine patterns of diversity across the *P. vivax* genome and find evidence for a recent selective sweep likely associated resistance to a drug that is not prescribed for treating *P. vivax* malaria. We also develop 18 new microsatellite loci for fine scale studies in Colombia.

Methods

Ethics statement

A passive surveillance study was conducted between 2011 and 2013 in outpatient clinics located in Tierralta [22]. The study protocol was approved by the Institutional Review Board (IRB) affiliated to the Malaria Vaccine and Drug Development Center (MVDC, Cali-Colombia). Patients with malaria infection as determined by microscopic examination of Giemsa-stained thick blood smears received oral and written explanations about the study and, after expressing their willingness to participate, were requested to sign an informed consent (IC) previously approved by the Institutional Review Board (IRB) affiliated to the MVDC. IC from each adult individual or from the parents or guardians of children under 18 years of age was obtained. Individuals between 7 and 17 years old were asked to sign an additional informed assent. A trained physician of the study staff completed a standard clinical evaluation and a physical examination in all malaria symptomatic subjects. The local health provider treated individuals as soon as the blood sample had been drawn, using national antimalarial therapy protocol of the Colombian Ministry of Health and Social Protection [22]. Specifically, patients infected with *P. vivax* were treated orally with chloroquine (25 mg per kg provided in three doses) and primaquine (0.25 mg per kg daily for 14 days).

Sample collection

We collected 10 mL of blood from each patient and stored each blood sample in EDTA. In order to eliminate as much human DNA as possible, each sample was diluted in one volume of PBS and filtered with a CF11 column ($\approx 3g$) that had previously been rinsed with PBS. A new column was used for each 5 mL of sample. Each filtered sample was centrifuged at 1,000 g for 10 minutes. The supernatant was discarded and the red blood cells (RBC) were kept at -20°C and sent to our laboratory in Cali for processing. The RBCs were suspended in one volume of PBS and aliquoted into 200 μL fractions. DNA was extracted from each aliquot by using a PureLink Genomic DNA kit (Invitrogen, USA) following specifications provided by the manufacturer.

Sequencing and alignment to reference

Depending on availability and quality of the sample, we used between 300 ng and 1 mg of DNA to construct sequencing libraries. These libraries were constructed using a Kapa Biosystems DNA Library Preparation Kit (Kapa Biosystems, USA). The resulting fragments were amplified in ten rounds of PCR, using a Kapa HiFi Library Amplification Kit (Kapa Biosytems, USA). Denaturation and clustering were performed using an Illumina cBot. Once the samples were clustered, the flow cell was loaded onto a HiSeq 2000. The run module used was a paired end 2x100 reads. All sequencing and library preperation stages were performed by the DNASU sequencing core at Arizona State University.

We used Bowtie version 2.1.0 [23] to map reads from each sample to a reference genome containing sequences derived from the Salvador I (SalI) strain's nuclear [24] (build ASM241v1) and apicoplast [25] genomes. The resulting alignments were processed using a modified version of the GATK project's best practice guidelines [26]. Specifically, we identified and marked potential PCR duplicates using the MarkDuplicates tool from Picard version 1.106 (http://picard.sourceforge.net) and performed local realignment around possible indels using GATK 2.8 [27, 28]. Finally, we adjusted the raw base quality scores by running GATK's BaseRecalibrator tool, treating set of putative Single Nucleotide Variants (SNVs) identified by SAMtools (0.1.18) [29] as known variants.

To investigate the possibility that our patient samples contain multiple distinct *P. vivax* strains [30, 31], we repeated this process using sequencing reads produced from a known single infection. We retrieved reads from the monkey-adapted SalI strain [30] from the NCBI Sequence Read Archive (accession SRS365051) We recorded the frequency of bases matching reference genome at each site for the patient-derived and single lineage alignments using a custom C++ program (http://dx.doi.org/10.5281/zenodo.18190) that makes use of the BamTools [32] library. We also calculated the overall proportion of all sequenced bases that produced a minority allele when mapped to reference for each sample.

Variant discovery

We called putative SNVs and small indels from the Colombian samples using the GATK UnifiedGenotyper [27]. Because we were able to establish that each patient was infected by a single lineage, we treated samples as haploid. Artifacts produced during the sequencing and mapping of reads to reference can lead to false positive variant calls [33]. Such false-positive variant calls may be particularity likely in *P. vivax* due the number multi-copy gene families and paralogs in this species [34]. In order to account for these potential artifacts, we took a conservative approach to variant calling and removed apparently variant sites that may have resulted mis-mapped reads. After performing an exploratory analysis comparing properties of our putative variants to a random sample of 100,000 non-variant sites, and using guidelines described by the GATK developers [28], we established the following set of criteria to identify likely false positive variants:

(a) Site within 50 kb of a chromosome end (which are dominated by repeats)

(b) Average mapping quality phred score <35

(c) More than one sample has multiple nucleotides called at the site, such that there are more than two reads that contain minor nucleotides

(d) P-value for a Fishers exact test of strand bias <0.001

(e) Absolute z-score for Mann-Whitney U-test of mapping quality difference between variant and reference allele containing reads >5

(f) Absolute z-score for Mann-Whitney U-test of read-position difference between variant and reference allele containing reads >5

(g) Total depth at site $\geq$ 165x (95th percentile across all sites)

To identify microsatellite loci that are segregating within Colombia and have relatively simple evolutionary histories, we filtered the indels labeled as STRs by UnifiedGenotyper by removing any matching following criteria:

(a) Locus has non-perfect repeats

(b) Repeat motif $>$8bp

(c) Locus is monomorphic among Colombian samples

We produced final variant sets for SNVs and microsatellites by removing all sites that matched at least one of the criteria listed above using PyVCF (http://pyvcf.readthedocs.org/). We produced a functional annotation for each polymorphic SNV with snpeff [35] using Ensembl functional annotation of the SalI reference (build ASM241v1.23) as input.

Validation of variants

We validated our SNV calling procedure by running the steps described above on a genome alignment generated from reads previously produced from the SalI strain [30] Because these reads represent an independent sequencing of the same strain that was used to produce the reference genome, we expect very few non-reference alleles. We also tested the effect of including low-coverage samples in our variant calling pipeline by repeating this procedure with the SalI alignment down-sampled to 2x coverage. We further tested the validity of filtered variant sites by searching for apparently singleton SNVs (those found only once in our population sample) in previously reported SNVs from other studies [18, 30].

Oligonucleotide primers for polymorphic microsatellites DNA markers

We designed PCR primers for 18 of the microsatellite loci identified using the above criteria. We first choose a subset of these putative microsatellites that were distributed across different *P. vivax* chromosomes, then developed markers using two strategies. First, nine loci were identified on alignments of conserved regions between *P. vivax* (SalI and the *P. vivax* genome data available in NCBI) and *P. cynomlogi* genomes. Second, nine loci were identified on conserved regions between SalI strain and Colombian samples from Tierralta.

All the alignments were made using ClustalX v2.0.12 and Muscle as implemented in SeaView v4.3.5. Dyes Hex and 6-FAM were used for labelling the forward primers. A complete characterization of the 18 microsatellites loci and the primers we used to amplify them is provided in S1 Text and S1 Table.

Population genetics

We calculated nucleotide diversity and Watterson's estimator of the population mutation rate $\hat{\theta}_w$ for the whole genome, distinct genomic features (i.e. sequences falling in exons, intron, untranslated regions and intergenic regions), and in 10 kb windows across each chromosome. We accounted for the varying levels of sequence coverage among our samples by using missing-data estimators for these measures [36]. Rather than setting an arbitrary coverage level at which a sample should be considered missing for a given site, we used our data and variant calling approach to identify the number of samples from which we could call a variant if it was present at each site in the genome. We first created new "reference genome" sequences by switching each unambiguous base in the SalI reference following Table 1. We then called variants against these "shifted" genomes using the same procedure described above (including filtering steps). At each site, a sample was considered missing if no variant could be called for that sample using the true reference genome or any of the shifted references.

We used PyBedtools [37, 38] to generate genomic windows, and extract polymorphisms from various sequence classes. We used diversity statistics calculated from genomic windows to identify genomic regions with unusually high or low genetic diversity. Specifically, we identified windows with values in the 1st or 99th percentile of either measure having first removed windows for which less than 85% of sites were callable. We discovered a region of particularly low diversity surrounding the *dhps* gene, so focused on this gene by calculating each statistic for set of over-lapping windows, each 10 kb wide and 500 bases apart from each other.

Table 1. Scheme used to produced alternative reference genomes

Reference	A	C	G	T
First shift	T	A	C	G
Second shift	G	T	A	C
Third shift	C	G	T	A

Results

Genome sequencing

We generated between 18 and 36 million paired-end reads from each samples (Table 2). Obtaining genomic sequences from clinical isolates of *P. vivax* is complicated by the presence of human DNA in parasite-containing blood samples. Although we took steps to remove leukocytes from each of our samples, the proportion of reads that could be mapped to the SalI reference genome differed markedly among samples, ranging from <1%–28%. These differences were reflected in the mean sequencing coverage achieved for each sample, which varies from less than one read per base in sample 500, to greater than 40 reads for sample 499.

Despite the presence of some poorly covered samples, our mapped reads allow comparison between multiple individuals for the vast majority of the *P. vivax* genome (Fig. 1). Because low-coverage samples still provide valuable data for variant calling at some sites and can often be reliably genotyped for those sites known to contain segregating variants we included all of our patient samples in subsequent analyses.

All sampled patients were infected by only one *P. vivax* strain

The presence of multiple distinct parasite lineages within a single host has presented a barrier to population genetic analysis in previous whole-genome studies of *P. vivax*. Because *Plasmodium* merozoites are haploid in the vertebrate host, the presence of such multiple infections in a patient can be inferred by the presence of multiple alleles at different loci. In the context of high-throughput sequencing, these additional alleles manifest as an excess of bases with intermediate frequencies at sites in a genome alignment. This result contrasts with the distribution expected from singly infected patients, where only rare sequencing errors will produce minority bases [30]. We tested our samples for multiple infections by comparing the distribution of minor bases frequencies in our alignments with the same distribution in an alignment produced from a known single infection (Fig. 2). In both the known single infection strain and our samples, minority bases were present only at low frequencies. This pattern contrasts distinctly with the high proportion of intermediate-frequency bases expected from mixed infections [30]. Because the shape of the distribution of minor base frequencies will be less informative for samples with low coverage, we also examined the proportion of all sequenced bases that were in the minority relative to other reads aligned to the same site. Again, these results are similar to those from SalI (Table 2). The proportion of minority bases in reads produced from a known single infection was 7.3×10^{-4}, while for our data this proportion was between 7.3×10^{-4} and 1.11×10^{-3}. The fact that no samples had the base frequency distributions characteristic of mixed-infections, and the low-coverage samples did no produce more minor bases than other samples, suggest all of our samples can be considered single infections.

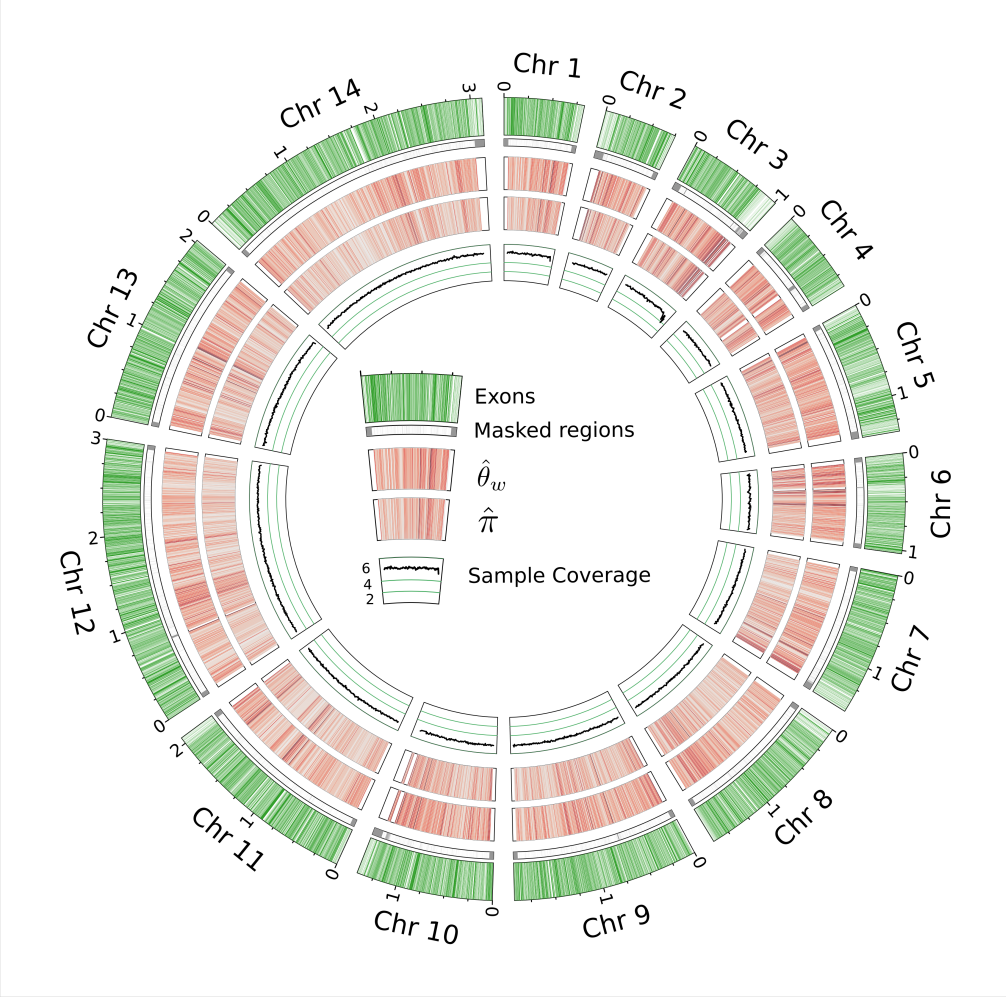

Figure 1. Summary of genomic data: Segments from outside to inside: *P. vivax* chromosomes with exonic regions shaded green; Regions excluded from variant calling in this study; heatmap of θ_w in 10kb windows, high-diversity regions are darker and the maximum value is 0.0018; heatmap of π in 10kb windows; Mean number of samples covered per base in 10kb windows

Variant discovery

We validated our SNV calling procedure by applying it to a set of sequencing reads independently produced from the same strain used to assemble the reference genome [30]. Using this procedure, we called a total of 34 SNVs from the >22 million base pair reference genome, none of which were called as polymorphic alleles in our patient-derived data. Repeating this procedure with a lower coverage (2x) dataset generated fewer SNVs (18) including only one new putative variant. The small number of variants called from the reference data confirms the conservative nature of our variant calling procedure.

In total, we identified 33,855 non-reference SNV alleles among the Colombian samples, 30,261 of which were polymorphic (Table 3). The total number of SNVs we detect is comparable to numbers found in studies of field isolates from other regions; samples from Madagascan and Cambodian populations contain 41,630 and 45,417 SNVs respectively in the genomic regions included in our study (Fig. 3) [30] . A recent study

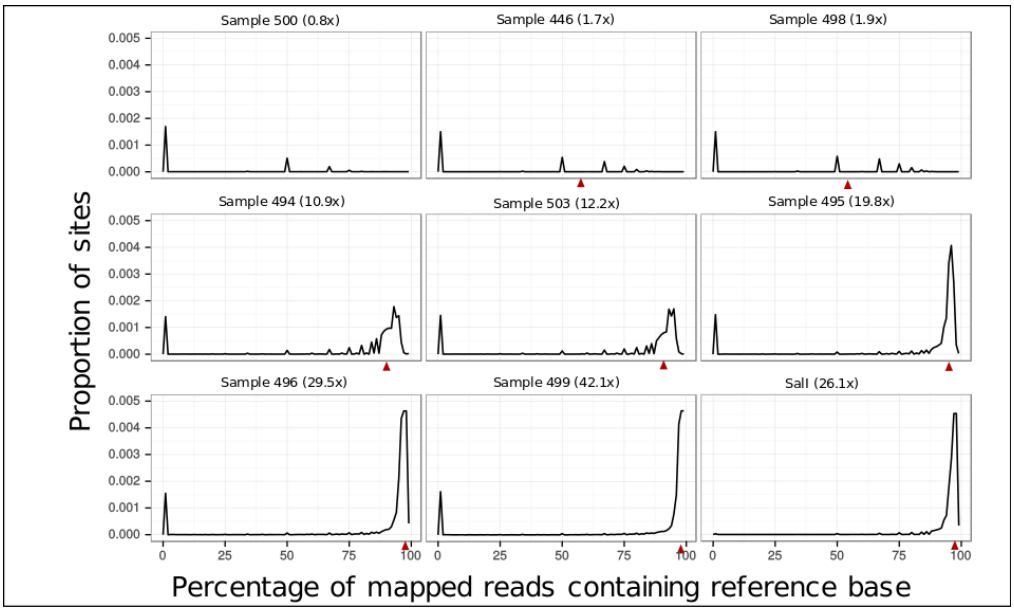

Figure 2. Reference base frequency spectra: The first eight panels represent the distribution of the frequency of reference bases in sequencing reads produced from our samples across all sites in the *P. vivax* genome. In the final panel the same distribution is graphed for reads generated from SalI, a known single infection. Red triangles represent frequency expected if one non-reference containing read was mapped to a site with average sequencing coverage for that sample (given in parenthesis after the sample name). The majority of sites from all samples contain only non-reference bases, these sites were removed to allow clearer visualisation of these distributions

.

using isolates from the Amazon basin in Peru [17] identified 10,989 SNVs in the same regions. Approximately two-thirds of the Peruvian SNVs (7,232) are also present in our Colombian samples. However, the Madagascan and Cambodian samples share many alleles that are absent in either South American population (11,618).

SNVs are relatively less common in exonic sites (1.2 SNVs per kb) than untranslated, intronic or intergenic sites (1.8 – 2.4 SNVs per kb) (Table 3). The ratio of non-synonymous to synonymous SNVs in exons is 1:1.51, substantially lower than the $\approx$ 1:4 ratio predicted for the *P. vivax* genome under neutrality [30]. This ratio, and the relative densities of SNVs in different sequence classes are very similar to those previously reported from Madagascan and Cambodian populations [30].

Among polymorphic SNVs, 12,913 (42.7%) were recorded in only one sample (Table 4). However, many of these apparent-singletons have been recorded in other studies. When we compare our SNVs to a catalogue or previously reported *P. vivax* SNVs [17, 18, 30] only 6,854 (20.2%) are unique to this study. Our low-coverage samples did not produce more singletons per callable-site than those with higher sequencing-coverage (Table 4), suggesting the remaining singletons are not simply a result of artifacts introduced by including these samples.

We identified 789 putative microsatellite loci that met our filtering criteria. To demonstrate the ability of whole genome studies to develop new markers, we designed PCR primers for 18 of these loci (choosing markers that well-spread among the 14 *P. vivax* chromosomes). We were able to generate PCR amplicons for each locus, and 16 were shown to be polymorphic within the validation panel, with between 2 and 4 loci segregating in this population (Table 5). The alleles of each locus could be determined

easily, as demonstrated by the electropherograms shown in S1 Fig.

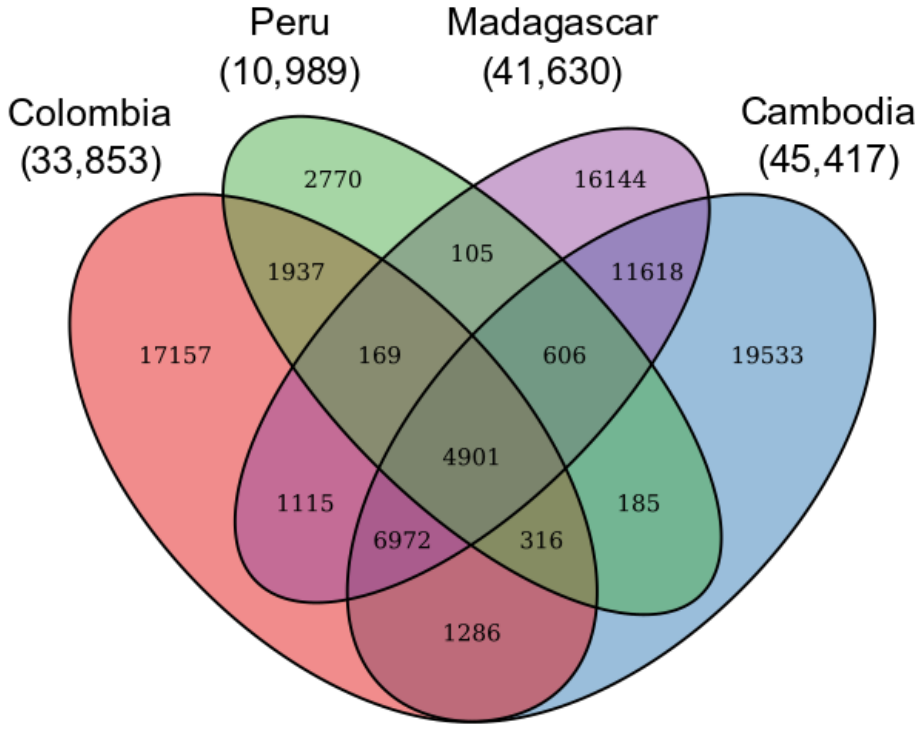

Figure 3. Allele sharing among *P. vivax* populations:. Note: Ellipses for Peruvian [17] Cambodian and Madagascan [30] populations include only those SNVs reported in the cited studies and not included in a region that was excluded in our study.

.

Population genetics

Because each of our patient samples represents a single parasite lineage, we can use standard population genetic analyses to investigate the evolutionary and demographic processes shaping *P. vivax* genomes in Colombia.

We calculated two measures of genetic diversity, nucleotide diversity (π) and Watterson's estimator ($\hat{\theta}_w$) of the population mutation rate [39, 40]. Across the whole genome, we estimate $\hat{\theta}_w$ to be 7.0×10^{-4}. The estimate for nucleotide diversity is slightly lower at 6.8×10^{-4}. For both diversity measures, genetic diversity is lowest in exonic regions, then increasingly higher in $3'$ and $5'$ untranslated regions of transcripts, intergenic regions and introns (Table 3).

A selective sweep around *dhps*

We identified regions of the genome with unusually high or low genetic diversity in this population (S2 Table). The most striking result of this analysis is an extended region of homozygosity on chromosome 14, which includes a 10kb window with no polymorphic

SNVs despite having a mean of 6.35 samples contributing data. When we narrowed our focus to this region by calculating $\hat{\theta}_w$ in overlapping windows, we found that the Dihydropteroate Synthetase gene (*dhps*) was at the centre of this region of low diversity (Fig. 4). Although there are no polymorphisms within this region, all eight of our samples contain a non-reference allele resulting from a G to C substitution in the second exon of *dhps*. The substitution is non-synonymous, and leads to the A383G amino acid substitution that has been associated with sulphadoxine resistance in numerous previous studies of *P. vivax* [41]. This pattern of low diversity surrounding a fixed substitution is the classic sign of a hard selective sweep [42, 43], in which a single mutant or migrant allele is rapidly fixed by selection. No other non-reference alleles were present in the *dhps* gene.

Most of the remaining genes that overlap with other high- or low-diversity windows encode proteins for which there is no functional annotation. Nevertheless, the high diversity windows include antigen and surface protein genes such as *msp7* and *vir* family proteins, which are known to be under balancing selection in *P. vivax* [18]. Because our conservative variant calling approach removed difficult-to-align multi-copy gene families, it is likely we excluded other genes under balancing selection.

We also examined other genes thought to be involved in drug resistance in *P. vivax* 6. Alleles of the dihydrofolate reductase (*dhfr*) gene are known to confer resistance to pyrimethamine, a drug administrated together with pyrimethamine (SP). We did not observe evidence of a selective sweep at the *dhfr* locus. However, all samples for which a genotype could be called reliably (six of eight) have non-synonymous SNVs leading to both S58R and S117N amino acid substitutions. Both of these substitutions have been associated with SP resistance in *P. vivax* [41]. There are two distinct nucleotide variants encoding the S58R substitutions in our population samples, with three parasites having AGC>AGA mutations in the 58th codon, and three others having an AGC>CGC mutation. We found a total of 13 additional non-synonymous variants in the ATP-cassette binding proteins 1 *pvmdr1* and PVX_124085 (a homolog of *P. falciparum mrp* proteins). There were no such variants in GTP cyclohydrase, Chloroquine resistance transporter ortholog, or Kelch 13, which are all considered possible drug resistance genes [17].

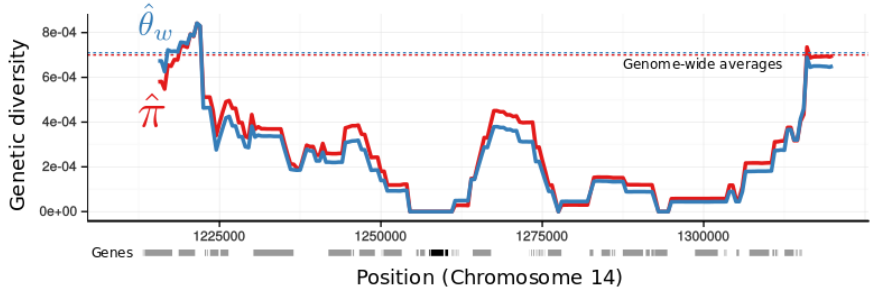

Figure 4. A selective sweep around *dhps*: Two measures of genetic diversity (y-axis), where calculated in overlapping windows across a portion of Chromosome 14 (positions in x-axis). The dashed path under the x-axis represents the position of exons in the current annotation of the *P. vivax* reference genome. Exons of *dhps* are shaded black, all others gray. Dashed lines represent the genome-wide average of each diversity measure.

Discussion

Genetic diversity in Colombia

The genetic diversity estimated from our Colombian *P. vivax* population is comparable to, though slightly lower than, estimates derived from a global samples of *P. falciparum* (where $\hat{\theta}_w$ has been estimated to be 1.03×10^{-3} using isolates from Africa, America, Asia and Oceania [44]). Considering that our samples come from a single department in Colombia, this result demonstrates that the relatively high genetic diversity reported for *P. vivax* can extend to small spatial scales.

This finding differs substantially from the perception that South American *P. vivax* populations in general have relatively low diversity due to a simple evolutionary history [18, 19]. Some populations, including the Peruvian population that was the focus of a recent whole genome study [17] do indeed have low genetic diversity, that may be the result of recent local introductions or expansions from a few founders following from malaria control programs [45]. On the other hand, the Colombian population studied here has substantial genetic diversity. It has been suggested that such diversity could be result of complex demographic processes involving multiple introductions and recombination among lineages in broad temporal and spatial scales [19]. A comprehensive population genomic study of South America would be required to understand the extent of such genetic polymorphism and the processes involved in its maintenance. Nevertheless, these results establish that malaria control programs can face a genetically diverse parasite population even in a relatively small spatial scales in South America. Such variation should be considered wherever control programs are evaluating the use molecular markers the context of surveillance

It is difficult to compare the genetic diversity of our sample to that of other *P. vivax* populations. Most studies that report estimates of genetic diversity for this species are focused on clinically important epitopes or a few markers. On the other hand, whole genome studies usually cannot report diversity statistics due to multiplicity of infection in their samples. However, we can make a crude comparison between our results and those from other whole genome studies [17, 30] by comparing the number of non-reference alleles found in each population. Despite our relatively small number of well-covered samples, restricted geographic range and the conservative approach to variant calling, we detected 33,855 SNVs. When we apply the same masking criteria used in this study to the variants reported form Cambodia and Madagascar we arrive at comparable number of SNVs (41,630 and 45,417 respectively). Our study discovered considerably more variants than a recent study of *P. vivax* isolates from the Amazon basin in Peru (10,989 variants). It is likely that the this difference reflects the low genetic diversity of the Peruvian population. Comparing these studies also highlights the degree of allele-sharing among populations (Fig. 3). Approximately two thirds of the SNVs detected from a *P. vivax* population in Peru where also detected in our Colombian samples, and 11,618 alleles are shared by the Cambodian and Madagascan populations but absent from both South American samples. This pattern may represent genetic differentiation between New World and Old World populations, although it is important to note that differences between these studies may also reflect different variant-calling procedures. A complete understanding of the global structure of *P. vivax* populations will require many more population samples and a consistent approach to variant calling

We can also compare our genetic diversity estimates to those produced from one large scale population genetic study. A study of using *P. vivax* isolates from across India [46] reported $\hat{\theta}_w$ values ranging (from 1.3×10^{-3} – 3×10^{-3}) from 5.6 kb of non-coding DNA. These values are somewhat higher than our diversity estimates from introns ($\hat{\theta}_w = 7.9 \times 10^{4}$) or intergenic regions ($\hat{\theta}_w = 7.5 \times 10^{-4}$) but within the range of values we calculate from 10kb windows.

Signatures of natural selection

We identified genomic windows with exceptionally high or low genetic diversity. High genetic diversity may be maintained by balancing selection, while regions of low-diversity may be subject to strong purifying selection or be the product of recent selective sweeps. The majority of genes contained within these windows encode proteins for which little is known. However, some of the high-diversity windows overlap with antigen and surface-protein genes that are thought to be subject to balancing selection in *P. vivax* globally [18] . It is possible that genes in other windows of exceptional diversity have likewise been subject to natural selection; this could be confirmed from a larger population sample and thus a more statistically powerful genome scan.

We also looked a patterns of diversity more broadly, comparing different genomic features. All measures of diversity were lowest in exon sequences, followed by untranslated regions, intergenic spaces, and introns. This pattern, along with the relative lack on non-synonymous SNVs, is consistent with earlier studies demonstrating purifying selection has a strong effect on on protein coding genes in *P. vivax* [47]. It is interesting to note our estimates of genetic diversity were higher for introns than intergenic regions. This pattern has been reported in *P. falciparum* [18] and may reflect the presence of conserved but unannotated genes in what are currently considered intergenic regions [44].

Drug resistance alleles in Colombia

Our analysis of low diversity regions revealed a selective sweep associated the A383G allele of *dhps*, which has previously been associated with resistance to Sulfadoxine [41]. Although resistance to SP treatment, and indeed resistance mediated by this particular mutation, is a well known phenomenon, this result is interesting for two reasons. First, by identifying a selective sweep around this mutation we are able to demonstrate that SP resistance has arisen locally within Colombia via the rapid fixation of a single allele. This finding, combined with the fact another *dhps* allele (A385G) is most commonly associated with SP resistance in Madagascar [48], French Guiana [48], India [49], Iran [50], Pakistan [51], Thailand [52] and China [53] suggests SP resistance has come about through multiple independent origins. Similar repeated evolution of *dhps* resistant mutations has been reported in *P. falciparum* [54–57]

The evolution of SP resistance is also interesting in an operational context because it demonstrates that this *P. vivax* population has been subject to drug pressure from SP, a drug that has not been part of the approved treatment for uncomplicated *P. vivax* malaria in Colombia (where the drugs of choice are still chloroquine-primaquine combination therapy). SP has been used to treat *P. falciparum* infections in Colombia, so it is possible this selective pressure has arisen from misdiagnosis of *P. vivax* infections or the use of SP to treat mixed *P. vivax*-*P. falciparum* infections. It is also possible that poor compliance with the national drug policies, including self-medication by some patients with access to antifolates, or the long half-life of antifolate drugs, could lead to *P. vivax* infections coming into contact with SP.

The region surrounding the *dhfr* does not show the pattern of decreased genetic diversity associated with a hard selective sweep. In this case, all genotyped samples have two SP-resistance alleles (S58R and S117N), with the first allele encoded by two distinct SNVs. Thus, SP-resistance alleles in each gene have somewhat different histories, with *dhps* A383G entering the population once and being rapidly driven toward fixation, but *dhfr* resistance arising from two separate alleles that have been maintained in the population. This pattern differs with the one found in *P. falciparum* where *dhfr* mutations associated with drug resistance are fixed, as the result of a selective sweep, whereas *dhps* mutations are still segregating with sensitive alleles in the

population [55, 57].

We detected non-synonymous variants in two other genes thought to be involved with drug resistance in *Plasmodium* species. There are eight non-synonymous variants in PVX_124085. The phenotypic effects of these variants are not known, but the changes to the *P. falciparum* ortholog of this gene have been associated with decreased sensitivity of primaqune [58] and antifolate drugs [59]. This study is the third time that an excess of non-synonymous mutations in this gene has been recorded from a population in South America [17, 60], and the clinical significance of this repeated finding should be investigated.

We identified five non-synonymous mutations in *pvmdr1*, three of which are known from populations in Asia and Madagascar [48, 61]. Mutations in the *P. falciparum* ortholog of this gene are associated with decreased sensitivity to chloroquine, and *pvmdr1* variants are thus considered putative chloroquine-resistance alleles. Among the variants we report only Y976F has been associated with decreased sensitivity to chloroquine and then only *in vitro* and with a modest effect size [62]. There is little evidence for drug failure with chloroquine in South America at present. Nevertheless, the presence alleles in a South American population warrants further investigation and presents an opportunity to test for an association between *pvmdr1* alleles and sensitivity to the drug in a clinical setting.

We also compare the variants we report from putative drug resistance genes with those reported from another South American population in the Amazon basin of Peru (Table 6). These populations share many alleles, including both SP resistance alleles in *dhfr* and five amino acid substitutions in PVX_124085. In contrast, four of the five variants we report form *pvmdr1* are not present in the Peruvian population and there are no non-synonymous substitutions in Peruvian *dhps* sequences. These results demonstrate the importance of local information in designing control programs, as each population contains distinct drug resistance alleles which may generate distinct responses to different treatments. In addition, the fact two populations separated by a considerable geographical distance and the Andes share multiple alleles that are identical by state at the nucleotide level suggests it is possible drug resistance alleles can spread by gene flow between distant populations in South America.

Microsatellite loci for *P. vivax* studies in the Americas

Although population genomic studies offer a unique view into the biology of *P.vivax*, smaller-scale studies that use genotypes from only a few loci will remain important in malaria research. Indeed, one important result from this study is a set of new microsatellite loci that can be used fine scale population genetic and molecular epidemiological studies in Colombia. Microsatellite loci are particularly useful for such studies, as their relatively high mutation rates can generate highly polymorphic loci. As a result, population genetic signals in microsatellite loci can reflect demographic events occurring at short time scales, including epidemiological events [45, 63]. These markers can also be used for population assignment and testing for multiple infection [20, 21, 64, 65].

Thus far, 160 microsatellites have been found in the genome of *P. vivax* [21], however many of these loci fail to amplify in some populations [20, 21]. It is not surprising that loci developed in one region are not necessarily informative in others: microsatellites have complex evolutionary histories [66] and high potential for homoplasy [67]. Thus, the widespread application of microsatellite loci to epidemiological problems will require the development of new markers known to amplify and be polymorphic within specific population. We found 789 putatively polymorphic microsatellite loci from our whole genome sequencing, demonstrating that the 160 markers currently used in *P. vivax* represent only a small proportion of loci available in this species. Moreover, we

demonstrated that loci we detect in our whole genome sequencing can be developed into useful makers. The 18 markers we developed yield patterns of repeats that are easy to score in populations in the Pacific Coast of Colombia and how high levels of polymorphism. Whether these markers will be useful at broader geographic scales remains to be seen, but the specific markers we developed will be useful for fine-scale studies in this region, where malaria elimination is currently been considered.

Conclusion

Our results add to growing evidence that *P. vivax* populations in the Americas are genetically more diverse than was previously proposed. Our study supports that the demographic history of malarial parasite populations in South America is far more complex than previously believed. Indeed, even at a small spatial scale, *P. vivax* populations could harbor extraordinary genetic diversity. Our study also demonstrates that genomic studies of natural populations of *P. vivax* can provide insights into how the parasite populations react to control strategies. Specifically, we identified a selective sweep associated with resistance for SP, a drug that is not used to treat *P. vivax* in Colombia. This resistance indicates a spillover effect of a drug that was primarily used to treat *P. falciparum*. The operational consequences of these results require additional investigations. Future studies with additional samples may detect additional regions under selection, and thus contribute to the identification of vaccine targets [68] or other clinically relevant phenotypes. Finally, we used our genomic data to develop a set of microsatellite markers that are both easy to genotype and known to be polymorphic within this population. These markers will aid future epidemiological studies and aid our understanding of malaria transmission and demography in Colombia.

Supporting Information Legends

S1 Text

Methods used for microsatellite devolpment

S1 Table

Characterization of 18 polymorphic *P. vivax* microsatellite loci. Size ranges of PCR products (in base pairs) are given for a small set of Colombian *P. vivax* (6) isolates analyzed. SalI was used as a positive control. Fluorescent dyes (Hex and 6-FAM) were used to label forward primers only. ML: Motif length and No.A: allele numbers.

S2 Table

Extreme diversity windows. 10kb windows with unusually high or low genetic diversity. Values for θ_w and π are $\times 10^{-3}$. The values in the "Start" column are the position of start of the genomic window in kb.

S1 Fig

Electropherograms showing peaks profiles for 18 polymorphic microsatellite loci. The y-axis correspond to fluorescence intensity (arbitrary units) and the x-axis is the PCR product length in base pairs (bp). The amplitude of the each peak in base pairs (bp) is shown in boxes underneath peaks. The allele range size for these small data set is also given for each locus.

Acknowledgments

References

1. WHO. World malaria report 2013. World Health Organization; 2014.
2. Galinski MR, Meyer EVS, Barnwell JW. *Plasmodium vivax*. In: The Epidemiology of *Plasmodium vivax*. Elsevier BV; 2013. p. 1–26. Available from: http://dx.doi.org/10.1016/B978-0-12-407826-0.00001-1.
3. Prajapati SK, Joshi H, Carlton JM, Rizvi MA. Neutral Polymorphisms in Putative Housekeeping Genes and Tandem Repeats Unravels the Population Genetics and Evolutionary History of *Plasmodium vivax* in India. PLOS Neglected Tropical Diseases. 2013 Sep;7(9):e2425. Available from: http://dx.doi.org/10.1371/journal.pntd.0002425.
4. Schneider KA, Escalante AA. Fitness components and natural selection: why are there different patterns on the emergence of drug resistance in *Plasmodium falciparum* and Plasmodium vivax? Malar J. 2013;12(1):15. Available from: http://dx.doi.org/10.1186/1475-2875-12-15.
5. Otten M, Aregawi M, Were W, Karema C, Medin A, Jima D, et al. Initial evidence of reduction of malaria cases and deaths in Rwanda and Ethiopia due to rapid scale-up of malaria prevention and treatment. Malar J. 2009;8(1):14. Available from: http://dx.doi.org/10.1186/1475-2875-8-14.
6. Escalante AA, Cornejo OE, Freeland DE, Poe AC, Durrego E, Collins WE, et al. A monkey's tale: The origin of *Plasmodium vivax* as a human malaria parasite. Proceedings of the National Academy of Sciences. 2005 Jan;102(6):1980–1985. Available from: http://dx.doi.org/10.1073/pnas.0409652102.
7. Prugnolle F, Rougeron V, Becquart P, Berry A, Makanga B, Rahola N, et al. Diversity, host switching and evolution of *Plasmodium vivax* infecting African great apes. Proceedings of the National Academy of Sciences. 2013 May;110(20):8123–8128. Available from: http://dx.doi.org/10.1073/pnas.1306004110.
8. Mueller I, Galinski MR, Baird JK, Carlton JM, Kochar DK, Alonso PL, et al. Key gaps in the knowledge of *Plasmodium vivax*, a neglected human malaria parasite. The Lancet infectious diseases. 2009;9(9):555–566.
9. Mu J, Myers RA, Jiang H, Liu S, Ricklefs S, Waisberg M, et al. *Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. Nature Genetics. 2010;42(3):268–271.
10. Ochola LI, Tetteh KKA, Stewart LB, Riitho V, Marsh K, Conway DJ. Allele Frequency-Based and Polymorphism-Versus-Divergence Indices of Balancing Selection in a New Filtered Set of Polymorphic Genes in *Plasmodium falciparum*. Molecular Biology and Evolution. 2010 May;27(10):2344–2351. Available from: http://dx.doi.org/10.1093/molbev/msq119.
11. Amambua-Ngwa A, Tetteh KKA, Manske M, Gomez-Escobar N, Stewart LB, Deerhake ME, et al. Population Genomic Scan for Candidate Signatures of Balancing Selection to Guide Antigen Characterization in Malaria Parasites. PLOS Genetics. 2012 Nov;8(11):e1002992. Available from: http://dx.doi.org/10.1371/journal.pgen.1002992.

12. Park DJ, Lukens AK, Neafsey DE, Schaffner SF, Chang HH, Valim C, et al. Sequence-based association and selection scans identify drug resistance loci in the *Plasmodium falciparum* malaria parasite. Proceedings of the National Academy of Sciences. 2012 Jul;109(32):13052–13057. Available from: http://dx.doi.org/10.1073/pnas.1210585109.

13. Jongwutiwes S, Putaporntip C, Iwasaki T, Ferreira MU, Kanbara H, Hughes AL. Mitochondrial genome sequences support ancient population expansion in*Plasmodium vivax*. Molecular Biology and Evolution. 2005;22(8):1733–1739.

14. Cornejo OE, Escalante AA. The origin and age of *Plasmodium vivax*. Trends in parasitology. 2006;22(12):558–563.

15. Imwong M, Nair S, Pukrittayakamee S, Sudimack D, Williams JT, Mayxay M, et al. Contrasting genetic structure in *Plasmodium vivax* populations from Asia and South America. International Journal for Parasitology. 2007;37(8):1013–1022.

16. Chenet SM, Branch OH, Escalante AA, Lucas CM, Bacon DJ. Genetic diversity of vaccine candidate antigens in *Plasmodium falciparum* isolates from the Amazon basin of Peru. Malar Journal. 2008;7(1):93. Available from: http://dx.doi.org/10.1186/1475-2875-7-93.

17. Flannery EL, Wang T, Akbari A, Corey VC, Gunawan F, Bright AT, et al. Next-Generation Sequencing of*Plasmodium vivax* Patient Samples Shows Evidence of Direct Evolution in Drug-Resistance Genes. ACS Infect Dis. 2015 Aug;1(8):367–379. Available from: http://dx.doi.org/10.1021/acsinfecdis.5b00049.

18. Neafsey DE, Galinsky K, Jiang RH, Young L, Sykes SM, Saif S, et al. The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. Nature genetics. 2012;44(9):1046–1050.

19. Taylor JE, Pacheco MA, Bacon DJ, Beg MA, Machado RLD, Fairhurst RM, et al. The evolutionary history of *Plasmodium vivax* as inferred from mitochondrial genomes: parasite genetic diversity in the Americas. Molecular Biology and Evolution. 2013;p. mst104.

20. Escalante A, Ferreira M, Vinetz J, Volkman S, Cui L, Gamboa D, et al. Malaria Molecular Epidemiology: Lessons from the International Centers of Excellence for Malaria Research Network. The American Journal of Tropical Medicine and Hygiene. In Press;.

21. Sutton PL. A call to arms: on refining *Plasmodium vivax* microsatellite marker panels for comparing global diversity. Malar J. 2013;12:447.

22. Arévalo-Herrera M, Lopez-Perez M, Medina L, Moreno A, Gutierrez JB, Herrera S. Clinical profile of *Plasmodium falciparum* and *Plasmodium vivax* infections in low and unstable malaria transmission settings of Colombia. Malar J. 2015 Apr;14(1). Available from: http://dx.doi.org/10.1186/s12936-015-0678-3.

23. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9(4):357–359.

24. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, et al. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature. 2008;455(7214):757–763.

25. Arisue N, Hashimoto T, Mitsui H, Palacpac NM, Kaneko A, Kawai S, et al. The *Plasmodium* apicoplast genome: conserved structure and close relationship of *P. ovale* to rodent malaria parasites. Molecular Biology and Evolution. 2012;29(9):2095–2099.

26. Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics. 2013;p. 11–10.

27. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;43(5):491–498.

28. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20(9):1297–1303.

29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–2079.

30. Chan ER, Menard D, David PH, Ratsimbasoa A, Kim S, Chim P, et al. Whole Genome Sequencing of Field Isolates Provides Robust Characterization of Genetic Diversity in *Plasmodium vivax*. PLOS Neglectd Tropical Diseases. 2012 Sep;6(9):e1811+. Available from: `http://dx.doi.org/10.1371/journal.pntd.0001811`.

31. Koepfli C, Ross A, Kiniboro B, Smith TA, Zimmerman PA, Siba P, et al. Multiplicity and diversity of *Plasmodium vivax* infections in a highly endemic region in Papua New Guinea. PLOS neglected tropical diseases. 2011;5(12):e1424.

32. Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;27(12):1691–1692.

33. Li H. Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples. arXiv preprint arXiv:14040929. 2014;.

34. Chan ER, Barnwell JW, Zimmerman PA, Serre D. Comparative Analysis of Field-Isolate and Monkey-Adapted *Plasmodium vivax* Genomes. PLOS Neglected Tropical Diseases. 2015 Mar;9(3):e0003566. Available from: `http://dx.doi.org/10.1371/journal.pntd.0003566`.

35. Cingolani P, Platts A, Coon M, Nguyen T, Wang L, Land SJ, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly. 2012;6(2):80–92.

36. Ferretti L, Raineri E, Ramos-Onsins S. Neutrality tests for sequences with missing data. Genetics. 2012;191(4):1397–1401.

37. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics. 2011;27(24):3423–3424.

38. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–842.

39. Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;105(2):437–460.

40. Watterson G. On the number of segregating sites in genetical models without recombination. Theoretical population biology. 1975;7(2):256–276.

41. Auliff A, Wilson DW, Russell B, Gao Q, Chen N, Le Ngoc A, et al. Amino acid mutations in *Plasmodium vivax* DHFR and DHPS from several geographical regions and susceptibility to antifolate drugs. The American Journal of Tropical Medicine and Hygiene. 2006;75(4):617–621.

42. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genetical research. 1974;23(01):23–35.

43. Kaplan NL, Hudson R, Langley C. The "hitchhiking effect" revisited. Genetics. 1989;123(4):887–899.

44. Nygaard S, Braunstein A, Malsen G, Van Dongen S, Gardner PP, Krogh A, et al. Long-and short-term selective forces on malaria parasite genomes. PLOS Genetics. 2010;6(9):e1001099.

45. Chenet SM, Schneider KA, Villegas L, Escalante AA. Local population structure of Plasmodium: impact on malaria control and elimination. Malar J. 2012;11(412):00576–10.

46. Gupta B, Srivastava N, Das A. Inferring the evolutionary history of Indian *Plasmodium vivax* from population genetic analyses of multilocus nuclear DNA fragments. Molecular Ecology. 2012;21(7):1597–1616.

47. Cornejo OE, Fisher D, Escalante AA. Genome-wide patterns of genetic polymorphism and signatures of selection in *Plasmodium vivax*. Genome biology and evolution. 2015;7(1):106–119.

48. Barnadas C, Tichit M, Bouchier C, Ratsimbasoa A, Randrianasolo L, Raherinjafy R, et al. *Plasmodium vivax dhfr* and *dhps* mutations in isolates from Madagascar and therapeutic response to sulphadoxine-pyrimethamine. Malar J. 2008;7(1):35. Available from: `http://dx.doi.org/10.1186/1475-2875-7-35`.

49. Prajapati SK, Joshi H, Dev V, Dua VK. Molecular epidemiology of *Plasmodium vivax* anti-folate resistance in India. Malar J. 2011;10(1):102. Available from: `http://dx.doi.org/10.1186/1475-2875-10-102`.

50. Afsharpad M, Zakeri S, Pirahmadi S, Djadid ND. Molecular assessment of *dhfr*/*dhps* mutations among *Plasmodium vivax* clinical isolates after introduction of sulfadoxine/pyrimethamine in combination with artesunate in Iran. Infection, Genetics and Evolution. 2012 Jan;12(1):38–44. Available from: `http://dx.doi.org/10.1016/j.meegid.2011.10.003`.

51. Raza A, Ghanchi NK, Khan M, Beg M. Prevalence of drug resistance associated mutations in *Plasmodium vivax* against sulphadoxine-pyrimethamine in southern Pakistan. Malar J. 2013;12(1):261. Available from: `http://dx.doi.org/10.1186/1475-2875-12-261`.

52. Thongdee P, Kuesap J, Rungsihirunrat K, Tippawangkosol P, Mungthin M, Na-Bangchang K. Distribution of dihydrofolate reductase (*dhfr*) and dihydropteroate synthase (*dhps*) mutant alleles in *Plasmodium vivax* isolates from Thailand. Acta Tropica. 2013 Oct;128(1):137–143. Available from: `http://dx.doi.org/10.1016/j.actatropica.2013.07.005`.

53. Huang F, Zhou S, Zhang S, Li W, Zhang H. Monitoring resistance of *Plasmodium vivax*: Point mutations in dihydrofolate reductase gene in isolates from Central China. Parasites & Vectors. 2011;4(1):80. Available from: http://dx.doi.org/10.1186/1756-3305-4-80.

54. Pearce RJ, Pota H, Evehe MSB, Bâ EH, Mombo-Ngoma G, Malisa AL, et al. Multiple Origins and Regional Dispersal of Resistant *dhps* in African *Plasmodium falciparum* Malaria. PLOS Med. 2009 Apr;6(4):e1000055. Available from: http://dx.doi.org/10.1371/journal.pmed.1000055.

55. McCollum AM, Basco LK, Tahar R, Udhayakumar V, Escalante AA. Hitchhiking and Selective Sweeps of *Plasmodium falciparum* Sulfadoxine and Pyrimethamine Resistance Alleles in a Population from Central Africa. Antimicrobial Agents and Chemotherapy. 2008 Sep;52(11):4089–4097. Available from: http://dx.doi.org/10.1128/AAC.00623-08.

56. Vinayak S, Alam MT, Mixson-Hayden T, McCollum AM, Sem R, Shah NK, et al. Origin and Evolution of Sulfadoxine Resistant *Plasmodium falciparum*. PLOS Pathogens. 2010 Mar;6(3):e1000830. Available from: http://dx.doi.org/10.1371/journal.ppat.1000830.

57. McCollum AM, Schneider KA, Griffing SM, Zhou Z, Kariuki S, Ter-Kuile F, et al. Differences in selective pressure on *dhps* and *dhfr* drug resistant mutations in western Kenya. Malar J. 2012;11(1):77. Available from: http://dx.doi.org/10.1186/1475-2875-11-77.

58. Raj DK, Mu J, Jiang H, Kabat J, Singh S, Sullivan M, et al. Disruption of a *Plasmodium falciparum* Multidrug Resistance-associated Protein (PfMRP) Alters Its Fitness and Transport of Antimalarial Drugs and Glutathione. Journal of Biological Chemistry. 2008 Dec;284(12):7687–7696. Available from: http://dx.doi.org/10.1074/jbc.M806944200.

59. Dahlstrom S, Veiga MI, Martensson A, Bjorkman A, Gil JP. Polymorphism in PfMRP1 (Plasmodium falciparum Multidrug Resistance rotein 1) Amino Acid 1466 Associated with Resistance to Sulfadoxine-Pyrimethamine Treatment. Antimicrobial Agents and Chemotherapy. 2009 Apr;53(6):2553–2556. Available from: http://dx.doi.org/10.1128/AAC.00091-09.

60. Dharia NV, Bright AT, Westenberger SJ, Barnes SW, Batalov S, Kuhen K, et al. Whole-genome sequencing and microarray analysis of ex vivo *Plasmodium vivax* reveal selective pressure on putative drug resistance genes. Proceedings of the National Academy of Sciences. 2010 Oct;107(46):20045–20050. Available from: http://dx.doi.org/10.1073/pnas.1003776107.

61. Lin JT, Patel JC, Kharabora O, Sattabongkot J, Muth S, Ubalee R, et al. Plasmodium vivax Isolates from Cambodia and Thailand Show High Genetic Complexity and Distinct Patterns of P. vivax Multidrug Resistance Gene 1 (pvmdr1) Polymorphisms. American Journal of Tropical Medicine and Hygiene. 2013 Mar;88(6):1116–1123. Available from: http://dx.doi.org/10.4269/ajtmh.12-0701.

62. Suwanarusk R, Russell B, Chavchich M, Chalfein F, Kenangalem E, Kosaisavee V, et al. Chloroquine Resistant*Plasmodium vivax*: In Vitro Characterisation and Association with Molecular Polymorphisms. PLOS ONE. 2007 Oct;2(10):e1089. Available from: http://dx.doi.org/10.1371/journal.pone.0001089.

63. Haasl RJ, Payseur BA. Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. Heredity. 2010 Mar;106(1):158–171. Available from: http://dx.doi.org/10.1038/hdy.2010.21.

64. Havryliuk T, an Marcelo U Ferreira POS. *Plasmodium vivax*: Microsatellite analysis of multiple-clone infections. Experimental Parasitology. 2008 Dec;120(4):330–336. Available from: http://dx.doi.org/10.1016/j.exppara.2008.08.012.

65. Orjuela-Sánchez P, Sá JM, Brandi MC, Rodrigues PT, Bastos MS, Amaratunga C, et al. Higher microsatellite diversity in *Plasmodium vivax* than in sympatric *Plasmodium falciparum* populations in Pursat, Western Cambodia. Experimental parasitology. 2013;134(3):318–326.

66. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004 Jun;5(6):435–445. Available from: http://dx.doi.org/10.1038/nrg1348.

67. Estoup A, Jarne P, Cornuet JM. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Molecular Ecology. 2002 Sep;11(9):1591–1604. Available from: http://dx.doi.org/10.1046/j.1365-294X.2002.01576.x.

68. Conway DJ. Paths to a malaria vaccine illuminated by parasite genomics. Trends in Genetics. 2015 Feb;31(2):97–107. Available from: http://dx.doi.org/10.1016/j.tig.2014.12.005.

Table 2. Summary of sequencing data.

Sample	Reads (millions)	Proportion of reads mapped to SalI	Mean Coverage	p5	p10	p20	Proportion of minority bases
446	27.754	0.007	1.7	0.041	0.001	0.000	7.3×10^{-4}
494	33.337	0.043	10.9	0.972	0.733	0.064	1.11×10^{-3}
495	29.129	0.084	19.8	0.993	0.979	0.636	8.5×10^{-4}
496	35.048	0.106	29.5	0.994	0.990	0.952	1.1×10^{-4}
498	24.365	0.010	1.9	0.075	0.001	0.000	7.8×10^{-3}
499	18.224	0.278	42.1	0.995	0.993	0.986	9.1×10^{-3}
500	25.050	0.004	0.8	0.005	0.000	0.000	7.6×10^{-3}
503	23.108	0.066	12.2	0.9830	0.825	0.100	9.1×10^{-3}

"p5", "p10" and "p20" refer to the proportion of sites covered by at least 5, 10 and 20 reads respectively."Proportion of minority bases" refers to the proportion of mapped-reads at a given site which contained a base that base that made up a minor fraction of all mapped bases.

Table 3. Summary of SNV data

	Total	Exonic	Intronic	3'UTR	5UTR	Intergenic
Number of SNVs	33855	16413	3107	808	1328	14333
Density	1.8	1.2	2.4	1.9	1.8	2.0
θ_w	7.0×10^{-4}	5.9×10^{-4}	7.9×10^{-4}	6.5×10^{-4}	6.4×10^{-4}	7.5×10^{-4}
π	6.8×10^{-4}	5.7×10^{-4}	7.5×10^{-4}	6.2×10^{-4}	6.1×10^{-4}	7.2×10^{-4}

Table 4. Breakdown of singletons

Sample	Singletons this study	Singletons all studies	Callable sites
446	660 (78.7)	242 (28.9)	38.1
498	10 (1)	8 (0.8)	45.4
499	2974 (158.5)	1623 (86.5)	85.3
494	3134 (168.0)	1606 (86.1)	84.8
495	837 (44.7)	520 (27.8)	85.1
496	2361 (126.0)	1291 (68.9)	85.2
500	122 (35.3)	32 (9.3)	15.7
503	2806 (150.1)	1523 (81.4)	85.0
Total	12913 (111.9)	6854 (59.4)	–

Values in parenthesis are number of singleton SNVs per million callable sites.
Values in Callable sites column is the proportion of all sites that were callable for a given sample.

Table 5. Summary of microsatellite loci discovered in this study

Name	Position	Motif	Repeats	Alleles (genomes)	Alleles (validation)
CLAIM1	2:96021	ATGC	5–7	4	4
CLAIM2	2:261105	AC	6–13	4	4
CLAIM3	5:280204	AACAGC	8–11	4	3
CLAIM4	5:1139547	AT	8–9	3	4
CLAIM5*	5:1195225	AAAT	6–10	4	4
CLAIM6*	5:1248584	AT	8–10	3	2
CLAIM7	6:436676	AT	8–9	3	1
CLAIM8*	7:761767	AT	8–9	3	2
CLAIM9	7:1343057	AG	8–12	3	1
CLAIM10*	8:439431	AT	8–10	4	4
CLAIM11	9:1048573	AAAT	6–7	3	4
CLAIM12*	9:1454922	AT	8–9	3	2
CLAIM13*	10:1348254	AT	8–9	3	2
CLAIM14	10:1368121	AT	3–8	3	3
CLAIM15	12:2076009	AC	7–11	4	3
CLAIM16*	12:2101147	AT	8–9	3	2
CLAIM17*	13:950389	AT	7–8	3	2
CLAIM18	14:1469652	AT	6–14	4	3

Loci marked with an asterisk were developed by comparing *P. vivax* sequences to *P. cynmolgi*. For the numbers of alleles, "genomes" refers to the number identified from genomic data and "validation" to the number detected from the validation panel. PCR primers for each locus are given in S1 Table:.

Table 6. Non-synonynous variants at drug resistance loci

Locus	Chromosome	Position	Codon	Codon	Number of samples
dhps	14	1257856	gCc/gGc	A383G	8 (8)
dhfr	5	964761	Agc/Cgc	S58R*	4 (7)
	5	964763	agC/agA*	S58R*	3 (7)
	5	964939	aGc/aAc	S117N*	6 (6)
pvmdr1	10	363169	tAc/tTc	Y976F	3 (8)
	10	363223	aCg/aTg*	T958M*	8 (8)
	10	363374	Atg/Ctg	M908L	8 (8)
	10	364598	Gat/Aat	D500N	4 (8)
	10	365435	Gtg/Ttg	V221L	5 (8)
PVX_124085	14	2043859	Caa/Gaa*	Q1419E*	1 (7)
	14	2044528	Ccg/Tcg	P1196S	3 (8)
	14	2044798	Gcg/Tcg*	A1106S*	2 (6)
	14	2044975	Ctg/Atg	L1047M	1 (7)
	14	2045050	Gtg/Atg*	V1022M*	7 (7)
	14	2046072	aGc/aTc	S681I	3 (6)
	14	2047233	aGg/aTg*	R294M*	7 (7)
	14	2047816	Gaa/Caa*	E100Q*	2 (6)

"Number of samples" refers to the number of genomic sequences that contained this allele in the present study, the number of samples from which any allele could be reliably called appears in parentheses. Variants marked with an asterisk were also recorded from a population in the Amazon basin in Peru [17]. Note the numbering of amino acid substitutions in PVX_124085 differs between these studies as a result of the different annotations used in each case.