

RESEARCH

Open Access



# Exploration of *Plasmodium vivax* transmission dynamics and recurrent infections in the Peruvian Amazon using whole genome sequencing

Annie N. Cowell<sup>1\*†</sup> , Hugo O. Valdivia<sup>2†</sup>, Danett K. Bishop<sup>2</sup> and Elizabeth A. Winzeler<sup>3</sup>

## Abstract

**Background:** *Plasmodium vivax* poses a significant challenge to malaria elimination due to its ability to cause relapsed infections from reactivation of dormant liver parasites called hypnozoites. We analyzed 69 *P. vivax* whole genome sequences obtained from subjects residing in three different villages along the Peruvian Amazon. This included 23 paired *P. vivax* samples from subjects who experienced recurrent *P. vivax* parasitemia following observed treatment with chloroquine and primaquine.

**Methods:** Genomic DNA was extracted from whole blood samples collected from subjects. *P. vivax* DNA was enriched using selective whole genome amplification and whole genome sequencing. We used single nucleotide polymorphisms (SNPs) from the core *P. vivax* genome to determine characteristics of the parasite population using discriminant analysis of principal components, maximum likelihood estimation of individual ancestries, and phylogenetic analysis. We estimated the relatedness of the paired samples by calculating the number of segregating sites and using a hidden Markov model approach to estimate identity by descent.

**Results:** We present a comprehensive dataset of population genetics of *Plasmodium vivax* in the Peruvian Amazonian. We define the parasite population structure in this region and demonstrate a novel method for distinguishing homologous relapses from reinfections or heterologous relapses with improved accuracy. The parasite population in this area was quite diverse with an estimated five subpopulations and evidence of a highly heterogeneous ancestry of some of the isolates, similar to previous analyses of *P. vivax* in this region. Pairwise comparison of recurrent infections determined that there were 12 homologous relapses and 3 likely heterologous relapses with highly related parasites. To the best of our knowledge, this is the first large-scale study to evaluate recurrent *P. vivax* infections using whole genome sequencing.

**Conclusions:** Whole genome sequencing is a high-resolution tool that can identify *P. vivax* homologous relapses with increased sensitivity, while also providing data about drug resistance and parasite population genetics. This information is important for evaluating the efficacy of known and novel antirelapse medications in endemic areas and thus advancing the campaign to eliminate malaria.

**Keywords:** Malaria, *Plasmodium vivax*, Relapse, Primaquine, Whole genome sequencing, Population genetics

\* Correspondence: [annie.cowell@gmail.com](mailto:annie.cowell@gmail.com)

†Annie N. Cowell and Hugo O. Valdivia contributed equally to this work.

<sup>1</sup>Division of Infectious Diseases, Department of Medicine, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA

Full list of author information is available at the end of the article



## Background

Malaria is a tropical disease caused by *Plasmodium* parasites which remains one of the most important public health problems worldwide [1]. The disease is endemic in more than 90 countries and poses a risk to nearly 2.5 billion people with an estimated 212 million cases and 429,000 deaths in 2015 [1]. Of the five species known to infect humans, *P. falciparum* and *P. vivax* stand out as the leading causes of malaria in endemic areas. Although *P. vivax* infections are not as deadly as those caused by *P. falciparum*, it is the most geographically widespread malaria species leading to enormous morbidity and severe disease [2, 3]. Implementation of malaria control strategies has significantly reduced malaria incidence and deaths between 2000 and 2015 [4]. This rapid reduction has been especially important for the control of *P. falciparum* in several endemic regions in the Americas; however, *P. vivax* has now replaced *P. falciparum* as the predominant species outside Africa [1]. The reasons for this change are related to the unique biological features of *P. vivax* including (i) the higher potential of transmission and infectivity to other species of mosquitoes compared to other *Plasmodium* species [5] and (ii) the ability to generate long lasting dormant hepatic parasites (hypnozoites) that can become active weeks, months, or years after the first infection, resulting in relapse [6].

Relapses represent a major threat to malaria elimination worldwide because hypnozoites are undetectable by current diagnostic tests [2] and present a new opportunity for malaria transmission once activated. Conventional medications used to treat blood-stage infections such as chloroquine are not efficacious against hypnozoites. Currently, the only treatment licensed to prevent *Plasmodium vivax* relapses through killing hypnozoites is primaquine, which frequently causes gastrointestinal side effects, poses a risk of hemolysis in people with G6PD deficiency [2, 7], and has decreased efficacy in people with mutations in the *cyp2d6* gene that encodes cytochrome P450 2D6 [8]. Furthermore, measuring primaquine efficacy is challenging in many endemic sites because administration is not adequately supervised. Additionally, because relapses can be due to activation of hypnozoites from the most recent infection (homologous relapse) or activation of hypnozoites from prior infections (heterologous relapse), it remains difficult to distinguish whether recurrent parasitemia is due to relapse or reinfection from a new mosquito bite.

Whole genome sequencing (WGS) can enable highly detailed comparisons of recurrent *P. vivax* infections [9] and thus can identify homologous relapses with greater accuracy. It also provides further information about parasite population structure, polymorphisms in drug resistance markers, and genomic regions under selection [10, 11]. Previous methods used to distinguish

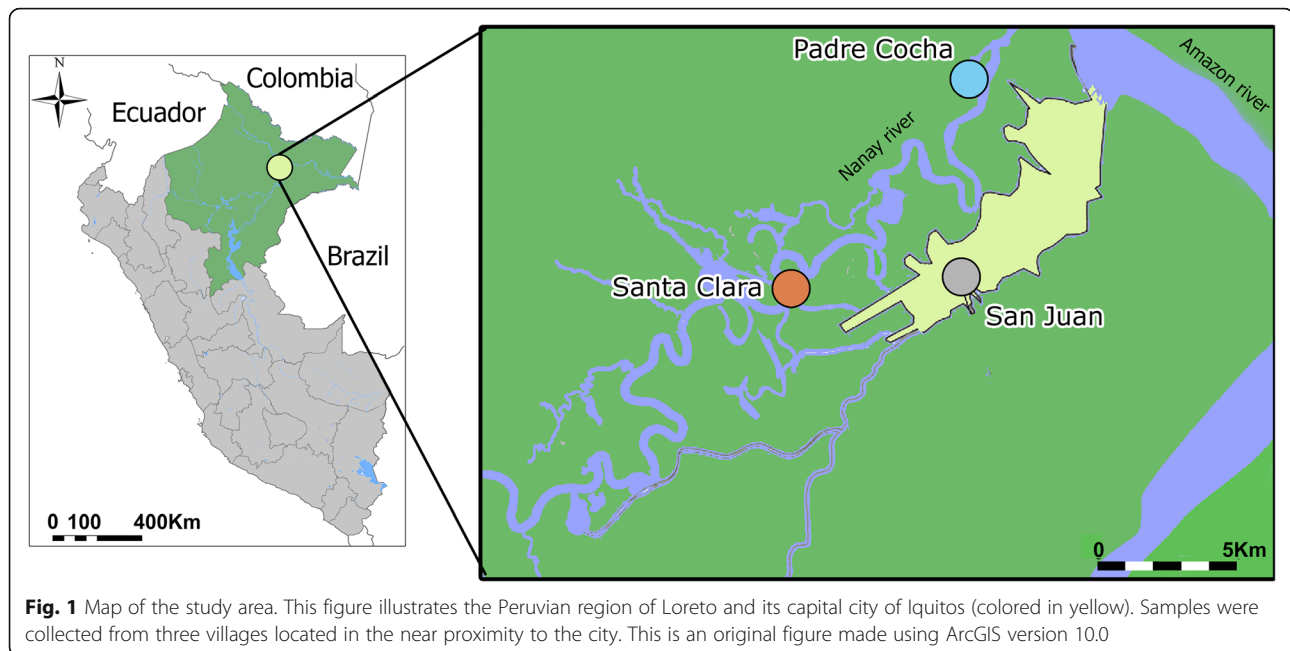
homologous relapses from reinfections include microsatellite marker comparison [12, 13] and deep sequencing of hypervariable genes such as merozoite surface protein 1 (*msp1*) [14]. However, these methods have a limited resolution that could affect an accurate differentiation of recurrent infections. For example, a prior study of a traveler who returned to a malaria non-endemic region compared *P. vivax* whole genome sequences from subsequent episodes of recurrent parasitemia and demonstrated that relapses could occur with a meiotic sibling, or closely related *P. vivax* parasite strain, likely representing recombination that had occurred in the mosquito midgut at the time of the initial infection [9]. In this case, using microsatellite markers alone to compare infections would have mistakenly determined that the recurrent infection was due to reinfection rather than relapse.

Here, we analyze 69 *P. vivax* whole genome sequences obtained from 46 subjects living in three villages near the city of Iquitos in the Peruvian Amazon region of Loreto (Fig. 1). This genomic set includes 46 *P. vivax* sequences that were derived from 23 paired samples obtained from the same subject collected before treatment with primaquine and chloroquine and at the time of recurrent *P. vivax* parasitemia following treatment. Genomic data was used to assess gene diversity, population structure, and drug resistance patterns within the population. In addition, we compared the 23 *P. vivax* paired samples to assess whether they represented homologous relapses or were more likely to be reinfections or heterologous relapses.

## Methods

### Subject sample collection and preparation

Whole blood samples were collected from subjects with symptomatic *P. vivax* infections from the endemic region of Iquitos in the northeastern Peruvian Amazon during a previous study conducted by the US Naval Medical Research Unit 6 (NAMRU-6) between 2006 and 2008 to evaluate three different primaquine regimens: 0.5 mg/kg  $\times$  5 days, 0.5 mg/kg  $\times$  7 days, and 0.25 mg/kg  $\times$  14 days [15]. This included 23 pairs of samples collected from subjects at two time points: initial infection prior to treatment with chloroquine and primaquine, and recurrent parasitemia between 36 to 210 days after observed treatment (Additional file 1: Table S1). Since one goal of the current study was to identify genetic markers or primaquine resistance, subjects with recurrent parasitemia between 17 and 35 days following treatment were considered to have potential chloroquine resistance and were excluded from this analysis [16]. Most subjects who received the shortest regimen (0.5 mg/kg  $\times$  5 days) were excluded from the present analysis since they had statistically higher rates of relapse compared to the other regimens [15]. Thick blood



smears were examined to identify the parasite species and to determine the level of parasitemia. Parasite density was calculated by counting the number of asexual parasites per 200 white blood cells in the thick smear (assuming an average of 6000 white blood cells per  $\mu\text{l}$ ). Two microscopists examined each blood smear independently, and a third microscopist gave confirmation in the event of a discrepancy. The final parasite density was calculated as the average of density readings from the two concordant microscopists. Microsatellite genotyping was performed using six neutral microsatellite markers as previously described [15]. Whole blood samples were collected in the field using EDTA-containing vacutainer tubes, and samples were frozen and transported to the central laboratory for further processing.

#### Selective whole genome amplification (SWGA)

DNA was isolated from thawed whole blood using QIAamp DNA Blood Mini Kit (Qiagen) following the manufacturer's recommendation and as described elsewhere [17]. Samples were subsequently resuspended in TE buffer, and genomic DNA was quantified using a Qubit 2.0 fluorometer (ThermoFisher). Thirty to 70 ng of input DNA was added to a 50- $\mu\text{l}$  reaction containing 3.5  $\mu\text{M}$  SWGA primers, 30 U phi29 DNA polymerase enzyme (New England Biolabs), phi29 DNA buffer (New England Biolabs), 1% bovine serum albumin, and water as previously described [18, 19]. The primer set used consists of 12 primers: 5'-AACGAAGC\*G\*A-3', 5'-ACGAAGCG\*A\*A-3', 5'-ACGACGA\*A\*G-3', 5'-ACGCGCA\*A\*C-3', 5'-CAACGCG\*G\*T-3', 5'-GACGAAA\*C\*G-3', 5'-GCGAAAAA\*G\*G-3', 5'-GCGAAGC\*G\*A-3',

5'-GCGGAAC\*G\*A-3', 5'-GCGTCGA\*A\*G-3', 5'-GGT TAGCG\*G\*C-3', and AACGAAT\*C\*G. The reaction was carried out on a thermocycler and consisted of a ramp down from 35 to 30  $^{\circ}\text{C}$  (10 min per degree), 16 h at 30  $^{\circ}\text{C}$ , 10 min at 65  $^{\circ}\text{C}$ , and hold at 4  $^{\circ}\text{C}$ . The samples were diluted 1:1 with DNase-free and RNase-free water and purified with Ampure XP beads (Beckman-Coulter) at a 1:1 ratio per the manufacturer's protocol.

#### Whole genome sequencing

Next-generation sequencing libraries of SWGA products were prepared using the Nextera XT DNA preparation kit (Illumina) per the manufacturer's protocol. These samples were pooled and clustered on a HiSeq 2500 (Illumina) in Rapid Run mode with 100 base pair paired end reads. Raw fastq files were aligned to the Sal-1 reference genome (PlasmoDB version 13, <http://plasmodb.org/common/downloads/release-13.0/PvivaxSal1/fasta/data/>) using the Burroughs-Wheeler Aligner (version 0.7.8) [20] and samtools (version 0.1.19) [21, 22] as previously described in the Platypus pipeline [23]. Picard (version 2.0.1) was used to remove unmapped reads, and the Genome Analysis Toolkit (GATK) [24] was used to realign the sequences around the indels.

#### Variant calling and analysis

We followed the GATK's best practices to call variants [25, 26]. The aligned sequences were run through GATK's HaplotypeCaller in "reference confidence" mode to create genomic GVCF files for each sample. This reference confidence model highlights areas of the genome that are likely to have variation and produces a

comprehensive record of genotype likelihoods and annotations for each site. The samples were joint genotyped using the GenotypeGVCFs tool. Variants were further filtered based on quality scores and sequencing bias statistics based on default parameters from GATK. SNPs were filtered out if they met any of the following criteria: quality depth (QD) < 2.0, mapping quality (MQ) < 50.0, phred-scaled  $p$  value using Fisher's exact test to detect strand bias (FS) > 60.0, symmetric odds ratio (SOR) > 4.0,  $Z$ -score from Wilcoxon rank sum test of alternative vs. reference read mapping qualities (MQRankSum) < -12.5, and ReadPosRankSum (RPRS) < -8.0. Variants were annotated using snpeff (version 4.2) [27]. SNP density was visualized in R for the detection of highly polymorphic regions. The core *P. vivax* genome, as defined by Pearson et al. [11], was used for further genome analysis.

$F_{ws}$  of samples with the highest genome coverage was estimated using *moimix* (<https://github.com/bahlolab/moimix>), a package available through R. The package calculates  $F_{ws}$  statistic using the equation  $F_{ws} = 1 - (Hw/Hs)$ , where  $Hw$  is the within-host heterozygosity and  $Hs$  is the population-level heterozygosity [28, 29].

#### Population structure analysis

The core *P. vivax* genome resulting from GATK was used for discriminant analysis of principal components using the adegenet package implemented in R [30]. DAPC constitutes a powerful tool for exploring the population structure without relying on a defined genetic model, Hardy-Weinberg equilibrium or linkage disequilibrium.

In order to provide an additional estimation of the parasite subpopulation in these foci, we performed maximum likelihood estimation of individual ancestries using ADMIXTURE [31]. This tool identifies the probability of membership of each individual to a cluster. We tested multiple runs by imputing successive values of  $K$  from 1 to 8 under a tenfold cross-validation procedure with 2000 pseudoreplicates under different initial seed values for each  $K$  and used ADMIXTURE's cross-validation to identify the most likely value of  $K$ . The pophelper R package and the software CLUMPP [32] were used to obtain the optimal alignments of the replicates for each  $K$  value and to make multiline plots.

#### Phylogenetic analysis

The core *P. vivax* genome was used to assess the phylogenetic relationship of the isolates collected in the Iquitos region. For this purpose, SNPs were used to generate genomic sequences for each isolate in GATK. Genomic sequences were subsequently aligned with MAFFT, and the resulting multiple sequence alignment was analyzed on jModelTest2 [33] for statistical selection of the best-fit models according to the Akaike information criterion (AIC), decision theory method (DT), and Bayesian

information criterion (BIC). A phylogenetic analysis was performed under a maximum likelihood approach on RAxML [34] using the general time reversible model selected by jModelTest with 1000 pseudoreplicates. Figtree v.1.4.24 was used to generate and visualize the resulting maximum likelihood tree.

#### Paired sample comparisons

We employed BioPerl to estimate the number of segregating sites between the baseline and post-treatment sample of each potential relapse and across all the permuted sample pairs in the population. In order to identify homologous relapses in the sampled population, we compared the number of segregating sites between the relapse pairs versus the average number of segregating sites on the permuted pairs minus 1.5 standard deviations. Potential relapse pairs were also screened under a hidden Markov model approach implemented in the glpsnort pipeline to detect genomic segments that could be identical by descent [35, 36]. In addition, we performed a direct pairwise comparison on filtered SNP data using custom scripts. SNPs were considered the same if they had the same read call. If the call at a locus was heterozygous for either samples, they were considered the same if  $\geq 80\%$  of reads at that locus for each sample were the same read call. Matlab was used to generate the comparison plots across the chromosomes. Polymorphisms of the cytochrome P450 2D6 gene in the homologous relapses were identified using the xTAG CYP2D6 kit (Luminex, USA) on a Luminex platform.

## Results

#### Sample collection and whole genome sequencing

The samples used in this study were collected during a clinical trial performed in three villages surrounding Iquitos, Peru, to assess the efficacy of three different primaquine regimens [15]. We obtained 69 high-quality *P. vivax* whole genome sequences directly from subject samples by selective whole genome amplification (SWGA) performed on genomic DNA (gDNA) extracted from whole blood samples [19]. We aligned these sequences to the *P. vivax* Salvador-1 reference genome and obtained an average of 24X coverage with  $61.1\% \pm 23.5$  of the genome covered by  $\geq 5$  reads (Additional file 2: Table S2). We identified a total of 24,571 high-quality single nucleotide polymorphisms (SNPs) in the core genome of this group of these sequences (Additional file 3: Table S3).

#### Chromosome SNP density

We investigated the density of SNPs in the 69 sequences at the chromosome level to identify regions of increased variability in the core genome. We identified regions with high SNP density on chromosomes 3, 6, 10, and 13. Genes located in these highly variable regions include

virulence factors such as members of the plasmodium interspersed repeat (*pir*) gene family (Chr 03) that mediate immune evasion and parasite host interaction [37], merozoite surface protein 8 (*msp8*) (Chr10) which is a potential *P. vivax* vaccine candidate [38], variant interspersed repeat 21 (*vir21*) (Chr 13) which participates in evasion via transcriptional switching [39], and several hypothetical proteins (PVX\_110960, PVX\_110955, PVX\_110950, PVX\_110945, PVX\_110940, and PVX\_110945) located on a SNP dense region on chromosome 6 (Additional file 4: Figure S1).

### Diversity of drug-resistant gene orthologs

Our understanding of the genetic changes that impart phenotypic drug resistance in *P. vivax* is significantly limited compared to *P. falciparum* mainly due to increased challenges with in vitro culture of the parasite and lack of validated drug susceptibility assays. Thus, drug resistance genes in *P. vivax* such as *pvmdr1* were previously identified based on orthologous genes in *P. falciparum*. In our sample set, we detected several SNPs in orthologs of known drug resistance genes in *P. falciparum* comprising up to 47 different haplotypes, consistent with prior whole genome sequencing studies of *P. vivax* from this region [40] (Table 1).

Although chloroquine resistance is common in *P. falciparum* in Peru, there is no current evidence of resistance in *P. vivax*, and thus, chloroquine remains the first-line treatment for infection [41]. We found several intronic changes in *pvcr1-0* (PVX\_087980), which encodes the chloroquine resistance transporter. These changes have previously been detected in a WGS study of *P. vivax* in Peru [40]; however, it is not currently well understood what functional changes these SNPs impart. We detected four missense SNPs in *pvmdr1* (PVX\_080100), which encodes the multidrug resistance-associated protein 1. Nonsynonymous SNPs in this gene have been associated with chloroquine resistance in prior in vitro assays, in particular a Y976F mutation [42]. While the T958M, M908L, and V221L mutations have been previously detected in Peru and other countries in South America [40, 43, 44], we report the F1070L mutation for the first time in Peru. This mutation is postulated to be a prerequisite for the subsequent acquisition of the Y976F [42] mutation in a two-step mutational path that leads to chloroquine resistance [45]. It has previously been detected in other regions of the world, including Thailand, Indonesia, Turkey, French Guyana, and Azerbaijan [45]. Other *P. falciparum* drug resistance-associated gene orthologs with nonsynonymous mutations were *pvdhfr* (PVX\_089950), which encodes the bifunctional dihydrofolate reductase-thymidylate synthase enzyme, and *dhps* (PVX\_123230), which encodes

**Table 1** Homozygous single nucleotide polymorphisms (SNPs) in *P. falciparum* drug resistance gene orthologs detected in the 69 *Plasmodium vivax* sequences

Locus	Chr	Position	Ref	Alt	Amino acid	Samples*
<i>pvcr1-0</i> (PVX_087980)	1	331,151	T	C	Intron	61 (67)
		331,819	G	A	Intron	29 (69)
		332,453	T	C	Intron	59 (68)
		332,874	A	C	Intron	67 (68)
<i>pvdhfr</i> (PVX_089950)	5	964,760	C	G	Phe57Leu	2 (66)
		964,762	G	A	Ser58Asn	4 (63)
		964,763	C	G,A	Ser58Arg	55 (59)
		964,939	G	A	Ser117Asn	58 (62)
<i>pvmdr1</i> (PVX_080100)	10	362,888	A	G	Phe1070Leu	6 (49)
		363,223	A	G	Thr958Met	44 (46)
		363,374	T	G	Met908Leu	46 (47)
		365,435	C	A	Val221Leu	9 (49)
Multidrug resistance protein 2 (PVX_118100)	12	2,413,549	A	T	Tyr514Phe	6 (58)
		2,415,530	G	T	Gln1174His	23 (57)
		2,415,666	G	A	Gly1220Ser	11 (60)
		2,416,390	C	T	Ala1461Val	1 (63)
		2,416,408	T	C	Val1467Ala	31 (41)
		2,416,420	T	C	Leu1471Pro	12 (62)
Multidrug resistance-associated protein 2 (PVX_124085)	14	2,043,859	G	C	Gln1407Glu	11 (13)
		2,045,050	C	T	Val1010Met	14 (15)
		2,047,090	G	A	Pro330Ser	1 (15)
		2,047,233	C	A	Arg282Met	13 (14)
		2,047,816	C	G	Glu88Gln	3 (18)
<i>dhps</i> (PVX_123230)	14	1,257,856	G	C	Ala383Gly	9 (15)
		1,258,389	C	T	Met205Ile	10 (12)

\*Samples column indicates the number of samples with the alternate allele, with the total number of samples genotyped in parentheses

the dihydropteroate synthetase enzyme. The role of these mutations in phenotypic drug resistance in *P. vivax* requires further exploration.

Among these genes, those encoding multidrug resistance protein 2 (PVX\_118100) and the multidrug resistance-associated protein 2 (PVX\_124085) stood out as having the highest number of SNPs. While the precise function of these genes in *P. vivax* has not been well studied, in *P. falciparum*, the multidrug resistance-associated protein 2 is considered the most diverse ABC transporter with a potential role in anti-malarial resistance and liver stage development [46, 47]. This gene has previously been found to have a high frequency of SNPs in *P. falciparum* isolates from Thailand and is thought to modulate the parasite's response to quinolone antimalarials, which include chloroquine [48].

**Population structure and diversity**

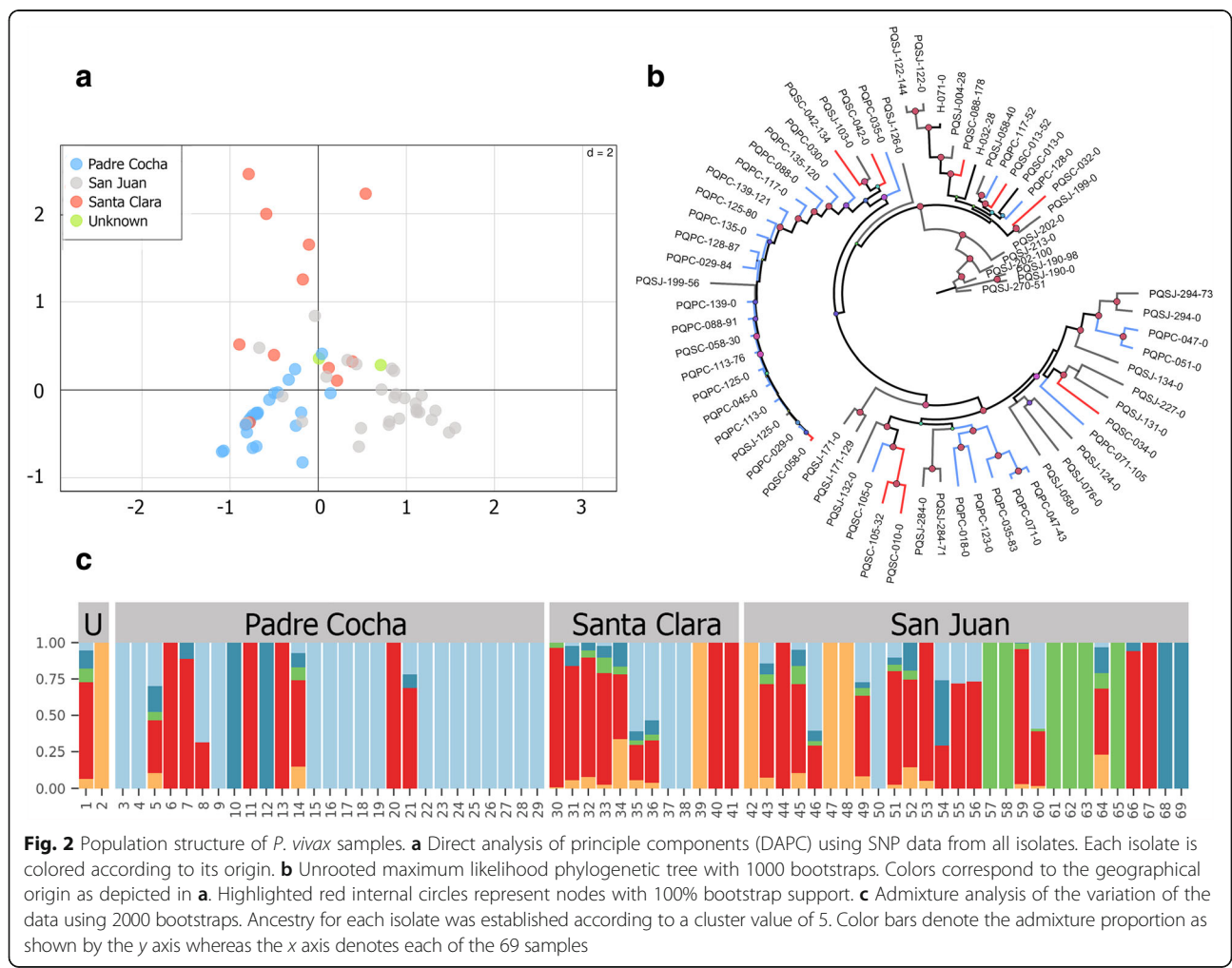
We expected that this *P. vivax* population would demonstrate high genetic similarity, and consist primarily of monoclonal infections, particularly in the villages of Padre Cocha and Santa Clara which are more remote, consistent with prior studies of *P. vivax* in this region [49, 50]. We used rates of heterozygous calls [11] and the  $F_{ws}$  statistic [28], which calculates the within-host heterozygosity, to determine infection clonality. The majority of the samples were monoclonal (97%), with only two samples (PQSC-105-32 and PQPC-018-0) considered multiclonal based on  $F_{ws} \leq 0.95$  and rate of heterozygous calls  $> 2 \times$  median (Additional file 5: Table S4).

We used discriminant analysis of principal components (DAPC) on the core SNPs to explore the population structure of *P. vivax* from the three sites. DAPC showed that the parasite population in this area is very diverse with some genetic differentiation according to collection site. In this regard, the community of Santa Clara appears to hold the most divergent *P. vivax* strains in this sample set (Fig. 2a). The maximum likelihood

phylogenetic analysis yielded a tree that was concordant with the DAPC results, emphasizing the high diversity of the parasite population and lack of geographical clustering (Fig. 2b). Given that parasites undergo sexual recombination within the mosquito, this lack of geographical clustering is indicative of gene flow between parasites from these villages. This is consistent with the frequent travel that occurs between the main city of Iquitos and the surrounding villages.

We next performed an ADMIXTURE analysis to assess the maximum likelihood estimation of individual ancestries in this population (Fig. 2c). Similar to the previous analyses of *P. vivax* in this region, there was evidence of a highly heterogeneous ancestry among the isolates with sampled genotypes derived from five ancestral populations.

Analysis of the resulting clusters showed that genotypes did not correlate with the geographical location where samples were collected, which could be a result of human population movements across the study sites. This becomes evident by inspecting the site of San Juan, which is



**Fig. 2** Population structure of *P. vivax* samples. **a** Direct analysis of principle components (DAPC) using SNP data from all isolates. Each isolate is colored according to its origin. **b** Unrooted maximum likelihood phylogenetic tree with 1000 bootstraps. Colors correspond to the geographical origin as depicted in **a**. Highlighted red internal circles represent nodes with 100% bootstrap support. **c** Admixture analysis of the variation of the data using 2000 bootstraps. Ancestry for each isolate was established according to a cluster value of 5. Color bars denote the admixture proportion as shown by the y axis whereas the x axis denotes each of the 69 samples

located in Iquitos city and is the most important trading center in the Peruvian Amazon. The parasite population of this site contained strains from all five different clusters including isolates with mixed genotypes that share characteristics of all these different populations.

On the contrary, Padre Cocha and Santa Clara, which are located 30 min by river from Iquitos, comprised only three out of the five clusters. These findings contrast with prior studies of *P. vivax* in the region, which show high inbreeding and a more clonal population structure [49, 51]. However, results of this study cannot be directly compared to these prior analyses since they were limited by the use of microsatellite data and most of their study sites were located in rural regions with different human migration patterns. This stresses the different epidemiological features of *P. vivax* in urban versus rural areas, where higher rates of migration in villages in closer proximity to a large city likely contribute to greater parasite heterogeneity.

#### Paired sample analysis

Comparisons between samples obtained from the same subject at the time of initial infection and at the time of recurrent infection revealed an overall high similarity between all isolates with a mean number of 489 segregating sites (Table 2). We used a hidden Markov model to determine regions of the genome that are identical by descent (IBD) [35]. We defined homologous relapses as having segregating sites equal to the mean number of segregating sites overall minus 1.5 standard deviations (segregating sites < 290) and IBD  $\geq$  99%. This identified a total of 12 homologous relapse pairs. The high similarity of several of the homologous relapse pairs, particularly those from the village of San Juan (PQSJ-122, PQSJ-171, PQSJ-190, PQSJ-284, and PQSJ-294) is reinforced by the maximum likelihood phylogenetic tree topology with bootstrap support values of 100% (Fig. 2b).

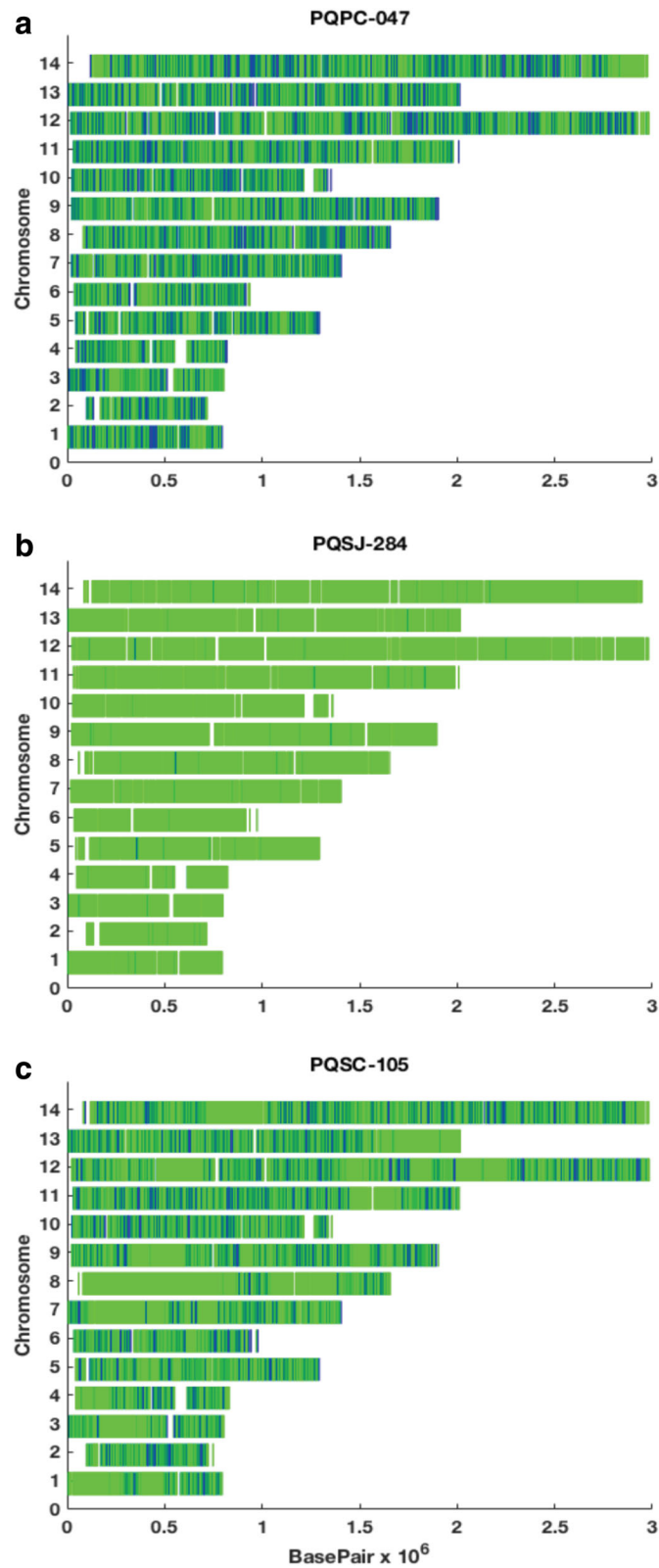
In addition, we sought to identify potential heterologous relapses caused by *P. vivax* meiotic siblings, which can occur due to recombination and outcrossing in the mosquito midgut during the initial infection [9]. We performed direct SNP comparisons across the core genome for all *P. vivax* pairs to help differentiate heterologous infections (Fig. 3a) from homologous relapses (Fig. 3b) and to identify highly related pairs, which share long blocks of concordant SNPs (Fig. 3c). These highly related pairs could be meiotic siblings or may represent heterologous relapses that represent reactivation of hypnozoites from the initial infection and another genetically different infection. We identified a total of three potential pairs: PQSC-042, PQSC-105, and PQSJ-199. These sample pairs had 52.0, 25.8, and 41.2% of their genomes that were IBD, in contrast with the homologous relapse pairs that had IBDs greater than 98%.

**Table 2** Pairwise comparison between samples obtained from the same subject at the time of initial infection and at the time of recurrent parasitemia

Initial infection	Recurrent parasitemia	Segregating sites	Core genome identical by descent (%)	Microsatellites (concordant/typed)
<i>PQPC-029-0</i>	<i>PQPC-029-84</i>	137	98.6	6/6
<i>PQPC-035-0</i>	<i>PQPC-035-83</i>	561	4.6	3/6
<i>PQPC-047-0</i>	<i>PQPC-047-43</i>	566	5.6	1/6
<i>PQPC-071-0</i>	<i>PQPC-071-105</i>	560	6.8	1/6
<i>PQPC-088-0</i>	<i>PQPC-088-91</i>	272	98.1	5/6
<i>PQPC-113-0</i>	<i>PQPC-113-76</i>	72	99.3	6/6
<i>PQPC-117-0</i>	<i>PQPC-117-52</i>	385	2.3	1/4
<i>PQPC-125-0</i>	<i>PQPC-125-80</i>	134	98.7	6/6
<i>PQPC-128-0</i>	<i>PQPC-128-87</i>	392	98.1	6/6
<i>PQPC-135-0</i>	<i>PQPC-135-120</i>	262	99.3	6/6
<i>PQPC-139-0</i>	<i>PQPC-139-121</i>	111	99.3	2/3
<i>PQSC-013-0</i>	<i>PQSC-013-52</i>	337	9.7	2/6
<i>PQSC-042-0</i>	<i>PQSC-042-134</i>	347	52.0	2/6
<i>PQSC-058-0</i>	<i>PQSC-058-30</i>	66	99.4	6/6
<i>PQSC-105-0</i>	<i>PQSC-105-32</i>	381	25.8	6/6
<i>PQSJ-058-0</i>	<i>PQSJ-058-40</i>	516	8.2	2/6
<i>PQSJ-122-0</i>	<i>PQSJ-122-144</i>	61	100	0/6
<i>PQSJ-171-0</i>	<i>PQSJ-171-129</i>	218	99.3	6/6
<i>PQSJ-190-0</i>	<i>PQSJ-190-98</i>	117	99.7	5/6
<i>PQSJ-199-0</i>	<i>PQSJ-199-56</i>	559	41.2	4/6
<i>PQSJ-202-0</i>	<i>PQSJ-202-100</i>	269	99.0	6/6
<i>PQSJ-284-0</i>	<i>PQSJ-284-71</i>	171	99.4	6/6
<i>PQSJ-294-0</i>	<i>PQSJ-294-73</i>	200	100	6/6

The mean number of segregating sites for all 69 samples included in the study was 489. We identified 12 pairs that are most likely to be homologous relapses, based on the mean number of segregating sites minus 1.5 standard deviations (< 290) and core genome IBD  $\geq$  99%, which are italicized

We compared our results to the microsatellite genotyping that was done during the original study (Table 2). Overall, results were concordant for 17 of the 23 pairs (73.9%). Microsatellite data was concordant for 9 of the 12 homologous relapses we identified with whole genome sequencing. There were two samples that were homologous by microsatellite markers but not by our data (PQPC-128 and PQSC-105). The PQSC-105 pair is one of the highly related pairs described above. Microsatellite markers may have misidentified this pair because the genomic areas genotyped with the markers could have been identical despite differences in the rest of the genome. The PQPC-128 pair had the lowest



**Fig. 3** (See legend on next page.)



(See figure on previous page.)

**Fig. 3** Direct SNP comparisons across the core genome of recurrent infections. Dark blue indicates sites where the paired samples are discordant while green demonstrates where they are concordant. **a** A recurrent infection with a heterologous *P. vivax* strain from paired samples from subject PQPC-047 (5.6% identical by descent (IBD)). **b** A homologous relapse from subject PQSJ-284 (99.4% IBD). **c** A possible heterologous relapse with large shared blocks of concordant SNPs from subject PQSC-105 (25.8% IBD)

number of informative sites for both samples on whole genome sequencing and thus may have been misidentified as a heterologous infection by our method. In addition, there were three pairs that were classified as heterologous infections by microsatellites, but were homologous relapses based on our data (PQPC-139, PQSJ-122, and PQSJ-190). The PQPC-139 pair could only be genotyped at three sites, PQSJ-122 had 0/6 markers concordant, and the PQSJ-190 pair was similar at 5/6 markers. This may represent microsatellite errors. Altogether, the comparison demonstrates how microsatellite markers may identify homologous relapses with high specificity, but also can demonstrate lower sensitivity compared to whole genome sequencing.

In the clinical trial, subjects who received the 5-day regimen had a significantly higher rate of homologous relapses, while subjects who received the 7- or 14-day regimens did not have significantly different rates of homologous relapses. From our paired sample comparison, we noted a trend towards a higher rate of relapse in subjects who received a shorter duration of treatment, with a relapse rate of 100% (1/1) with the 5-day regimen, 66.7% (6/9) in the 7-day regimen group, and 41.7% (5/12) with the 14-day regimen, although the sample size was not large enough to reach statistical significance.

The presence of these homologous relapses underscored the need to assess the host genetics to identify if they occurred due to alterations in primaquine metabolism. It is known that the human CYP2D6 enzyme, encoded by the highly polymorphic *cyp2d6* gene, is important in the metabolism of many drugs, including primaquine. In this regard, poor or intermediate activity CYP2D6 phenotypes have been associated with an increased risk for *P. vivax* relapse following treatment with primaquine [52, 53]. CYP2D6 phenotypes were assessed in all homologous relapse pairs in our sample set. Eight of these ten pairs were classified as extensive metabolizers (at least one allele coding for an enzyme with normal activity), and four were classified as intermediate metabolizers (heterozygous for one null and one active allele). Thus, the CYP2D6 poor or intermediate phenotype did not explain the majority of homologous relapses in our study.

We further analyzed each homologous relapse pair to identify SNPs that emerged after treatment but were not present in the initial infection to identify genetic changes that arose as a result of drug or immune system pressures. Three homologous relapses had missense mutations found in a sporozoite and liver-stage asparagine-rich

protein (PVX\_092945) that were not present in the initial infection: PQPC-029 (N647I, A646T), PQPC-125 (N647I, A646T), and PQSJ-171 (A654G). The protein encoded by this gene is specifically expressed in sporozoites and during liver stage development and may function as a regulator of gene expression during liver stage replication [54]. We identified two heterozygous mutations in multidrug resistance protein 2 (PVX\_118100) in the relapse sample for PQSJ-122 (V1467A, L1471P) which were not present in the initial infection.

## Discussion

This study provides an extensive dataset of population genetics of *Plasmodium vivax* in the Peruvian Amazon. We define the parasite population structure in this region using whole genome sequencing and highlight a novel method for distinguishing homologous relapses from reinfections or heterologous relapses. To the best of our knowledge, this is the first large-scale study to evaluate recurrent *P. vivax* infections using whole genome sequencing.

Our paired sample analysis and comparison to prior microsatellite genotyping demonstrate that whole genome sequencing has increased sensitivity for detecting homologous relapses and relapses due to highly related meiotic siblings. Upon comparison of our data to the prior microsatellite genotyping, we found that microsatellites failed to detect some homologous relapses. In addition, microsatellites misidentified a highly related pair that likely represented a heterologous relapse with meiotic siblings. We cannot definitively determine how using our method could have changed the outcomes of the clinical trial without performing analysis on a larger number of samples. However, using whole genome sequencing likely would lead to the identification of an increased number of homologous relapses and thus possible primaquine failures. While the original study did not find a significant difference between the 0.5 mg/kg  $\times$  7 days and 0.25 mg/kg  $\times$  14 days primaquine dosing regimens, we did note a trend in our data towards a higher rate of relapse with the shorter duration of treatment. Thus, the method of comparing recurrent *P. vivax* infections implemented in our study would improve the assessment of antirelapse therapy efficacy during clinical trials performed in endemic settings. Further evaluation should be done using this method in future clinical trials of known and novel antirelapse therapies, especially as the cost of whole genome sequencing continues to

decline and new methods such as SWGA are developed to enrich *P. vivax* DNA directly from subject samples.

We identified several mutations in many genes orthologous to multidrug resistance genes in *P. falciparum*, with a particularly high SNP rate in *pvmrp2*, and several novel alleles noted in others, although it remains unclear at this time what functional changes these impart. Although none of the resulting genotypes have been associated with resistance in *P. vivax*, the high diversity in drug resistance genes in the core genome underscores the potential risk for the emergence and spread of resistance. These findings highlight how little is known about the genetic basis of drug resistance in *P. vivax*. One major reason is the lack of a robust in vitro culture system for *P. vivax* compared to *P. falciparum*. In this study, we demonstrate that SWGA is a useful tool for enriching the amount of *P. vivax* DNA in unprocessed samples to improve the efficiency of WGS.

Our assessment of the population structure of *P. vivax* parasites in this region reveals a high level of diversity with evidence of recombination among isolates across these villages. In addition, genetic clustering by DAPC suggested very little differentiation according to the sampling sites. This finding was also confirmed by genetic clustering using ADMIXTURE that revealed at least five parasite clusters on our population with no separation by geographic location. The low level of differentiation among parasites according to site could be due to high mobilization of people between villages and within the city of Iquitos. It is important to note that a series of campaigns aimed at preventing and controlling malaria were executed during the period of sample collection. These activities were funded under the Global Fund initiative to control malaria in the border areas of the Andean region (PAMAFRO project). These campaigns were successful at lowering malaria incidence up to nearly 50% until 2011 when the project ended [41]. Therefore, it is possible that reduced gene flow and diversity could be a result of the impact of the intervention on the parasite population in the region. Further studies are needed to evaluate the dynamics of the parasite population and the effects of this major intervention on the evolution of malaria in this setting, especially given the sustained increase in malaria rates after PAMAFRO.

Our study had several limitations. Performing whole genome sequencing on *P. vivax* directly from subject samples currently remains costly and inefficient without enrichment techniques such as SWGA. It is important to obtain high-quality whole genome sequences to perform paired sample comparisons since one of the pairs with a low number of informative sites may have been misclassified as a heterologous infection. However, due to uneven amplification across the genome with SWGA,

it is more difficult to detect copy number variants, and thus, we were not able to perform this analysis. In addition, SWGA may amplify the majority clone in a multiclonal sample, thus potentially increasing the number of monoclonal samples [19]. Our finding of a majority of monoclonal samples, with only one multiclonal sample in Santa Clara and one in Padre Cocha, was not entirely consistent with other studies of this region in the Peruvian Amazon. Since the population of San Juan has the highest mobility, it would be expected that multiclonal samples would be more common at that site.

In addition, despite the high sensitivity of whole genome sequencing, it remains challenging to distinguish reinfections from relapses in a malaria-endemic area with genetically similar *P. vivax* isolates. It also remains impossible to distinguish a heterologous relapse from reinfection without being able to genotype all the hypnozoites that a person carries in their liver. However, we identified homologous relapses based on pairwise similarity compared to the similarity of the entire population and utilized a stringent cutoff. Finally, due to small sample size, we were unable to perform a genome-wide association study to identify SNPs associated with relapse. We were thus unable to identify particular SNPs that were associated with homologous relapses and thus could be implicated as an underlying genetic mechanism of primaquine resistance.

## Conclusions

Overall, our study shows that whole genome sequencing is a highly sensitive tool for gathering information about potential drug resistance, identifying homologous relapses with improved accuracy, and analyzing population structure and gene flow, especially as the cost of this technology continues to decline rapidly. Despite the significant reduction of malaria prevalence worldwide, the changing epidemiology of *P. vivax* malaria due to the presence of asymptomatic infections that can still transmit the disease and the risk of relapses challenges sustainable progress towards its elimination. These limitations require research that can help us to elucidate the changing landscape of *P. vivax* transmission, better understand the genetic diversity of *P. vivax* and allow us to monitor the efficacy of antirelapse treatment.

## Additional files

**Additional file 1: Table S1.** Study information for 46 of the *P. vivax* whole genome sequences obtained from 3 villages surrounding Iquitos, Peru. Data for 23 subjects was unavailable. (XLSX 11 kb)

**Additional file 2: Table S2.** Sequencing statistics for 69 *Plasmodium vivax* sequences collected in 3 villages in the Peruvian Amazon after alignment to the *P. vivax* reference genome. (XLSX 13 kb)

**Additional file 3: Table S3.** High quality single nucleotide polymorphisms in the core genome of the 69 *P. vivax* sequences. Samples are listed starting at column M and are labeled with the subject code followed by the day of collection. (XLSX 17419 kb)

**Additional file 4: Figure S1.** Genome wide SNP density. Mean number of single nucleotide polymorphisms (SNPs) across each of the 14 *Plasmodium vivax* chromosomes. SNP dense regions were identified in chromosomes 3, 6, 10 and 13. (TIF 3188 kb)

**Additional file 5: Table S4.** Clonality calculations for the 69 *P. vivax* samples using the Fws (within-host heterozygosity) statistic and number of heterozygous calls in the core single nucleotide polymorphisms (SNPs). As expected, the majority of samples were monoclonal (97%) and only two samples (highlighted) were multiclonal based on Fws  $\leq 0.95$  and number of heterozygous calls greater than two times the median. (XLSX 11 kb)

### Abbreviations

CYP2D6: Cytochrome P450 2D6; DAPC: Discriminant analysis of principal components; IBD: Identical by descent; SNP: Single nucleotide polymorphism; SWGA: Selective whole genome amplification

### Acknowledgements

The authors would like to acknowledge Dolores Rimarachin, Greys Braga, Gerson Guedes, Rocio del Rio, Mayra Uraco, and Salomón Durand for their support with the sample collection and execution of the initial study.

### Funding

This work was partially supported by the US DoD Armed Forces Health Surveillance Center and its Global Emerging Infections Surveillance and Response System section (AFHSC/GEIS), PROMIS ID 20160390151 in addition to National Institutes of Health (NIH) Grant U19 AI 089681. EAW was supported by NIH Grant R01 AI 103058. ANC was supported by a NIH T32 AI 007036 grant. The support for sample preparation and sequencing at the University of California, San Diego, was provided by NIH Grant P50 GM 085764. Sequencing was conducted at the IGM Genomics Center, UC San Diego. Two authors of this manuscript are military service members or employees of the US Government. This work was prepared as part of their duties. Title 17 U.S.C. § 105 provides that Copyright protection under this title is not available for any work of the US Government. Title 17 U.S.C. § 101 defines a US Government work as a work prepared by a military service member or employee of the US Government as part of that person's official duties. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the US Government.

### Availability of data and materials

The *P. vivax* genome Illumina sequencing reads of the 69 samples used for variant analysis in this study are available on the National Center for Biotechnology Information's Short Read Archive with the accession number SRP095853 for 18 previously published samples (PQPC-113-0 (prior name sample 1), PQPC-125-0 (sample 2), PQPC-139-0 (sample 3), PQPC-117-0 (sample 4), PQPC-029-0 (sample 5), PQSC-058-0 (sample 6), PQSC-042-0 (sample 7), PQPC-071-0 (sample 8), PQPC-051-0 (sample 9), PQPC-123-0 (sample 10), PQPC-135-0 (sample 11), PQPC-088-0 (sample 12), PQPC-018-0 (sample 13), PQPC-128-0 (sample 14), PQPC-047-0 (sample 15), PQSC-013-0 (sample 16), PQPC-035-0 (sample 17), PQPC-030-0 (sample 18)), and accession number SRP132126 for the remaining samples.

### Authors' contributions

ANC, HOV, and EAW designed the experiments, analyzed the data, and wrote the manuscript. DKB participated in the data interpretation and critical revision of manuscript. All authors read and approved the final manuscript.

### Ethics approval and consent to participate

The protocol for the collection of field samples was approved by the Institutional Review Board of the US Naval Medical Research Center (Protocol NMRCD.2005.0005) and the National Institutes of Health of Peru (Protocol 009-2004) in compliance with all applicable Federal regulation governing the protection of human subjects and with the Declaration of Helsinki. All adult subjects provided written informed consent; all children 8–17 years old

provided verbal assent to participate in the study. For children under the age of 17, informed consent was obtained from their parent or legal guardian.

### Competing interests

The authors declare that they have no competing interests.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Author details

<sup>1</sup>Division of Infectious Diseases, Department of Medicine, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA. <sup>2</sup>U.S. Naval Medical Research No. 6, Venezuela Ave, Block 36, Bellavista, Callao, Peru. <sup>3</sup>Division of Host-Microbe Systems & Therapeutics, Department of Pediatrics, UC San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA.

Received: 9 February 2018 Accepted: 25 June 2018

Published online: 04 July 2018

### References

- World Health Organization. World malaria report. Geneva: World Health Organization; 2016. p. volumes.
- Bassat Q, Velarde M, Mueller I, Lin J, Leslie T, Wongsrichanalai C, Baird JK. Key knowledge gaps for *Plasmodium vivax* control and elimination. *Am J Trop Med Hyg.* 2016;95:62–71.
- The World Health Organization. Severe malaria. *Tropical Med Int Health.* 2014;19(Suppl 1):7–131.
- Cibulskis RE, Alonso P, Aponte J, Aregawi M, Barrette A, Bergeron L, Ferguson CA, Knox T, Lynch M, Patouillard E, et al. Malaria: global progress 2000–2015 and future challenges. *Infect Dis Poverty.* 2016;5:61.
- Mueller I, Galinski MR, Baird JK, Carlton JM, Kochar DK, Alonso PL, del Portillo HA. Key gaps in the knowledge of *Plasmodium vivax*, a neglected human malaria parasite. *Lancet Infect Dis.* 2009;9:555–66.
- White NJ, Imwong M. Relapse. *Adv Parasitol.* 2011;80:113–50.
- Baird JK, Valecha N, Duparc S, White NJ, Price RN. Diagnosis and treatment of *Plasmodium vivax* malaria. *Am J Trop Med Hyg.* 2016;95:35–51.
- Ribeiro V, Cavaco I. Pharmacogenetics of cytochromes P450 in tropical medicine. *Curr Drug Targets.* 2006;7:1709–19.
- Bright AT, Manary MJ, Tewhey R, Arango EM, Wang T, Schork NJ, Yanow SK, Winzeler EA. A high resolution case study of a patient with recurrent *Plasmodium vivax* infections shows that relapses were caused by meiotic siblings. *PLoS Negl Trop Dis.* 2014;8:e2882.
- Hupalo DN, Luo Z, Melnikov A, Sutton PL, Rogov P, Escalante A, Vallejo AF, Herrera S, Arevalo-Herrera M, Fan Q, et al. Population genomics studies identify signatures of global dispersal and drug resistance in *Plasmodium vivax*. *Nat Genet.* 2016;48(8):953–8.
- Pearson RD, Amato R, Auburn S, Miotto O, Almagro-Garcia J, Amarantunga C, Suon S, Mao S, Noviyanti R, Trimarsanto H, et al. Genomic analysis of local variation and recent evolution in *Plasmodium vivax*. *Nat Genet.* 2016;48(8):959–64.
- McCollum AM, Soberon V, Salas CJ, Santolalla ML, Udhayakumar V, Escalante AA, Graf PC, Durand S, Cabezas C, Bacon DJ. Genetic variation and recurrent parasitaemia in Peruvian *Plasmodium vivax* populations. *Malar J.* 2014;13:67.
- Orjuela-Sanchez P, da Silva NS, da Silva-Nunes M, Ferreira MU. Recurrent parasitemias and population dynamics of *Plasmodium vivax* polymorphisms in rural Amazonia. *Am J Trop Med Hyg.* 2009;81:961–8.
- Lin JT, Hathaway NJ, Saunders DL, Lon C, Balasubramanian S, Kharabora O, Gosi P, Sriwichai S, Kartchner L, Chuor CM, et al. Using amplicon deep sequencing to detect genetic signatures of *Plasmodium vivax* relapse. *J Infect Dis.* 2015;212:999–1008.
- Durand S, Cabezas C, Lescano AG, Galvez M, Gutierrez S, Arrospe N, Alvarez C, Santolalla ML, Bacon DJ, Graf PC. Efficacy of three different regimens of primaquine for the prevention of relapses of *Plasmodium vivax* malaria in the Amazon Basin of Peru. *Am J Trop Med Hyg.* 2014;91:18–26.
- Graf PC, Durand S, Alvarez Antonio C, Montalvan C, Galves Montoya M, Green MD, Santolalla ML, Salas C, Lucas C, Bacon DJ, Fryauff DJ. Failure of supervised chloroquine and primaquine regimen for the treatment of *Plasmodium vivax* in the Peruvian Amazon. *Malar Res Treat.* 2012;2012:936067.

17. Baldeviano GC, Okoth SA, Arrospe N, Gonzalez RV, Sanchez JF, Macedo S, Conde S, Tapia LL, Salas C, Gamboa D, et al. Molecular epidemiology of *Plasmodium falciparum* malaria outbreak, Tumbes, Peru, 2010–2012. *Emerg Infect Dis.* 2015;21:797–803.
18. Sundararaman SA, Plenderleith LJ, Liu W, Loy DE, Learn GH, Li Y, Shaw KS, Ayoub A, Peeters M, Speede S, et al. Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. *Nat Commun.* 2016;7:11078.
19. Cowell AN, Loy DE, Sundararaman SA, Valdivia H, Fisch K, Lescano AG, Baldeviano GC, Durand S, Gerbasi V, Sutherland CJ, et al. Selective whole-genome amplification is a robust method that enables scalable whole-genome sequencing of *Plasmodium vivax* from unprocessed clinical samples. *MBio.* 2017;8.
20. Li H: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM ArXiv e-prints.2013 <http://adsabs.harvard.edu/abs/2013arXiv1303.3997L>.
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25:2078–9.
22. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* 2011;27:2987–93.
23. Manary MJ, Singhakul SS, Flannery EL, Bopp SE, Corey VC, Bright AT, McNamara CW, Walker JR, Winzler EA. Identification of pathogen genomic variants through an integrated pipeline. *BMC Bioinformatics.* 2014;15:63.
24. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20:1297–303.
25. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43:491–8.
26. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013, 43:11.10.11–11.10.33.
27. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 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 (Austin).* 2012;6:80–92.
28. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, O'Brien J, Djimde A, Doumbo O, Zongo I, et al. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature.* 2012;487:375–9.
29. Auburn S, Campino S, Miotto O, Djimde AA, Zongo I, Manske M, Maslen G, Mangano V, Alcock D, MacInnis B, et al. Characterization of within-host *Plasmodium falciparum* diversity using next-generation sequence data. *PLoS One.* 2012;7:e32891.
30. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics.* 2011;27:3070–1.
31. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 2009;19:1655–64.
32. Jakobsson M, Rosenberg NA. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics.* 2007;23:1801–6.
33. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods.* 2012;9:772.
34. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 2014;30:1312–3.
35. Wong W, Griggs AD, Daniels RF, Schaffner SF, Ndiaye D, Bei AK, Deme AB, MacInnis B, Volkman SK, Hartl DL, et al. Genetic relatedness analysis reveals the cotransmission of genetically related *Plasmodium falciparum* parasites in Thies, Senegal. *Genome Med.* 2017;9:5.
36. Schaffner SF, Taylor AR, Wong W, Wirth DF, Neafsey DE. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. *Malaria J.* 2018; 17:196. <https://www.ncbi.nlm.nih.gov/pubmed/?term=hmmIBD%3A+software+to+infer+pairwise+identity+by+descent+between+haploid+genotypes>.
37. Frech C, Chen N. Variant surface antigens of malaria parasites: functional and evolutionary insights from comparative gene family classification and analysis. *BMC Genomics.* 2013;14:427.
38. Cheng Y, Wang B, Changrob S, Han JH, Sattabongkot J, Ha KS, Chootong P, Lu F, Cao J, Nyunt MH, et al. Naturally acquired humoral and cellular immune responses to *Plasmodium vivax* merozoite surface protein 8 in patients with *P. vivax* infection. *Malar J.* 2017;16:211.
39. Bernabeu M, Lopez FJ, Ferrer M, Martin-Jaular L, Razaname A, Corradin G, Maier AG, Del Portillo HA, Fernandez-Becerra C. Functional analysis of *Plasmodium vivax* VIR proteins reveals different subcellular localizations and cytoadherence to the ICAM-1 endothelial receptor. *Cell Microbiol.* 2012;14:386–400.
40. Flannery EL, Wang T, Akbari A, Corey VC, Gunawan F, Bright AT, Abraham M, Sanchez JF, Santolalla ML, Baldeviano GC, et al. Next-generation sequencing of *Plasmodium vivax* patient samples shows evidence of direct evolution in drug-resistance genes. *ACS Infectious Diseases.* 2015;1:367–79.
41. Rosas-Aguirre A, Gamboa D, Manrique P, Conn JE, Moreno M, Lescano AG, Sanchez JF, Rodriguez H, Silva H, Llanos-Cuentas A, Vinetz JM. Epidemiology of *Plasmodium vivax* malaria in Peru. *Am J Trop Med Hyg.* 2016;95:133–44.
42. Suwanaruk R, Russell B, Chavchich M, Chalfein F, Kenangalem E, Kosaisavee V, Prasetyorini B, Piers KA, Barends M, Brockman A, et al. Chloroquine resistant *Plasmodium vivax*: in vitro characterisation and association with molecular polymorphisms. *PLoS One.* 2007;2:e1089.
43. Gomes LR, Almeida-de-Oliveira NK, de Lavigne AR, de Lima SR, de Pina-Costa A, Brasil P, Daniel-Ribeiro CT, Menard D, Ferreira-da-Cruz Mde F. *Plasmodium vivax* mdr1 genotypes in isolates from successfully cured patients living in endemic and non-endemic Brazilian areas. *Malar J.* 2016;15:96.
44. Schousboe ML, Ranjitkar S, Rajakaruna RS, Amerasinghe PH, Morales F, Pearce R, Ord R, Leslie T, Rowland M, Gadalla NB, et al. Multiple origins of mutations in the mdr1 gene—a putative marker of chloroquine resistance in *P. vivax*. *PLoS Negl Trop Dis.* 2015;9:e0004196.
45. Brega S, Meslin B, de Monbrison F, Severini C, Gradoni L, Udomsangpetch R, Sutanto I, Peyron F, Picot S. Identification of the *Plasmodium vivax* mdr-like gene (pvmr1) and analysis of single-nucleotide polymorphisms among isolates from different areas of endemicity. *J Infect Dis.* 2005;191:272–7.
46. Rijpma SR, van der Velden M, Gonzalez-Pons M, Annoura T, van Schaijk BC, van Gemert GJ, van den Heuvel JJ, Ramesar J, Chevalley-Maurel S, Ploemen IH, et al. Multidrug ATP-binding cassette transporters are essential for hepatic development of *Plasmodium* sporozoites. *Cell Microbiol.* 2016;18:369–83.
47. Mok S, Liong KY, Lim EH, Huang X, Zhu L, Preiser PR, Bozdech Z. Structural polymorphism in the promoter of pfmpr2 confers *Plasmodium falciparum* tolerance to quinoline drugs. *Mol Microbiol.* 2014;91:918–34.
48. Veiga MI, Osorio NS, Ferreira PE, Franzen O, Dahlstrom S, Lum JK, Nosten F, Gil JP. Complex polymorphisms in the *Plasmodium falciparum* multidrug resistance protein 2 gene and its contribution to antimalarial response. *Antimicrob Agents Chemother.* 2014;58:7390–7.
49. Delgado-Ratto C, Soto-Calle VE, Van den Eede P, Gamboa D, Rosas A, Abatih EN, Rodriguez Ferrucci H, Llanos-Cuentas A, Van Geertruyden JP, Erhart A, D'Alessandro U. Population structure and spatio-temporal transmission dynamics of *Plasmodium vivax* after radical cure treatment in a rural village of the Peruvian Amazon. *Malar J.* 2014;13:8.
50. Delgado-Ratto C, Gamboa D, Soto-Calle VE, Van den Eede P, Torres E, Sanchez-Martinez L, Contreras-Mancilla J, Rosanas-Urgell A, Rodriguez Ferrucci H, Llanos-Cuentas A, et al. Population genetics of *Plasmodium vivax* in the Peruvian Amazon. *PLoS Negl Trop Dis.* 2016;10:e0004376.
51. Van den Eede P, Van der Auwera G, Delgado C, Huysse T, Soto-Calle VE, Gamboa D, Grande T, Rodriguez H, Llanos A, Anne J, et al. Multilocus genotyping reveals high heterogeneity and strong local population structure of the *Plasmodium vivax* population in the Peruvian Amazon. *Malar J.* 2010;9:151.
52. Bennett JW, Pybus BS, Yadava A, Tosh D, Sousa JC, McCarthy WF, Deye G, Melendez V, Ockenhouse CF. Primaquine failure and cytochrome P-450 2D6 in *Plasmodium vivax* malaria. *N Engl J Med.* 2013;369:1381–2.
53. Pybus BS, Marcsisin SR, Jin X, Deye G, Sousa JC, Li Q, Caridha D, Zeng Q, Reichard GA, Ockenhouse C, et al. The metabolism of primaquine to its active metabolite is dependent on CYP 2D6. *Malar J.* 2013;12:212.
54. Silvie O, Goetz K, Matuschewski K. A sporozoite asparagine-rich protein controls initiation of *Plasmodium* liver stage development. *PLoS Pathog.* 2008;4:e1000086.