

Genomic analysis of local variation and recent evolution in

Plasmodium vivax

Pearson RD^{1,2}, Amato R^{1,2*}, Auburn S^{3*}, Miotto O^{1,2,4}, Almagro-Garcia J², Amaratunga C⁵, S Seila⁶, S Mao⁷, Noviyanti R⁸, Trimarsanto H⁸, Marfurt J³, Anstey NM³, William T⁹, Boni MF¹⁰, Dolecek C¹⁰, Hien TT¹⁰, White NJ⁴, Michon P^{11,12}, Siba P¹¹, Tavul L¹¹, Harrison G^{13,14}, Barry A^{13,14}, Mueller I^{13,14}, Ferreira MU¹⁵, Karunaweera N¹⁶, Randrianarivelojosia M¹⁷, Qi G¹⁸, Hubbart C², Hart L², Jeffery B², Drury E¹, Mead D¹, Kekre M¹, Campino S¹, Manske M¹, Cornelius V^{1,2}, MacInnis B¹, Rockett KA^{1,2}, Miles A^{1,2}, Rayner JC¹, Fairhurst RM⁵, Nosten F^{4,19}, Price RN^{3,20}, Kwiatkowski DP^{1,2}

1. Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK
2. MRC Centre for Genomics and Global Health, Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, UK
3. Global and Tropical Health Division, Menzies School of Health Research and Charles Darwin University, Darwin, Northern Territories 0811, Australia
4. Mahidol-Oxford Tropical Medicine Research Unit, Mahidol University, Bangkok 10400, Thailand
5. National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, Maryland 20852, USA
6. National Centre for Parasitology, Entomology, and Malaria Control, Phnom Penh, Cambodia
7. Sampov Meas Referral Hospital, Pursat, Cambodia
8. Eijkman Institute for Molecular Biology, Jakarta 10430, Indonesia
9. Infectious Diseases Society Sabah-Menzies School of Health Research Clinical Research Unit and Queen Elizabeth Hospital Clinical Research Centre, Kota Kinabalu, Sabah, Malaysia
10. Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam
11. Papua New Guinea Institute of Medical Research, Madang, Papua New Guinea
12. Faculty of Medicine and Health Sciences, Divine Word University, Madang, Papua New Guinea
13. Division of Population Health and Immunity, The Walter and Eliza Hall Institute for Medical Research, Parkville, Victoria, Australia
14. Department of Medical Biology, University of Melbourne, Parkville, Victoria, Australia
15. Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil
16. Department of Parasitology, Faculty of Medicine, University of Colombo, Sri Lanka
17. Institut Pasteur de Madagascar, Antananarivo, Madagascar
18. Jiangsu Institute of Parasitic Diseases, Key Laboratory of Parasitic Disease Control and Prevention (Ministry of Health), Jiangsu Provincial Key Laboratory of Parasite Molecular Biology, Wuxi, Jiangsu, People's Republic of China
19. Shoklo Malaria Research Unit, Mae Sot, Tak 63110, Thailand
20. Centre for Tropical Medicine and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, OX3 7LJ, UK

* Equal contribution

The widespread distribution and relapsing nature of *Plasmodium vivax* infection present major challenges for malaria elimination. To characterise the genetic diversity of this parasite within individual infections and across the population, we performed deep genome sequencing of >200 clinical samples collected across the Asia-Pacific region, and analysed data on >300,000 SNPs and 9 regions of the genome with large structural variations. Individual infections showed complex patterns of genetic structure, with variation not only in the number of dominant clones but also in their level of relatedness and inbreeding. At the population level, we observed strong signals of recent evolutionary selection both in known drug resistance genes and at novel loci, and these varied markedly between geographical locations. These findings reveal a dynamic landscape of local evolutionary adaptation in *P. vivax* populations, and provide a foundation for genomic surveillance to guide effective strategies for control and elimination.

P. vivax is the main cause of malarial illness in many parts of the world and it is estimated that over 2.5 billion people are at risk of infection.¹⁻³ It is notably absent from most of sub-Saharan Africa, where the species appears to have originated, because most of the human population is protected from infection by the Duffy negative blood group, suggesting that *P. vivax* has been a strong force for human evolutionary selection.^{4,5} *P. vivax* is a particularly challenging problem for malaria elimination because of its broad geographical range and its ability to produce hypnozoites, dormant forms of the liver-stage parasite that cause relapsing infection and that are refractory to most classes of antimalarial drugs.⁶ *P. vivax* is becoming increasingly resistant to chloroquine, the first-line treatment, and the molecular mechanisms of resistance remain unknown.⁷

Efforts to control *P. vivax* require an understanding of how the parasite population varies between locations and how it is evolving. Microsatellite approaches have yielded useful insights into its epidemiology, population structure and transmission dynamics⁸⁻¹⁰ but analysis of genome variation has so far been restricted to relatively small numbers of samples.¹¹⁻¹⁸ Two practical obstacles to large-scale sequencing are low levels of parasitaemia in clinical samples and the difficulty of culturing *P. vivax ex vivo* for more than a few days. In this study, we collected blood samples from patients with *P. vivax* malaria and performed leucocyte depletion to reduce human DNA content prior to parasite genome sequencing.¹⁹ Using the Illumina Genome Analyzer and Hi-Seq platforms, we generated paired-end sequence reads for 292 clinical samples from 13 endemic countries (Supplementary Table 1). Our sampling frame focused on Southeast Asia (Thailand, Cambodia, Vietnam, Laos, Myanmar and Malaysia) and Oceania (Papua Indonesia and Papua New Guinea), with smaller numbers of samples from China, India, Sri Lanka, Brazil and Madagascar.

In the first stage of analysis, we aligned sequence reads against the Salvador 1 (Sal 1) reference genome¹¹ and used the GATK UnifiedGenotyper to discover 726,077 putative single nucleotide polymorphisms (we refer to these as unfiltered SNPs). Analysis of mapping quality and coverage revealed poor confidence of alignment and a high rate of genotyping errors in subtelomeric regions and in three internal chromosomal regions containing members of the serine repeat antigen family and merozoite surface proteins 3 and 7. The removal of these ‘inaccessible’ regions defines a core genome of 21.4 Mb, comprising 11.1 Mb coding and 10.3 Mb non-coding sequence (Supplementary Figures 1 and 2, Supplementary Table 2). We excluded samples with <50% of genomic positions at $\geq 5x$ coverage (our threshold for genotype calling) and also excluded SNPs with a high risk of genotyping errors based on quality control metrics including discordance between technical replicates. After applying these filters we retained 228 samples and 303,616 high-quality SNPs. To determine replicability of these genotyping data by an independent method, we used the Sequenom primer-extension mass spectrometry platform to type 111 SNPs, giving a concordance rate of 99.98% for homozygous calls and 93.6% when mixed-allele calls were included. For detailed population genetic analyses we used 148 samples from western Thailand, western Cambodia and Papua Indonesia that had genotype calls for >80% of the high-quality SNPs: these had a median read depth of 43x (Supplementary Table 1).

The high-quality SNPs were divided approximately equally between coding and non-coding regions (150,739 vs 152,877), and 58% (87,877) of the coding SNPs were non-synonymous. The allele frequency spectrum was dominated by low frequency variants, with over 50% of high-quality SNPs being at $\leq 1\%$ minor allele frequency (MAF) (Supplementary Figures 3 and 4). Estimated nucleotide diversity (π) based on all unfiltered SNPs was 1.5×10^{-3} , consistent with previously reported values¹⁵; whereas analysis restricted to high quality SNPs gives a more conservative estimate of 5.6×10^{-4} (Supplementary Table 3). Rates of nucleotide diversity (π), Tajima’s D and the ratio of non-synonymous to synonymous variants (N/S ratio) were estimated for individual genes (Table 1, Supplementary Table 4). A striking finding was that π , D and N/S ratio are highly significantly elevated among the >200 genes that lack a known ortholog in *P. falciparum*, *P. yoelii* or both. High levels of diversity were also observed in genes expressed in late schizonts, those containing signal peptides or transmembrane domains, *vir* genes (i.e. those in the accessible genome) and genes encoding reticulocyte binding proteins.

Nine regions of the core genome were found to have large structural variations by analysis of read depth after normalisation by GC-content (Figure 1, Supplementary Tables 5 and 6). The four most common structural variants showed significant differences in frequency between Papua Indonesia

(ID), Western Cambodia (WKH) and Western Thailand (WTH), suggesting different selective pressures. The first was a 9 kb deletion on chromosome 8 (present in 73% ID, 6% WKH, and 3% WTH samples) that includes the first three exons of a gene encoding a cytoadherence-linked asexual protein. The second was a 7 kb duplication on chromosome 6 (5%ID, 35% WKH, 25% WTH) encompassing *pvdbp*, the gene that encodes the Duffy binding protein which mediates *P. vivax* invasion of erythrocytes.² *Pvdbp* duplications have been shown to be common in Malagasy strains of *P. vivax* infecting Duffy-negative individuals²¹, and these findings show they can also reach relatively high frequency in places where nearly all individuals are Duffy-positive²². The third common structural variation was a 37 kb duplication on chromosome 10 that includes *pvmdr1*. Duplication of *pvmdr1* duplication has previously been associated with resistance to mefloquine²³ and is homologous to the *pfmdr1* amplification responsible for mefloquine resistance in *P. falciparum*. Mefloquine has never been a recommended treatment for *P. vivax*; it is therefore of considerable interest that *pvmdr1* duplication is present in 19% of WTH samples, but not in WKH or ID samples. In Western Thailand, mefloquine has been used extensively as the first-line treatment for *P. falciparum*, either as a monotherapy or in combination with artesunate, and likely induces high selective pressure on relapsing *P. vivax* infections, which occur frequently following *P. falciparum* infection²⁴. The fourth common structural variant was a 3kb duplication on chromosome 14 that includes the gene PVX_101445 and was seen only in Papua Indonesia. Notably, this locus also shows signals of recent selection and is discussed further below.

Levels of linkage disequilibrium (LD) were extremely low, e.g. r^2 decayed to <0.1 within <200 bp in WTH and WKH samples, and within <500 bp in ID samples, after correcting for population structure and other confounders (Figure 2). Sexual recombination between parasites occurs in the mosquito shortly after it has ingested blood from an infected individual, so one determinant of LD is the prevalence of genetically mixed infections. This is of particular interest in *P. vivax*, since the hypnozoite phase allows a cohort of parasites from a single mosquito bite to persist within a person for many years,⁶ such that mixed infection can comprise unrelated parasites from separate mosquito bites, or meiotic siblings inoculated by the same mosquito, or a combination of these.^{16,25,26} As a starting point in characterising mixed infections, we calculated the F_{WS} metric²⁷ which showed that many samples were essentially clonal but a significant number had mixed infection (Figure 3A). In each sample, we examined the distribution of non-reference allele frequencies (NRAF) across all SNPs (Figure 3B, Supplementary Figure 5). In 55% of samples the vast majority of SNPs were homozygous, indicating an infection with a single, dominant clone. In 28% of samples, there were many heterozygous SNPs, and their NRAF distribution was bimodal and symmetrical, indicating the presence of two dominant clones. In the remaining 16% of samples,

there were many heterozygous SNPs and the NRAF distribution showed more than 2 peaks (or had no clear pattern), indicating that these samples had more complex patterns of mixed infection. The above proportions are averaged across WTH, WKH and ID, but broadly similar patterns were observed in each population (Supplementary Table 7).

To investigate whether mixed infections were due to genetically related or unrelated parasites, we examined how heterozygosity varied across the genome (Figure 3B and Supplementary Figure 5). Among samples containing two dominant clones, 60% had long runs of homozygosity (RoH), implying that the clones were closely related, while the remaining 40% displayed heterozygosity across the genome, implying that the clones were unrelated. These RoH are equivalent to the long blocks of haplotype-sharing that have been observed by single cell genome sequencing of meiotic sibling parasites isolated from the same infected individual.²⁸ RoH extending across ~50% of the genome suggests that the two clones are meiotic siblings, while less extensive RoH indicates more a distant relationship and more extensive RoH is indicative of inbreeding over multiple generations. Among samples with 2 dominant clones and significant RoH, the RoH extended over 40-60% of the genome in 11 samples, <40% in 9 samples and >60% in 5 samples. A few samples with >2 dominant clones also displayed RoH, suggesting that these infections were dominated by a group of closely related parasites.

Taken together, these data reveal complex patterns of mixed infection that cannot be simply summarised with a single metric (Figure 3). There is variation not only in the number of dominant clones in a sample, but also in their relative proportions and their relatedness. F_{WS} , which is analogous to an inbreeding coefficient²⁷, is determined both by the number of clones and their relatedness. However, by aggregating these different types of information, deep sequencing could provide a useful epidemiological tool to differentiate mixed infections that are due to separate mosquito bites from those that are due to sibling parasites inoculated by the same mosquito.^{6,16,25,26,28}

At the level of the whole population, principal components analysis revealed a population structure that was primarily determined by geographic location (Figure 4). These data are consistent with previous microsatellite and SNP barcode studies^{10,29}, but provide a higher level of resolution. Using a more formal model-based approach (ADMIXTURE) we found that the best fit was obtained for three putative populations, and these showed perfect separation between our three main sampled populations of western Thailand, western Cambodia and Papua Indonesia (Figure 4, Supplementary Figure 6).³⁰ A neighbour-joining tree, used to visually represent the genetic distance matrix, also shows three distinct branches separating Western Southeast Asia (Western Thailand, Myanmar and

China), Eastern Southeast Asia (Cambodia, Vietnam, Eastern Thailand and Laos) and Southeast Asian and Pacific Islands (Malaysia, Papua Indonesia and Papua New Guinea). The separation of the *P. vivax* population of Southeast Asia into distinct Western and Eastern groups is consistent with observations in *P. falciparum*³¹ and reflects the malaria-free corridor that has been established through central Thailand. Samples from outside Southeast Asia were too disparate and small in numbers to be reliably assigned to specific groups of population structure by this analysis.

We searched for evidence of recent evolutionary selection in western Thailand, western Cambodia and Papua Indonesia. These locations represent different branches in the population genetic structure analysis, with an estimated mean pairwise F_{ST} of 0.031 for WTH-WKH, 0.078 for WTH-ID, and 0.069 for WKH-ID. Six regions of the genome showed strong evidence of recent selection based on the cross-population extended haplotype homozygosity test (XP-EHH), with P values of 10^{-8} to 10^{-18} , supported by other evidence such as the integrated haplotype score (iHS) and highly differentiated SNPs (Figure 5, Supplementary Table 8 and Supplementary Figure 7). Each of these regions encompasses multiple genes, such that specific genes under selection cannot be identified from these data alone. Nonetheless, there are several noteworthy candidate genes whose key features are summarised.

Signals of recent selection on chromosomes 5, 13 and 14 were observed in western Thailand relative to western Cambodia. The chromosome 5 and 14 regions contain well-known genes for resistance to pyrimethamine (*pvdhfr*) and sulfadoxine (*pvdhps*)^{32,33}. Although chloroquine has been the main treatment for *P. vivax* malaria, sulfadoxine-pyrimethamine was introduced to Thailand in 1973 as first-line treatment for *P. falciparum*³⁴, and selective pressure on *P. vivax* may have been considerable because of its widespread use in the private sector and the high frequency of *P. vivax* relapses following treatment of *P. falciparum* infection²⁴. Selective sweeps at *pvdhfr* and *pvdhps* have also been observed in South America.^{17,18} The chromosome 13 region includes PVX_084940, which encodes a putative voltage-dependent anion-selective channel containing a porin domain proposed to be implicated in antibiotic resistance³⁵.

The strongest signals of recent selection, on chromosome 10 and 14, were observed in ID relative to WTH and WKH. This is of particular interest since high-grade chloroquine resistant *P. vivax* is firmly established in Papua Indonesia⁷ and the molecular mechanism is unknown. Moreover, there was no signal of selection in *pvcr1-o*, the orthologue of the main gene responsible for chloroquine resistance in *P. falciparum*.³⁶ The chromosome 14 region under selection encompasses 22 genes, among which the strongest candidate appears to be PVX_101445, a hypothetical membrane protein. PVX_101445 possesses duplications seen in Indonesia but not elsewhere, with up to four copies in a single

sample, as well as two non-synonymous SNPs (I81L and R97G) which both have a high derived allele frequency in ID (1.0) but not in WTH (0.11) or WKH (0.0). Of 13 ID samples with gene amplification in this region, three encompass multiple genes, while the remaining contain solely PVX_101445; four samples were found to carry three or four ($n=2$) copies of the gene (Figure 1, Supplementary Tables 5 and 6). The chromosome 10 region under selection encompasses 29 genes and lies within 150 kb of *pvm-dr1*, which has been implicated in chloroquine resistance in *ex-vivo* studies in Papua Indonesia³⁷. The three most highly differentiated SNPs in this region ($\text{NRAF} \leq 0.05$ in WTH and WKH, $\text{NRAF} \geq 0.95$ in ID) are non-synonymous mutations (E2699K, P486S and E81K) in PVX_079910, a conserved protein of unknown function, which produced the highest XP-EHH scores in this region.

Another signal of selection, observed in WTH/WKH relative to ID, was in a region of chromosome 2 that contains four genes including *pvm-rp1* (PVX_097025). This gene encodes an ABC transporter that has been implicated as a drug resistance candidate^{12,18} and whose *P. falciparum* homologues are associated with resistance to multiple anti-malarial drugs^{38,39}. The most highly differentiated SNP in this region causes a T234M mutation in *pvm-rp1* and has derived allele frequencies of 0.91, 0.95 and 0.00 in WTH, WKH and ID respectively. It is possible that the chromosome 2 signal reflects resistance to an antimalarial drug, such as mefloquine, that has been used to treat *P. falciparum* in Cambodia and Thailand but not in Indonesia.^{40,41}

SNPs that are highly differentiated between populations can provide additional evidence of evolutionary selection. Pairwise comparisons between WTH, WKH and ID identified 40 SNPs with $F_{ST} > 0.9$ (Supplementary Table 9), half of which were in the regions of recent positive selection. The remaining 20 SNPs showed a significantly higher proportion of non-synonymous changes than the genome-wide average (12/20 vs 87,877/303,616; $P=3.3 \times 10^{-4}$ by Fisher's exact test), and included SNPs in a putative drug/metabolite transporter gene (PVX_122995) with 8 predicted transmembrane domains, and in a putative transporter gene with 13 predicted transmembrane domains (PVX_003935), providing additional candidate genes for drug resistance studies (Supplementary Note). We note that the signals of selection highlighted here are those with the strongest statistical support, but there are likely to be many other examples of recent evolution in this data. More generally, this study provides a rich resource of data on the population diversity of *P. vivax*, which can be explored at www.malariagen.net/apps/pvgv. [Note to reviewers: the link is a draft web application that will be updated prior to publication.] This web application provides summary data on all of the SNPs analysed in this study, and allows users to undertake their own analyses of allele frequencies and differentiation between populations.

This study demonstrates the feasibility of population-level genome sequencing of *P. vivax*, despite the low levels of parasitaemia in clinical samples and the lack of an effective culture method. As well as characterising common patterns of genome variation that are the result of ancient events, the present findings reveal a dynamic evolutionary landscape, in which the parasite population is adapting to local selective pressures that reflect ongoing epidemiological processes. The difficulty of investigating *P. vivax* in the laboratory provides a strong incentive to exploit genomics to address gaps in knowledge of parasite phenotype. Genomic signals of recent selection could help identify local emergences of resistance, both to the drugs used specifically to treat *P. vivax* and to those that are targeted at *P. falciparum*. Knowledge of the genetic structure of individual infections is an important step towards understanding local patterns of malaria transmission, the epidemiology of relapsing infection, and the dynamics of genetic recombination in natural populations of *P. vivax*. Taken together, these findings point to various ways in which genomic analyses might be integrated into future clinical and epidemiological studies of *P. vivax*, and highlight the importance of translating this information into more effective strategies for malaria control and elimination.

Table 1

Gene categories enriched for high N/S ratio, nucleotide diversity, and Tajima's *D*. Each metric is represented by its median and *P* value by Mann-Whitney test, comparing genes in a given category versus all others, with bold font indicating significant values ($P < 0.05$ after Bonferroni correction). Rows are ordered by π . N/S=non-synonymous/synonymous ratio. π =nucleotide diversity per base. *D*=Tajima's *D*. No Pf/Py ortholog=genes that lack a known ortholog in *P. falciparum*/*P. yoelii*. TM domain=genes containing a transmembrane domain. Max schizont=maximum expression during the intraerythrocytic cycle was in late schizont stage⁴². Max sporozoite/zygote/ookinete=maximum expression in the sporozoite/zygote/ookinete⁴³. These estimates are based on high-quality SNPs in genes with ≥ 10 SNPs in the subset of 148 samples used for detailed population comparisons as described in Methods. Estimates for individual genes, including all SNPs or restricted to high-quality SNPs, are given in Supplementary Table 4.

Comparison	Genes	N/S	<i>P</i> (N/S)	π	<i>P</i> (π)	<i>D</i>	<i>P</i> (<i>D</i>)
No Pf ortholog	97	2.23	6.9x10⁻¹⁸	7.3x10 ⁻⁴	7.5x10⁻⁹	-1.86	2.6x10⁻⁴
No Py ortholog	251	1.86	1.1x10⁻²⁰	6.7x10 ⁻⁴	7.1x10⁻¹¹	-1.92	5.3x10⁻⁸
Max schizont	844	1.60	2.0x10⁻¹³	6.1x10 ⁻⁴	5.1x10⁻⁷	-2.04	1.7x10⁻⁴
Max sporozoite	422	1.43	3.6x10 ⁻¹	6.0x10 ⁻⁴	3.2x10 ⁻²	-2.03	6.9x10 ⁻²
Signal peptide	569	1.46	6.5x10 ⁻²	6.0x10 ⁻⁴	6.1x10⁻⁶	-1.95	1.3x10⁻¹²
TM domain	646	1.50	1.9x10 ⁻²	5.9x10 ⁻⁴	1.2x10⁻⁴	-1.98	1.7x10⁻¹³
Max ookinete	230	1.40	6.1x10 ⁻¹	5.8x10 ⁻⁴	2.0x10 ⁻¹	-2.08	8.6x10 ⁻¹
Has paralog	206	1.38	3.4x10 ⁻¹	5.7x10 ⁻⁴	6.4x10 ⁻²	-2.01	5.8x10 ⁻³
Max zygote	339	1.35	2.6x10 ⁻²	5.4x10 ⁻⁴	7.4x10 ⁻¹	-2.10	8.9x10 ⁻¹
All genes	3062	1.43		5.5x10 ⁻⁴		-2.07	

References

1. Gething, P. W. *et al.* A long neglected world malaria map: Plasmodium vivax endemicity in 2010. *PLoS Negl. Trop. Dis.* **6**, e1814 (2012).
2. Miller, L. H., Mason, S. J., Clyde, D. F. & McGinniss, M. H. The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. *N. Engl. J. Med.* **295**, 302–4 (1976).
3. Ménard, D. *et al.* Plasmodium vivax clinical malaria is commonly observed in Duffy-negative Malagasy people. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5967–71 (2010).
4. Price, R. N. *et al.* Vivax malaria: neglected and not benign. *Am J Trop Med Hyg* **77**, 79–87 (2007).
5. Battle, K. E. *et al.* The global public health significance of Plasmodium vivax. *Adv Parasitol* **80**, 1 (2012).
6. White, N. J. Determinants of relapse periodicity in Plasmodium vivax malaria. *Malar. J.* **10**, 297 (2011).
7. Price, R. N. *et al.* Global extent of chloroquine-resistant Plasmodium vivax: a systematic review and meta-analysis. *Lancet. Infect. Dis.* **14**, 982–91 (2014).
8. Karunaweera, N. D. *et al.* Extensive microsatellite diversity in the human malaria parasite Plasmodium vivax. *Gene* **410**, 105–112 (2008).
9. Barry, A. E., Waltmann, A., Koepfli, C., Barnadas, C. & Mueller, I. Uncovering the transmission dynamics of Plasmodium vivax using population genetics. *Pathog. Glob. Health* **109**, 142–152 (2015).
10. Koepfli, C. *et al.* Plasmodium vivax Diversity and Population Structure across Four Continents. *PLoS Negl. Trop. Dis.* **9**, e0003872 (2015).
11. Carlton, J. M. *et al.* Comparative genomics of the neglected human malaria parasite Plasmodium vivax. *Nature* **455**, 757–763 (2008).
12. Dharia, N. V *et al.* Whole-genome sequencing and microarray analysis of ex vivo Plasmodium vivax reveal selective pressure on putative drug resistance genes. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 20045–50 (2010).

13. Hester, J. *et al.* De novo assembly of a field isolate genome reveals novel *Plasmodium vivax* erythrocyte invasion genes. *PLoS Negl. Trop. Dis.* **7**, e2569 (2013).
14. Chan, E. R. *et al.* Whole Genome Sequencing of Field Isolates Provides Robust Characterization of Genetic Diversity in *Plasmodium vivax*. *PLoS Negl. Trop. Dis.* **6**, e1811 (2012).
15. Neafsey, D. E. *et al.* The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. *Nat. Genet.* **44**, 1046–1050 (2012).
16. Bright, A. T. *et al.* 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.* **8**, e2882 (2014).
17. Winter, D. J. *et al.* Whole Genome Sequencing of Field Isolates Reveals Extensive Genetic Diversity in *Plasmodium vivax* from Colombia. *PLoS Negl. Trop. Dis.* **9**, e0004252 (2015).
18. Flannery, E. L. *et al.* Next-Generation Sequencing of *Plasmodium vivax* Patient Samples Shows Evidence of Direct Evolution in Drug-Resistance Genes. *ACS Infect. Dis.* **1**, 367–379 (2015).
19. Auburn, S. *et al.* Characterization of within-host *Plasmodium falciparum* diversity using next-generation sequence data. *PLoS One* **7**, e32891 (2012).
20. Miles, A. *et al.* Genome variation and meiotic recombination in *Plasmodium falciparum*: insights from deep sequencing of genetic crosses. *bioRxiv* 024182 (2015).
doi:10.1101/024182
21. Menard, D. *et al.* Whole genome sequencing of field isolates reveals a common duplication of the Duffy binding protein gene in Malagasy *Plasmodium vivax* strains. *PLoS Negl. Trop. Dis.* **7**, e2489 (2013).
22. Howes, R. E. *et al.* The global distribution of the Duffy blood group. *Nat. Commun.* **2**, 266 (2011).
23. Suwanarusk, R. *et al.* Amplification of *pvm-dr1* associated with multidrug-resistant *Plasmodium vivax*. *J. Infect. Dis.* **198**, 1558–1564 (2008).
24. Douglas, N. M. *et al.* *Plasmodium vivax* recurrence following *falciparum* and mixed species malaria: risk factors and effect of antimalarial kinetics. *Clin. Infect. Dis.* **52**, 612–20 (2011).

25. Imwong, M. *et al.* The first *Plasmodium vivax* relapses of life are usually genetically homologous. *J. Infect. Dis.* **205**, 680–3 (2012).
26. Lin, J. T. *et al.* Using Amplicon Deep Sequencing to Detect Genetic Signatures of *Plasmodium vivax* Relapse. *J. Infect. Dis.* **212**, 999–1008 (2015).
27. Manske, M. *et al.* Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature* **487**, 375–379 (2012).
28. Nair, S. *et al.* Single-cell genomics for dissection of complex malaria infections. *Genome Res.* **24**, 1028–38 (2014).
29. Baniecki, M. L. *et al.* Development of a Single Nucleotide Polymorphism Barcode to Genotype *Plasmodium vivax* Infections. *PLoS Negl. Trop. Dis.* **9**, e0003539 (2015).
30. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–20 (2005).
31. Miotto, O. *et al.* Genetic architecture of artemisinin-resistant *Plasmodium falciparum*. *Nat. Genet.* **47**, 226–34 (2015).
32. Korsinczky, M. *et al.* Sulfadoxine resistance in *Plasmodium vivax* is associated with a specific amino acid in dihydropteroate synthase at the putative sulfadoxine-binding site. *Antimicrob. Agents Chemother.* **48**, 2214–2222 (2004).
33. Imwong, M. *et al.* Novel point mutations in the dihydrofolate reductase gene of *Plasmodium vivax*: evidence for sequential selection by drug pressure. *Antimicrob. Agents Chemother.* **47**, 1514–1521 (2003).
34. Alam, M. T. *et al.* Tracking origins and spread of sulfadoxine-resistant *Plasmodium falciparum* dhps alleles in Thailand. *Antimicrob. Agents Chemother.* **55**, 155–164 (2011).
35. Pagès, J.-M., James, C. E. & Winterhalter, M. The porin and the permeating antibiotic: a selective diffusion barrier in Gram-negative bacteria. *Nat. Rev. Microbiol.* **6**, 893–903 (2008).
36. Pava, Z. *et al.* Expression of *Plasmodium vivax* crt-o is related to parasite stage but not ex vivo chloroquine susceptibility. *Antimicrob. Agents Chemother.* (2015). doi:10.1128/AAC.02207-15
37. Suwanarusk, R. *et al.* Chloroquine resistant *Plasmodium vivax*: in vitro characterisation and

- association with molecular polymorphisms. *PLoS One* **2**, e1089 (2007).
38. Mu, J. *et al.* Multiple transporters associated with malaria parasite responses to chloroquine and quinine. *Mol Microbiol* **49**, 977–989 (2003).
 39. Raj, D. K. *et al.* 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 (2009).
 40. WWARN. History Of Resistance. (2015). at <http://www.wwarn.org/resistance/malaria/history>
 41. Maguire, J. D. & Marwoto, H. Mefloquine is highly efficacious against chloroquine-resistant Plasmodium vivax malaria and Plasmodium falciparum malaria in Papua, Indonesia. *Clin. Infect. {...}* **2197**, 1067–1072 (2006).
 42. Bozdech, Z. *et al.* The transcriptome of Plasmodium vivax reveals divergence and diversity of transcriptional regulation in malaria parasites. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 16290–16295 (2008).
 43. Westenberger, S. J. *et al.* A systems-based analysis of Plasmodium vivax lifecycle transcription from human to mosquito. *PLoS Negl. Trop. Dis.* **4**, e653 (2010).
 44. Tao, Z.-Y., Xia, H., Cao, J. & Gao, Q. Development and evaluation of a prototype non-woven fabric filter for purification of malaria-infected blood. *Malar. J.* **10**, 251 (2011).
 45. Auburn, S. *et al.* Effective preparation of Plasmodium vivax field isolates for high-throughput whole genome sequencing. *PLoS One* **8**, e53160 (2013).
 46. Li, H. & Durbin, R. *Fast and accurate short read alignment with Burrows-Wheeler transform.* *Bioinformatics* **25**, 1754–1760 (2009).
 47. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
 48. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
 49. DePristo, M. a *et al.* A framework for variation discovery and genotyping using next-

- generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
50. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. **6**, 80–92
 51. Logan-Klumpler, F. J. *et al.* GeneDB--an annotation database for pathogens. *Nucleic Acids Res.* **40**, D98–108 (2012).
 52. Tachibana, S.-I. *et al.* Plasmodium cynomolgi genome sequences provide insight into Plasmodium vivax and the monkey malaria clade. *Nat. Genet.* **44**, 1051–1055 (2012).
 53. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–64 (2009).
 54. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–75 (2007).
 55. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
 56. Voight, B. F., Kudravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72 (2006).

Methods

Ethics statement

All samples used in this study were derived from patient blood samples obtained with informed consent from the patient or a parent or guardian. At each location, sample collection was approved by the appropriate local ethics committee: Eijkman Institute Research Ethics Committee, Jakarta, Indonesia; Human Research Ethics Committee of NT Department of Health and Families and Menzies School of Health Research, Darwin, Australia; Oxford Tropical Research Ethics Committee, Oxford, UK; Ethics Committee, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand; Research Review Committee of the Institute for Medical Research and the Medical Research Ethics Committee (MREC), Ministry of Health Malaysia; Review Board of Jiangsu Institute of Parasitic Diseases, Wuxi, China; National Ethics Committee for Health Research, Phnom Penh, Cambodia; Institutional Review Board, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland, USA; National Ethics Committee for Health Research, Lao Peoples' Democratic Republic; The Government of the Republic of the Union of Myanmar, Ministry of Health, Department of Medical Research (Lower Myanmar); Institutional Review Board of the Institute of Biomedical Sciences, University of São Paulo, Brazil; Scientific and Ethical Committee of the Hospital for Tropical Diseases in Ho Chi Minh City, Vietnam; Ethics Review Committee, Faculty of Medicine, University of Colombo, Sri Lanka; Papua New Guinea Institute of Medical Research Institutional Review Board, the Medical Research Advisory Committee of Papua New Guinea and the Walter and Eliza Hall Institute Human Research Ethics Committee; National Ethics Committee of Madagascar.

Sample preparation

Samples were collected from patients presenting at hospitals or health centres with symptomatic, uncomplicated *P. vivax* malaria as determined by microscopy. Venous blood was drawn into tubes coated with ethylenediaminetetraacetic acid (EDTA) or lithium heparin, and leukocyte depletion was carried out to minimise the amount of human DNA in the sample to be sequenced. Methods for leukodepletion included magnetic cell separation technology and filtration using non-woven fabric filters or cellulose-based constructs^{44,45}. Some samples were also cultured *ex vivo* for up to 48 h to enrich for schizonts⁴⁵. DNA extraction was typically performed using the QIAamp Blood Midi or Maxi kits (Qiagen) according to the manufacturer's instructions. Total DNA concentration was measured using the Quant-iT™ dsDNA HS assay (Invitrogen) as per the manufacturer's protocol, and the proportion of human DNA in each sample was determined by RT-qPCR.⁴⁵

DNA derived from leukocyte-depleted monkey blood was obtained for the Sal1 reference strain from the Malaria Research and Reference Reagent Resource Center (<https://www.mr4.org>). Owing to the limited quantity of Sal1 DNA available, a genomic DNA aliquot (10 ng) was subject to multiple-displacement whole-genome amplification using the REPLI-g kit according to the manufacturer's instructions (Qiagen).

DNA sequencing

Sequencing was performed on the Illumina GA II or HiSeq 2000 platform at the Wellcome Trust Sanger Institute. Paired-end multiplex or non-multiplex libraries were prepared using the manufacturer's protocol, with the exception that genomic DNA was fragmented using Covaris Adaptive Focused Acoustics rather than nebulisation. Multiplexes comprised 12 tagged samples. Cluster generation and sequencing were undertaken according to the manufacturer's protocol for paired-end 75 bp, 76 bp or 100 bp sequence reads. We initially used 272 samples from confirmed cases of *P. vivax* malaria that had at least 50 ng total gDNA with $\leq 80\%$ human DNA. At the analysis stage, we included 20 additional samples from presumed cases of *P. falciparum* malaria, which were found by sequencing to have substantial proportions of reads mapping to the *P. vivax* reference genome (Supplementary Table 1). Illumina sequence reads have been submitted to the European Nucleotide Archive (study accessions ERP000194 and ERP003694).

Read mapping and coverage

Reads mapping to the human reference genome were removed before all analyses, and the remaining reads were mapped to the *P. vivax* Sal1 reference genome¹¹ (http://plasmodb.org/common/downloads/release-10.0/PvivaxSal1/fasta/data/PlasmoDB-10.0_PvivaxSal1_Genome.fasta) using bwa⁴⁶ version 0.5.9-r16 with default parameters. Standard alignment metrics were generated for each sample using the bamcheck utility from samtools⁴⁷.

Various "bam improvement" steps were applied to the bwa outputs before further analyses. The Picard (<http://picard.sourceforge.net>) tools CleanSam, FixMateInformation and MarkDuplicates were successively applied to the bam files of each sample, using Picard version 1.110. GATK version 3.1-1 indel realignment⁴⁸ was applied using default parameters and no list of known indels. The output of this stage was a set of 292 "improved" bam files, one for each sample.

We ran GATK's CallableLoci⁴⁹ on each improved sample bam file to determine the proportion of genomic positions callable in each sample using parameters `--minDepth 5 --minBaseQuality 27 --`

minMappingQuality 27. This identifies a site as callable if there are ≥ 5 reads with base and mapping quality of ≥ 27 and if $\leq 10\%$ of reads have mapping quality 0.

The *P. vivax* Sal1 reference genome¹¹ consists of 14 large chromosomal sequences ranging in size from 0.76-3.12 Mbp, and 2,733 shorter contigs ranging in size from 200-101,928 bases. It is assumed that these shorter contigs are sequences from the subtelomeric ends of the autosomal chromosomes. In all subsequent analyses, we have analysed only those reads that mapped to the 14 large chromosomal sequences, which are named Pv_Sal1_chr01 - Pv_Sal1_chr14.

A total of 247 samples were identified as having at least 50% of Pv_Sal1_chr01 - Pv_Sal1_chr14 positions whose genotypes could be reliably called. After trimming the dataset to remove instances of multiple samples from the same individual, , we were left with 228 samples for further analysis (Supplementary Table 1).

SNP discovery and annotation

We discovered potential SNPs by running GATK's UnifiedGenotyper⁴⁹ across all 247 sample-level bam files with parameters `--downsampling_type NONE --sample_ploidy 1 --output_mode EMIT_VARIANTS_ONLY --min_base_quality_score 17 --genotyping_mode DISCOVERY --genotype_likelihoods_model SNP -stand_emit_conf 4.0 -stand_call_conf 4.0 --p_nonref_model EXACT_GENERAL_PLOIDY -contamination 0 --computeSLOD`

SNPs were annotated using a number of different methods. Functional annotations were applied using snpEff version 2.0.5⁵⁰, with gene annotations downloaded from GeneDB⁵¹ at ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/P_vivax/2014/May_2014. GATK VariantAnnotator was used to create the following standard annotation metrics: BaseQRankSum, DP, Dels, FS, HaplotypeScore, HRun, MQ, MQRankSum, MQ0, QD and ReadPosRankSum.

Because GATK's UnifiedGenotyper outputs unfiltered allele depths at each SNP for each sample, we created custom Python scripts based on the pyvcf (<https://github.com/jamescasbon/PyVCF>) and pysam (<https://code.google.com/p/pysam>) modules to calculate filtered allele depths. Low quality reads with either mapping quality < 27 or base quality < 27 at the SNP were removed from allele depth calculations, giving allele depths at high-quality reads only. We created a "NonUniqueness" score (UQ)²⁷ for each position in the reference genome and annotated each SNP with this score. Under Hardy–Weinberg equilibrium, it is expected that heterozygosity at a given SNP (the probability of observing multiple alleles in the same sample) is related to its allele frequency in the population and to the inbreeding coefficient of that population by the relationship $h =$

$2(1 - f)p(1 - p)$, where p is the frequency of the SNP in the population, h its expected heterozygosity, and f the inbreeding coefficient of the population. A substantial divergence from this relationship is likely to arise from alignment artefacts, such as systematic incorrect mappings of reads from paralogous regions. Given that f is unknown and can be influenced by various epidemiological factors, we estimated a surrogate from the data as follows. We used the set of all discovered SNPs with MAF >0.05 to fit a quadratic model of the form $y = mx(1 - x)$, where x represents the allele frequency and y the observed heterozygosity. We obtained a robust estimate of $m = \dots$ by using the `rq` implementation in the R `quantreg` package and using a median regression (which is more robust to outliers than standard mean regression).

For all the SNPs, we calculated the “residuals”, i.e. the difference between the heterozygosity predicted from the model given the allele frequency and the actual heterozygosity observed in the population. The residuals were used as a HyperHeterozygosity score, which was subsequently used in variant filtering. We annotated each SNP with the names of any gene, coding sequence, transfer RNA, small nuclear RNA or repeat region, based on annotations available in GeneDB⁵¹ (ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/P_vivax/2014/May_2014/). We imputed the ancestral allele at SNPs by comparison with the closely related species *P. cynomolgi*. Illumina reads from this species generated in a recent study⁵² were downloaded from ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/DRA000/DRA000196/DRX000265. We mapped the reads against the *P. vivax* reference using `bwa`⁴⁶ version 0.6.2-r126. We then selected the SNPs discovered in our *P. vivax* samples and genotyped (with respect to the *P. vivax* reference) these positions in the *P. cynomolgi* data using GATK’s UnifiedGenotyper (version 3.1-1). Where the genotype in the *P. cynomolgi* was the same as one of the alleles seen in our *P. vivax* data, the allele was defined as ancestral. In this way we were able to impute ancestral alleles for 30% of the *P. vivax* SNPs.

Determining SNP genotype

Because many of our samples exhibit evidence of mixed infection, we did not use the GATK genotype calls, as these are made under an assumption of clonality. Instead, genotypes were defined based on filtered allele depths. Genotypes were called as:

- *homozygous reference* if there were ≥ 5 high-quality reference allele reads and ≤ 1 high-quality alternative allele read.
- *homozygous alternative* if there were ≥ 5 high-quality alternative allele reads and ≤ 1 high-quality reference allele read

- *heterozygous* if there were ≥ 2 high-quality reference allele reads, ≥ 2 high-quality alternative allele reads and ≥ 5 high-quality reads in total
- *missing* if there were < 5 high-quality reads.

For each SNP, we created a Missingness score, which was the number of samples from all 247 samples that had a missing genotype based on the above rules.

Variant filtering

Filtering of variants derived from sequencing data generally uses either (1) a machine learning approach where a set of known high-quality SNPs is available *a priori* or (2) a hard filtering approach where thresholds are set on annotation metrics for which extreme values are known or thought to be enriched for artefactual SNPs. Given that a large set of known high-quality SNPs is not available for *P. vivax* we used the second approach.

The filtering strategy proceeds by first identifying a set of SNPs for which there is evidence of genotyping errors. We then identify regions of the genome that are enriched for errors. After masking these regions, we identify a set of thresholds on the annotation metrics described in the previous section that identify sets of SNPs in the unmasked regions that are enriched for genotyping errors.

A SNP was considered to have a genotyping error if it had a discordant genotype call in at least one of twelve pairs of technical replicate samples (two sequencing runs based on DNA extracted from the same blood sample). All twelve technical replicate sample pairs had < 200 homozygous genotype call differences between samples. For each discovered SNP we calculate the technical replicate discordance rate as:

$$\frac{\# \text{ discordant genotypes between technical replicates}}{\# \text{ technical replicate pairs where both sample have nonmissing genotypes}}$$

When analysing the genome-wide mapping quality and position coverage, we observed that certain regions, such as those near the ends of chromosomal sequences containing *vir* genes, and some internal chromosomal regions, had much lower mapping quality and higher levels of missingness (Supplementary Figure 2A). Subsequent analyses also showed these regions to have higher densities of variants, and as such we term these regions hypervariable regions. In addition to subtelomeric regions, we also declared the following three internal chromosomal regions as hypervariable:

1. A region on Pv_Sal1_chr04 containing 13 genes of the SERA family.

2. A region on Pv_Sal1_chr10 containing 11 genes of the msp3 family.
3. A region on Pv_Sal1_chr12 containing 11 genes of the msp7 family.

We name hypervariable regions at chromosome ends *SubtelomericHypervariable*, and hypervariable regions within chromosomes *InternalHypervariable*. All other regions are named *Core*.

Supplementary Table 2 shows the start and end coordinates of each of these regions. The hypervariable regions defined make up a total of 1.26/22.62 Mb (5.6%) of the chromosomal genome sequences.

The hypervariable regions were identified on the basis of low mapping quality and high missingness. For all hypervariable regions we see at least two consecutive 10-kb windows where both the mapping quality and missingness are in the most extreme 5% of the distribution across all windows. There are no other regions of the genome for which this is the case.

In order to convince ourselves that the lower mapping qualities and higher levels of missingness in the hypervariable were due to intrinsic properties of the reads mapping to those regions, rather than being due to some artefact of the SNPs we discovered in those regions, we analysed the qualities of the reads at all genomic positions using the outputs of the GATK's CallableLoci. We restricted this analysis to 187 samples which had the highest levels of genome-wide coverage (those having at least 97% of genomic positions callable). Within the Core, InternalHypervariable and SubtelomericHypervariable regions, we determined the proportions of positions that were not callable for different reasons (Supplementary Figure 2B). This analysis showed that the lower mapping qualities and higher levels of missingness were seen at all genomic positions within the hypervariable regions, and not only at variable positions. The lower mapping quality in hypervariable regions is likely to be largely due to paralogy between different members of gene families, while the higher missingness is likely to be largely due to greater divergence from the reference genome in these regions.

The hypervariable regions also have higher discordance rate of technical replicates, which we assume to be due to a greater proportion of artefactual SNPs, most likely driven by mapping problems. Although the hypervariable regions defined make up a total of only 5.6% of the chromosomal genome sequences, they contain 18.9% discovered SNPs, and these SNPs contain 57,944/191,256 (30.3%) of all discordant genotypes among technical replicate. For all subsequent analyses, SNPs discovered in hypervariable regions are masked and not used. To simplify analysis we also filter out 21,367 SNPs that are not biallelic, and 59,459 SNPs for which our genotyping method results in reference or missing genotypes for all samples. After removing 137,652 SNPs that

fall in the masked regions defined above we are left with 529,048 SNPs with a mean discordance rate of 2.5% between technical replicates.

For each of our 14 variant annotation metrics (11 standard GATK variant annotations, plus HyperHeterozygosity, Missingness and NonUniqueness scores) we calculate the mean technical replicate duplicate discordance ratio for SNPs within each percentile of the metric (Supplementary Figure 8). We remove SNPs at extreme values of each metric. We define extreme values as those beyond which the mean replicate duplicate discordance ratio is greater than double the mean value across all SNPs. The SNPs retained are those that meet all of the following criteria:

- BaseQRankSum \geq -14.975
- BaseQRankSum \leq 6.706
- DP \geq 11844
- DP \leq 17798
- Dels = 0.0
- FS \leq 14.63418
- HRun \leq 4.0
- HaplotypeScore \leq 14.739201
- MQ \geq 51.6
- MQ0 \leq 1
- MQRankSum \geq -7.924
- MQRankSum \leq 13.387
- QD \geq 12.43
- ReadPosRankSum \geq -6.149
- ReadPosRankSum \leq 3.791
- QuadFitRes_AllQCplus \leq 0.036298
- MISS_AllQCplus \leq 0.158
- UQ \leq 23

Our final set of 303,616 SNPs passed all the above filters and had a mean discordance rate of 0.53% among technical replicates.

Sequenom analysis of genotyping concordance

The Sequenom[®] primer-extension mass spectrometry genotyping platform was used to validate SNP genotype calls made by Illumina sequencing. Two separate validation experiments were performed using laboratory procedures described previously²⁷. In the first validation experiment we developed Sequenom assays for 164 genome-wide SNPs and tested these on 142 of the *P. vivax* samples described above. In the second validation experiment we developed Sequenom assays for 107 SNPs in erythrocyte invasion and putative drug resistance genes, and tested these on 220 of the above *P. vivax* samples. We then applied quality control filters to the Sequenom data, removing samples with missing genotypes in \geq 50% of assays and SNPs that gave artefactual genotype calls in blank control samples containing water or human DNA. We were left with 111 SNPs that could be reliably

typed by Sequenom and were also among the set of 303,616 high quality SNPs typed by Illumina sequencing. This gave a concordance rate of 93.6% overall and of 99.98% for homozygous calls (Supplementary Table 10). This was considered to be satisfactory since Sequenom assays are difficult to calibrate for use in mixed infections, and previous work on *P. falciparum* has shown Illumina sequencing to be generally more reliable than Sequenom for heterozygous calls (see supplementary material to ref²⁷).

Copy number variation

Coverage in non-overlapping 300bp bins was calculated using pysamstats (<https://github.com/alimanfoo/pysamstats>). Normalisation was undertaken within each sample by dividing the coverage by the median coverage across all bins with the same integer percentage GC content. Copy number variants (CNVs) were called using a hidden Markov model with the Python package sklearn.hmm.GaussianHMM using a similar procedure to that used previously for *P. falciparum* genetic crosses²⁰. Two samples (PH0914-Cx and PH0177-C) were removed from this analysis as they had excessive variation in read coverage. Our analysis focused on CNVs >3kbp and those detected by read-depth analysis were further validated by assessment of read pair orientation in the breakpoint regions

Samples used for population genetic analyses

For population genetic analyses we selected samples that were typable at >80% of the 303,616 high-quality SNPs. They included 88 samples from Western Thailand, 19 from Western Cambodia and 41 from Indonesia. All other locations had <10 eligible samples which was considered too few for detailed population genetic comparisons. This sample size was not pre-determined, but was the largest that we were able to achieve in the timeframe of this study. Supplementary Table 1 identifies the origin of the 148 samples that were used for all population genetic analyses (excepting the PCA and neighbour-joining tree for which we used all 228 samples).

Diversity, Tajima's D and N/S ratio amongst gene classes

We classified genes using annotations from PlasmoDB (http://www.plasmodb.org/common/downloads/release-13.0/PvivaxSal1/txt/PlasmoDB-13.0_PvivaxSal1Gene.txt).

Nucleotide diversity, Tajima's D and N/S ratio were calculated using custom Python scripts.

Statistical analyses were performed using the SciPy stats package

(<http://docs.scipy.org/doc/scipy/reference/stats.html>)

Population structure

For a given population P , we estimated the *non-reference allele frequency (NRAF)* at a given SNP as the mean of the within-sample allele frequency (f_w) for all samples in P which have a valid genotype at that SNP. The *minor allele frequency (MAF)* is then computed as $\min(\text{NRAF}, (1 - \text{NRAF}))$.

To investigate the global population structure, we started by computing an $N \times N$ pairwise distance matrix, where N is the number of samples. Each element of the matrix contained an estimate of genetic distance between the relevant pair of samples, obtained by summing the pairwise distance at each SNP estimated from within-sample allele frequency (f_w). When comparing a pair of samples s_A and s_B at a single SNP i where a genotype could be called in each sample, with within-sample allele frequencies f_A and f_B respectively, the distance d_{AB} was estimated as $d_{AB} = f_A (1 - f_B) + f_B (1 - f_A)$. The genome-wide distance D_{AB} between the two samples is then calculated as

$$D_{AB} = \frac{1}{n_{AB}} \sum_i w_i d_{AB}$$

where n_{AB} is the number of SNPs where both samples could be genotyped, w_i is an LD weighting factor. The LD weighting factor, which corrects for the cumulative contribution of physically linked polymorphisms, was computed at each SNP i with $\text{MAF} \geq 0.1$ in our sample set, by considering a window of m SNPs ($j = 0.. m$) centred at i . For each j , we computed the squared correlation coefficient r^2_{ij} between SNPs i and j . Ignoring positions j where $r^2_{ij} < 0.1$, the weighting w_i was computed by

$$w_i = \frac{1}{1 + \sum_j r^2_{ij}}$$

Principal coordinate analysis (PCoA) was performed using the same pairwise distance matrices using the Classical Multidimensional Scaling (MDS) method. PCoA is a computationally efficient variant of principal component analysis (PCA) in which a pairwise distance matrix is used as input, rather than a table of genotypes. The matrix was supplied as input to the MDS algorithm, using the R language `cmdscale` implementation. A neighbour-joining tree was then produced using the `nj` implementation

in the R ape package. To explore the effects on population structure of using a different reference genome, we aligned the same samples to the Papua Indonesia P01 genome assembly (www.genedb.org/Homepage/PvivaxP01) using GATK Best Practices. As shown in Supplementary Figure 9, the neighbour-joining tree was very similar to that obtained with the Sal1 reference genome.

We performed admixture analysis using ADMIXTURE.⁵³ As the ADMIXTURE model assumes perfect linkage equilibrium between markers (i.e. they are independent of each other), we excluded SNP pairs that appeared to be linked. We discarded SNPs according to the observed correlation coefficients by using the PLINK tool set.⁵⁴ We scanned the genome with a sliding window of 60 SNPs in size, advanced in steps of 10 SNPs, and removed any SNP with a correlation coefficient ≥ 0.1 with any other SNP within the window. Additionally, we removed all SNPs with extremely low minor allele frequency (MAF ≤ 0.005), as these SNPs are less informative for the inference process. We then ran ADMIXTURE 1.3, in haploid mode, using the 76,544 remaining SNPs with 5-fold cross-validation and several K values (i.e. the number of putative populations) ranging from 1 to 12. In order to avoid fluctuations in the likelihood due to the stochasticity of the optimization process we repeated the process 5 times with different random seeds. We assessed the plausible choice for the number of populations by using the delta ΔK metric developed by Evanno and colleagues (Supplementary Figure 6).³⁰

To estimate the F_{ST} between two populations at a given SNP, we used $F_{ST} = 1 - (\hat{\pi}_s / \hat{\pi}_t)$ where π_s is the average probability that two samples chosen at random from the same population will carry different allele at the SNP, and π_t is the average probability that two samples chosen at random from the joint population will carry different allele. Estimates for F_{ST} were obtained by using the $NRAFs$ in the two populations (p_1 and p_2) to compute

$$\hat{\pi}_s = \frac{1}{2}(2p_1(1 - p_1) + 2p_2(1 - p_2)) \quad \text{and} \quad \hat{\pi}_t = 2 \cdot \frac{(p_1 + p_2)}{2} \cdot \left(1 - \frac{(p_1 + p_2)}{2}\right)$$

Within host diversity

F_{WS} metrics were calculated as previously described for *P. falciparum*²⁷. Analysis of heterozygosity within mixed samples was performed using custom Python scripts.

Recombination

We analyzed the decay of LD with genomic distance for each population separately. LD was measured by computing two commonly used measures (D' and r^2) for pairs of SNPs of varying distance. After categorizing SNPs into equally spaced MAF intervals, LD calculations were conducted separately for each frequency bin, and later combined. We accounted for offsets due to population structure by a sample rotation method, and by measuring "random" LD between SNPs on different chromosomes. Complete details are given in Manske *et al.*²⁷.

Signatures of selection

XP-EHH and iHS scores were calculated using previously described methods as per Sabeti *et al.*⁵⁵ and Voight *et al.*⁵⁶. As described in these studies, the distributions of scores follow an approximately normal distribution and, hence, P values were based on this distribution. Where genotypes exhibited heterozygous calls, the calls were converted to a homozygous call for the allele with the largest number of reads at that position. As a consequence, in mixed samples, haplotype-based analysis was essentially conducted on the majority strain present within each infection.

Data access and URLs

Illumina sequence reads have been submitted to the European Nucleotide Archive with study accessions ERP000194 (<http://www.ebi.ac.uk/ena/data/view/ERP000194>) and ERP003694 (<http://www.ebi.ac.uk/ena/data/view/ERP003694>). Additional metadata and genotype calls on individual samples will be released at www.malariagen.net/data prior to publication. Further details of all SNPs reported in this dataset including their genome coverage, mapping quality and allele frequencies in different populations, together with tools for querying the data, can be explored at www.malariagen.net/apps/pvgv. (*Note to reviewers: this link is to a draft web application that will be updated prior to publication.*)

Acknowledgements

We wish to thank the patients and communities that provided samples for this study, and our many colleagues who supported this work in the field. Sequencing, data analysis and project coordination were funded by the Wellcome Trust (098051, 090770/Z/09/Z), the Medical Research Council (G0600718) and the UK Department for International Development (M006212). AEB and IM acknowledge the Victorian State Government Operational Infrastructure Support and Australian

Government NHMRC IRIISS. SA and RNP are funded by the Wellcome Trust (Senior Fellowship in Clinical Science awarded to RNP, 091625). This study was supported in part by the Intramural Research Program of the National Institute of Allergy and Infectious Diseases, National Institutes of Health.

Figure legends

Figure 1. Copy number variation

Common forms of copy number variation in a region of chromosome 8 with a deletion of the first three exons of PVX_094265; in regions of chromosome 6 and 14 with copy number variations of *pvdbr* and PVX_101445 respectively; and in a region of chromosome 10 region where multiple genes including *pvmbr1* are duplicated. Top panel shows an illustrative sample for each genomic region: upper trace shows GC-normalised coverage with inferred copy number marked by red line; lower trace shows the proportion of read pairs mapping in opposing directions, indicating the presumptive breakpoints of a duplication (note that not all samples have identical breakpoints, Supplementary Table 5). Lower panel shows number of samples in each population having a copy number other than one: western Thailand (WTH, n=88), western Cambodia (WKH, n=19) and Papua Indonesia (ID, n=41).

Figure 2. Patterns of linkage disequilibrium

Genome-wide values for r^2 were calculated between pairs of SNPs over a range of distances, and corrected for the inflation caused by population structure and other confounders as described in Methods. Median values of linkage disequilibrium decay over short distances, e.g. r^2 falls to <0.1 within 200 bp in western Thailand (green) and western Cambodia (blue), and within 500 bp in Papua Indonesia (red).

Figure 3. Genetic structure of mixed infections

Top panels show distribution of F_{WS} across all samples. Each dot represents an individual sample. F_{WS} is analogous to an inbreeding coefficient²⁷ and a value of 1 indicates a perfect clone.

Panel A1: Distribution of F_{WS} in western Thailand (WTH), western Cambodia (WKH) and Papua Indonesia (ID), showing median (thick line) and inter-quartile range (thin line).

Panel A2: Distribution of F_{WS} stratified by the number of dominant clones in a sample and by whether they are related to each other.

Panel A3: Distribution of F_{WS} (vertical axis) and the proportion of heterozygous genotype calls (horizontal axis) in samples with different numbers of dominant clones.

Each row of Panel B shows an illustrative sample. Left: non-reference allele frequency (NRAF) distribution across all heterozygous SNPs. Right: horizontal axis is chromosomal position; vertical axis is heterozygosity calculated in 20kb bins with the scale truncated (0 to 0.03) to highlight runs of homozygosity (RoH). Sample *a* is near-clonal as evidenced by $F_{WS} = 1$ and lack of heterozygous SNPs. Samples *b-e* each contain two dominant clones as evidenced by the bimodal NRAF distribution. Sample *b* contains two unrelated clones (no RoH). Sample *c* contains two partially related clones (RoH across minority of the genome). Sample *d* contains two meiotic siblings (RoH extending over ~50% of the genome). Sample *e* contains two clones that are the product of inbreeding over multiple generations (RoH extending over ~80% of the genome). Sample *f* appears to contain a complex mixture of related parasites (relatively flat NRAF distribution indicates multiple dominant clones but there is substantial RoH).

Figure 4. Parasite population structure.

Population structure is evident by principal components analysis (panel A), ADMIXTURE (panel B) and on a neighbour joining tree (panel C). ADMIXTURE analysis identifies three major components of population structure which correspond to the three largest groups of samples, i.e. western Thailand (n=88), western Cambodia (n=37) and Papua Indonesia (n=55). The neighbour-joining tree shows how these three major components encompass the Southeast Asian and Pacific Islands (Malaysia, Papua Indonesia, Papua New Guinea), the western part of mainland Southeast Asia (Western Thailand, Myanmar, and China) and the eastern part of the mainland (Cambodia, Vietnam, Eastern Thailand, and Laos). Samples from other parts of the world (India, Sri Lanka, Madagascar, and Brazil) are separated from Southeast Asian samples by long branches.

Figure 5. Population-specific signatures of recent positive selection

Metrics of extended haplotype homozygosity were estimated in 88 samples from western Thailand (WTH), 19 from western Cambodia (WKH) and 41 from Papua Indonesia (ID). The strongest evidence for recent selection was identified by XP-EHH (i.e. by comparing populations) and in most cases this was supported by iHS tests within individual populations. Horizontal axis represents genome position with chromosomes 1-14 shown in alternating colours. Vertical axis shows the results of XP-EHH and iHS tests represented by $-\log_{10} P$ values on a scale of 0 to 15. Dashed line shows the Bonferroni-corrected threshold for genome-wide significance, red points mark significant P values. Loci with ≥ 2 SNPs with significant P values within 80 kb of each other are marked by red lines in the tracks labelled 'Selected regions'. The iHS signal on chromosome 13 in WKH was confined to two

adjacent SNPs and is therefore not marked as significant. These signatures are described in more detail in Supplementary Table 8 and Supplementary Figure 7.