Whole-genome sequencing and microarray analysis of ex vivo *Plasmodium vivax* reveal selective pressure on putative drug resistance genes

Neekesh V. Dharia[a], A. Taylor Bright[a], Scott J. Westenberger[a], S. Whitney Barnes[b], Serge Batalov[b], Kelli Kuhen[b], Rachel Borboa[b], Glenn C. Federe[b], Colleen M. McClean[c], Joseph M. Vinetz[c], Victor Neyra[d], Alejandro Llanos-Cuentas[d], John W. Barnwell[e], John R. Walker[b], and Elizabeth A. Winzeler[a,b,1]

[a]Department of Cell Biology, Institute for Childhood and Neglected Diseases 202, The Scripps Research Institute, La Jolla, CA 92037; [b]Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121; [c]Department of Medicine, Division of Infectious Diseases, University of California San Diego School of Medicine, La Jolla, CA 92037; [d]Alexander von Humboldt Institute of Tropical Medicine, Lima 01, Peru; and [e]Division of Parasitic Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30341

Edited* by Anthony A. James, University of California, Irvine, CA, and approved September 24, 2010 (received for review March 22, 2010)

***Plasmodium vivax* causes 25–40% of malaria cases worldwide, yet research on this human malaria parasite has been neglected. Nevertheless, the recent publication of the *P. vivax* reference genome now allows genomics and systems biology approaches to be applied to this pathogen. We show here that whole-genome analysis of the parasite can be achieved directly from ex vivo-isolated parasites, without the need for in vitro propagation. A single isolate of *P. vivax* obtained from a febrile patient with clinical malaria from Peru was subjected to whole-genome sequencing (30× coverage). This analysis revealed over 18,261 single-nucleotide polymorphisms (SNPs), 6,257 of which were further validated using a tiling microarray. Within core chromosomal genes we find that one SNP per every 985 bases of coding sequence distinguishes this recent Peruvian isolate, designated IQ07, from the reference Salvador I strain obtained in 1972. This full-genome sequence of an uncultured *P. vivax* isolate shows that the same regions with low numbers of aligned sequencing reads are also highly variable by genomic microarray analysis. Finally, we show that the genes containing the largest ratio of nonsynonymous-to-synonymous SNPs include two AP2 transcription factors and the *P. vivax* multidrug resistance-associated protein (PvMRP1), an ABC transporter shown to be associated with quinoline and antifolate tolerance in *Plasmodium falciparum*. This analysis provides a data set for comparative analysis with important potential for identifying markers for global parasite diversity and drug resistance mapping studies.**

malaria | *pvmrp*

Malaria research has tended to neglect *Plasmodium vivax*, the cause of 25–40% of malaria cases worldwide (1). Global eradication of malaria will depend on having effective therapies against this parasite, but experimental work is hampered by the lack of in vitro cultivation methods necessary for propagation and experimental manipulation (2).

Recent developments have increased the priority of studying *P. vivax*. First, recent reports indicate that severe vivax malaria occurs and is understudied (3–5). Second, as happened previously with *Plasmodium falciparum*, chloroquine-resistant and multidrug-resistant *P. vivax* strains are emerging (3, 6–9). *P. vivax* also appears increasingly resistant to, or tolerant of, primaquine, the only licensed antimalarial that can eliminate the latent liver-stage hypnozoites (10). The latent hypnozoites prevent clearing of *P. vivax* malaria by schizonticidal drugs and result in recurrent relapses. Loss of primaquine as an effective hypnozoite treatment will hamper eradication efforts; understanding and monitoring emerging resistance is imperative.

Because *P. vivax* cannot be easily propagated in vitro, obtaining genomic DNA sufficient for whole-genome studies is difficult and in the past has required primate infections. Genome-sequencing efforts required *P. vivax* isolate Salvador I (SalI) to be amplified in primates to generate enough genomic DNA for sequencing at 10× coverage. The current assembly includes 14 chromosomes (22.6 Mb) and ~2,700 small contigs (4.3 Mb), mostly consisting of unassembled subtelomeric sequences (11). Additional sequencing of worldwide isolates of *P. vivax* has been approved at sequencing centers (12), but the necessity of maintaining parasites in infected primates greatly limits the number of samples available. Although the availability of the genome sequence provides new opportunities to discover drug and vaccine targets and to perform gene expression (13, 14) and proteomic studies, the lack of worldwide genetic diversity data has hampered population studies. At present, many *P. vivax* studies still rely on a small set of polymorphic antigens such as circumsporozoite protein, merozoite surface proteins, and apical membrane antigen (AMA-1) to assess diversity (15–20). Although these single-gene approaches are often sufficient for determining polyclonal infections, they are not the ideal markers for studying parasite genetic diversity because they are highly variable. For this reason, microsatellite markers have been developed for population genetics studies (21–23) although a suitable density of markers is not available to infer genes involved in drug resistance.

Cost-effective and rapid whole-genome sequencing technologies are available that allow 30–100× coverage of the haploid *P. vivax* genome. The central question is whether such methods can be used directly on patient samples with a high degree of accuracy. Here, we report whole-genome sequencing results with 30× coverage of a *P. vivax* isolate obtained directly ex vivo without further propagation in vitro or in monkeys. In addition to whole-genome sequencing, the parasite isolate was analyzed using a whole-genome tiling microarray (14, 24). Our studies provide thousands of cross-validated polymorphisms in this Peruvian isolate of *P. vivax*, which are applicable to analysis of genetic diversity in *P. vivax* and to the identification of highly variable genes, such as a *P. vivax* multidrug resistance protein (*pvmrp1*, PVX_097025), which may be under immune or drug pressure. Further characterization of these highly variable genes could help to identify drug resistance genes as well as new vaccine candidates.

Author contributions: N.V.D., A.T.B., S.J.W., J.M.V., V.N., A.L.-C., J.R.W., and E.A.W. designed research; N.V.D., A.T.B., S.J.W., S.W.B., R.B., and G.C.F. performed research; K.K., C.M.M., and J.W.B. contributed new reagents/analytic tools; N.V.D., A.T.B., S.J.W., S.W.B., and S.B. analyzed data; and N.V.D., A.T.B., and E.A.W. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

Data deposition: Reads from the sequencing of IQ07 are available on the National Center for Biotechnology Information Sequence Read Archive (accession SRP003406.1). Only reads that aligned to the *P. vivax* reference genome have been deposited to protect the privacy of the anonymous patient donor. Microarray data from this study have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE23982).

[1]To whom correspondence should be addressed. E-mail: winzeler@scripps.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1003776107/-/DCSupplemental.

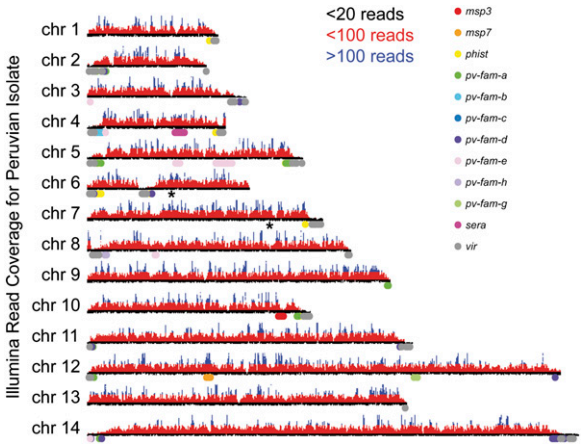

Fig. 1. Pileup alignment of Illumina sequencing reads across the *P. vivax* genome. Regions are colored by read depth, with regions covered by less than 20 reads in black, 20–100 reads in red, and more than 100 reads in blue. Subtelomeric and internal regions containing repetitive regions or multicopy highly variable genes indicated by circles correlate with lower read depth. Asterisks indicate regions with loss of hybridization by microarray that contain highly variable surface protein genes (Fig. 2).

Results

Full-Genome Analysis of a Patient Isolate. The initial objective of this study was to determine the feasibility of analyzing the complete genome of a *P. vivax* isolate obtained directly ex vivo from an infected human. Infected erythrocytes from a 5-mL blood sample were obtained from a smear-positive acutely febrile patient with uncomplicated *P. vivax* malaria in Iquitos, Peru, in April 2007; parasitemia was estimated at 1.9% by light microscopy, which is high for *P. vivax* malaria. The blood sample was passed over a CF11 leukocyte depletion filter, and total genomic DNA was extracted from the patient isolate (IQ07). Nested PCR showed no coinfection with *P. falciparum*, and *merozoite surface protein 3* (*msp3*)-α PCR-restriction fragment length polymorphism (RFLP) genotyping confirmed the infection to be monoallelic at this locus (25). Total genomic DNA was subjected to whole-genome amplification and was sequenced using the Illumina genome analyzer platform (26). We obtained 58 million reads with read lengths of 40 bases. Average genome coverage was 30× although coverage was variable in subtelomeric regions and in members of multigene families due to poor current assembly of these areas (Fig. 1). Analysis of the sequencing results indicated that the leukocyte depletion step increased the number of parasites per leukocyte 10-fold (Table 1), thereby reducing human DNA contamination and allowing analysis of the *P. vivax* genome.

Sequencing Reveals Tens of Thousands of Mutations. A major aim of resequencing in malaria parasites is to find single-nucleotide polymorphisms (SNPs) that can be used to reveal genes under immune or drug selection or to find markers for population genetic studies. IQ07 isolate reads were aligned to the 14 assembled *P. vivax* chromosomes using Bowtie (27), and SNPs relative to the SalI reference strain were called using Maq (28). We limited our analysis to those SNPs with unambiguous base change calls and that were covered by three or more reads. The resulting high-confidence SNP set consisted of 18,261 SNPs (Dataset S1), which had a SNP rate of 0.81 SNPs per 1,000 bases, which is similar to rates observed between clones of *P. falciparum* isolated from different regions (26). Using Sanger sequencing, we confirmed all SNPs in a subset of high-confidence calls by Illumina sequencing as polymorphisms in IQ07; although the majority of SNPs (21/24) were correctly identified, three of the Illumina SNPs were instead confirmed as deletions in IQ07 relative to SalI (*SI Materials and Methods*, Table S6, and *SI Appendix*).

We next sought to cross-validate the SNPs predicted by sequencing using a high-density *P. vivax* tiling microarray (14). Microarrays are less expensive and simpler to use than whole-genome sequencing and thus may represent a valuable alternative approach to direct sequencing for assessing diversity in field isolates. To assess genetic diversity by microarray, we used the same whole-genome amplified DNA from IQ07 and SalI described above. The patient-derived isolate showed mean and median hybridization intensities for human-specific probes at only slightly higher levels than background but with a significantly different distribution (Kolmogorov–Smirnov test p value 7.4×10^{-40}) (Fig. S1). This result, along with our sequencing data, indicated that leukocyte depletion, although not 100% effective, resulted in a large reduction of contaminating human DNA in the patient-derived sample.

The algorithms designed for analysis of our *P. falciparum* microarray (24) were extended to analyze hybridizations of genomic DNA to the *P. vivax* array. Despite the lower probe density of 6-bp spacing compared with our *P. falciparum* array with 2- to 3-bp spacing, we observed clear signals for polymorphisms. The appearance of SNPs in the genomic DNA of the patient-derived isolate compared with the reference strain resulted in hybridization intensities that were lower in the patient isolate than in the reference. The log of the ratio of intensities of the patient isolate IQ07 to the reference strain SalI was analyzed using a one-tailed z-test, and an F-test was used to predict the base-pair position of the polymorphism as described previously (24). We classified probes with a z-test p value of less than 1×10^{-15} as highly likely to contain mutations in the isolate, resulting in a total of 8,118 predicted polymorphisms (Dataset S2). A lower threshold of 1×10^{-5} detected 31,018 polymorphisms. Using the high-stringency cutoff, our microarray was able to achieve a SNP detection rate of 33.6% with a false discovery rate of 4.5% (Table S1 and *SI Materials and Methods*), thus validating thousands of SNPs that can be used in genetic diversity studies. The 18,261 SNPs identified with high confidence from sequencing and the 8,118 polymorphisms identified by microarray were used for further analysis (presented below).

Using a stringent p value of 1×10^{-15} allowed us to detect polymorphisms with high confidence but likely underestimated the number of polymorphisms per gene as evidenced by more SNPs than microarray polymorphisms being identified by sequencing for most genes (Dataset S3). It is interesting to note, however, that for highly variable genes, we found more polymorphisms by microarray than by sequencing. For example, we found 51 polymorphisms by microarray but only 7 SNPs by sequencing in one *msp3* gene (PVX_097685), 107 but only 4 in the *reticulocyte binding protein 2 precursor* (PVX_090325), and 41 but only 4 in *reticulocyte-binding protein 1* (PVX_098585), respectively. These genes often contain low-complexity regions, which are more likely to bear small insertion/deletion events within their sequences (29) and may compromise alignments accounting for a lower number of reads (average coverage of 1.6×, 2.6×, and 11.1×, respectively, versus average coverage of 40.0× for all genes; Dataset S3 and Dataset S4). Thus, our data demonstrate that our dual approaches for polymorphism detection are complementary and that SNPs are likely to be systematically underestimated for the most variable genes.

Table 1. Evaluation of human leukocyte contamination by sequencing analysis

Cell	DNA content of cell	Total reads	Reads per Mb	Average fold coverage	Relative no. of cells	Relative gene copy number
Human nucleated cell	6.4 Gb (diploid)	25,774,830	4,027	0.16	1	1
P. vivax parasite	26.8 Mb (haploid)	22,427,291	836,839	33.47	208	104

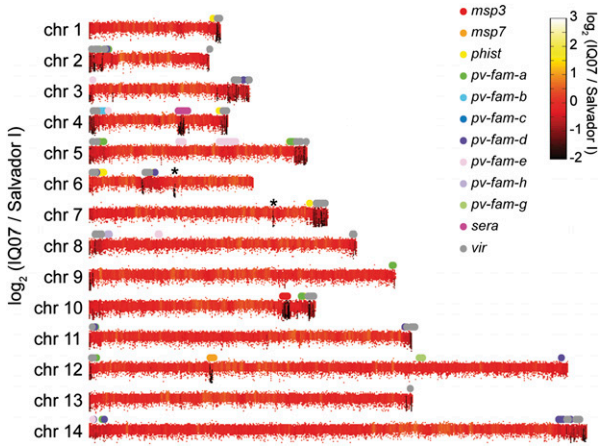

Fig. 2. Genome-wide detection of polymorphisms in a patient-derived isolate of *P. vivax*. The $\log_2$ ratios of intensities of IQ07 versus SalI were displayed along each chromosome in the *P. vivax* genome and are colored by the running mean over 500 bp. Loss of hybridization (black) in IQ07 relative to the reference isolate indicates highly variable genes and correlates with subtelomeric and internal loci of multigene families that are involved in host–pathogen interactions such as *msp3*, *msp7*, *sera*, and *vir*. The loss of hybridization on chromosomes 6 and 7 (marked with asterisks) that contain hypothetical genes of unknown function corresponds to syntenic regions in *P. falciparum* that contain highly variable surface protein genes, including *msp* genes.

High Level of Mutations in Multigene Families. In *P. falciparum*, members of multigene families play roles in antigenic variation, as only some members of a multigene family may be expressed whereas the rest are transcriptionally silenced (30, 31). In both *P. vivax* and *P. falciparum*, these genes are clustered in specific regions of the genome (11, 32). On the basis of work with *P. falciparum* (33) and with *P. vivax* (34) it is known that members of multigene families are genetically variable, and, as expected, both sequencing and microarray analysis identified exceptionally high levels of genetic variability in members of multigene families in our *P. vivax* patient-derived isolate (Fig. 2 and Table 2). In isolate IQ07, 8.13%, 2.14%, and 6.16% of unique probes that map to annotated chromosomal *msp3*-a (12 genes), *msp7* (11 genes), and *vir* (87 genes with probes) gene families, respectively, were called as polymorphic by our algorithm using a *p*-value cutoff of 1×10^{-15}. In contrast, only 0.25% of probes that map to genes that fall outside of recognizable gene families were considered polymorphic using the same *p*-value cutoff. Similarly, by sequencing, the *msp3*, *msp7*, and *vir* gene families had SNP rates of 13.42%, 3.73%, and 2.53%, respectively; other genes had a SNP rate of 0.06%. Other multigene families such as the serine-repeat antigen (SERA) family and the *P. vivax* tryptophan-rich antigen family (Pv-fam-a) showed high levels of sequence variability with both methods (Table 2).

Genes Under Selection. Our work with laboratory-evolved *P. falciparum* has shown that in vitro drug pressure can rapidly select for multiple independent coding mutations specifically in genes encoding enzymes or transporters involved in drug resistance (35). Although one mutation may help the parasite evade a drug, a compensatory mutation at an unrelated site in the same protein may also be needed to help to stabilize the protein's structure or improve its catalytic efficiency. This pattern has been observed by sequencing field isolates as well (36, 37). There are more SNPs (12 nonsynonymous and 0 synonymous) in the chloroquine resistance transporter (*pfcrt*) in assorted *P. falciparum* isolates relative to what would be expected by chance without considering the resistance or sensitivity phenotype (36, 37). Indeed, after excluding nonconserved, membrane-localized or antigenic proteins, *pfcrt* remains one of the most variable in the *P. falciparum* genome (33, 36, 37). Thus, well-characterized genes conserved across species with higher rates of nonsynonymous SNPs will often have roles in drug resistance. We therefore examined the group of genes that showed the highest ratios of nonsynonymous substitutions (d_N) per site to synonymous (d_S) substitutions per site (d_N/d_S) in our isolate relative to SalI. Altogether, 87 genes had d_N/d_S greater than one with at least five nonsynonymous mutations (Table S2). The group included a reticulocyte-binding protein, *ama-1*, several transcription factors, and *msp5*. In addition, the group contained *pvmrp1*, an ABC transporter that bears five nonsynonymous SNPs and no synonymous SNPs (Fig. 3*E* and Table S2). The syntenic *P. falciparum* gene, also showing the greatest sequence homology, is the multidrug resistance protein 1 (PFA0590w, PfMRP1, BlastP $P < 10^{-300}$), which has been associated with quinoline resistance (38–40) and resistance to antifolates (41). Most of the substitutions are in sequences that are conserved across *Plasmodium* species (Fig. S2), indicating their importance. In addition, homology modeling based on the *Staphylococcus aureus* ABC transporter SAV1866 showed that half of the mapped SNPs were located in the transmembrane pore region of the transporter (Fig. S3). Notably, *P. falciparum* knockout strains are also more sensitive to primaquine (40). Given the tremendous need for molecular markers that

Table 2. Classification of genes

Class description	No.of genes	SNPs per gene	Average SNPs per base (%)	Microarray polymorphisms per gene	Average polymorphisms per probe (%)
Merozoite surface protein 3	12	13.42	0.57	32.58	8.13
SERA family	13	14.69	0.47	13.46	2.70
AP2 transcription factors	26	4.38	0.07	1.61	0.15
Vir genes	87	2.53	0.20	11.62	6.16
Pv-fam-d proteins	13	1.92	0.13	3.85	1.24
Merozoite surface protein 7	11	3.73	0.36	3.09	2.14
Tryptophan-rich antigens (Pv-fam-a)	21	2.29	0.13	2.19	0.81
Membrane proteins	208	2.29	0.07	1.18	0.25
Phist proteins (Pf-fam-b)	18	1.28	0.10	1.06	0.51
None	4,421	1.53	0.06	0.91	0.25
Pv-fam-b proteins	5	0.60	0.05	0.80	0.53
Early transcribed membrane proteins	9	1.00	0.15	0.78	1.02
RAD proteins (Pv-fam-e)	40	1.10	0.12	0.75	0.53
Pv-fam-h proteins	4	1.25	0.10	0.50	0.41
Ribosomal proteins	132	0.27	0.04	0.15	0.13
Pv-fam-g proteins	3	0.00	0.00	0.00	0.00

Numbers are calculated for SNPs covered by at least three reads and the stringent cutoff of 1×10^{-15} for polymorphisms. "No. of genes" indicates the number of genes of a particular class.

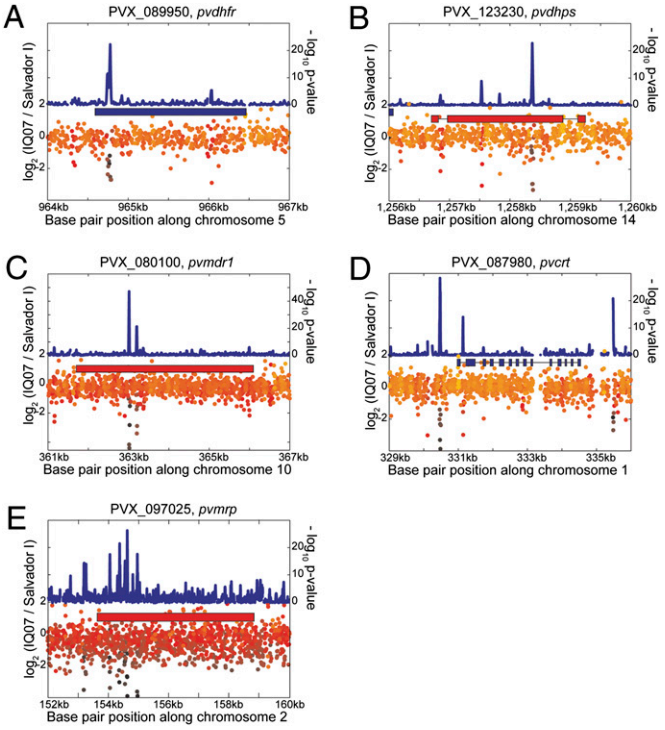

Fig. 3. Predicted polymorphisms in drug resistance genes in a Peruvian patient-derived *P. vivax* isolate. The result from the microarray-based prediction algorithm is plotted in blue above the gene models. The $\log_2$ ratios of the intensities of IQ07 and Salvador I were displayed below and are colored by the running mean over 25 bp using the same scale as displayed in Fig. 2. (*A*) We detected a mutation near codon 58 in *pvdhfr*. (*B*) We detected a strong mutation signal near codon 205 in *pvdhps*, which has not been previously described as polymorphic. (*C*) IQ07 contained mutations near codons 958 and 1022 in *pvmdr1*. (*D*) We detected a mutation ~500 bp upstream of the start codon that was in the 5′ UTR of *pvcrt*. We detected a mutation ~100 bp into the first intron of the *pvcrt* gene but did not detect any mutations in exons. (*E*) We detected multiple mutations by microarray in the putative ABC transporter *pvmrp*.

predict quinoline resistance in *P. vivax*, these associations are worth investigating in more detail.

Positive Selection on Transcription Factors. Although mutations in the chloroquine resistance transporter (*pvcrt*, PVX_087980) are not thought to be associated with chloroquine resistance (23, 42, 43), increased expression of *pvcrt* has been found in chloroquine-resistant *P. vivax* (44). In addition, the level of dihydrofolate reductase (*pvdhfr*, PVX_089950) expression is higher in some *P. vivax* isolates relative to *P. falciparum* (14). Thus, transcription may play a role in drug resistance in *P. vivax*. Interestingly, we find that six AP2 transcription factors (45) (PVX_083040, PVX_091065, PVX_123750, PVX_122680, PVX_090110, PVX_114260) as well as the ADA2 transcription factor (PVX_094945) have d_N/d_S greater than one with multiple nonsynonymous SNPs (Table S2). Changes in the amino acid sequence in these transcription factors could give increased expression of downstream targets, which might include drug resistance genes, as some of the identified SNPs fall within AP2 domains or putative regulatory sequences (Table S3). Indeed, the AP2 family of transcription factors was among the most variable classes of proteins in our analysis with an average of 4.38 SNPs per gene (Table 2) although their large size means that the number per base is not exceptional. Similar patterns are observed in *P. falciparum* in genes with AP2 domains: PF11_0442 bears 21 nonsynonymous SNPs and 1 synonymous SNP, PFL1075w (ortholog of PVX_123750) has 28 and 3, and PF13_0235 (ortholog of PVX_083040) has 53 and 9. Data from mammalian genomes have indicated that evolution may be driven through the expansion and modification of transcription factors (46), and it is not unreasonable to postulate that evolution to drug or host immune responses in malaria parasites may be driven by changes in transcription factors and their DNA-binding domains. The genetic variability of transcription factors demonstrated here through whole-genome analysis warrants further investigation of this hypothesis.

Drug Resistance. Our Peruvian isolate is likely to have been subjected to 35 additional years of intensive, modern drug pressure relative to the Salvador I reference strain, which was isolated from a human in the La Paz region of El Salvador in 1972 (47). We therefore examined microarray- and sequencing-based variation in known resistance genes (Fig. 3 and Table 3). Both methods identified a common mutation in South American *pvdhfr*, S58R (48). We detected a strong signal for a mutation near codon 205 (M205I by sequencing) of dihydropteroate synthetase (*pvdhps*, PVX_123230), a location not previously described as polymorphic. We also found that IQ07 contains described mutations M908L, T958M, and a synonymous mutation at codon 1022 in the multidrug resistance gene 1 (*pvmdr1*, PVX_080100) (23). Finally, we detected a known mutation ~500 bp upstream of the start codon in the 5′ untranslated region of *pvcrt* (48) and a mutation ~100 bp into the first intron of this gene.

Pre-erythrocytic Vaccine Candidates. In *P. falciparum*, genes that produce the most abundant proteins in sporozoites have been investigated as vaccine candidates (49, 50). In addition, these pre-erythrocytic vaccine candidates show evidence of higher-than-expected variability in *P. falciparum* (33). We thus sought to classify the most highly variable and highly expressed transcripts for the *P. vivax* sporozoite stage (Table S4). Using *P. vivax* sporozoite expression data (14), we identified abundantly transcribed genes (>500 units) whose expression peaked in sporozoites with nonsynonymous SNPs and d_N/d_S greater than 1 (34 genes). Included in the 34 genes were *ama-1*, *circumsporoziote protein* (PVX_119355), and *sporozoite surface protein 2* (PVX_082735), which are antigens that have been considered for pre-erythrocytic or blood stage vaccines in *P. vivax* (51). Further investigation of the other genes in this short list might reveal proteins useful as vaccine candidates, but further sequencing of additional isolates is needed to distinguish between genes variable because of immune selection versus random chance.

Discussion

We present here the second genome sequence of *P. vivax*, using methodological advances that make such genomic analysis available to small laboratories. The primary disadvantage of whole-genome sequencing is that the downstream data analysis is not yet streamlined. Although sequencing provides the precise base-pair change of a polymorphism, our microarray-based method can produce a snapshot of polymorphisms in the *P. vivax* genome in under an hour of data analysis after an overnight hybridization. The techniques demonstrated here have utility in the field for answering fundamental, previously unaddressable questions about *P. vivax* population genetics and diversity.

Developing tools for studying genetic diversity directly from field samples in *P. vivax* is critical because it allows for the discovery and monitoring of genes involved in drug resistance as well as for identifying potential vaccine candidates. Genetic diversity studies have been useful in *P. falciparum* for rapidly identifying genomic regions in linkage disequilibrium following selection for drug-resistant parasites (33, 36). In addition, microarray-based genetic diversity studies of *P. falciparum* identified the amplification of GTP cyclohydrolase I (*pfgch1*, PFL1155w) and suggested it as a marker for antifolate resistance (33). These types of studies are particularly needed in *P. vivax* as there is evidence of the emergence of resistance to chloroquine, which is widely used as a first-line treatment, and the mechanisms of chloroquine action and resistance are fundamentally

Table 3. Polymorphism detection in *P. vivax* drug resistance genes by microarray and sequencing analysis

Gene	Chromosome	Microarray polymorphic range (bp) (prediction)	*P* value cutoff	Illumina sequencing (SNP reads/total reads)
pvdhfr	5	964,772–964,814 (964,797)	10^{-15}	T964796C [Y69Y] (13/13)
		964,742–964,820 (964,797)	10^{-10}	C964763G [S58R] (31/31)
pvdhps	14	1,258,370–1,258,412 (1,258,389)	10^{-15}	C1258389T [M205I] (12/12)
pvmdr1	10	363,008–363,055 (363,031)	10^{-15}	G363032A [L1022L] (12/13)
		363,200–363,242 (363,221)	10^{-15}	G363223A [T958M] (18/19)
		363,362–363,404 (363,383)	10^{-05}	T363374G [M908L] (14/15)
pvcrt	1	330,470–330,518 (330,489)	10^{-15}	T330482C [5′ upstream region] (2/2)
		330,470–330,518 (330,489)	10^{-15}	G330484T [5′ upstream region] (1/1)
		330,470–330,518 (330,489)	10^{-15}	C330495A [5′ upstream region] (4/5)
		331,138–331,172 (331,152)	10^{-12}	T331151C [intronic] (58/58)
pvmrp	2	154,052–154,088 (154,069)	10^{-15}	C154067T [H1586Y] (13/13)
		154376–154,412 (154,393)	10^{-15}	G154391A [V1478I] (6/6)
		154,556–154,592 (154,575)	10^{-15}	G154567C [G1419A] (8/8)
		154,628–154,670 (154,647)	10^{-15}	T154646G [Y1393D] (11/11)
		154,964–155,000 (154,983)	10^{-15}	T154979A [L1282I] (10/10)

Microarray polymorphisms are presented as base-pair ranges with the predicted SNP position in parentheses. *p* value indicates the cutoff for detection of polymorphisms by microarray. Sequencing polymorphisms are presented as SNP base-pair position and nucleotide change. For sequencing results, the predicted amino acid change resulting from the SNP is indicated in brackets.

different from those in *P. falciparum* (7). Once identified, the global geographic distribution of drug resistance genes can help inform treatment policy. In addition, genetic diversity studies can provide information on the variability of possible vaccine candidates as has been shown in *P. falciparum* (33, 36, 37, 52).

In our study, *pvmrp1* has a d_N/d_S of infinity, suggesting positive selection. Studies of PfMRP1 have shown that one of its roles is to pump out glutathione (40). The killing effect of primaquine is thought to be through introducing oxidative damage to mitochondria, and thus if parasites showed an improved ability to export glutathione adducts, tolerance might be expected. In addition, knockouts of the gene in *P. falciparum* have rendered the parasites more sensitive to primaquine (40). It should be noted, however, that PfMRP1 has also been associated with artemisinin resistance in *P. falciparum* (53), and, therefore, PvMRP1 may be a general drug resistance protein. The data presented here coupled with data from *P. falciparum* suggest that *pvmrp* is a promising candidate for further evaluation.

Along with identifying drug resistance genes and potential vaccine candidates, future studies using the polymorphisms identified here can help monitor the emergence and spread of drug resistance. Although many believe that primaquine tolerance in *P. vivax* is widespread (26), quantifying the extent of resistance is difficult, and we have no information on our isolate's sensitivity to primaquine, chloroquine, or quinine. There are currently no cell-based efficacy assays or genetic markers for primaquine resistance in *P. vivax* (42), and the main issue hindering the study of primaquine resistance is the inability to identify a relapse from a reinfection. The set of genetic markers from our study can potentially be used to distinguish a relapse from a new infection, and once true relapses are identified, polymorphisms that associate with resistance can be ascertained. If molecular markers that could be used to predict primaquine tolerance were discovered, it might allow the development of appropriate treatment plans that could be used to extend the life of this drug.

New whole-genome technologies have the ability to rapidly accelerate research of the neglected malaria parasite *P. vivax*. Research has been restricted due to the difficulty of in vitro cultivation, but the ability to analyze genomic DNA directly from patient-derived isolates promises to overcome this limitation. We show here how microarray-based polymorphism detection and whole-genome sequencing can be used to create genome-wide maps of genetic diversity in *P. vivax* and how this may lead to new drug-resistant candidate genes, new vaccine candidates, and new tools in developing treatment policy.

Materials and Methods

Ethics Statement. The protocol used to collect human blood samples for this work was approved by the Human Subjects Protection Program of The Scripps Research Institute and the University of California, San Diego, and by the Ethical Committees of Universidad Peruana Cayetano Heredia and the Asociacion Benefica PRISMA, Iquitos, Peru. Written informed consent was obtained from each subject or a parent in the case of minors. The consent form states in English and Spanish that samples may be used for any scientific purpose involving this or any other project, now or in the future, and that the samples may be shared with other researchers.

Sample Collection. Two sources of DNA were used for our initial experiments. First, several hundred nanograms of pure *P. vivax* SalI DNA [the genome-sequencing project reference strain (11)] was obtained from John Barnwell at the Centers for Disease Control and Prevention. Second, we obtained *P. vivax* DNA isolated from a Peruvian patient known to have *P. vivax* malaria (IQ07). The material was derived from 5 mL of whole blood that had been obtained with informed consent. White blood cells were removed using a Whatman CF11 column (Whatman plc). DNA purified from the SalI sample in duplicate and the patient-derived sample was subjected to a whole-genome amplification using the Qiagen RepliG kit. The amplified DNA was used for Illumina sequencing and microarray analysis.

Microarray Analysis. For these experiments, a custom Affymetrix whole-genome tiling microarray based on the current assembly of the *P. vivax* genome sequence was used. Twenty micrograms of genomic DNA from each amplification reaction (SalI in replicate and IQ07) and 2.5 ng each of Bio B, Bio C, Bio D, and Cre Affymetrix control plasmids were fragmented with DNaseI and end-labeled with biotin (54). The samples were hybridized to the microarrays overnight at 45 °C using Affymetrix kits, washed, and scanned following manufacturer's instructions.

The polymorphism detection was performed as previously described for *P. falciparum* (24). Briefly, using a sliding window of three overlapping probes, we scanned for sets that had significantly lower hybridization in IQ07 when compared with the SalI reference that were indicative of a polymorphism as determined by *z*-test using an empirically derived SD. Combining the data from the sliding windows of three probes enabled us to establish the boundaries within which the polymorphisms were contained. To predict the precise position of SNPs, we used an empirically determined model of the loss of hybridization based on the SNP position in probes. An *F*-test was then performed with the model, on the basis of the null hypothesis of the mean equaling zero, to predict the position of the polymorphism.

Illumina Sequencing and Analysis. The *P. vivax* isolate was sequenced on three lanes of an Illumina Genome Analyzer Flow Cell with a paired-end module, resulting in 57,917,344 total reads. The reads were aligned to the SalI genome (PlasmoDB version 5.5) and the human genome (National Center for Biotechnology Information build 36.3) using Bowtie (27). The alignment was

MICROBIOLOGY

performed with singleton reads using option –v 3, which emulated Short Oligonucleotide Alignment Program-style alignment (55) in which three mismatches were allowed in a 40-bp read. The Bowtie output was converted into Maq maps, and Maq (28) was used to generate the list of polymorphisms between IQ07 and the reference genome for the 14 assembled chromosomes. SNPs that were unambiguous (no mixed SNP calls) and covered with more than three reads were used to validate the results of the microarray analysis. SNPs that mapped to the predicted coding sequences of core chromosomal genes were used to delineate synonymous and nonsynonymous SNPs. Maximum-likelihood estimates were obtained for the numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site between the SalI reference and IQ07 and for the ratio d_N/d_S. Calculations were performed using the *codeml* program of the PAML package version 4.4b (56) (*SI Materials and Methods*).

ACKNOWLEDGMENTS. We thank S. E. R. Bopp for insightful discussions. E.A.W. was supported by grants from the W. M. Keck Foundation and by National Institutes of Health Grant R01AI059472. Work on *P. vivax* drug discovery at the Genomics Institute of the Novartis Research Foundation is funded by a grant from the Medicines for Malaria Venture and by the Wellcome Trust to the NGBS consortium. US Public Health Service Grants D43TW007120, K24AI068903, and R01AI067727 (to J.M.V.) supported the work done in Peru by C.M.M., J.M.V., V.N., A.L.-C. and A.T.B. was supported in part by the UCSD Genetics Training Program through an institutional training grant from the National Institute of General Medical Sciences (T32 GM008666).

1. Mendis K, Sina B, Marchesini P, Carter R (2001) The neglected burden of Plasmodium vivax malaria. *Am J Trop Med Hyg* 64(1–2 Suppl):97–106.
2. Price RN, et al. (2007) Vivax malaria: Neglected and not benign. *Am J Trop Med Hyg* 77(6 Suppl):79–87.
3. Tjitra E, et al. (2008) Multidrug-resistant Plasmodium vivax associated with severe and fatal malaria: A prospective study in Papua, Indonesia. *PLoS Med* 5:e128.
4. Genton B, et al. (2008) Plasmodium vivax and mixed infections are associated with severe malaria in children: A prospective cohort study from Papua New Guinea. *PLoS Med* 5: e127.
5. Kochar DK, et al. (2009) Severe Plasmodium vivax malaria: A report on serial cases from Bikaner in northwestern India. *Am J Trop Med Hyg* 80:194–198.
6. Baird JK (2008) Real-world therapies and the problem of vivax malaria. *N Engl J Med* 359:2601–2603.
7. Sharrock WW, et al. (2008) Plasmodium vivax trophozoites insensitive to chloroquine. *Malar J* 7:94.
8. de Santana Filho FS, et al. (2007) Chloroquine-resistant Plasmodium vivax, Brazilian Amazon. *Emerg Infect Dis* 13:1125–1126.
9. Vinetz JM (2006) Emerging chloroquine-resistant Plasmodium vivax (benign tertian) malaria: The need for alternative drug treatment. *Clin Infect Dis* 42:1073–1074.
10. Krudsood S, et al. (2008) High-dose primaquine regimens against relapse of Plasmodium vivax malaria. *Am J Trop Med Hyg* 78:736–740.
11. Carlton JM, et al. (2008) Comparative genomics of the neglected human malaria parasite Plasmodium vivax. *Nature* 455:757–763.
12. Carlton JM, Escalante AA, Neafsey D, Volkman SK (2008) Comparative evolutionary genomics of human malaria parasites. *Trends Parasitol* 24:545–550.
13. Bozdech Z, et al. (2008) The transcriptome of Plasmodium vivax reveals divergence and diversity of transcriptional regulation in malaria parasites. *Proc Natl Acad Sci USA* 105:16290–16295.
14. Westenberger SJ, et al. (2010) A systems-based analysis of Plasmodium vivax lifecycle transcription from human to mosquito. *PLoS Negl Trop Dis* 4:e653.
15. Zakeri S, et al. (2010) Molecular characterization of Plasmodium vivax clinical isolates in Pakistan and Iran using pvmsp-1, pvmsp-3alpha and pvcsp genes as molecular markers. *Parasitol Int* 59:15–21.
16. Aresha M, et al. (2008) Genotyping of Plasmodium vivax infections in Sri Lanka using Pvmsp-3alpha and Pvcs genes as markers: A preliminary report. *Trop Biomed* 25:100–106.
17. Moon SU, et al. (2009) High frequency of genetic diversity of Plasmodium vivax field isolates in Myanmar. *Acta Trop* 109:30–36.
18. Kim JR, et al. (2006) Genetic diversity of Plasmodium vivax in Kolkata, India. *Malar J* 5:71.
19. Cui L, et al. (2003) Genetic diversity and multiple infections of Plasmodium vivax malaria in Western Thailand. *Am J Trop Med Hyg* 68:613–619.
20. Mueller I, Kaiok J, Reeder JC, Cortés A (2002) The population structure of Plasmodium falciparum and Plasmodium vivax during an epidemic of malaria in the Eastern Highlands of Papua New Guinea. *Am J Trop Med Hyg* 67:459–464.
21. Gunawardena S, et al. (2010) Geographic structure of Plasmodium vivax: Microsatellite analysis of parasite populations from Sri Lanka, Myanmar, and Ethiopia. *Am J Trop Med Hyg* 82:235–242.
22. Van den Eede P, et al. (2010) High complexity of Plasmodium vivax infections in symptomatic patients from a rural community in central Vietnam detected by microsatellite genotyping. *Am J Trop Med Hyg* 82:223–227.
23. Orjuela-Sánchez P, et al. (2009) Analysis of single-nucleotide polymorphisms in the crt-o and mdr1 genes of Plasmodium vivax among chloroquine-resistant isolates from the Brazilian Amazon region. *Antimicrob Agents Chemother* 53:3561–3564.
24. Dharia NV, et al. (2009) Use of high-density tiling microarrays to identify mutations globally and elucidate mechanisms of drug resistance in *Plasmodium falciparum*. *Genome Biol* 10:R21.
25. Bruce MC, et al. (2000) Cross-species interactions between malaria parasites in humans. *Science* 287:845–848.
26. Zhou Y, et al. (2008) Evidence-based annotation of the malaria parasite's genome using comparative expression profiling. *PLoS ONE* 3(2):e1570.
27. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
28. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18:1851–1858.
29. Amodu OK, Hartl DL, Roy SW (2008) Patterns of polymorphism in genomic regions flanking three highly polymorphic surface antigens in Plasmodium falciparum. *Mol Biochem Parasitol* 159:1–6.
30. Su XZ, et al. (1995) The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. *Cell* 82:89–100.
31. Scherf A, et al. (1998) Antigenic variation in malaria: In situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in Plasmodium falciparum. *EMBO J* 17:5418–5426.
32. Gardner MJ, et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* 419:498–511.
33. Kidgell C, et al. (2006) A systematic map of genetic variation in Plasmodium falciparum. *PLoS Pathog* 2:e57.
34. Merino EF, et al. (2006) Multi-character population study of the vir subtelomeric multigene superfamily of Plasmodium vivax, a major human malaria parasite. *Mol Biochem Parasitol* 149:10–16.
35. Rottmann M, et al. (2010) Spiroindolones, a potent compound class for the treatment of malaria. *Science* 329:1175–1180.
36. Volkman SK, et al. (2007) A genome-wide map of diversity in Plasmodium falciparum. *Nat Genet* 39:113–119.
37. Jeffares DC, et al. (2007) Genome variation and evolution of the malaria parasite Plasmodium falciparum. *Nat Genet* 39:120–125.
38. Klokouzas A, et al. (2004) Plasmodium falciparum expresses a multidrug resistance-associated protein. *Biochem Biophys Res Commun* 321:197–201.
39. Mu J, et al. (2003) Multiple transporters associated with malaria parasite responses to chloroquine and quinine. *Mol Microbiol* 49:977–989.
40. Raj DK, et al. (2009) Disruption of a Plasmodium falciparum multidrug resistance-associated protein (PfMRP) alters its fitness and transport of antimalarial drugs and glutathione. *J Biol Chem* 284:7687–7696.
41. Dahlström S, Veiga MI, Mårtensson A, Björkman A, Gil JP (2009) Polymorphism in PfMRP1 (Plasmodium falciparum multidrug resistance protein 1) amino acid 1466 associated with resistance to sulfadoxine-pyrimethamine treatment. *Antimicrob Agents Chemother* 53:2553–2556.
42. Baird JK (2009) Resistance to therapies for infection by Plasmodium vivax. *Clin Microbiol Rev* 22:508–534.
43. Barnadas C, et al. (2008) Plasmodium vivax resistance to chloroquine in Madagascar: Clinical efficacy and polymorphisms in pvmdr1 and pvcrt-o genes. *Antimicrob Agents Chemother* 52:4233–4240.
44. Fernández-Becerra C, et al. (2009) Increased expression levels of the pvcrt-o and pvmdr1 genes in a patient with severe Plasmodium vivax malaria. *Malar J* 8:55.
45. De Silva EK, et al. (2008) Specific DNA-binding by apicomplexan AP2 transcription factors. *Proc Natl Acad Sci USA* 105:8393–8398.
46. Levine M, Tjian R (2003) Transcription regulation and animal diversity. *Nature* 424: 147–151.
47. Collins WE, Contacos PG, Krotoski WA, Howard WA (1972) Transmission of four Central American strains of Plasmodium vivax from monkey to man. *J Parasitol* 58: 332–335.
48. Cui L, et al. (2005) Gene discovery in Plasmodium vivax through sequencing of ESTs from mixed blood stages. *Mol Biochem Parasitol* 144:1–9.
49. Florens L, et al. (2002) A proteomic view of the Plasmodium falciparum life cycle. *Nature* 419:520–526.
50. Le Roch KG, et al. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* 301:1503–1508.
51. Herrera S, Corradin G, Arévalo-Herrera M (2007) An update on the search for a Plasmodium vivax vaccine. *Trends Parasitol* 23:122–128.
52. Mu J, et al. (2007) Genome-wide variation and identification of vaccine targets in the Plasmodium falciparum genome. *Nat Genet* 39:126–130.
53. Dahlström S, et al. (2009) Plasmodium falciparum multidrug resistance protein 1 and artemisinin-based combination therapy in Africa. *J Infect Dis* 200:1456–1464.
54. Winzeler EA, et al. (1998) Direct allelic variation scanning of the yeast genome. *Science* 281:1194–1197.
55. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: Short oligonucleotide alignment program. *Bioinformatics* 24:713–714.
56. Yang Z, Nielsen R, Hasegawa M (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol* 15:1600–1611.