Contents

Data shown in Figure 1

Due to the fragmented nature of several of the nuclear genome sequences represented in **Figure 1**, contigs that could not be assigned to a chromosome are not shown. Details concerning the number of sequences and number of genes in gene families that are shown are indicated below:

PLASMODIUM FALCIPARUM: *sequences in figure*: 14; total sequence length: 23,264,339 bp; longest sequence: 3,291,871 bp; median sequence length: 1,614,689 bp; number of genes: 5,440; *sequences NOT in figure*: number of sequences: 2; total sequence length: 35,388 bp; longest sequence: 29,422 bp; median sequence length: 17,694 bp; number of genes: 32.

PLASMODIUM VIVAX: *sequences in figure*: number of sequences: 14; total sequence length: 22,620,986 bp; longest sequence: 3,120,417 bp; median sequence length: 1,977,566 bp; number of genes: 5,058; VIR: 92; MSP-3: 12; Pv-fam-d: 13; Pv-fam-h: 3; Pv-fam-a: 23; PHIST: 18; Pv-fam-b: 6; PST-A: 9; RBP: 7; etramp: 9; SERA: 13; MSP-7: 11; Pv-fam-e: 42; *sequences NOT in figure*: number of sequences: 2,745; total sequence length: 4,262,317 bp; longest sequence: 101,928 bp; median sequence length: 4,280 bp; number of genes: 375.

PLASMODIUM KNOWLESI: *sequences in figure*: number of sequences: 14; total sequence length: 23,462,190 bp; longest sequence: 3,159,096 bp; median sequence length: 2,286,589 bp; number of genes: 5107; Pk-fam-a: 10; PHIST: 13; Pv-fam-a: 14; PST-A: 3; etramp: 7; SERA: 7; Pv-fam-e: 12; Pk-fam-d: 3; Pk-fam-c: 5; MSP-3: 3; Pk-fam-b: 9; Pv-fam-d: 2; SICAvar: 190; kir: 67; Pv-fam-h: 1; RBP: 1; *sequences NOT in figure*: number of sequences: 511; total sequence length: 636,470 bp; longest sequence: 29,816 bp; median sequence length: 1,245bp ; number of genes: 79.

PLASMODIUM YOELII YOELII: *sequences in figure*: number of sequences: 14; total sequence length: 14,839,937 bp; longest sequence: 2,071,994 bp; median sequence length: 1,252,829 bp; number of genes: 5,216; *sequences NOT in figure*: number of sequences: 1,136; total sequence length: 5,512,276 bp; longest sequence: 34,875 bp; median sequence length: 3,073 bp; number of genes: 2,645.

Expanded Table 2

A comparison of the sequences and binding sites of *P. vivax* and *P. falciparum* orthologs involved in antimalarial drug interactions. A portion of this table describing our results with artemisinin and atovaquone drug interaction genes is shown in the main text (**Table 2**). Here we show an expanded table including the results of others concerning six front-line drugs/classes of drug and the six genes associated with resistance to the drugs.

Antimalarial drug	Gene	Gene PID*	Protein PID*	X-ray (X) structure or homology (H) model	Drug binding site residues (red: polymorphism between *P. vivax* and *P. falciparum*)	Polymorphisms associated or suspected with resistance (blue: unique to *P. falciparum*; green: unique to *P. vivax*)	Conclusions
Artemisinin derivatives	*P. vivax atpase6*	60.7	70.0	H*	(I261 A263[1] F264 Q267 L268 I271 I275 A313 I942 I946 N949 I950 V953 A954 F957 S1008 L1009 I1010 L1015 Y1018 I1019)*	None reported	*P. vivax* predicted to be more susceptible to artemisinin than *P. falciparum*
	P. falciparum atpase6			H*	(I261 L263[1]† F264 Q267 L268 I271 I275 A313 I973 I977 N980 I981 V984 A985 F988 N1039 L1044 I1045 L1050 Y1053 I1054)*	(E432K A623E S769N**)[2]	
Atovaquone	*P. vivax cyt b*	86.0	89.5	H*	(I119 F123 Y126 M133 V140 I141 L144 I258 P260 F264 F267 Y268 L271 V284 L285 L288)*	None reported	Resistance mutations in *P. falciparum* likely to affect the same residues in *P. vivax*
	P. falciparum cyt b			H[3]	(I119 F123 Y126 M133 V140 I141 L144 I258 P260 F264 F267 Y268†† L271 V284 L285 L288)[3]	M133I[3]‡ T142I[3]‡ L144F[3]‡ I258M[3]‡ F267I[3]‡ Y268S[3]/N[4] L271V[3]‡ K272R[3]‡ P275T[3]‡ G280D[3]‡ V284K[3]‡	

Antimalarial drug	Gene	Gene PID*	Protein PID*	X-ray (X) structure or homology (H) model	Drug binding site residues (red: polymorphism between *P. vivax* and *P. falciparum*)	Polymorphisms associated or suspected with resistance (blue: unique to *P. falciparum*; green: unique to *P. vivax*)	Conclusions
Cycloguanil Pyrimethamine WR99210	*P. vivax dhfr-ts*	65.5	70.0	X[5]	(I13‡‡ C14 D53 M54 F57 P122 L128 I173‡‡ Y179)[5, †]	I13L[6] P33L[7] C49R[8] N50K[9] F57I[10]/L[7] S58R[7] T61M[6] V111L[11] S117N[7]/T[10] I173F[7] A199V[11]	*P. vivax* F57L/T61M gives increased resistance to WR99210 and cycloguanil but not to pyrimethmine
	P. falciparum dhfr-ts			X[12]	(I14 C15 D54 M55 F58 P113 L119 I164‡‡ Y170)[12]	A16V[13]/S[14] S46L[15] C50R[16] N51I[13,17] C59R[13,17] 108N[13,17] /T[13,17] V140L[18] I164L[17]/M[19] N188K[20] S189R[20] F223S[21]	
Sulphadoxine Dapsone	*P. vivax pppk-dhps*	58.5	56.1	H[22]	(S382 A383†† P384 F552 R580 K581 V585† Y688 N691 H713)[22]	A383G[22] A553G[22] A647S/P[23]	*P. vivax* V585 may explain innate resistance to sulphadoxine
	P. falciparum pppk-dhps			H[22]	(S436 A437†† P438 F580 R608 K609 A613†† Y663 N666 H688)[22]	S436A[24]/F[24,25] A437G[24,25] K540E[26] A581G[24,25] A613S[24,25] /T[24]	

Antimalarial drug	Gene	Gene PID*	Protein PID*	X-ray (X) structure or homology (H) model	Drug binding site residues (red: polymorphism between *P. vivax* and *P. falciparum*)	Polymorphisms associated or suspected with resistance (blue: unique to *P. falciparum*; green: unique to *P. vivax*)	Conclusions
Chloroquine	*P. vivax cg10*	66.6	73.2	-	Not known	Q34H[27] P38L[27] L384F[27]	-
	P. falciparum crt			-	Not known	C72S[28]/R[29] M74I[28] N75E[28]/D[30]/K[31]/I[29] K76T[28]/N[32] /I[32] I77T[30] H97L[33]/Q[28] A144F[34]/T[35] Y[36] L148I[34] L160Y[35] I194T[34] A220S[28] Q271E[28] N326S[28]/D[28] T333S[34] I356T[28]/L[28] R371I[28]/T[28]	
Mefloquine Chloroquine	*P. vivax mdr1*	68.4	76.0	-	Not known	N196S[37] S237P[37] I443T[37] D500N[37] R511H[37] V556A[37] G698S[37] A783V[37] K823E[37] M908L[37] T958M[37] Y976F[38,39] F1076L[38] L1104F[37] E1261K[37] Q1440R[37] E1447G[37]	-
	P. falciparum mdr1			-	Not known	(N86Y Y184F S1034C N1042D D1246Y)[40]	

* Determined during this study
‡ Identified from *in vitro* drug selection studies
† Predicted to be located in a drug active site and mutations predicted to cause resistance
†† Predicted to be located in a drug active site and mutations known to cause resistance
‡‡ Known to be located in the drug active site and mutations known to cause resistance
** Association with resistance not conclusive

Abbreviations: *atpase6*, adenosine triphosphatase-6; *cyt b*, cytochrome b; *dhfr-ts*, dihydrofolate reductase-thymidylate synthase; *pppk-dhps*, dihydropterin pyrophosphokinase-dihydropteroate synthase; *cg10*, candidate gene 10; *crt*, chloroquine resistance transporter; *mdr1*, multiple drug resistance; PID, percent identity.

Notes

Chloroquine has been the antimalarial drug of choice to treat both *P. falciparum* and *P. vivax* for more than 50 years, but *P. falciparum* resistance is widespread and reports of *P. vivax* resistance are increasing. Although two key molecules involved in chloroquine resistance in *P. falciparum*, *pfcrt* and *mdr1*, have orthologs in *P. vivax*, there are no crystal structures of the proteins and weak sequence similarity to other proteins prevented homology modelling during this study.

References

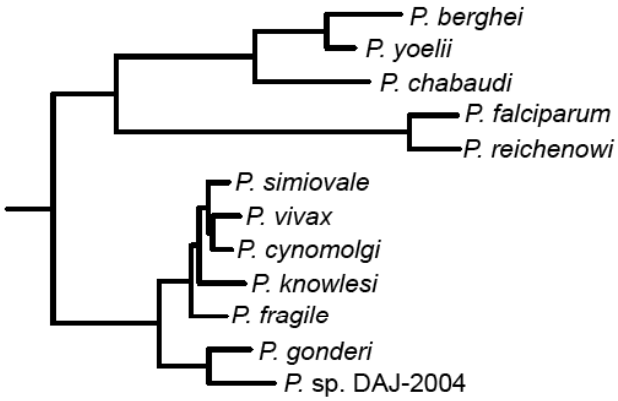
1. Uhlemann, A.C. et al. A single amino acid residue can determine the sensitivity of SERCAs to artemisinins. *Nat Struct Mol Biol* **12**, 628-9 (2005).
2. Jambou, R. et al. Resistance of Plasmodium falciparum field isolates to in-vitro artemether and point mutations of the SERCA-type PfATPase6. *Lancet* **366**, 1960-3 (2005).
3. Korsinczky, M. et al. Mutations in Plasmodium falciparum cytochrome b that are associated with atovaquone resistance are located at a putative drug-binding site. *Antimicrob Agents Chemother* **44**, 2100-8 (2000).
4. Fivelman, Q.L., Butcher, G.A., Adagu, I.S., Warhurst, D.C. & Pasvol, G. Malarone treatment failure and in vitro confirmation of resistance of Plasmodium falciparum isolate from Lagos, Nigeria. *Malar J* **1**, 1 (2002).
5. Kongsaeree, P. et al. Crystal structure of dihydrofolate reductase from Plasmodium vivax: pyrimethamine displacement linked with mutation-induced resistance. *Proc Natl Acad Sci U S A* **102**, 13046-51 (2005).
6. Imwong, M. et al. Association of genetic mutations in Plasmodium vivax dhfr with resistance to sulfadoxine-pyrimethamine: geographical and clinical correlates. *Antimicrob Agents Chemother* **45**, 3122-7 (2001).
7. Eldin de Pecoulas, P., Basco, L.K., Tahar, R., Ouatas, T. & Mazabraud, A. Analysis of the Plasmodium vivax dihydrofolate reductase-thymidylate synthase gene sequence. *Gene* **211**, 177-85 (1998).
8. Hastings, M.D. et al. Novel Plasmodium vivax dhfr alleles from the Indonesian Archipelago and Papua New Guinea: association with pyrimethamine resistance determined by a Saccharomyces cerevisiae expression system. *Antimicrob Agents Chemother* **49**, 733-40 (2005).

9. Tjitra, E., Baker, J., Suprianto, S., Cheng, Q. & Anstey, N.M. Therapeutic efficacies of artesunate-sulfadoxine-pyrimethamine and chloroquine-sulfadoxine-pyrimethamine in vivax malaria pilot studies: relationship to Plasmodium vivax dhfr mutations. *Antimicrob Agents Chemother* **46**, 3947-53 (2002).
10. 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**, 1514-21 (2003).
11. Hastings, M.D. et al. Dihydrofolate reductase mutations in Plasmodium vivax from Indonesia and therapeutic response to sulfadoxine plus pyrimethamine. *J Infect Dis* **189**, 744-50 (2004).
12. Yuvaniyama, J. et al. Insights into antifolate resistance from malarial DHFR-TS structures. *Nat Struct Biol* **10**, 357-65 (2003).
13. Peterson, D.S., Walliker, D. & Wellems, T.E. Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine in falciparum malaria. *Proc Natl Acad Sci U S A* **85**, 9114-8 (1988).
14. Wang, P. et al. Resistance to antifolates in Plasmodium falciparum monitored by sequence analysis of dihydropteroate synthetase and dihydrofolate reductase alleles in a large number of field samples of diverse origins. *Mol Biochem Parasitol* **89**, 161-77 (1997).
15. Paget-McNicol, S. & Saul, A. Mutation rates in the dihydrofolate reductase gene of Plasmodium falciparum. *Parasitology* **122**, 497-505 (2001).
16. Plowe, C.V. et al. Mutations in Plasmodium falciparum dihydrofolate reductase and dihydropteroate synthase and epidemiologic patterns of pyrimethamine-sulfadoxine use and resistance. *J Infect Dis* **176**, 1590-6 (1997).
17. Cowman, A.F., Morry, M.J., Biggs, B.A., Cross, G.A. & Foote, S.J. Amino acid changes linked to pyrimethamine resistance in the dihydrofolate reductase-thymidylate synthase gene of Plasmodium falciparum. *Proc Natl Acad Sci U S A* **85**, 9109-13 (1988).
18. Zindrou, S., Nguyen, P.D., Nguyen, D.S., Skold, O. & Swedberg, G. Plasmodium falciparum: mutation pattern in the dihydrofolate reductase-thymidylate synthase genes of Vietnamese isolates, a novel mutation, and coexistence of two clones in a Thai patient. *Exp Parasitol* **84**, 56-64 (1996).
19. Thaithong, S. et al. Pyrimethamine resistant mutations in Plasmodium falciparum. *Mol Biochem Parasitol* **52**, 149-57 (1992).
20. Bates, S.J. et al. Rare, highly pyrimethamine-resistant alleles of the Plasmodium falciparum dihydrofolate reductase gene from 5 African sites. *J Infect Dis* **190**, 1783-92 (2004).
21. Tanaka, M., Gu, H.M., Bzik, D.J., Li, W.B. & Inselburg, J.W. Dihydrofolate reductase mutations and chromosomal changes associated with pyrimethamine resistance of Plasmodium falciparum. *Mol Biochem Parasitol* **39**, 127-34 (1990).
22. 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**, 2214-22 (2004).
23. Menegon, M., Majori, G. & Severini, C. Genetic variations of the Plasmodium vivax dihydropteroate synthase gene. *Acta Trop* **98**, 196-9 (2006).
24. Brooks, D.R. et al. Sequence variation of the hydroxymethyldihydropterin pyrophosphokinase: dihydropteroate synthase gene in lines of the human malaria parasite, Plasmodium falciparum, with differing resistance to sulfadoxine. *Eur J Biochem* **224**, 397-405 (1994).
25. Triglia, T. & Cowman, A.F. Primary structure and expression of the dihydropteroate synthetase gene of Plasmodium falciparum. *Proc Natl Acad Sci U S A* **91**, 7149-53 (1994).

26. Triglia, T., Menting, J.G., Wilson, C. & Cowman, A.F. Mutations in dihydropteroate synthase are responsible for sulfone and sulfonamide resistance in Plasmodium falciparum. *Proc Natl Acad Sci U S A* **94**, 13944-9 (1997).
27. Nomura, T. et al. Evidence for different mechanisms of chloroquine resistance in 2 Plasmodium species that cause human malaria. *J Infect Dis* **183**, 1653-61 (2001).
28. Fidock, D.A. et al. Mutations in the P. falciparum digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol Cell* **6**, 861-71 (2000).
29. Best Plummer, W., Pinto Pereira, L.M. & Carrington, C.V. Pfcrt and pfmdr1 alleles associated with chloroquine resistance in Plasmodium falciparum from Guyana, South America. *Mem Inst Oswaldo Cruz* **99**, 389-92 (2004).
30. Lim, P. et al. pfcrt polymorphism and chloroquine resistance in Plasmodium falciparum strains isolated in Cambodia. *Antimicrob Agents Chemother* **47**, 87-94 (2003).
31. Nagesha, H.S. et al. New haplotypes of the Plasmodium falciparum chloroquine resistance transporter (pfcrt) gene among chloroquine-resistant parasite isolates. *Am J Trop Med Hyg* **68**, 398-402 (2003).
32. Cooper, R.A. et al. Alternative mutations at position 76 of the vacuolar transmembrane protein PfCRT are associated with chloroquine resistance and unique stereospecific quinine and quinidine responses in Plasmodium falciparum. *Mol Pharmacol* **61**, 35-42 (2002).
33. Chen, N. et al. Sequence polymorphisms in pfcrt are strongly associated with chloroquine resistance in Plasmodium falciparum. *J Infect Dis* **183**, 1543-5 (2001).
34. Durrand, V. et al. Variations in the sequence and expression of the Plasmodium falciparum chloroquine resistance transporter (Pfcrt) and their relationship to chloroquine resistance in vitro. *Mol Biochem Parasitol* **136**, 273-85 (2004).
35. Chen, N. et al. pfcrt Allelic types with two novel amino acid mutations in chloroquine-resistant Plasmodium falciparum isolates from the Philippines. *Antimicrob Agents Chemother* **47**, 3500-5 (2003).
36. Yang, Z. et al. Molecular analysis of chloroquine resistance in Plasmodium falciparum in Yunnan Province, China. *Tropical medicine and international health* **12**, 1-10 (2007).
37. Sa, J.M. et al. Plasmodium vivax: allele variants of the mdr1 gene do not associate with chloroquine resistance among isolates from Brazil, Papua, and monkey-adapted strains. *Exp Parasitol* **109**, 256-9 (2005).
38. Brega, S. et al. Identification of the Plasmodium vivax mdr-like gene (pvmdr1) and analysis of single-nucleotide polymorphisms among isolates from different areas of endemicity. *J Infect Dis* **191**, 272-7 (2005).
39. Suwanarusk, R. et al. Chloroquine Resistant Plasmodium vivax: In Vitro Characterisation and Association with Molecular Polymorphisms. *PLoS ONE* **2**, e1089 (2007).
40. Foote, S.J. et al. Several alleles of the multidrug-resistance gene are closely linked to chloroquine resistance in Plasmodium falciparum. *Nature* **345**, 255-8 (1990).

Supplementary Methods and Results

***Plasmodium* phylogeny.** As an aid to understanding the comparative analysis between the four species (*P. vivax, P. knowlesi, P. y. yoelii* and *P. falciparum*) presented in this paper, the figure below shows a phylogenetic tree of several mammalian *Plasmodium* species inferred from the three mitochondrial protein-coding genes. Each node in the phylogeny shown is supported by 100% posterior probability. Rodent malaria parasites (*P. berghei*, *P. yoelii*, and *P. chabaudi*) are sister to the human malaria parasite *P. falciparum*, and its closest relative *P. reichenowi*, a parasite of chimpanzees. The remaining malaria parasite species infect primates, including the human species *P. vivax* and its closest monkey relative *P. cynomolgi*, and *P. knowlesi*, which infects both humans and non-human primates. In contrast to these Asian primate-infecting species, *P. gonderi* and *P.* sp. DAJ-2004 (mandrill) are parasites that have been isolated from African primates.



Genome sequencing and assembly. Eight naïve splenectomized *Saimiri boliviensis boliviensis* (squirrel) monkeys were infected with the Salvador I strain of *P. vivax* isolated from a natural infection in the Department of La Paz, El Salvador [1]. Infected blood was depleted of platelets and leukocytes by passage through activated acid-washed glass beads, Plasmodipur filters (Euro-Diagnostica B.V., The Netherlands) followed by Whatman CF-11 cellulose powder column. Parasite DNA was purified and small (2-3 kb) and medium (10-12 kb) genomic DNA libraries constructed using techniques perfected from the *P. falciparum* genome project [2]. Approximately 0.38 million paired sequence reads were obtained from random clones using the whole-genome shotgun approach [3], providing ~10-fold sequence redundancy. Reads were assembled using the Celera Assembler [4]. Very few contigs (~75, the largest contig size 11.6 kb) were identified as contaminating *S. b. boliviensis* through BLASTN searches (>95% identity over >100 nt) with a custom database of sequences from multiple monkey species, and excluded from further analysis. This low level of contaminating host DNA indicates the success of monkey leukocyte removal by the methods mentioned above. Using AutoEditor [5], a second-generation base-caller which scans the multi-alignment of reads in contigs, the rate of apparent single nucleotide polymorphisms (SNPs) in assembly multi-alignments was calculated. We found 1,784 SNPs in all contigs in the assembly, a rate of 0.068 SNPs/kb. Approximately 8% of the SNPs occurred in contigs identified as repetitive, *i.e.,* in places where the quality of the alignments was expected to be poor. In unique contig regions, we calculated a rate of 0.071 SNPs/kb or 1 SNP per ~14 kb. This demonstrates that a significant degree of sequence homogeneity exists, indicating that the Salvador I strain used for the genome project did not contain a mixture of parasite genotypes.

Genome closure and chromosome mapping. A combination of automated gap closure and manual closure [2] strategies were used to improve sequence coverage and close gaps in the assembly. In the absence of linking information between contigs, alignment of the *P. vivax* and *P. falciparum* genomes using MUMmer [6] identified putatively linked *P. vivax* contigs through synteny, which were then confirmed by PCR across the intervening sequence gap using primers designed to neighbouring contigs. To assign contigs to a chromosome, unique probes generated by PCR from each contig were hybridized to *P. vivax* Salvador I chromosomes separated by pulsed-field gradient gel electrophoresis using optimized electrophoresis conditions [7]. The majority of large contigs with a total length of ~22.57 Mb and high G+C content (mean ~45.0%) could be assigned to one of the 14 *P. vivax* chromosomes, and the telomeres of three chromosomes were completed. An additional ~4.26 Mb of small subtelomeric contigs with low G+C content (mean ~28.4%) could not be assigned to a chromosome due to their repetitive nature. At 10-fold sequence coverage and with partial gap closure, the genome is at a 'comparative' grade of finishing.

Gene prediction and manual annotation. *P. vivax* contigs > 1 kb were processed through TIGR's eukaryotic annotation pipeline. The gene prediction algorithm JIGSAW, a program designed to use the output from gene finders, splice site prediction programs and sequence alignments, was used to predict gene models. Jigsaw was trained using datasets from the automated gene predictors GlimmerHMM, TIGRscan and Phat, which in turn had been trained with a dataset of 295 *P. vivax* genes manually modeled based on high confidence EST (expressed sequence tag) and protein alignments. The AAT package [8] was used to compute sequence similarity to nucleotide and protein datasets, such as the predicted protein sequences from *P. falciparum, P. y. yoelii and P. knowlesi*, and a non-redundant amino acid database filtered from public sources. After excluding genes coding for proteins < 50 amino acids in length and with no sequence similarity to other ESTs or proteins, a total of 5,433 gene models were predicted including 42 pseudogenes and 90 truncated genes (genes missing their 5' or 3' regions). EST and cDNA sequences were aligned against the genomic sequence, assembled into non-redundant alignments and compared against the annotation data set using PASA [9]. A total of 22,219 EST sequences were downloaded from GenBank dbEST, and 94.4% passed the validation criteria for EST assembly producing 5,732 assemblies. When compared to the annotated gene models, a total of 3,169 EST-assemblies mapped to 2,641 genes. Of these, 919 PASA-based updates resulted in changes to the gene structure, 735 updates resulted in extended UTRs, 101 corrected protein structure and 93 updates resulted in proper addition of the EST assembly into the gene structure. All gene models were manually inspected at least twice, once for curation of gene structures and once for curation of functional annotation. Subsets of the genes such as gene families and genes of the exportome, were also manually checked during analysis. Finally, orthologs between *P. vivax* and *P. falciparum* were identified using a bidirectional best hit analysis [10]. Detailed comparisons between orthologous genes were used to identify annotation inconsistencies between these closely related genomes. Statistical summaries comparing exon number, protein length and alignment fidelity highlighted ortholog clusters with inaccurate gene predictions. Using a web-based comparative genomics environment called Sybil (http://sybil.sf.net), the suspect orthologous proteins were investigated in their genomic context. The results of the comparative genome analyses were used to target gene structures for manual update, resulting in the revised annotation of hundreds of *P. vivax* genes.

Putative signal peptide sequences were identified using SignalP [11], and transmembrane regions predicted by tmHMM [12]. tRNAscan-SE [13] was run on the assembled genomic sequence to identify tRNAs. rRNA sequences were manually annotated based on homology to previously identified rRNA sequences (GenBank accession numbers: U17510, AF202181, L29561 and U86613).

Gene Ontology assignments to the *P. vivax* proteome was performed by automatic transference of pre-existing GO annotation of *P. falciparum*. In addition, remaining GO assignments for genes lacking annotation was done by correspondence between InterPro and Pfam entries and GO terms, using Interpro2GO (http://www.geneontology.org/external2go/interpro2go) and Pfam2GO (http://www.geneontology.org/external2go/pfam2go). To give a broad overview of the ontology content, GO slims were created using map2slim.pl (http://www.geneontology.org/GO.slims.shtml#script) which maps a set of annotations up to their parent GO slim terms. A total of 9,200 GO terms based upon *P. falciparum* (covering 1,121 different GO terms on 2,514 genes), and 243 GO terms based upon Pfam entries (covering 82 different GO terms on 122 distinct genes), were assigned.

Public database submission. The following datasets have been submitted to GenBank: 387,977 traces are in the Trace Archive; 2,770 WGS contigs and annotation, Accession No. AAKM00000000; 14 pseudochromosomes, Accession Nos. CM000442-CM000455. All *P. vivax* genes have a unique locus_tag associated with them, indicated as follows: PVX_xxxxxx.

Genome sequence analysis.

Comparison of ortholog lengths. Approximately 3,725 ortholog pairs between *P. falciparum* and *P. vivax* (see below) were compared. The lengths of the orthologs (coding sequence only) are not Normally distributed, so a nonparametric, paired-values approach was used to test for significant difference. *P. falciparum* coding sequences were significantly longer than their *P. vivax* orthologs (Wilcoxon rank-sum test, $z = 13.8$, $p < 0.0001$). The median CDS lengths were 1,638 bp and 1,602 bp respectively, and the median difference was 3 bp (bootstrapped 95% confidence interval: 3-6 bp).

Low complexity regions. Low complexity regions (LCRs) are defined as regions of biased composition containing simple sequence repeats, and often represent rapidly diverging, exposed, hydrophilic, non-globular protein domains; tandem repeats are a subset of low complexity regions. The SEG algorithm [14], which detects regions of biased amino acid composition independently of the physicochemical properties of the amino acids involved, was used to determine LCRs in the complete proteomes of *P. vivax* and *P. falciparum* using the following parameters: window length 45: trigger complexity: 3.4, extension complexity: 3.75. For analysis of orthologous proteins, 3,865 orthologs (identified using Jaccard-filtered orthologous gene analysis, see below) were aligned using Clustalx and the sequences in the gaps (either *P. falciparum* sequence if a gap existed in the *P. vivax* ortholog, or vice versa) searched for low complexity. Due to the SEG parameters used, LCRs above 44 amino acids only were identified, and all gap sequences below that threshold were discarded.

GC skew. GC skew is defined as a preference for guanine (G) versus cytosine (C) in the leading strand of DNA replication. GC skew was computed as (G-C)/(G+C), where G and C represent the number of Gs and Cs in a fixed-size window, for all chromosomes of *P. vivax* (**Figure 1**). In all 14 chromosomes, the opposite ends (extending well beyond any telomeric and subtelomeric regions) show clear patterns of opposing skew; *i.e.*, one chromosome end shows highly negative skew and the opposite end shows positive skew. These patterns suggest that an unknown process is driving G-to-C mutations on one strand. Alternatively, due to the paucity of purines (adenine and guanine) which cannot be made *de novo* by the parasite but are scavenged from the host, periodic scarcity during DNA replication could occur resulting in a change from incorporation of Gs to incorporation of Cs as the replication bubble moves away from the replication origin. GC skew across *P. falciparum* chromosomes has also been noted [15].

Codon usage. CodonW was used to identify the codon and amino acid usage of *P. vivax* and *P. falciparum* orthologs (http://codonw.sourceforge.net/).

Comparison of transcription-associated protein (TAP) domains. Approximately 158 hidden–Markov models (HMMs) of domains present in the proteins that constitute the Transfac Database of well-characterised eukaryotic transcriptional factors [16], in addition to those of DNA–binding domains present in organisms across the tree of life [17], were used to search the proteomes of *P. falciparum*, *P. knowlesi* and *P. vivax* using HMMER version 2.3.2. Any sequences that matched an HMM with a score greater than the lowest score for sequences included in the Pfam full alignment of the family, were considered as hits. *Plasmodium* transcription–associated protein (TAP) families were detected by querying the proteomes with a set of reference proteins identified as being associated with all aspects of the transcriptional process, and identified by their associated GO terms. The searches were performed using BLASTP, and sequence similarities with an *E*-value $\leq 10^{-6}$ were considered significant. In conjunction, the genomes from budding and fission yeast, two filamentous fungi (*Aspergillus nidulans* and *Neurospora crassa*), the plant *Arabidopsis thaliana*, and three animal genomes (*Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus*) were queried with the reference set. The sequences that showed homology to the TAPs in the reference set, and the TAPs matching these sequences, were then clustered on the basis of sequence similarity using TRIBE-MCL [18] with an inflation value of 2.0 [19].

Putative promoter motif prediction. Computational studies suggest that putative promoter elements are conserved across genomes [20]. To identify putative transcription factor binding sites, the upstream regions of *P. vivax* genes belonging to clusters of potentially co-regulated genes were searched for over-represented nucleotide motifs ('words') and the resulting words compared across species to determine conservation. Since little expression data exists for *P. vivax*, 62 clusters of *P. falciparum* genes identified as co-expressed in the same life-cycle stages [21-24] were used and the corresponding orthologous genes were identified in *P. vivax*. Orthology was established using blast to identify reciprocal best hits with an E-value of 10^{-5} or less. Implicit in our analysis is the assumption that expression patterns of gene clusters are largely conserved between *Plasmodium* species. The regions ~2 kb upstream of the translation start site, or to the nearest upstream gene if that was closer, of the genes in each *P. vivax* cluster were masked to remove low complexity repeats with dust (http://blast.wustl.edu/pub/dust), and searched for over-represented words with the enumerative word finder Weeder [25]. Concurrently, word finding was independently performed in the orthologous upstream regions in *P. falciparum, P. y. yoelii* and *P. knowlesi.*

Weeder was used to search for over-represented words satisfying the following criteria: (1) hexamers with at most one substitution; (2) octamers with at most two substitutions; and (3) decamers with at most three substitutions. Weeder ranks over-represented words using the score

$$\sum_{i=1}^{k} I(p,i)\log\left(\frac{1}{Exp(p,b_i)\cdot|S_i|}\right)$$

where k is the number of sequences in the cluster, $I(p,i)$ is an indicator variable equaling 1 if a word p occurs in the i^{th} input sequence with a permissible number of substitutions or 0 otherwise, $Exp\ (p,b_i)$ is the expected number of occurrences of pattern p with at most b_i substitutions computed from all upstream regions in the genome and $|S_i|$ is the length of upstream sequence i.

The top five most over-represented words of each length found upstream of genes in orthologous clusters were compared by manual inspection and local alignment using Matcher from the EMBOSS

suite of bioinformatics programs [26] and the resulting over-represented and conserved words are presented in **Supplementary Table 4**. Position weight matrices (PWMs) and sequence logos were constructed for the *P. vivax* instances of each putative binding site from the best instance of the Weeder consensus sequence in the upstream regions of each respective cluster. The best case is the instance with the lowest number of substitutions satisfying the Weeder word substitution criteria mentioned above. If two or more words had the lowest number of substitutions, the most upstream one was selected. To determine how specific the seven putative conserved binding sites are to the particular upstream regions used for their discovery, their PWMs were scanned against all upstream regions using PWM_SCAN [27] at *p-value* cutoff of 1/5000 and then their group specificities [28] were computed using the formula:

$$\text{Group specificity} = \sum_{i=x}^{\min(s_1,s_2)} \frac{\binom{s_1}{i}\binom{N-s_1}{s_2-i}}{\binom{N}{s_2}}$$

where s_1 is the number of genes considered in the discovery of the site, s_2 is the number of genome-wide hits for the putative site at the *p-value* cutoff of 1/5000, x is the intersection of s_1 and s_2 and N is the number of upstream regions considered in the analysis (5409). Sequence logos and group specificity scores are given in **Supplementary Table 4**.

Identification of polymorphic microsatellites. A total of 933 scaffolds from the 9.22.2003 genome assembly (excluding 1.9 Mb of degenerate sequences which could not be assembled) were searched for the presence of tandem repeats using Tandem Repeats Finder [29]. Repeats were classified according to repeat unit length: between 2-6 bases the repeat was classified as a microsatellite; >7 bases and the repeat was classified as a minisatellite. Repeats consisting of tracts of a single base were excluded from further analysis. A Perl script extracted 333 microsatellite repeats from the output of Tandem Repeats Finder, and AutoPrimer software (http://www.autoprimer.com/) was used to design primers for each microsatellite that would produce an amplicon of length ~650 bp. DNA from eight *P. vivax* laboratory lines (Brazil I, Miami II, Pakchong, Panama I, Nica, Thai II, Vietnam IV and Indonesia XIX) was used in a 50 ul PCR reaction containing 5 µl 10x Buffer (Applied Biosystems), 3 µl 25mM MgCl (Applied Biosystems), 0.5 µl Taq Gold enzyme (5u/ul; Applied Biosystems), 1 µl dNTP (New England Biolabs), 31.5 µl water, 5 µl of 1:100 diluted DNA, 2 µl of forward primer (Illumina), 2 µl of reverse primer (Illumina), with the following reaction conditions: initial denaturation Step 1: 94° C for 5 minutes, followed by Step 2: 94° C for 20 seconds, Step 3: 55° C for 10 seconds, Step 4: 50° C for 10 seconds, Step 5: 65° C for 2 minutes, with Steps 2-5 repeated for 35 cycles; then Step 6: 65° C for 5 minutes. PCR products were separated on a 2% agarose gel and scored for polymorphism.

Computation of ortholog groups. Ortholog groups (OGs) for *P. vivax, P. falciparum, P. knowlesi* and *P. y. yoelii*, as shown in **Figure 1**, were computed as described [30]. Briefly:
1. All-vs-all BLASTP: an all-vs-all protein similarity matrix was computed by searching each of the predicted polypeptides from the four *Plasmodium* species against all of the others. WashU-BLASTP was used for these searches, with the following parameter settings: -E 1e-5 -matrix BLOSUM62 -wordmask none -B 150 -V 150 -gspmax --shortqueryok --novalidctxok -cpus 1

2. Jaccard clustering: polypeptides from each *Plasmodium* species were clustered independently using only BLASTP GSPs from step 1 that had a sequence identity score of at least 80% and an E-value of at most 1e-5. The BLASTP matches that meet these criteria were used to compute a Jaccard similarity coefficient for each distinct pair of polypeptides in the same genome. Given two polypeptides P1 and

P2 the Jaccard similarity coefficient is defined as the ratio of: num_matches_to_P1_AND_P2 / num_matches_to_P1_OR_P2. For any pair of polypeptides where the Jaccard coefficient is >0.6, the coefficients were used to determine the edges of a graph. In other words, the nodes of the graph are the polypeptides in the genome, and the edges were drawn between two polypeptides if and only if their Jaccard coefficient is 0.6 or higher. The connected components of this graph are referred to as "Jaccard clusters."

3. Bidirectional best hit clustering: the second phase of the clustering algorithm starts with the Jaccard clusters computed in step 2. This phase of the algorithm identifies pairs of Jaccard clusters such that: a. The clusters are from different genomes; b. The highest-scoring BLASTP match of at least one polypeptide in each of the clusters is to a polypeptide in the other cluster. (Note that in this phase of the clustering analysis ALL of the BLASTP matches from step 1 are used.) As in step 2, a graph is constructed, the nodes of which are the Jaccard clusters produced in step 2. An edge is drawn between two nodes (Jaccard clusters) if and only if they are bidirectional best BLASTP matches of each other, as described above. The connected components of this graph are the ortholog groups described in this paper.

Of the 23,884 proteins from the four species, 15,111 (63%) belong to OGs that have at least one representative from each of the four species (4-way OGs) representing a total of 3,736 4-way OGs. Of these, 3,632 OGs have a 1:1:1:1 relationship (14,528 proteins; 61%). There are 18,230 genes (~77%) that belong to any type of OG, including 2-, 3- and 4-way OGs. The remaining 23% (5,654 genes) are species-specific gene families and/or gene expansions.

Computation of orthologs between *P. vivax, P. falciparum, P. knowlesi*, *P. y. yoelii, P. chabaudi* and *P. berghei,* as shown in **Supplementary Table 6**, used the cluster algorithm OrthoMCL [31]. This program was run on all datasets using default parameters, and the output manually curated with reference to a previous synteny map for the rodent malaria parasites [32]. This table identifies genes that are likely to be orthologous based upon synteny but whose sequence has diverged so much as to not be identifiable using sequence similarity methods.

Computation of synteny blocks. Syntenic "blocks" of genes whose relative locations are conserved in two *Plasmodium* species were computed and the results shown in **Figure 1**. The syntenic block algorithm defines a "match" as a pair of genes (from different *Plasmodium* species) that belong to the same ortholog group. A "block" is defined as a set of matches in which the genes are located on one of two sequences (one of which is designated the "reference" sequence, and the other of which is designated a "query" or "target" sequence.) For each reference sequence (i.e., *P. vivax* chromosome) the algorithm enumerates all possible query sequences. An initially empty list of partial syntenic blocks is created and the matches are sorted according to the positions of the reference genes, ignoring orientation. The algorithm iterates over the ordered matches and for each match the current list of partial syntenic blocks is searched to determine whether 0, 1, or 2 blocks are "near" the match (see below). The next action depends on the result of this test:

0: the match is used to create a new partial syntenic block
1: the match is added to the single nearby partial syntenic block
2: the two nearby blocks are merged and the match is then added to the newly-merged block

Once all the matches have been examined, the partial syntenic blocks are filtered to remove any blocks with fewer than five distinct matches. The remaining partial blocks are the syntenic blocks for the given reference sequence. A key part of this algorithm is how it determines whether a match is "near" a

partial syntenic block. In this analysis parameters were chosen such that a match is considered to be "near" a block if:

1. the genes in the match and those in the partial syntenic block are located on the same pair of sequences
2. in each of the two sequences, the interval between the gene in the match and the nearest gene in the partial block may contain no more than 5 orthologous genes (i.e., genes that are in ortholog groups but were not themselves added to the partial block).

Chromosome evolution analysis. A total dataset of 38 interspecies, conserved syntenic blocks were selected from a pool of 3,884 three-way Jaccard-filtered ortholog groups (the number of blocks [38] differs from the number presented in **Supplemental Figure 3** [36] due to slight differences in computational methods). Clusters were initially obtained by intersection of pairwise synteny maps with a minimum requirement of five orthologous genes per syntenic block. First, the minimal number of rearrangements required to transform one *Plasmodium* genome into another was computed using the GRIMM algorithm (http://www-cse.ucsd.edu/groups/bioinformatics/GRIMM). Then, the genomic configuration of the *Plasmodium* common ancestor was calculated using the MGR program (http://www-cse.ucsd.edu/groups/bioinformatics/MGR). Five independent runs were performed to assure consistency in the results, due to a requirement by MGR for the maximum allowed number of syntenic blocks per genome to bet set at 30. In each run, the same genomic configuration for the common ancestor was obtained.

Whole genome molecular evolution analysis. Approximately 61% of the protein-coding genes in *P. vivax* (3,322 of 5,433) were assigned as 'high-quality' orthologous groups (OGs) between *P. vivax* and *P. falciparum*, consisting of gene pairs with ≥ 80% identity at the amino acid level, similar length (ratio of shorter to longer gene > 0.9) and without paralogs in either of the two species (Jaccard filter cutoff at 0.6). For each high-quality OGs, maximum likelihood (ML) estimates were obtained for the numbers of synonymous (d_S) and non-synonymous (d_N) substitutions per site between *P. vivax* and *P. knowlesi*, and for ω (the ratio d_N/d_S). The GC composition at third codon positions of 4-fold degenerate amino acids was also determined for both species. All calculations were performed using the codon-based substitution model of Yang et al. [33] implemented in the program *codeml* of the PAML package version 3.15 [34,35], running in pairwise mode (runmode = –2), and with codon equilibrium frequencies estimated from average nucleotide frequencies at each codon position (codonFreq = 2). The values of d_N and d_S were also estimated according to the Nei-Gojobori (NG) method [36] implemented in PAML, and very similar estimates were obtained, with subtle differences relative to the ML estimates reflecting the tendency of the NG to overestimate d_S and underestimate d_N [37]. High-quality OGs were partitioned according to GO (Gene Ontology) classifications and the presence of specific motifs, and averages and standard deviations of ω for each of the resulting groups were calculated. A complete list of the values for d_S, d_N and for ω for the 3,322 *P. vivax* and *P. knowlesi* orthologs is available upon request.

Differences among chromosomes in d_S, d_N, ω and GC composition were tested with analyses of variance (ANOVAs), as implemented in Microsoft Excel. The presence of local, intra-chromosomal variation in d_S was investigated by testing whether substitution rates in neighboring genes were correlated: for each chromosome, the original dataset consists of an ordered list of n values of d_S, one for each of the n high quality OGs in the chromosome, where the order is determined by increasing value of the chromosome coordinates of the OG from which the d_S value was estimated. The correlation between d_S of gene at position X, $d_S(X)$, and the d_S of gene preceding X in the chromosome, $d_S(X\text{-}1)$ was then estimated. To determine the statistical significance of the correlation coefficient (CC)

of this analysis we compared it to a distribution of expected CCs given a random distribution of d_S values in the chromosome. For each chromosome, these random distributions were generated as follows: starting with the original ordered list of d_S values for a chromosome, the d_S for each gene was, in turn, swapped with the d_S from another, randomly chosen, gene in the chromosome. After this randomization stage, the CC between $d_S(X)$ and $d_S(X\text{-}1)$ was estimated, given the new order of d_S values. The process was repeated 5,000 times to generate a distribution of expected CCs in a randomly arranged list of d_S values. Finally, the CC of the original sample was compared against this distribution to determine the probability P of obtaining this value by chance. P is approximately equal to the fraction of randomized datasets with a smaller or equal CC value. The randomization routine was written in *Perl*, and the correlation analyses in the *R* statistical language.

In order to determine if the substitutions in 226 apicoplast-targetted genes are preferentially located at the 5' end, which encompasses the spliced signal peptide, the average d_N/d_S was calculated over each gene, and compared with the average d_N/d_S in the first 20 amino acids and with the average d_N/d_S for the last 20 amino acids. All start and stop codons were excluded as well as gaps, so that all fields had 20 aligned codons. All comparisons for which d_N or $d_S > 4$ (error due to small length of alignment) or $d_N/d_S = 99$ (default dN/dS assignment by PAML when $d_S = 0$ or very, very small) were excluded (36/226 comparisons at the 3' end, 18/226 comparisons at the 5' end.) The average d_N/d_S for the complete genes is 0.21, for the 5' end of genes is 0.70 and for the 3' end of genes is 0.38. Thus the results show that the 5' end is three times more variable than the gene as a whole, and two times more variable than the 3' end, and that substitutions are disproportionately located at the ends of apicoplast-targetted genes, with the 5' end being the most rapidly evolving segment.

Metabolome and transportome annotation. The SHARKhunt protocol in the metaSHARK pipeline [38] was used to search predicted gene models and raw genome sequence data for assignation of EC numbers. Metabolic pathways were built using Pathway Tools [39], and the EC annotations associated with *P. falciparum* and *P. vivax* COGs were manually compared and curated for differences. For identification of membrane transporters, the *P. vivax* predicted protein sequences were searched against a curated database of transport proteins [40] for similarity to known or putative transport proteins using BLASTP. All of the query proteins with significant hits (E-value < 0.001) were collected and searched against the NCBI non-redundant protein database and Pfam database [41]. Transmembrane protein topology was predicted by TMHMM [12]. A web-based interface was implemented to facilitate the annotation processes, which incorporates number of hits to the transporter database, BLAST and HMM search E-value and score, number of predicted transmembrane segments, and the description of top hits to the non-redundant protein database. The results can be viewed at the TransportDB web site (http://www.membranetransport.org/).

Protein family clustering. The *P. vivax* proteome was searched against Pfam [42] and TIGRfam [43] HMM profiles using HMMER2 (http://hmmer.wustl.edu) to identify known domains. Any sequence region scoring above the trusted cutoff assigned to the domain profile was designated as representing that domain. Domain sequences were then removed from the protein sequences and the remaining peptides searched against each other using BLASTP in order to identify potential novel domains not represented in the domain databases [44]. Similar peptide sequences were clustered by creating a link between any two peptide sequences having an identity above 30% over an amino acid span of at least 50 aa. and an Expect value < 0.001. The Jaccard coefficient of community [45] was calculated for each linked pair of peptide sequences a and b as follows:

$$J_{a,b} = \frac{\#\text{ distinct accessions matching } a \text{ and } b \text{ including } (a,b)}{\#\text{ distinct accessions matching either } a \text{ or } b}$$

with the Jaccard coefficient (Ja,b), which we also refer to as the link score, providing a measure of similarity between the two proteins. The associations between peptides that had an insufficient link score were dissolved, and the remaining links were used to generate single linkage clusters. The clustered peptides were then aligned using ClustalW and used to develop conserved protein domains not present in the Pfam and TIGRfam databases. *P. vivax*-specific domain alignments containing five or more members were considered true domains for the purpose of building families. The peptides in alignments were searched back against the *P. vivax* proteome to seek out additional members that may have been excluded during earlier stages due to the parameters employed. Full-length protein sequences were then grouped on the basis of the presence of Pfam/TIGRfam domains and potential novel domains. Proteins with exactly the same domain composition were classified into putative protein families. To determine the phylogenetic relationship among the members of each protein family, phylogenetic trees were constructed by the Neighbor–Joining method with ClustalX and a bootstrap value of 1,000.

***Vir* gene family annotation, clustering and analysis.** Due to the poor performance of gene prediction algorithms to identify *vir* genes (60 *vir* genes were predicted *in silico*), all small contigs < 10 kb were searched against previously identified *virs*, and the motif-finder RepeatScout [46]. An additional 286 *vir* genes were identified and manually curated, to give a total of 346, including 80 that are either gene fragments or pseudogenes (*i.e.,* they may not be expressed). A majority of the *virs* (221) are located on contigs <10 kb that could not be assigned to specific chromosomes. All previosuly identified *vir* sequences were searched against the newly identified *P. vivax* genome *virs* using BLASTP (cutoff e-10), and 165 identified as novel. These were clustered using Jaccard analysis and six new subfamilies G-L established; a total of 84 *virs* could not be clustered into a family due to their extreme divergence. Conserved motifs in VIR proteins were identified using MEME [47] with parameters as follows: 0 or 1 occurrence of each motif per sequence (-mod zoops); maximum number of motifs to identify (nmotifs) = 10; minimum motif width (-minw)= 5; maximum motif motif width (-maxw)=25. Motif 2 (VVGVMMTFFFLYKFTP) encodes hydrophobic amino acids typical of transmembrane domains; motif 3 contains a PEXEL-like sequence (KERKDLHDYFKNYDTIKC); remaining motifs are predicted to be exposed globular domains (1: CIYLNYWLYDQI, 4: KRKGKIFEHNYYEEYEKELAMYGSE, 5: VGAFFRGGRGRVHRIPRSFHGQFPG, 6: CEKYCTYVTYIKSLYE, 7: IADSPGTLGTVHEELDSNFFRNIIM, 8: TFLDSQMDRYYLNYQPDQDSYY, 9: VKELCKKLVRNLKKIS, 10: YDPKDLLSKLDC). The overall organization and order of the motifs is maintained, with the central core motifs 9, 1, 3, 6 and 10 followed by C-terminus motifs 7, 2, 4, 8 and 5 embedded in a variant-sized portion of the molecule. Noticeable exceptions to this organization are subfamilies D and H which lack most of the central core motifs and seem unrelated to remaining VIR subfamilies. Twenty-five introns were found to contain a conserved motif 5'-GTAATTGCACATTCGA[T]$_4$ATGAAATATGA[ATGTT]$_2$AGTATCTAT-3' proximal to the donor splice site with the majority (16 of the 25) belonging to subfamily E. The promoter regions of 23 of these 25 genes also contain a highly conserved motif 5'-TTTTATG/TTAGACAATATTA-3' which is highly position-specific, with 60% of the motifs occurring 500-600 bp upstream of *vir* gene translation start sites. RT-PCR of subfamilies A, B, C, D, E, H, K and L was performed using cDNA from the Salvador I laboratory isolate and from three patient isolates from Brazil (subfamilies F, G, I and J were too degenerate for primer design) using methods described previously [48,49]. A total of 317 VIR proteins, 135 CIR proteins and 245 BIR proteins were aligned and clustered using a Neighbour-Joining method as implemented in PAUP* [50].

GPI-anchor prediction. *P. vivax* predicted proteins were assessed for GPI-anchoring with GPI-HMM, an HMM predictor trained with validated *P. falciparum* GPI-anchored proteins (GPI-APs; [51]). The signals required for GPI-anchoring reside at the both ends of protein sequences but because these regions are sometimes missed by automated gene predictors *P. vivax* orthologs of *P. falciparum* GPI-APs were identified and alignments of the orthologs used to identify *P. vivax* proteins missing terminal regions which could be then recovered by manual curation. The *P. vivax* proteins with positive GPI-HMM scores were manually curated to remove false positives following the criteria of Gilson *et al*. [51].

Apicoplast target prediction. All peptides were examined for the presence of an N-terminal leader consisting of a signal sequence followed by a transit peptide-like region, slightly enriched in basic amino acids and highly depleted in acidic residues, using SignalP [11] and a version of PlasmoAP modified to adjust for differences in amino acid usage between *P. vivax* and *P. falciparum* [52]. Candidates with close *Cryptosporidium* matches (the closest known relative of *Plasmodium* that lacks an apicoplast) were excluded as false positives. Orthologs of predicted *P. falciparum* proteins were identified by BLASTing IDA (Inferred from Direct Assay) and ISS (Inferred from Sequence or Structural Similarity) predicted *P. falciparum* apicoplast proteins against all predicted *P. vivax* proteins. In most cases, best hits were already predicted to be apicoplast targeted by identification of targeting presequences. Where this was not the case, alignments were manually inspected and examined for apicoplast-targeting leaders that were missed.

The initial *P. falciparum* predicted apicoplast proteome included ~175 genes with predicted functions and another ~340 with hypothetical proteins. In the new *P. vivax* list, this is now reduced to 157 genes with predicted functions and another 159 with unknown functions. The reduction in proteome size is for several reasons. The availability of more complete or near-complete genomes (particularly other *Plasmodium* species as well as *Theileria* and *Cryptosporidium*), means that the N-termini can be more accurately determined by comparison, and this has weeded out numerous falsely predicted apicoplast proteins. The availability of the *Cryptosporidium* genomes also allows false-positives with apicoplast-like leaders to be identified. Some of the *P. falciparum* genes originally annotated as apicoplast-targeted now have apparent PEXEL erythrocyte-targeting motifs, and these can also now be considered as false-positives. The elimination of doubtful apicoplast proteins from the IEA (Inferred from Electronic Annotation) list was also performed more strictly than for the original *P. falciparum* IEA annotation, and that stringency also contributes to the reduction in size for this *P. vivax* dataset.

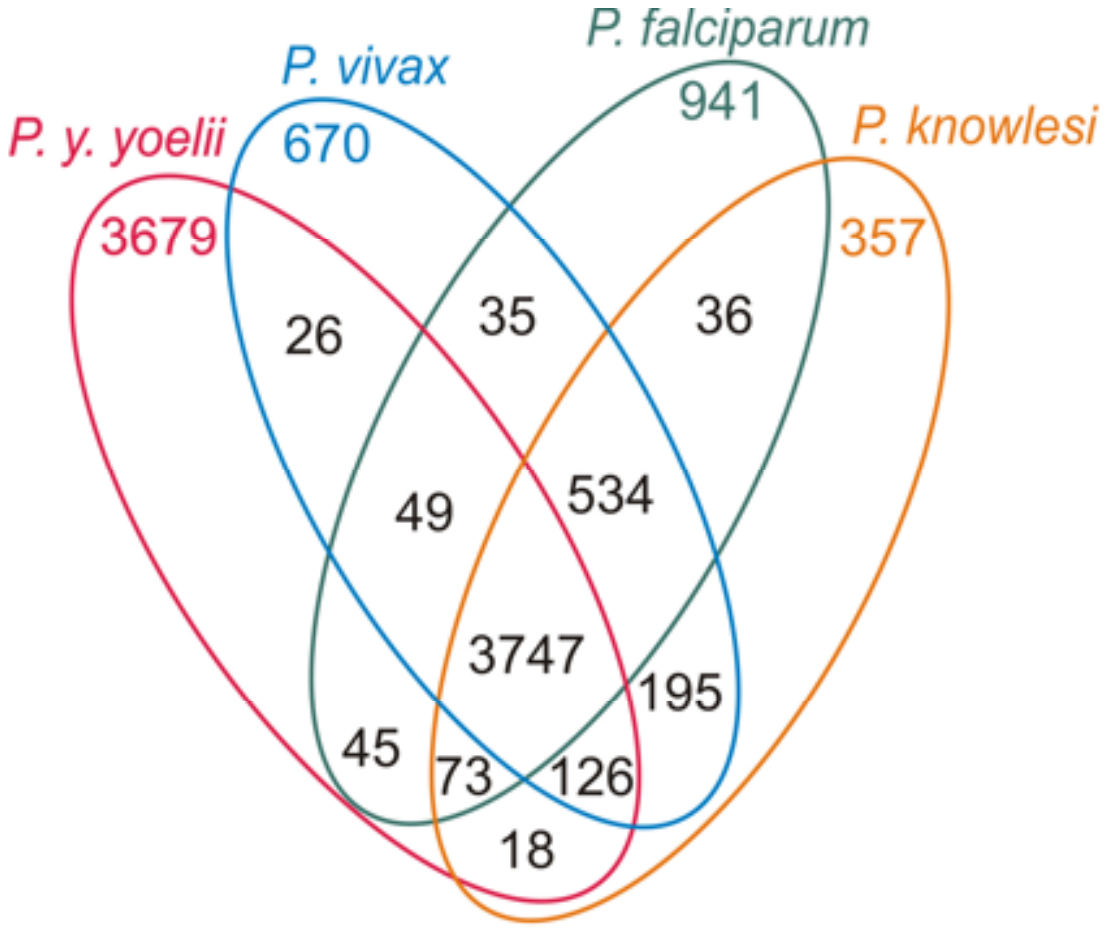
Screening the proteome for candidate hypnozoite genes. There are two likely scenarios for exoerythrocytic hypnozoite formation: (1) a *P. vivax*-specific pathway exists, consisting of one or more *P. vivax*-specific proteins that are uniquely encoded in the genome; or (2) regulation of exoerythrocytic form development is different in *P. vivax* compared to other *Plasmodium* species, and involves genes that are not unique to *P. vivax* that could, for example, be differentially transcribed in sporozoite or early liver stages. Orthologous gene clustering (see above) identified ~ 150 unique *P. vivax* genes that did not cluster into a *Plasmodium* orthologous group (**Supplementary Table 11**) and that might be involved in scenario (1). However, in the absence of functional data these genes provide little insight. To identify genes involved in scenario (2), we created a fasta file of protein sequences extracted from GenBank (Build 02/10/2006) using 'dormancy' as a key word. A total of 173 sequences were identified, and PFAM and BLAST searches were undertaken against the *P. vivax* predicted proteome using BLAST cut-off of < E-value 10^{-10}. A list of *P. vivax* homologs is shown in **Supplementary Table 12**.

Homology modeling. Homology models of *P. vivax* ATPASE6 and *P. falciparum* ATPASE6 based on

the crystal structure of 1IWO [53], and of *P. vivax* CYTB based on the crystal structures of 1BCC [54] and 1BE3 [55] were constructed using the Homology software program of the Insight II molecular modeling system from Molecular Simulations Inc. (MSI). The models were minimised using Discover (MSI) to an RMS deviation of less than 10^{-6} using steepest descents and conjugate gradients taking into account morse and cross terms and charges. Docking of artemisinin to both *P. falciparum* and *P. vivax* ATPASE6 models and atovaquone to the *P. vivax* CYTB model was performed using Genetic Optimisation of Ligand Binding (GOLD) from the Cambridge Crystallographic Data Centre. Contact residues were defined as being located within 3Å of artemisinin or atovaquone when docked to the *Plasmodium* species ATPASE6 and CYTB models, respectively.

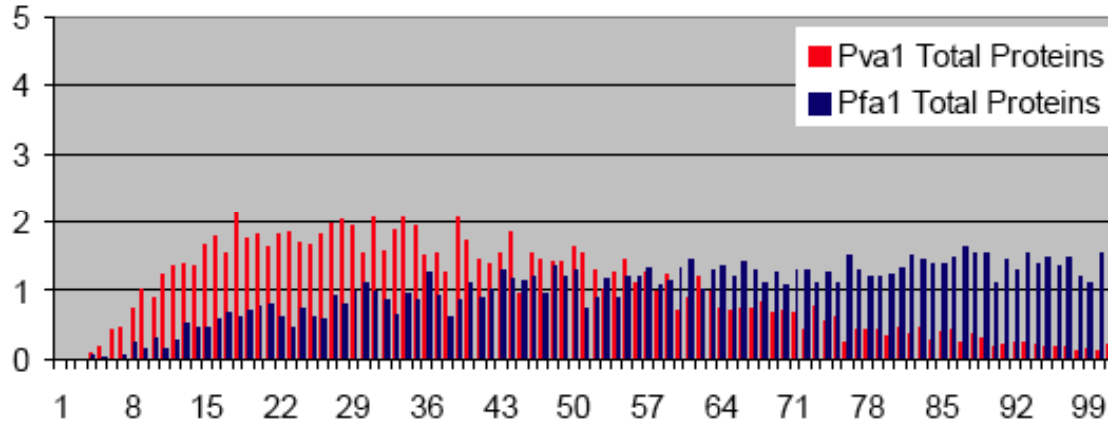
Supplementary Figures

Supplementary Figure 1. Orthologs shared by four *Plasmodium* species. From a total of 4,884 ortholog groups analyzed, 3,747 are shared by all four species, representing ~77% of the total number of genes. *P. vivax* and *P. knowlesi* share the most number of ortholog groups (4,337), whereas *P. knowlesi* and *P. y. yoelii* share the least (3,958), a result that mirrors the phylogenetic relationships between the species. *P. y. yoelii* appears to contain the most singletons; however this value is inflated due to the large number of partial gene predictions, a function of the fragmented nature of the genome sequence.

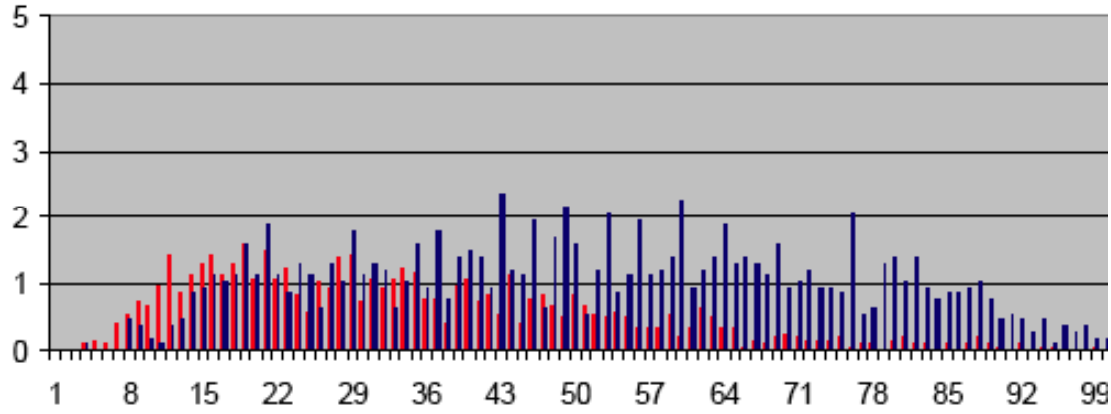


Supplementary Figure 2. Coverage of *P. vivax* and *P. falciparum* proteins with low complexity regions (LCR) for (**A**) all proteins; (**B**) 1,070 housekeeping orthologs; and (**C**) 2,033 conserved hypothetical orthologs. The percent LCR coverage is indicated on the x-axis, percentage of proteins with that coverage on the y-axis, blue bars represent *P. falciparum* and red bars indicate *P. vivax*. LCRs constitute a smaller proportion of *P. vivax* proteins on average (39%) than *P. falciparum* proteins (60%). Similarly for orthologs between the two species, average LCR coverage was lower in *P. vivax* for housekeeping orthologs (32% for *P. vivax* versus 50% for *P. falciparum*) and for hypothetical orthologs (40% for *P. vivax* versus 69% for *P. falciparum*). Conserved hypothetical genes are more likely to show errors in their structural annotation which may explain the difference between results shown in **B** and **C**.

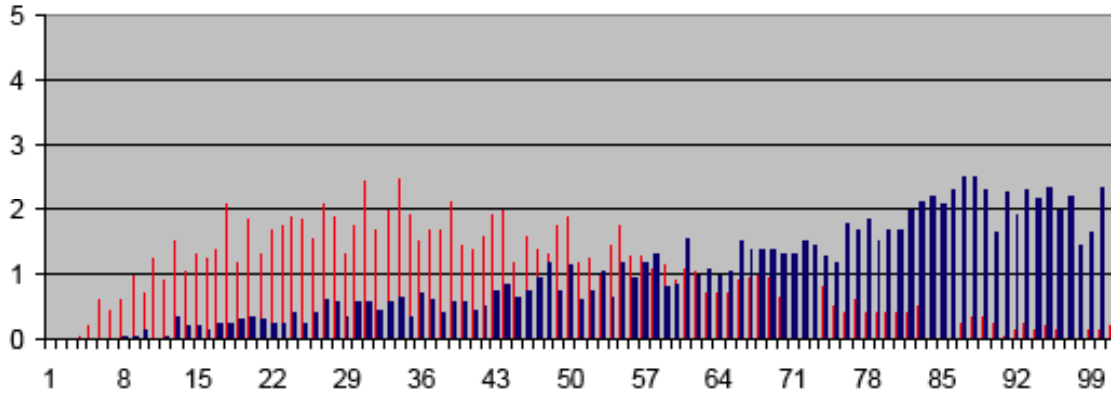
A.



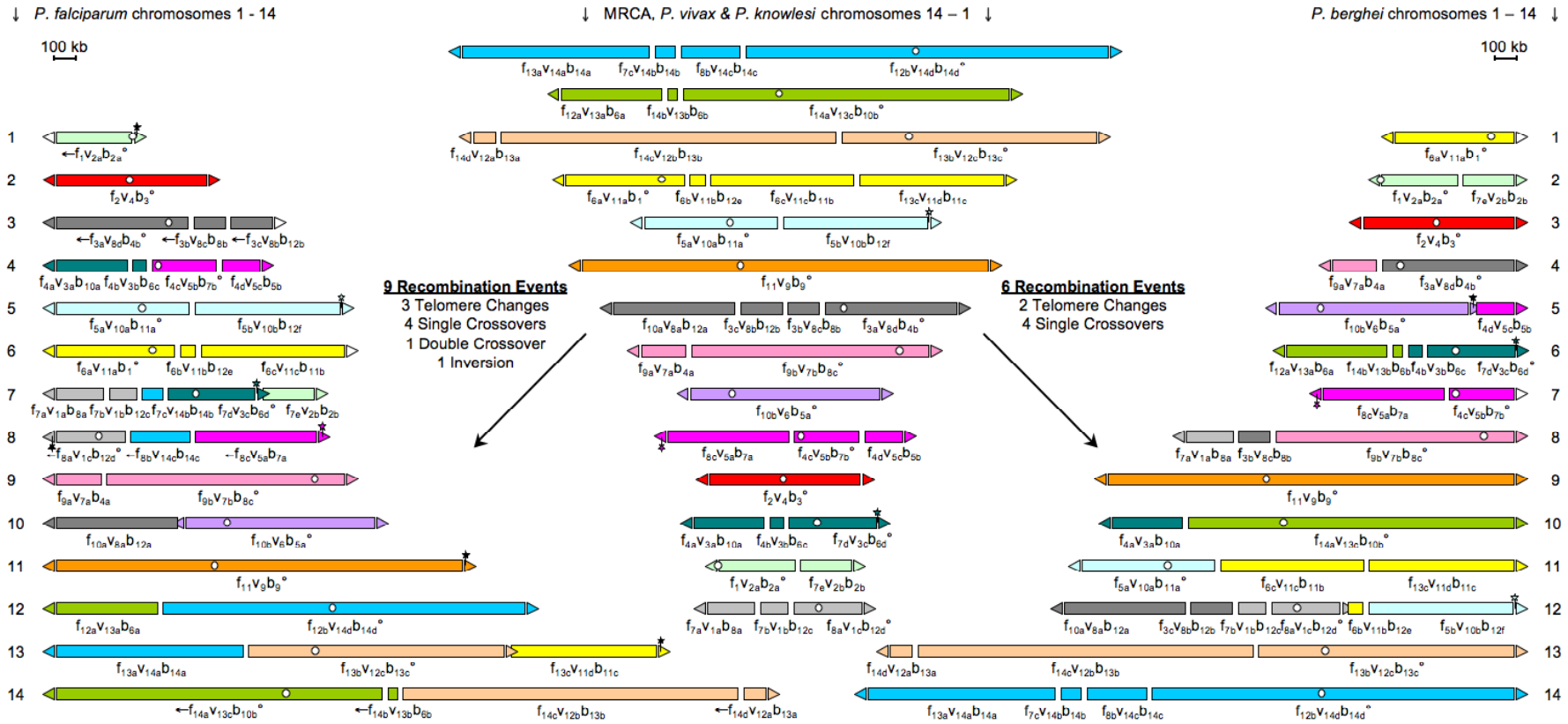
B.



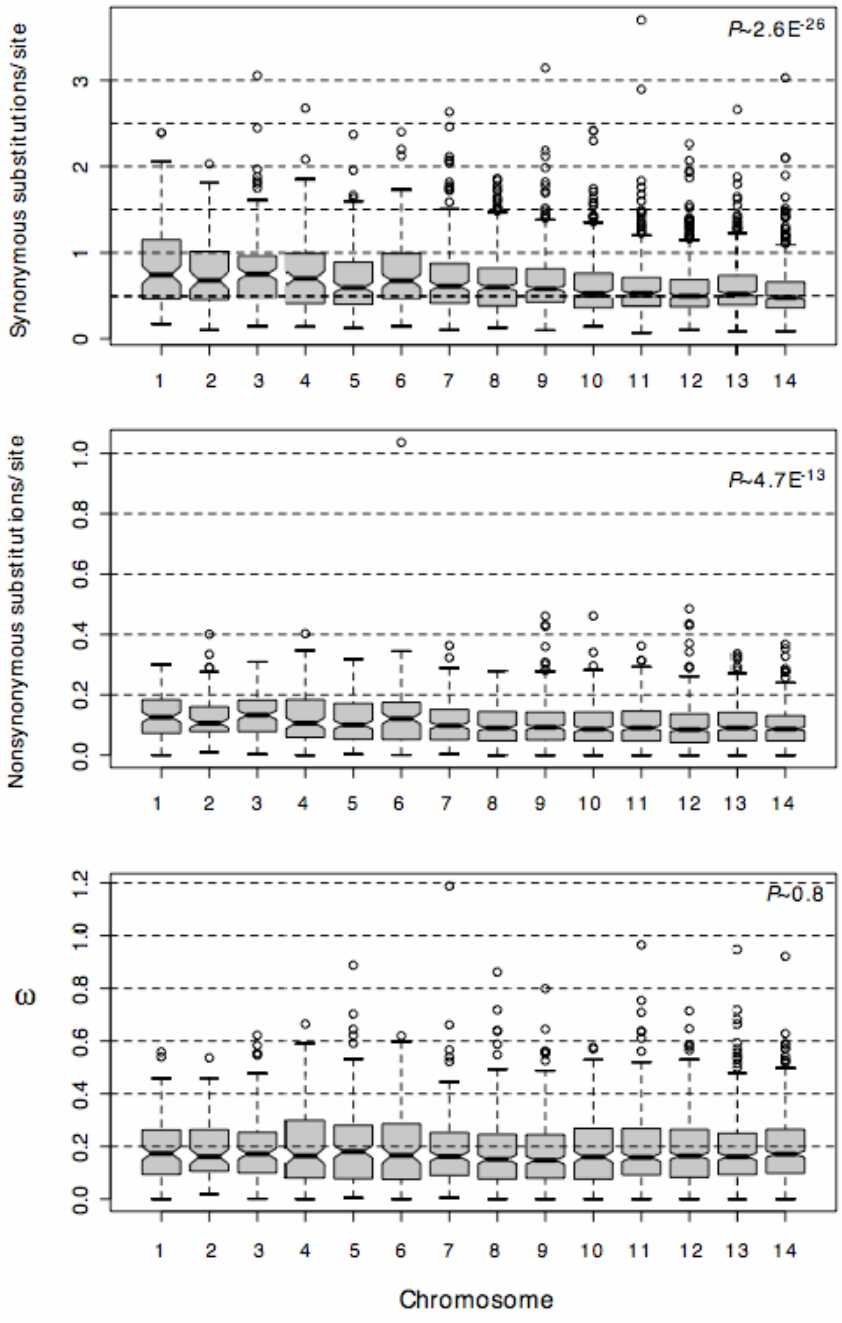
C.



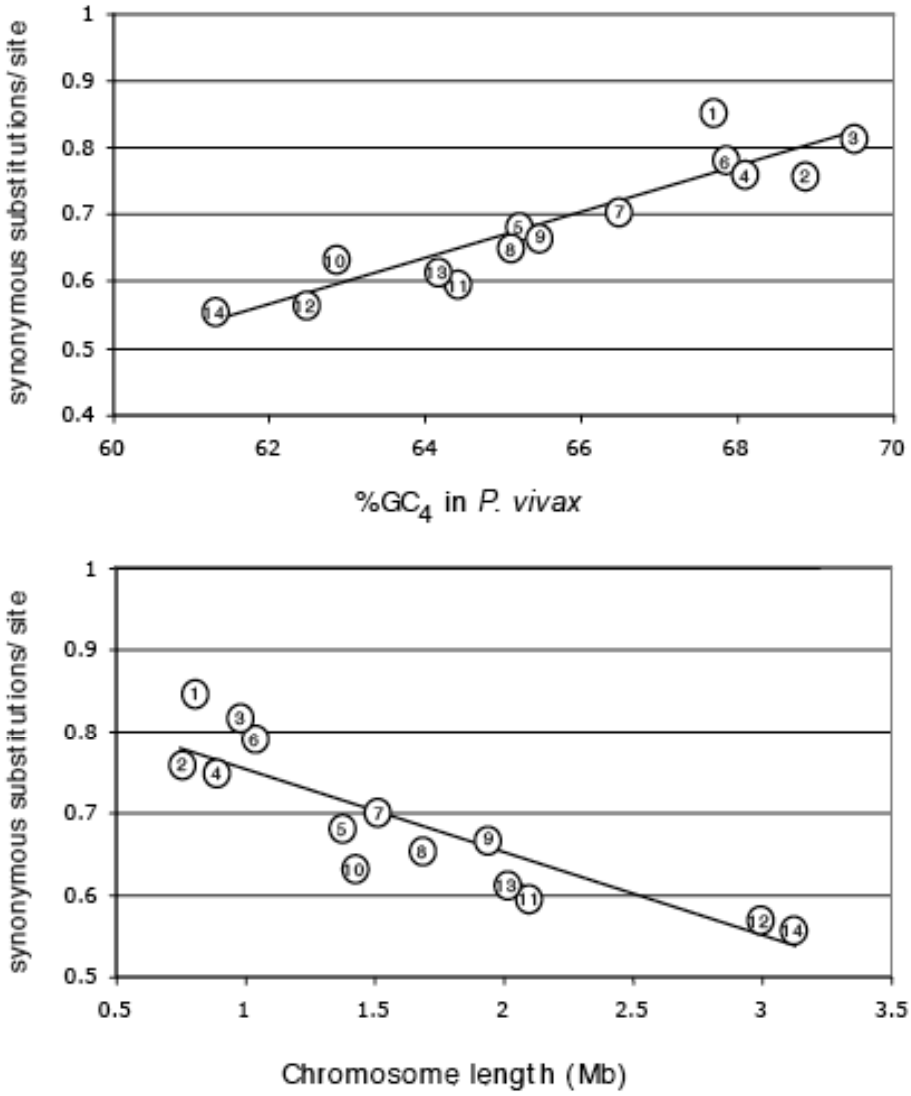
Supplemental Figure 3. A schematic chromosome map of *P. vivax*, *P. knowlesi*, *P. falciparum* and the rodent malaria parasites (here represented by *P. berghei*). The genome organization of *P. vivax* and *P. knowlesi* (center, also identified as the most recent common ancestral [MRCA] form) is shown next to the organization of *P. falciparum* (left) and *P. berghei* (representing three rodent malaria species; right); the karyotype of the latter two species can be reconstructed through nine and six chromosomal rearrangements, respectively, from the MRCA genome. The 36 synteny blocks are coloured according to their chromosomal location in the MRCA genome, and are identified as $f_{\#x}v_{\#x}b_{\#x}$, where f, v and b stand for *P. falciparum*, *P. vivax* and *P. berghei*, # represents the chromosome number, x the synteny block order and ° the presence of a conserved centromere. The following symbols are used: ← inverted orientation of ten synteny blocks in *P. falciparum*; ○ predicted centromeres; ▷ telomeric/subtelomeric regions in the MRCA, coloured according to their chromosomal location (three new regions in *P. falciparum* and two new regions in the rodent malaria parasites are indicated as white arrowheads); ✦ *rrna* gene units (black and coloured stars represent non-syntenic and three syntenic *rrna* gene units, respectively).



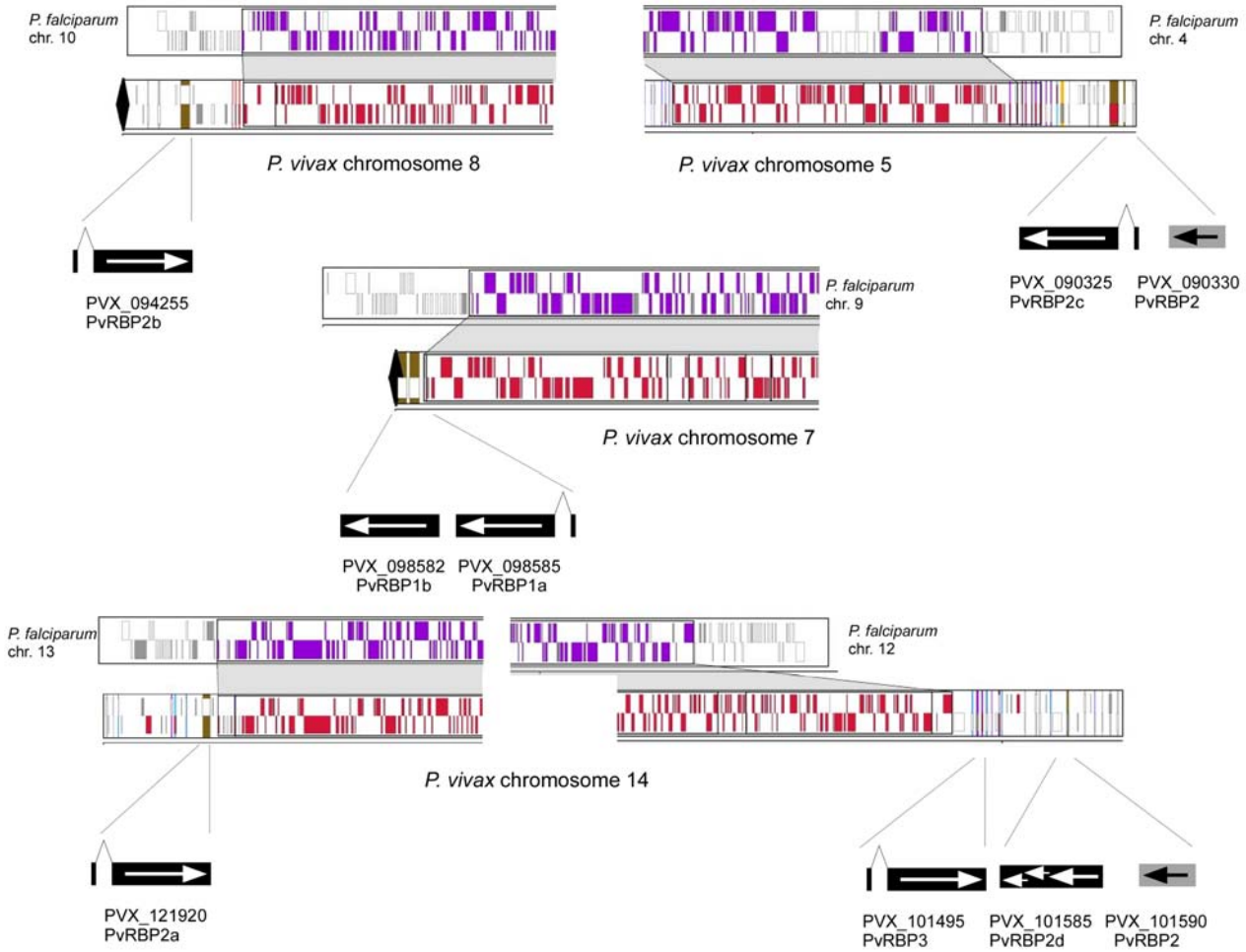
Supplementary Figure 4. Per chromosome rate of synonymous (d_S) and nonsynonymous (d_N) substitutions and omega (ω, ~ d_N/d_S), estimated between a set of 3,324 high-quality orthologous gene pairs from *P. vivax* and *P. knowlesi*. Each hourglass plot (interquartile range, IQR) contains the sample's 25% to 75% range (quartiles Q1 to Q3, respectively), with the bottleneck placed at the sample median. Horizontal tick marks show the range of all elements within Q1-1.5 X IQR and Q3+1.5 X IQR (equivalent to the 99.3% interval of a normal distribution). Open circles represent outliers (data points outside this range). The width of the bottleneck (*i.e.*, the length of the V-shaped notch) is an indication of the confidence of the median; a lack of overlap of the bottleneck between samples implies that the samples are statistically different. Chromosomes differ significantly in both d_S and d_N values, but not in ω (ANOVA probability values in top right corner). Mean d_S and d_N per chromosome are highly correlated (r^2~90%; probability of F~2.3e^{-7}).



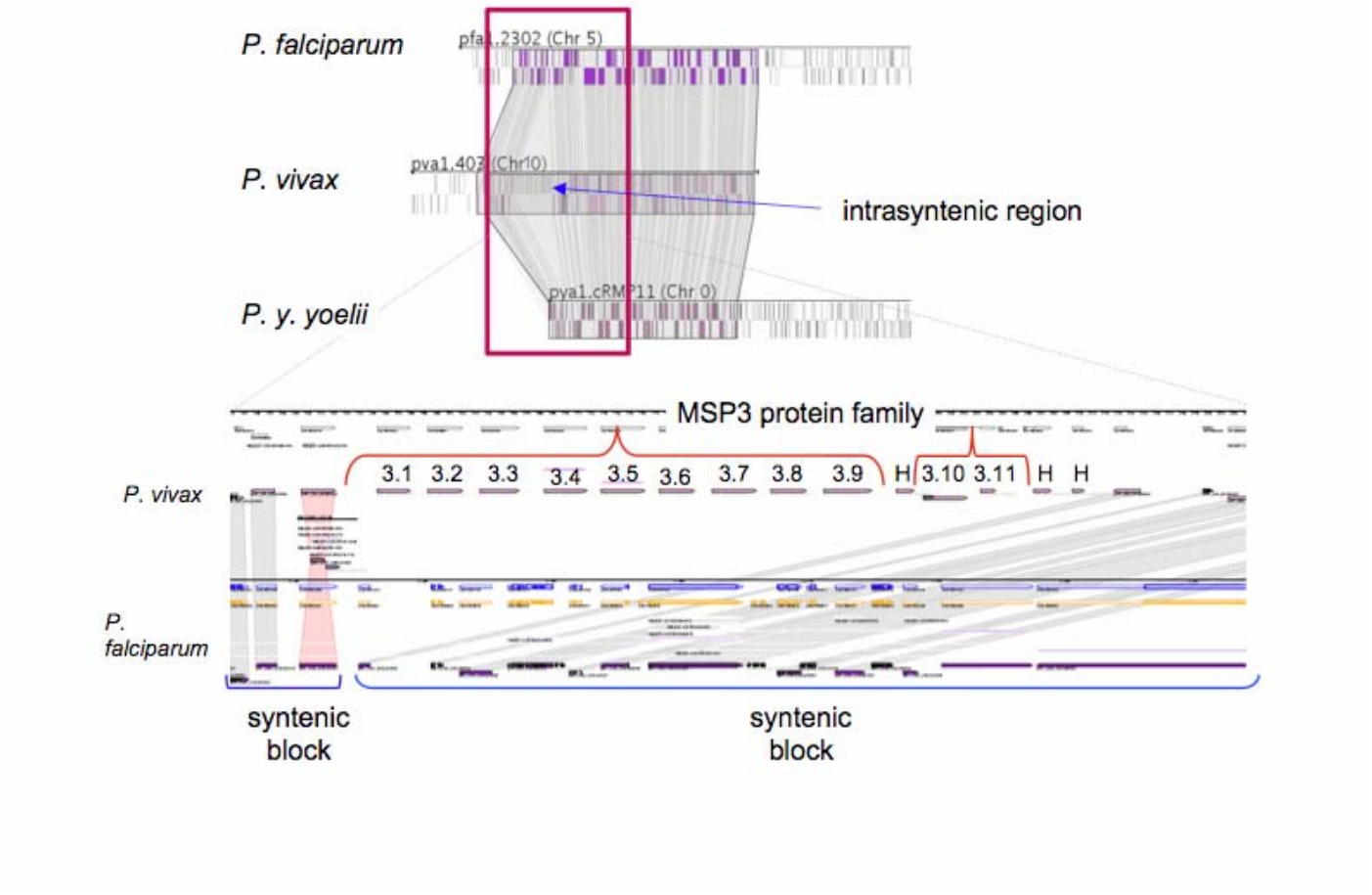
Supplementary Figure 5. Relationship between the average number of synonymous substitutions per synonymous site (d_S) per chromosome and the G+C content in four-fold degenerate sites (%GC_4) in *P. vivax* genes (top panel), and between average d_S per chromosome and chromosome length (bottom panel). The chromosome number corresponding to each data point is shown. Average d_S is strongly and positively correlated with % GC_4 (r^2 ~ 87%; prob. F ~ 1.2E^{-6}) and negatively correlated with chromosome length (r^2 ~76%; prob. F ~4.9e^{-5}). Synonymous sites of genes on the smallest chromosomes (~1 Mb) evolve ~1.5 times faster than genes on the two largest (~3 Mb) chromosomes. Thus, genes in the most G+C-rich, centromere-proximal, locations evolve faster than average; G+C content decreases very gradually with distance from the centromere, and so does the rate of sequence evolution, as measured by d_N and d_S. Since regions of high G+C content make up a larger proportion of the small chromosomes than they do of the larger ones, average d_N and d_S are higher in smaller chromosomes.



Supplementary Figure 6. Chromosomal locations of nine *P. vivax rbp* genes. All *rbp* genes are subtelomerically located within synteny breakpoints. PVX_101585 (*rbp2d*) is an apparently truncated gene with two internal frame shifts (similar to *PfRh3*) that is most likely a duplication of *rbp2c* and is missing the characteristic intron. *rbp2* loci are inverted in location relative to the single copy *rbp3* and two *rbp1* genes, which are orientated towards the telomeres. Such orientation of the *rbp2* type genes would allow for the potential of recombination with other *rbp2* loci, although there is no evidence of frequent recombination between the genes. Not shown is the partial gene *rbp1* (PVX_125738), which could not be localized to a chromosome.



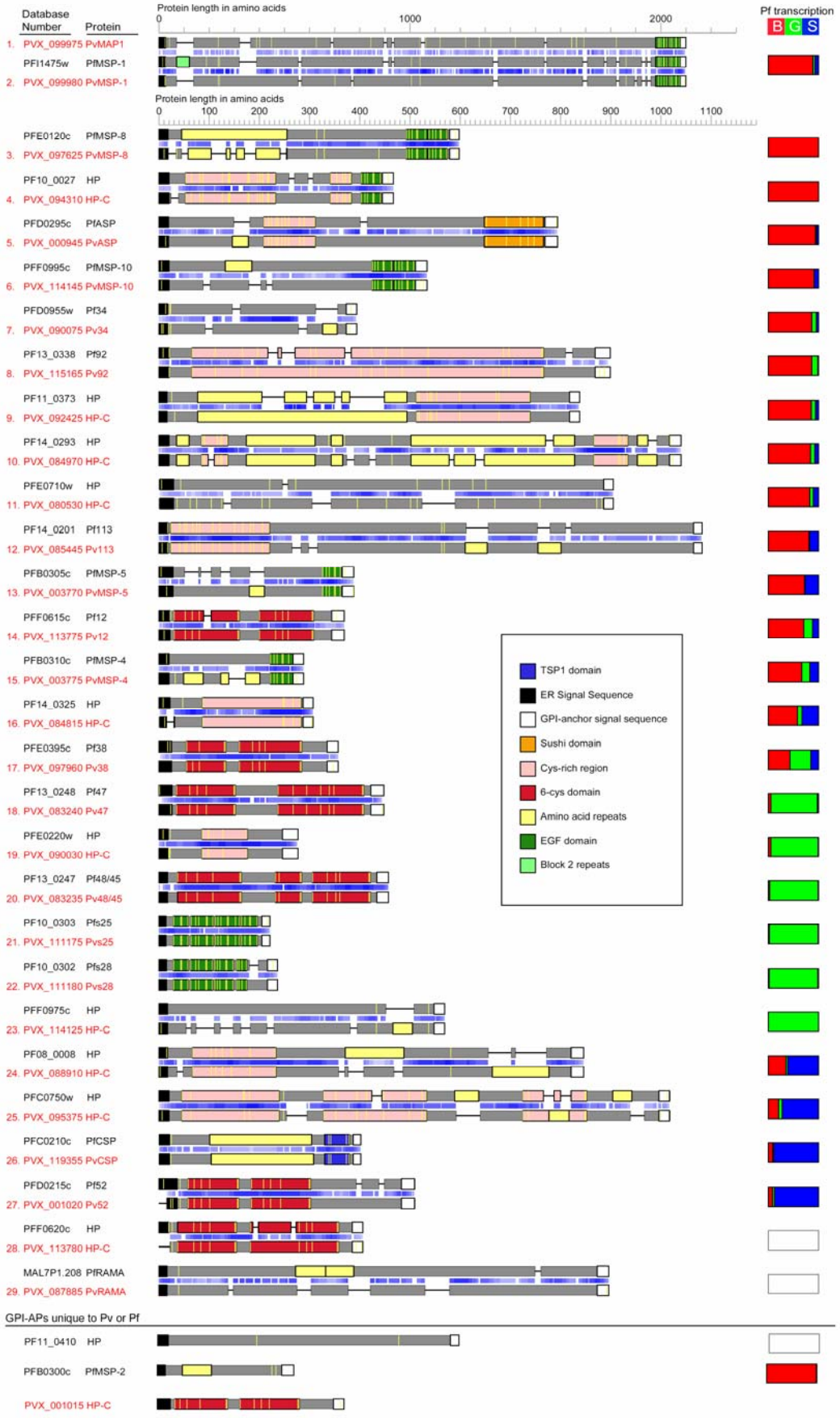
Supplementary Figure 7. Schematic of *P. vivax msp3* gene family expansion and syntenic regions in *P. falciparum* and *P. y. yoelii*. Eleven *msp* gene family members, labeled 3.1 to 3.11, constitute the majority of the 14 genes located within an intrasyntenic indel found on chromosome 10 of *P. vivax*. Three *P. vivax msp3* genes (*PvMSP3α* [3.10], *PvMSP3β* [3.3], and *PvMSP3γ* [3.1] [56]) were originally identified as homologs of *P. falciparum msp3* genes, with their characteristic central alanine-rich domain with heptad repeats predicted to form α-helical secondary and coiled-coil tertiary structures, a C-terminal acidic domain and the absence of a transmembrane domain or GPI anchor [56,57]. The added defining characteristic encoded in each *msp3* gene family member, not present in three of the genes in the indel (PVX_097715, PVX_097735 and PVX_097730 all indicated as H for hypothetical), is the amino acid motif NLNRG, located immediately after the signal peptide sequence. One interesting question concerns whether all members of the family are expressed simultaneously, or if different members are expressed at different times during the cell cycle, or under certain physiological conditions.



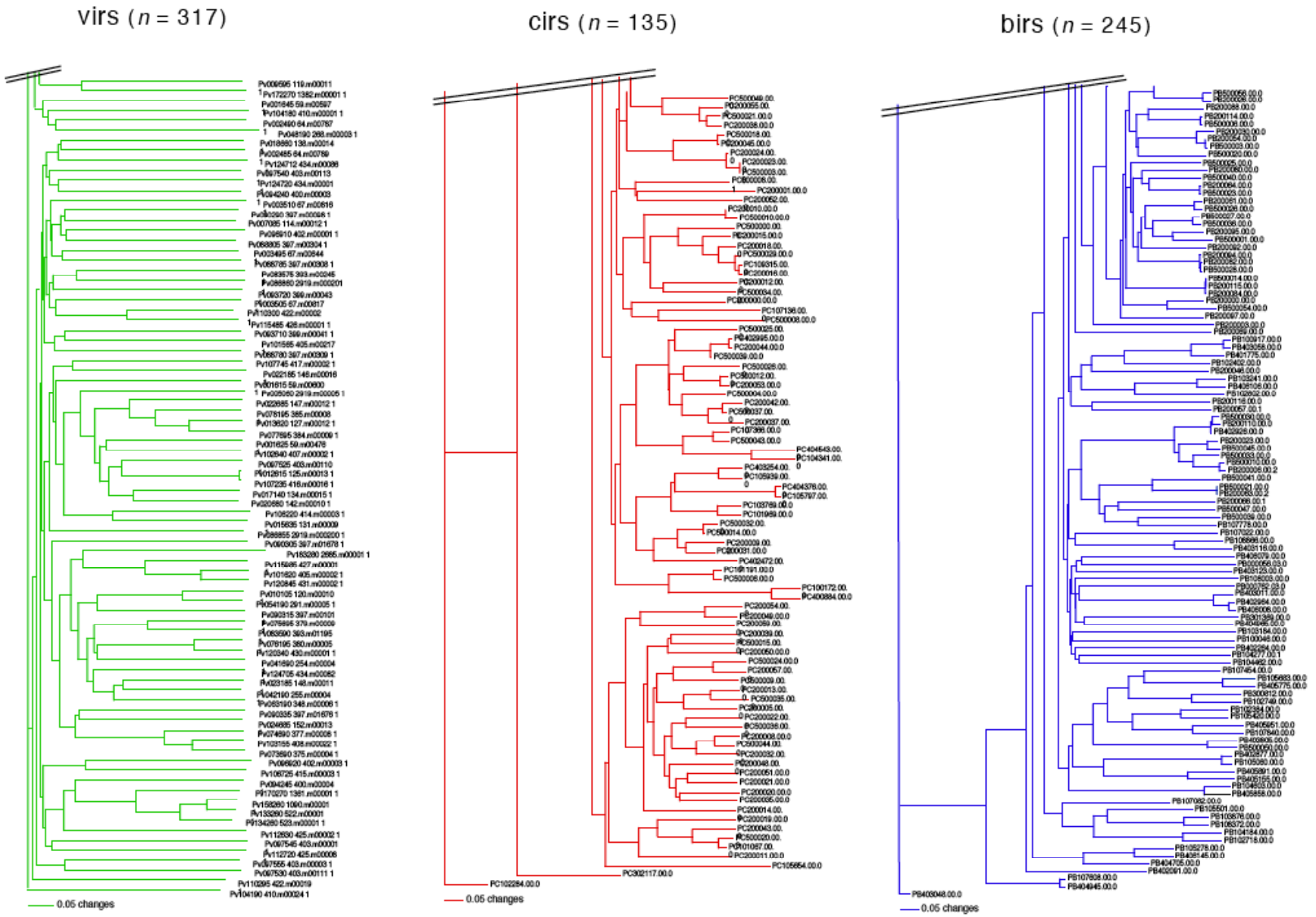
Supplementary Figure 8. Comparison of the predicted GPI-anchored proteomes of *P. vivax* and *P. falciparum*. Alignments of orthologous GPI-anchored proteins are shown schematically, with the degree of amino acid similarity proportional to the intensity of blue shading. Database numbers and protein names of *P. falciparum* (black text) and *P. vivax* (red text) GPI-anchored proteins are shown along with their predicted protein structure. Cysteine residues (yellow bars) and recognisable domains are indicated (colored boxes). Affymetrix microarray expression data for *P. falciparum* [22] are shown on the right. HP: hypothetical; HP-C: hypothetical conserved.

Interestingly, both the *P. vivax* and *P. knowlesi* genomes encode an apparently paralogous gene next to *msp1*, the largest and most abundant protein on the *P. falciparum* merozoite surface. The gene encoding the paralog, *Pvmap1*, is found immediately upstream of the characterised *Pvmsp1*. *Pvmap1* is not closely related to *Pvmsp1*, displaying only 11% identity and 22% similarity although its overall size and major structural features, such as a C-terminal double EGF module, are similar. Similar to PfMSP1, the first EGF domain in PvMAP1 includes a complete set of three disulphide bridges. One of these canonical bonds is absent in PvMSP1 [58] and in MSP1 proteins in rodent malaria species [59]. As yet there is no evidence for expression of *Pvmap1* although, only limited gene expression data is available for *P. vivax* at this stage. Functional redundancy of the paralogs, although unlikely, would have a significant impact on blood-stage vaccine development in this species since much research currently focuses on PvMSP1 [60].

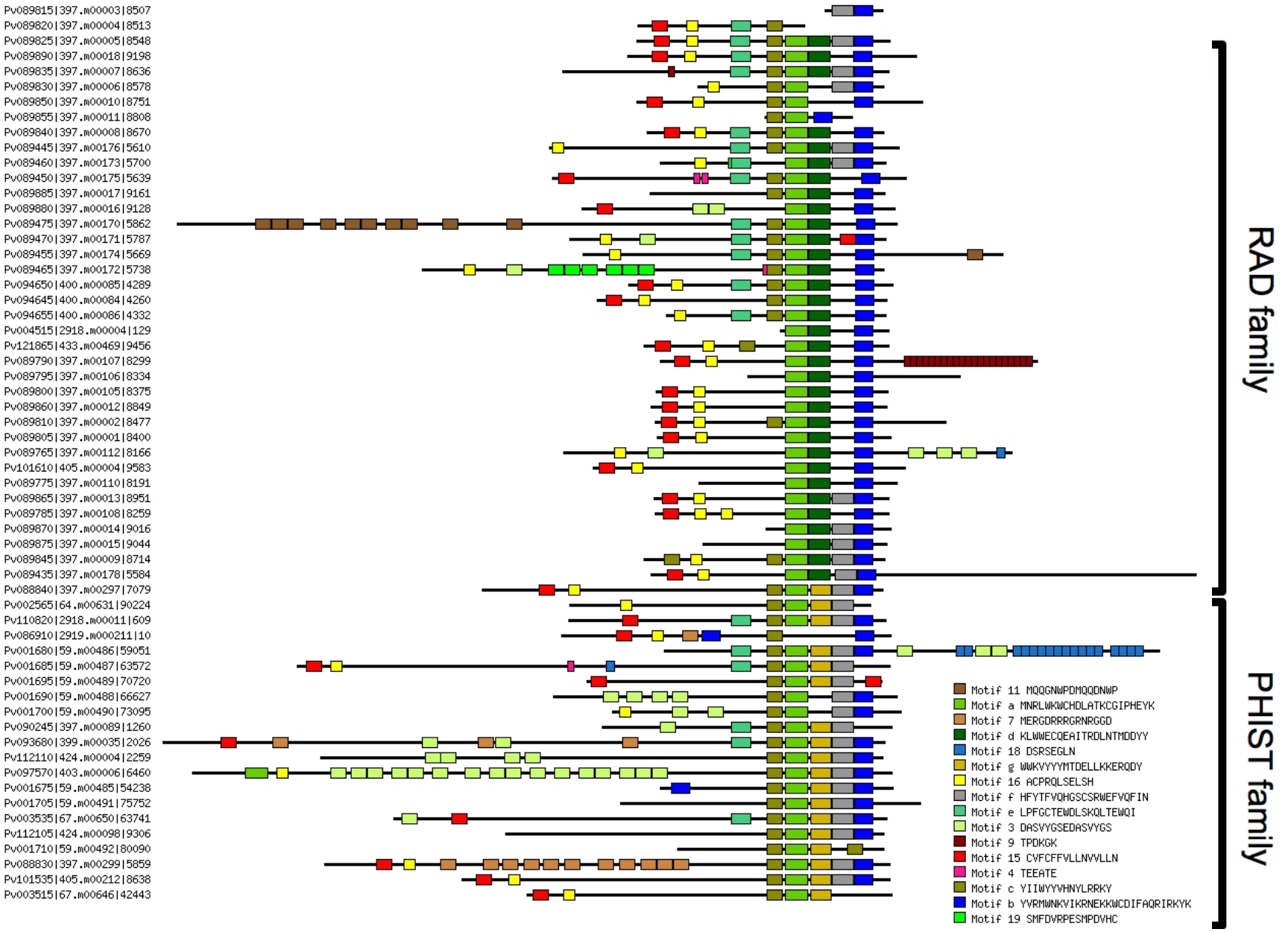
The six cysteine protein family, shown below as those proteins that contain the '6-cys domain' and including Pf/Pv12, Pf/Pv38, Pf/Pv47, Pf/Pv48/45, Pf/Pv52 and HP/HP-C, appear to be restricted to phylum apicomplexa. These Cys_6 proteins are related to the mostly GPI-anchored surface antigen protein family of *Toxoplasma gondii* thought to bind the sulfated proteoglycans on host cells [61]. Pf12 and Pf38 are natural antigens, localized to the plasma membrane of mature merozoites, although Pf38 also localizes to apical organelles [62]. *P. vivax* appears to have an additional member of the Cys_6 family not present in *P. falciparum* (PVX_001015), which resides in a microsyntenic break point on *P. vivax* chromosome 3 close to the subtelomere.


Database Number
Protein
Protein length in amino acids
0
1000
2000
Pf transcription
B G S
1. PVX_099975 PvMAP1
PFI1475w PfMSP-1
2. PVX_099980 PvMSP-1
Protein length in amino acids
0 100 200 300 400 500 600 700 800 900 1000 1100
PFE0120c PfMSP-8
3. PVX_097625 PvMSP-8
PF10_0027 HP
4. PVX_094310 HP-C
PFD0295c PfASP
5. PVX_000945 PvASP
PFF0995c PfMSP-10
6. PVX_114145 PvMSP-10
PFD0955w Pf34
7. PVX_090075 Pv34
PF13_0338 Pf92
8. PVX_115165 Pv92
PF11_0373 HP
9. PVX_092425 HP-C
PF14_0293 HP
10. PVX_084970 HP-C
PFE0710w HP
11. PVX_080530 HP-C
PF14_0201 Pf113
12. PVX_085445 Pv113
PFB0305c PfMSP-5
13. PVX_003770 PvMSP-5
PFF0615c Pf12
14. PVX_113775 Pv12
PFB0310c PfMSP-4
15. PVX_003775 PvMSP-4
PF14_0325 HP
16. PVX_084815 HP-C
PFE0395c Pf38
17. PVX_097960 Pv38
PF13_0248 Pf47
18. PVX_083240 Pv47
PFE0220w HP
19. PVX_090030 HP-C
PF13_0247 Pf48/45
20. PVX_083235 Pv48/45
PF10_0303 Pfs25
21. PVX_111175 Pvs25
PF10_0302 Pfs28
22. PVX_111180 Pvs28
PFF0975c HP
23. PVX_114125 HP-C
PF08_0008 HP
24. PVX_088910 HP-C
PFC0750w HP
25. PVX_095375 HP-C
PFC0210c PfCSP
26. PVX_119355 PvCSP
PFD0215c Pf52
27. PVX_001020 Pv52
PFF0620c HP
28. PVX_113780 HP-C
MAL7P1.208 PfRAMA
29. PVX_087885 PvRAMA
TSP1 domain
ER Signal Sequence
GPI-anchor signal sequence
Sushi domain
Cys-rich region
6-cys domain
Amino acid repeats
EGF domain
Block 2 repeats
GPI-APs unique to Pv or Pf
PF11_0410 HP
PFB0300c PfMSP-2
PVX_001015 HP-C

Supplementary Figure 9. Snapshot of a portion of the phylogenetic trees for three of the *pir* gene superfamilies. A total of 82 VIRs, 84 CIRs and 104 BIRs are shown, a subsection of the total phylogenetic tree for each family. Of note are the long branch lengths of the VIR tree as opposed to the shorter CIR and BIR branch lengths, indicating more significant structuring in the latter and greater divergence of members in the former. Only one representative section of the tree is shown for each species.


virs (n = 317)
cirs (n = 135)
birs (n = 245)
0.05 changes
0.05 changes
0.05 changes

Supplementary Figure 10. The *P. vivax* Pv-fam-e (RAD) gene family is closely related to the PHIST gene family. Both RAD and PHIST share similar motifs at C-terminal regions, but differ in that RAD members contain motif D whereas PHIST members contain motif G. Most RAD and PHIST proteins contain a PEXEL motif (Motif 16) and a hydrophobic region (Motif 15) that may be a signal peptide or signal anchor at N-terminal regions. There are 44 *P. vivax* RAD proteins, located on chromosomes 3, 4, 5, 6, 8 and 14 almost all of which are internally located; 39 of these located on chromosomes 5 and 8 are shown. The 21 PHIST genes are primarily subtelomerically located with one tandem locus of eight genes on chromosome 6; all are shown below. Gene identifiers, indicated as Pvxxxxxx, refer to GenBank identifiers PVX_xxxxxx (other identifiers are from internal databases at TIGR).


Pv089815|397.m00003|8507
Pv089820|397.m00004|8513
Pv089825|397.m00005|8548
Pv089890|397.m00018|9198
Pv089835|397.m00007|8636
Pv089830|397.m00006|8578
Pv089850|397.m00010|8751
Pv089855|397.m00011|8808
Pv089840|397.m00008|8670
Pv089445|397.m00176|5610
Pv089460|397.m00173|5700
Pv089450|397.m00175|5639
Pv089885|397.m00017|9161
Pv089880|397.m00016|9128
Pv089475|397.m00170|5862
Pv089470|397.m00171|5787
Pv089455|397.m00174|5669
Pv089465|397.m00172|5738
Pv094650|400.m00085|4289
Pv094645|400.m00084|4260
Pv094655|400.m00086|4332
Pv004515|2918.m00004|129
Pv121865|433.m00469|9456
Pv089790|397.m00107|8299
Pv089795|397.m00106|8334
Pv089800|397.m00105|8375
Pv089860|397.m00012|8849
Pv089810|397.m00002|8477
Pv089805|397.m00001|8400
Pv089765|397.m00112|8166
Pv101610|405.m00004|9583
Pv089775|397.m00110|8191
Pv089865|397.m00013|8951
Pv089785|397.m00108|8259
Pv089870|397.m00014|9016
Pv089875|397.m00015|9044
Pv089845|397.m00009|8714
Pv089435|397.m00178|5584
Pv088840|397.m00297|7079
Pv002565|64.m00631|90224
Pv110820|2918.m00011|609
Pv086910|2919.m00211|10
Pv001680|59.m00486|59051
Pv001685|59.m00487|63572
Pv001695|59.m00489|70720
Pv001690|59.m00488|66627
Pv001700|59.m00490|73095
Pv090245|397.m00089|1260
Pv093680|399.m00035|2026
Pv112110|424.m00004|2259
Pv097570|403.m00006|6460
Pv001675|59.m00485|54238
Pv001705|59.m00491|75752
Pv003535|67.m00650|63741
Pv112105|424.m00098|9306
Pv001710|59.m00492|80090
Pv088830|397.m00299|5859
Pv101535|405.m00212|8638
Pv003515|67.m00646|42443
Motif 11 MQQGNWPDMQQDNWP
Motif a MNRLWKWCHDLATKCGIPHEYK
Motif 7 MERGDRRRGRNRGGD
Motif d KLWWECQEAITRDLNTMDDYY
Motif 18 DSRSEGLN
Motif g WWKVYYYMTDELLKKERQDY
Motif 16 ACPRQLSELSH
Motif f HFYTFVQHGSCSRWEFVQFIN
Motif e LPFGCTEWDLSKQLTEWQI
Motif 3 DASVYGSEDASVYGS
Motif 9 TPDKGK
Motif 15 CVFCFFVLLNVVLLN
Motif 4 TEEATE
Motif c YIIWYYVHNYLRRKY
Motif b YVRMWNKVIKRNEKKWCDIFAQRIRKYK
Motif 19 SMFDVRPESMPDVHC
RAD family
PHIST family

Supplementary Tables

Supplementary Table 1. *P. vivax* Salvador I genome sequence statistics. The mitochondrial genome of the *P. vivax* Salvador I isolate has previously been sequenced and published (GenBank Acc. No. AF055587). Approximately 15% of the estimated 35 kb apicoplast genome of Salvador I was sequenced during the genome project.

Data class	Value
Nuclear genome	
Total no. contigs	2,775
Total length contigs (bp)	26,835,503
Total no. contigs assigned to a chromosome	30
Total size contigs assigned to a chromosome	22,573,186
Mean G+C content contigs assigned to a chromosome (%)	45.0
Total no. small contigs	2,745
Total size small contigs	4,262,317
Mean G+C content small contigs (%)	27.0
Number complete telomeric chromosome ends	3
Mitochondrial genome	
Number mitochondrial contigs	1
Total length mitochondrial contig (bp)	6,850
Mean G+C (%)	30.6
Apicoplast genome	
Number apicoplast contigs	2
Total length apicoplast contigs (bp)	5,064
Mean G+C (%)	17.1

Supplementary Table 2. The proportion of sequences in 3,865 *P. vivax* and *P. falciparum* orthologs that contain low complexity regions (LCRs) as determined using the SEG algorithm. Recent studies have suggested that LCR abundance depends partly upon genomic A+T content [63], in accordance with our findings that the more G+C-rich *P. vivax* genome contains less LCRs, but how this relates to differences in immune evasion mechanisms between *P. vivax* and *P. falciparum* is unclear.

A. Results showing the proportion of LCRs in the gaps (non-aligned regions, *i.e.,* present in one species but not the other) and conserved (aligned) regions of orthologs for the two species. A total of 520 *P. vivax* proteins contain gaps (*i.e.,* the sequence is absent in *P. falciparum*), and 44.6% of these contain LCRs. *P. falciparum* has a greater number of proteins with gaps (*i.e.,* the sequence is absent in *P. vivax*; 33% of total orthologs), and 49.3% of these gaps contain LCRs. The results suggest that the size difference between pairs of orthologous genes may in part be due to amplification of LCRs.

	P. vivax	***P. falciparum***
No. orthologous protein pairs	3,865	3,865
Gaps >45 amino acids (% of orthologs)	520 (13%)	1296 (33%)
Gaps containing LCRs (% of gaps)	232 (44.6%)	643 (49.3%)
Conserved regions containing LCRs (% of orthologs)	2,826 (73%)	3,272 (84%)

B. Further details regarding the aligned and non-aligned regions between *P. vivax* and *P. falciparum*. The results show that *P. falciparum* has both more unique (gap) sequences in orthologs than *P. vivax* and that these contain more LCRs, and that the species also contains conserved regions that contain a higher proportion of LCRs. aa: amino acid.

	Total length of gaps with LCRs (aa)	**Total length of gaps (aa)**	**%**	**Total length of conserved regions with LCRs (aa)**	**Total length of conserved regions (aa)**	**%**	**Percent of proteins with no LCRs**
P. vivax	29,616	186,449	15.9	1,043,034	2,867,450	36.4	43.7%
P. falciparum	115,647	341,797	33.8	1,980,963	2,861,034	69.2	48.9%

Supplementary Table 3. Transcription-associated protein (TAP) analysis in *Plasmodium*.

A. Comparative TAP analysis in three *Plasmodium* species. A total of 102 protein family clusters were identified across the eukaryotic tree of life that contain *Plasmodium* sequences. Some 138, 146 and 137 TAP homologs in *P. knowlesi*, *P. falciparum* and *P. vivax* respectively were identified in these clusters. Over 90% of the TAP families (total 93) contain sequences originating from all three species, with the vast majority of clusters containing equivalent proportions of sequences from the three species. No clusters were observed which contained sequences only in *P. knowlesi* or *P. vivax*. The majority of the TAP homologs are present in families containing sequences from the other *Plasmodium* species (*P. knowlesi*: 132/138; *P. falciparum*: 135/146; *P. vivax*: 135/137), strongly suggesting that transcriptional components are closely related in all three.

Species	***P. knowlesi***	***P. vivax***	***P. falciparum***	**No. families**	**%**
P. falciparum	-	-	4	2	2.0
Pf:Pk	5	-	6	5	4.9
Pf:Pv	-	1	1	1	1.0
Pk:Pv	1	1	-	1	1.0
Pf:Pv:Pk	132	135	135	93	91.2
Total	**138**	**137**	**146**	**102**	

B. Comparison of TA-HMMs in three *Plasmodium* species. 158 HMMs of domains present in well-characterised transcription factors (TA-HMMs) were used to query the three *Plasmodium* proteomes; only 19 of the TA-HMMs matched sequences in these genomes, corresponding to 57 genes in *P. knowlesi*, 58 genes in *P. falciparum* and 56 genes in *P. vivax*. Both in absolute counts and after normalising for genome size, each of the 19 TA-HMMs recognised very similar numbers of proteins. A high abundance of genes encoding proteins possessing the CCCH-zinc finger (PF00642) are observed in all three genomes (16 copies in *P. knowlesi* and *P. falciparum*, 15 copies in *P. vivax*) implying that post-transcriptional mechanisms may dominate over gene activation. Pfam ID: Pfam identifier; P_k / P_v / P_f : number of sequences observed in the *P. knowlesi*, *P. vivax* and *P. falciparum* genomes matching the HMM, normalised by genome size and expressed as number per 10,000 genes; the number of genes is shown in parentheses; Av.±sd: average of the normalised counts with a standard deviation; HMM description: Pfam description of the hidden-Markov model.

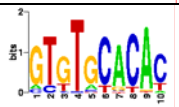
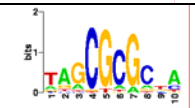
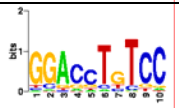
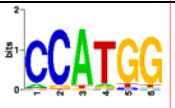
Pfam ID	P_k	P_v	P_f	Av.±sd	HMM description
PF00642	32.0 (16)	27.6 (15)	29.2 (16)	29.6±2.2	Zinc finger, C-x8-C-x5-C-x3-H type (and similar)
PF00249	14.0 (7)	12.9 (7)	12.8 (7)	13.2±0.7	Myb-like DNA-binding domain
PF00439	12.0 (6)	11.0 (6)	12.8 (7)	11.9±0.9	Bromodomain
PF00096	6.0 (3)	9.2 (5)	9.1 (5)	8.1±1.8	Zinc finger, C2H2 type
PF00505	6.0 (3)	7.4 (4)	7.3 (4)	6.9±0.8	HMG (high mobility group) box
PF00808	6.0 (3)	7.4 (4)	5.5 (3)	6.3±1.0	Histone-like transcription factor (CBF/NF-Y) and archaeal histone
PF00352	4.0 (2)	3.7 (2)	3.7 (2)	3.8±0.2	Transcription factor TFIID (or TATA-binding protein, TBP)
PF00382	4.0 (2)	3.7 (2)	3.7 (2)	3.8±0.2	Transcription factor TFIIB repeat
PF02146	4.0 (2)	3.7 (2)	3.7 (2)	3.8±0.2	Sir2 family
PF04054	4.0 (2)	3.7 (2)	3.7 (2)	3.8±0.2	CCR4-Not complex component, Not1
PF02178	6.0 (3)	1.8 (1)	1.8 (1)	3.2±2.4	AT hook motif
PF01167	2.0 (1)	1.8 (1)	1.8 (1)	1.9±0.1	Tub family
PF01381	2.0 (1)	1.8 (1)	1.8 (1)	1.9±0.1	Helix-turn-helix
PF01388	2.0 (1)	1.8 (1)	1.8 (1)	1.9±0.1	ARID/BRIGHT DNA binding domain
PF01754	2.0 (1)	1.8 (1)	1.8 (1)	1.9±0.1	A20-like zinc finger
PF02891	2.0 (1)	1.8 (1)	1.8 (1)	1.9±0.1	MIZ zinc finger
PF03343	2.0 (1)	1.8 (1)	1.8 (1)	1.9±0.1	SART-1 family
PF05669	2.0 (1)	1.8 (1)	1.8 (1)	1.9±0.1	SOH1
PF00313	2.0 (1)	–	1.8 (1)	1.9±0.1	'Cold-shock' DNA-binding domain

Supplementary Table 4. Transcription factor binding site analysis in *Plasmodium*.

A. Consensus sequences of putative conserved transcription factor binding sites. The table includes sites which are over-represented upstream of genes in a particular cluster and conserved across at least two *Plasmodium* species. Regions of similarity between the species are highlighted in gray. All sites have a Weeder score of 20 or more except for the *P. y. yoelii* sporozoite-associated site which has a score of 18.96. For each predicted binding site, the number of genes within the cluster with the predicted site is given followed by the total number of genes in the co-regulated cluster. Clusters identified in previous studies include: sporozoite cluster [22]; merozoite invasion cluster [21]; de-phosphorylation, phosphorylation, heterophilic cell adhesion and glucose catabolism clusters [24]. NI: not identified.

Cluster	*P. vivax*	No.	*P. falciparum*	No.	*P. y. yoelii*	No.	*P. knowlesi*	No.
Sporozoite	5'-TGCATGCA-3'	15/15	5'-TGCATGCA-3'	20/23	5'-TGCATGCA-3'	10/12	5'-TGCATGCA-3'	16/16
Merozoite invasion	5'-GTGTGCACAC-3'	35/35	5'-GTGCAC-3'	38/49	5'-GTGTACAC-3'	29/38	5'-TGTGCACA-3'	41/41
Merozoite invasion	5'-TAGCGCGCTA-3'	35/35	5'-AGCATGCT-3'	40/49	NI	-	5'-CGAGCGCTCG-3'	32/41
De-phosph-orylation	5'-GCACGCGTGC-3'	39/39	5'-GTGCGTGC-3'	32/43	5'-CACGTGTG-3'	26/42	5'-GCACGTGC-3'	45/45
Phosph-orylation	5'-GGACCTGTCC-3'	64/64	5'-GACATGTC-3'	57/75	NI	-	NI	-
Hetero-philic cell adhesion	5'-CCATGG-3'	37/37	5'-CCATGG-3'	48/89	NI	-	NI	-
Glucose catabolism	5'-GCGTACGC-3'	55/55	5'-TCGTACGA-3'	46/65	NI	-	NI	-

B. Sequence logos of the seven conserved putative binding sites. These logos are constructed from the *P. vivax* sequence only. The cluster statistics column contains the number of genes in the cluster with at least one PWM hit for the site at a *p-value* cutoff of 1/5000 followed by the total number of genes in the co-regulated cluster (discrepancies between **A** and **B** in the number of genes results from the use of different *p-value*s). The genome statistics column contains the number of genes not in the cluster that have at least one PWM hit of the site in their upstream regions followed by the number of upstream regions that are not in the cluster. In all instances, the ratio of genes with hits in the cluster to cluster size is greater than the ratio of genes with hits not in the cluster to the number of genes not in the cluster. The group specificity is a hypergeometric *p-value* (Bonferroni-corrected) showing how specific a site is to the genes used in its discovery.

Cluster	Logo	Cluster statistics	Genome statistics	Group specificity
Sporozoite		12/15	1912/5394	0.0037
Merozoite invasion		26/35	1885/5374	$1.5*10^{-5}$
Merozoite invasion		22/35	1938/5374	0.0070
De-phosphorylation		22/39	1574/5370	0.0022
Phosphorylation		44/64	1389/5345	0.00020
Heterophilic cell adhesion		12/37	856/5372	0.064
Glucose catabolism		28/55	1321/5354	0.00013

C. Details of genes within the sporozoite co-expressed cluster with the conserved motif shown in **A**. The cluster consists of *P. vivax, P. knowlesi* and *P. y. yoelii* orthologs of *P. falciparum* proteins that have been found to be differentially expressed in sporozoites [22], the stage of the parasite in the mosquito salivary glands. The proteins in the cluster include thrombospondin-related sporozoite protein (TRSP), which also localizes to the micronemes of the merozoite stage in the erythrocytic cycle; two proteins which are suspected to be involved in sporozoite-hepatocyte and/or sporozoite-salivary gland interactions in *P. falciparum*; and two early transcribed membrane proteins (ETRAMPs). ETRAMPs are localized to the parasitophorous vacuole membrane surrounding the intracellular parasite and have been detected in both pre-erythrocytic and erythrocytic parasites. Five *P. falciparum* and *P. y. yoelii* genes in red font do not contain the motif identified in **A**.

***P. vivax* genes**	***P. falciparum* genes/orthologs**	***P. knowlesi* genes/orthologs**	***P. y. yoelii* genes/orthologs**	***Plasmodium* gene name/function**	**Reference**
PVX_001000	PFD0235c	PK4_0310c	PY01067	hypothetical protein	
PVX_001020	PFD0215c	PK4_0260c	PY01341	pbs36-related protein	
PVX_000815	PFD0425w	PK4_0640w	PY00455	S5, sporozoite-hepatocyte or sporozoite-salivary gland interactions, signal peptide	64
PVX_002775	PFB0690w	PK2_1400w	PY00196	hypothetical protein	
PVX_003790	PFB0325c	PK2_0590c	PY02063	cysteine protease, putative	
PVX_080300	PFE0950c	-	-	hypothetical protein	
PVX_080665	PFE0565w	PK5_1050w	PY00913	hypothetical protein	
PVX_081560	PFA0200w	PK1_0190w	PY07092	TRSP	64
PVX_082585	PF13_0181	PK13_3370c	-	hypothetical protein	
PVX_083025	MAL13P1.212	PK13_4170c	-	hypothetical protein	
PVX_085335	PF14_0220	PK14_1005w	PY00020		
PVX_118680	PF14_0729	PK2_0060c	-	ETRAMP	65
PVX_123155	PF08_0088	PK8_1170c	PY01796	S23, sporozoite-hepatocyte or sporozoite-salivary gland interactions, signal peptide, transmemebrane domain(s)	64
PVX_121950	PF13_0012	PK13_0070c	PY03011	ETRAMP	
PVX_111390	PF10_0261	PK10_2720c	PY07496	hypothetical protein	
-	PF14_0722	PK14_3487c	PY00150	hypothetical protein	
-	PF10_0393	-	-	rifin	
-	PF10_0401	-	-	rifin	

-	PFI0080w	-	-	stevor	
-	MAL8P1.7	-	-	hypothetical protein	
-	PFC1100w	-	-	rifin	
-	PFL0995c	PK12_1860c	-	hypothetical protein	
-	PFD0015c	-	-	rifin	

Supplementary Table 5. Polymorphic microsatellites (MS) identified in the *P. vivax* genome and their nearest genes. Each microsatellite has been assigned a unique number according to the order in which it was first identified in the genome; these numbers are used to indicate the position of each MS on the *P. vivax* chromosomes shown in **Figure 1**. The unique identifier is composed of the MS chromosome position followed by the unique number; identifiers in **bold** were used in population genetic studies [66,67]. The number of alleles identified for each MS is indicated, followed by the number of isolates that were successfully amplified, *e.g.*, 3/8 indicates three alleles identified out of a total of eight *P. vivax* lines that were successfully genotyped. Forward and reverse sequences indicate primers which can be used to amplify regions of ~650 bp around each MS with subsequent electrophoretic separation of PCR products on 2% agarose gels. These primers were designed for studies in the field where access to capillary electrophoresis equipment may not be possible.

The table is appended as a separate file to the supplementary information.

Supplementary Table 6. Comparative analysis of all orthologous genes from six *Plasmodium* species, *P. falciparum* (Pf), *P. vivax* (Pv), *P. knowlesi* (Pk), *P. berghei* (Pb), *P. chabaudi* (Pc), and *P. y. yoelii* (Py), using *P. falciparum* as the reference genome and generated using OrthoMCL. The chromosomal location and contig identifier are indicated for each gene in the six genomes, including archaic gene and contig identifiers for clarity. Colours correspond to those used in **Supplementary Figure 3**. All genes have been assigned into one of the following categories:

Orthologous: 4,450 genes with a clear ortholog in *P. vivax, P. knowlesi* and the rodent malaria parasites
Orthologous n.a.: 16 genes with significant sequence similarity to a syntenic region but for which the ortholog is not annotated
Positionally conserved: 40 genes with little sequence similarity but with a conserved location, direction and syntenic context
Pf-specific, PfPv-specific or PfPvPk-specific: 50, 3 and 59 genes found respectively in the indicated species
Species-specific: genes that have expanded differentially in the six species (*e.g.* the *sera* and *msp7* families)
Intersyntenic: 42 genes located within synteny breakpoints *i.e.* between synteny blocks
Subtelomeric: 706 genes located outside the 36 conserved synteny blocks
Unclear: 90 predominantly small and multi-exon genes for which corresponding sequence is missing in either *P. vivax* and P. *knowlesi*, or the rodent malaria parasites, or both

The table is appended as a separate file to the supplementary information.

Supplementary Table 7. Chromosome and genome averages for d_N, d_S, ω and %G+C$_4$ (percent G+C in third position of four-fold degenerate amino acids) and the amount of the local variation in d_S, of all 14 *P. vivax* chromosomes. Intra-chromosomal variation in d_S is present, as shown by the correlation between the d_S value of adjacent genes. This correlation is particularly strong for chromosome 10, in which genes that are located within the first 400 kb of the centromere have on average a higher d_S than those distal to the centromere. The relationship between the value of d_S and the distance to the centromere mirrors that of G+C content, and is present in all chromosomes. The correlation between d_S and d_N leads to fairly homogeneous ω values, which do not differ statistically among chromosomes.

	n^a	d_N (±std dev)	d_S (±std dev)	ω (±std dev)	%G+C$_4$	$r^{2\,b}$
Chrom 1	113	0.13 (±0.077)	0.85 (±0.480)	0.19 (±0.122)	67.7 (±10.98)	0.15***
Chrom 2	80	0.13 (±0.077)	0.77 (±0.404)	0.19 (±0.122)	68.8 (±11.44)	0.06***
Chrom 3	132	0.13 (±0.077)	0.82 (±0.457)	0.20 (±0.135)	69.5 (±10.26)	0.13***
Chrom 4	126	0.13 (±0.089)	0.76 (±0.432)	0.21 (±0.159)	68.1 (±11.24)	0.08***
Chrom 5	188	0.11 (±0.079)	0.68 (±0.378)	0.20 (±0.151)	65.2 (±11.04)	0.07***
Chrom 6	125	0.13 (±0.117)	0.78 (±0.430)	0.19 (±0.145)	67.8 (±9.65)	0.06***
Chrom 7	213	0.11 (±0.068)	0.71 (±0.432)	0.19 (±0.142)	66.5 (±9.77)	0.06***
Chrom 8	263	0.10 (±0.068)	0.66 (±0.363)	0.18 (±0.140)	65.1 (±10.91)	0.11***
Chrom 9	301	0.10 (±0.074)	0.67 (±0.381)	0.18 (±0.130)	65.4 (±9.14)	0.11***
Chrom 10	219	0.10 (±0.072)	0.64 (±0.403)	0.18 (±0.127)	62.8 (±11.85)	0.24***
Chrom 11	319	0.10 (±0.069)	0.61 (±0.375)	0.19 (±0.141)	64.2 (±9.65)	0.07***
Chrom 12	472	0.10 (±0.072)	0.57 (±0.299)	0.19 (±0.127)	62.5 (±9.84)	0.13***
Chrom 13	312	0.10 (±0.065)	0.61 (±0.338)	0.19 (±0.138)	64.2 (±10.59)	0.13***
Chrom 14	460	0.10 (±0.062)	0.56 (±0.313)	0.19 (±0.124)	61.3 (±10.11)	0.17***
Genome	**3322**	**0.11(±0.074)**	**0.65 (±0.381)**	**0.19 (±0.135)**	**64.6 (±10.54)**	

[a] Number of high quality orthologous gene pairs.
[b] Proportion of the variance in d_S explained by local variation: square of the correlation coefficient of the relationship between d_S in adjacent genes X and X-1, *i.e.,* d_S (X) *vs.* d_S (X-1). Significance level determined with a randomization test, with 5000 replicates (not significant: $P > 0.05$; *: $0.01 < P \leq 0.05$; **: $0.001 < P \leq 0.01$; ***: $P \leq 0.001$).

Supplementary Table 8. Average d_N/d_S between *P. vivax* and *P. knowlesi*, for protein-coding genes grouped according to three GO classifications and to the presence of specific motifs. The breakdown according to cellular component revealed a high d_N/d_S ratio for genes targeted to the apicoplast relative to genes with other known cellular locations. Since a hallmark of the amino terminal transit peptide found in bipartite apicoplast targeting motifs is variability of primary sequence [68], we analyzed the 226 apicoplast-targeted genes to identify which regions of the genes appeared to be evolving the most rapidly (Supplementary Methods). Substitutions were found to be disproportionately located at the 5' end of the genes, indicating that the bipartite leader is evolving fast in this group of proteins rather than the apicoplast-targeted proteins.

Classification	*n*	Mean	Std dev
Biological Process			
Metabolism			
carbohydrate	41	0.08	0.060
nucleic acid and precursors	271	0.13	0.090
lipid	32	0.17	0.101
protein	420	0.11	0.104
primary	21	0.13	0.067
Cell cycle	21	0.10	0.081
Cell organization and biogenesis	32	0.12	0.086
Cell invasion	2	0.11	0.049
Cell adhesion	16	0.21	0.075
Signal transduction	41	0.12	0.086
Transport	207	0.10	0.095
Response to stress	24	0.09	0.071
Physiological process	116	0.13	0.102
Other	2316	0.22	0.137
Cellular Component			
Cytoplasm	55	0.09	0.100
Nucleus	143	0.10	0.108
Mitochondrion	99	0.09	0.070
Apicoplast	226	0.21	0.115
Endoplasmic reticulum	10	0.08	0.068
Golgi apparatus	8	0.06	0.045
Ribosome	103	0.07	0.075
Cytoskeleton	20	0.11	0.086
Plasma membrane	13	0.12	0.063
Organelle (other)	6	0.09	0.070
Intracellular (other)	127	0.09	0.089
Other	2883	0.20	0.135
Molecular Function			
Motor	19	0.14	0.091
Catalytic	694	0.14	0.100
Structural molecule	117	0.07	0.078
Transporter	94	0.09	0.074
Binding	702	0.14	0.113

Antioxidant	5	0.08	0.031
Enzyme regulator	11	0.08	0.059
Transcription regulator	27	0.11	0.080
Translation regulator	36	0.10	0.084
Other	2169	0.22	0.138
Motif-specific classification			
Trans-membrane (TMM)	578	0.20	0.143
Signal peptide (SP)	363	0.23	0.148
TMM and SP	147	0.26	0.171
GPI-anchored	18	0.28	0.129
Exportome	37	0.39	0.186
All high quality COGs	3322	0.19	0.135

Supplementary Table 9. Comparative transporter analysis between *P. falciparum*, *P. vivax* and select other parasitic protists and metazoans. Each box indicates the number of transporters, and is color coded from orange (1-10 proteins), to yellow (11-40 proteins), to green (51-150 proteins) according to their abundance.

	P. fal 3D7	*P. vivax Sal I*	*T. parva*	*C. parvum genotype 2*	*E. histolytica HM1:IMSS*	*S. cerevisiae S288C*	*D. discoideum AX4*	*A. thaliana*	*C. elegans*	*D. melano-gaster*	*H. sapiens*
Genome Size (Mb):	23	23	8.4	9.1	23.8	13	34	125	97	120	3150
Total Transporter Proteins:	85	87	73	74	98	303	247	922	655	599	762
ATP-Dependent	**30 (35%)**	**28 (32%)**	**28 (38%)**	**23 (31%)**	**39 (40%)**	**43 (14%)**	**89 (36%)**	**162 (18%)**	**72 (11%)**	**72 (12%)**	**82 (11%)**
The ATP-binding Cassette (ABC) Superfamily	14	13	17	13	18	24	61	108	48	51	47
The Arsenite-Antimonite (ArsAB) Efflux Family	1	1	1	0	1	1	2	0	0	1	1
The H+- or Na+-translocating F-type, V-type and A-type ATPase (F-ATPase) Superfamily	2	2	2	1	1	2	2	5	2	1	2
The H+-translocating Pyrophosphatase (H+-PPase) Family	2	2	1	0	0	0	0	3	0	0	0
The P-type ATPase (P-ATPase) Superfamily	11	10	7	9	19	16	24	46	22	19	32
Ion Channels	**6 (7%)**	**7 (8%)**	**4 (5%)**	**5 (7%)**	**10 (10%)**	**24 (8%)**	**22 (9%)**	**128 (14%)**	**229 (35%)**	**158 (26%)**	**326 (43%)**
ATP-gated Cation Channel (ACC) Family	0	0	0	0	0	0	0	0	0	0	7
The Bcl-2 (Bcl-2) Family	0	0	0	0	0	0	0	0	0	1	12
The Anion Channel-forming Bestrophin (Bestrophin) Family	0	0	0	0	0	0	0	0	21	4	4
The CD20 Ca2+ Channel (CD20) Family	0	0	0	0	0	0	0	0	0	0	9
The Chloride Channel (ClC) Family	0	0	0	0	3	1	7	7	6	3	10
The Gap Junction-forming Connexin (Connexin) Family	0	0	0	0	0	0	0	0	0	0	18
The gp91phox Phagocyte NADPH Oxidase-associated Cytochrome b558 (CytB) H+-channel Family	0	0	0	0	0	9	3	18	0	10	13
The Epithelial Chloride Channel (E-ClC) Family	0	0	0	0	0	0	0	0	0	0	4
The Epithelial Na+ Channel (ENaC) Family	0	0	0	0	0	0	0	0	20	25	8
The Glutamate-gated Ion Channel (GIC) Family of Neurotransmitter Receptors	0	0	0	0	0	0	1	19	9	27	20
The Intracellular Chloride Channel (ICC) Family	0	0	0	0	0	0	0	0	0	0	1
The Gap Junction-forming Innexin (Innexin) Family	0	0	0	0	0	0	0	0	22	8	0
The Animal Inward Rectifier K+ Channel (IRK-C) Family	0	0	0	0	0	0	0	0	1	3	22

The Ligand-gated Ion Channel (LIC) Family of Neurotransmitter Receptors	0	0	0	0	0	0	0	0	69	23	45
The Yeast Stretch-Activated, Cation-Selective, Ca2+ Channel, Mid1 (Mid1) Family	0	0	0	0	0	1	0	0	0	0	0
The Major Intrinsic Protein (MIP) Family	1	1	0	0	0	4	4	38	7	7	11
The CorA Metal Ion Transporter (MIT) Family	2	2	0	0	5	5	3	12	0	0	1
The Large Conductance Mechanosensitive Ion Channel (MscL) Family	0	0	0	0	0	0	0	0	0	0	0
The Small Conductance Mechanosensitive Ion Channel (MscS) Family	1	1	2	2	1	0	1	8	0	0	0
The Non-selective Cation Channel-2 (NSCC2) Family	1	1	1	1	0	1	0	1	1	1	1
The Organellar Chloride Channel (O-ClC) Family	0	0	0	0	0	0	0	0	0	1	6
The Polycystin Cation Channel (PCC) Family	0	0	0	0	0	0	0	0	0	5	6
The Phospholemman (PLM) Family	0	0	0	0	0	0	0	0	0	0	7
The Ryanodine-Inositol 1,4,5-triphosphate Receptor Ca2+ Channel (RIR-CaC) Family	0	0	0	0	0	0	0	0	5	3	6
The Chloroplast Envelope Anion Channel-forming Tic110 (Tic110) Family	0	0	0	0	0	0	0	1	0	0	0
The Transient Receptor Potential Ca2+ Channel (TRP-CC) Family	0	0	0	0	0	1	0	0	5	7	23
The Urea Transporter (UT) Family	0	0	0	0	0	0	0	0	0	0	2
The Voltage-gated Ion Channel (VIC) Superfamily	1	2	1	2	1	2	3	35	63	31	90
Secondary Transporter	**49 (58%)**	**52 (60%)**	**41 (56%)**	**42 (57%)**	**49 (50%)**	**227 (75%)**	**134 (54%)**	**636 (69%)**	**349 (53%)**	**359 (60%)**	**337 (44%)**
The ATP:ADP Antiporter (AAA) Family	0	0	0	0	0	0	0	2	0	0	0
The Amino Acid/Auxin Permease (AAAP) Family	2	1	0	7	7	7	1	43	11	15	13
The Arsenical Resistance-3 (ACR3) Family	0	0	0	0	0	1	0	0	0	0	0
The Anion Exchanger (AE) Family	0	0	0	0	0	1	1	7	4	2	10
The Auxin Efflux Carrier (AEC) Family	0	0	0	0	0	4	0	8	0	0	0
The Alanine or Glycine:Cation Symporter (AGCS) Family	0	0	0	0	0	0	0	0	0	0	0
The Ammonium Transporter (Amt) Family	0	0	0	0	0	3	5	6	6	2	4
The Amino Acid-Polyamine-Organocation (APC) Family	0	0	0	0	3	24	8	12	11	11	14
The Aromatic Acid Exporter (ArAE) Family	0	0	0	0	0	0	0	11	0	0	0
The Arsenite-Antimonite (ArsB) Efflux Family	0	0	0	0	0	0	0	0	0	5	1
The Bile Acid:Na+ Symporter (BASS) Family	0	0	0	0	0	1	0	5	0	2	5

The Betaine/Carnitine/Choline Transporter (BCCT) Family	0	0	0	0	0	0	0	0	0	0	0
The Ca2+:Cation Antiporter (CaCA) Family	1	1	0	2	2	4	1	12	8	11	8
The Cation-Chloride Cotransporter (CCC) Family	0	0	0	0	0	1	0	1	6	5	9
The Cation Diffusion Facilitator (CDF) Family	1	1	0	1	6	5	5	8	8	7	10
The Chromate Ion Transporter (CHR) Family	0	0	0	0	0	0	0	0	0	0	0
The Concentrative Nucleoside Transporter (CNT) Family	0	0	0	0	0	0	0	0	2	2	3
The Monovalent Cation:Proton Antiporter-1 (CPA1) Family	1	1	1	0	2	2	2	8	11	5	3
The Monovalent Cation:Proton Antiporter-2 (CPA2) Family	0	0	0	1	0	1	2	32	0	0	1
The Dicarboxylate/Amino Acid:Cation (Na+ or H+) Symporter (DAACS) Family	0	0	1	0	0	0	0	0	6	2	7
The Divalent Anion:Na+ Symporter (DASS) Family	0	0	0	0	5	3	0	4	4	3	5
The Drug/Metabolite Transporter (DMT) Superfamily	6	8	4	11	3	9	12	121	15	14	18
The Equilibrative Nucleoside Transporter (ENT) Family	2	2	0	1	2	1	3	8	5	3	4
The Folate-Biopterin Transporter (FBT) Family	2	2	3	1	0	0	0	7	0	0	0
The Formate-Nitrite Transporter (FNT) Family	1	1	1	0	1	1	0	0	0	0	0
The Glycoside-Pentoside-Hexuronide (GPH):Cation Symporter Family	1	0	0	0	4	0	3	9	1	1	5
The Glycerol Uptake (GUP) Family	0	0	0	0	0	2	1	0	0	0	0
The Hydroxy/Aromatic Amino Acid Permease (HAAAP) Family	0	0	0	0	0	0	0	1	0	0	0
The K+ Uptake Permease (KUP) Family	0	0	0	0	0	0	1	13	0	0	0
The Lysosomal Cystine Transporter (LCT) Family	0	0	0	1	1	1	2	0	0	2	2
The Lactate Permease (LctP) Family	0	0	0	0	0	0	0	0	0	0	0
The Mitochondrial Carrier (MC) Family	10	11	9	5	1	34	32	52	34	45	44
The 4 TMS Multidrug Endosomal Transporter (MET) Family	0	0	0	0	0	0	0	0	0	0	3
The Chloroplast Maltose Exporter (MEX) Family	0	0	0	0	0	0	0	1	0	0	0
The Major Facilitator Superfamily (MFS)	15	16	16	8	4	85	27	92	137	144	82
The Multidrug/Oligosaccharidyl-lipid/Polysaccharide (MOP) Flippase Superfamily	0	0	0	0	0	3	2	56	0	0	2
The Malonate:Na+ Symporter (MSS) Family	0	0	0	0	0	0	0	0	0	0	0

The Mitochondrial Tricarboxylate Carrier (MTC) Family	0	0	0	0	0	1	1	0	6	2	3
The Nucleobase:Cation Symporter-1 (NCS1) Family	0	0	0	0	0	10	0	1	0	0	0
The Nucleobase:Cation Symporter-2 (NCS2) Family	0	0	0	0	0	0	2	12	5	1	4
The NhaD Na+:H+ Antiporter (NhaD) Family	0	0	0	0	0	0	0	2	0	0	0
The Ni2+-Co2+ Transporter (NiCoT) Family	0	0	0	0	0	0	0	2	0	0	0
The Metal Ion (Mn2+-iron) Transporter (Nramp) Family	1	1	0	0	0	3	2	7	2	1	2
The Neurotransmitter:Sodium Symporter (NSS) Family	1	2	1	0	0	0	0	0	12	21	18
The Organo Anion Transporter (OAT) Family	0	0	0	0	0	0	0	0	3	8	11
The Oligopeptide Transporter (OPT) Family	0	0	0	0	0	3	1	15	0	0	0
The Organic Solute Transporter (OST) Family	0	0	0	0	0	0	0	0	0	0	2
The Cytochrome Oxidase Biogenesis (Oxa1) Family	1	1	1	0	0	1	2	4	0	1	1
The Inorganic Phosphate Transporter (PiT) Family	1	1	1	0	0	1	0	1	5	1	2
The Phosphate:Na+ Symporter (PNaS) Family	0	0	0	0	0	0	0	0	1	0	2
The Proton-dependent Oligopeptide Transporter (POT) Family	0	0	0	0	0	1	3	49	3	3	4
The Reduced Folate Carrier (RFC) Family	0	0	0	0	0	0	0	0	3	3	4
The Resistance to Homoserine/Threonine (RhtB) Family	0	0	0	0	0	0	0	0	0	0	0
The Resistance-Nodulation-Cell Division (RND) Superfamily	1	1	0	2	4	1	5	2	24	4	7
The Na+-dependent Bicarbonate Transporter (SBT) Family	0	0	0	0	0	0	0	0	0	0	0
The Silicon Transporter (Sit) Family	0	0	0	0	0	0	0	0	0	0	0
The Solute:Sodium Symporter (SSS) Family	0	0	0	0	0	1	0	1	3	19	11
The Sulfate Permease (SulP) Family	1	1	1	0	0	4	4	11	7	9	11
The Twin Arginine Targeting (Tat) Family	0	0	0	0	0	0	0	3	0	0	0
The Telurite-resistance/Dicarboxylate Transporter (TDT) Family	0	0	0	0	0	1	0	4	0	0	0
The Threonine/Serine Exporter (ThrE) Family	0	0	0	0	0	2	0	0	0	0	0
The K+ Transporter (Trk) Family	0	0	0	0	2	2	1	1	0	0	0
The Zinc (Zn2+)-Iron (Fe2+) Permease (ZIP) Family	1	1	2	2	2	3	5	13	6	5	2
Unclassified	**0 (0%)**	**0 (0%)**	**0 (0%)**	**4 (5%)**	**0 (0%)**	**6 (2%)**	**3 (1%)**	**4 (0%)**	**4 (1%)**	**9 (2%)**	**14 (2%)**
The Copper Transporter-1 (Ctr1) Family	0	0	0	0	0	1	0	0	0	0	0

The Copper Transporter-2 (Ctr2) Family	0	0	0	0	0	2	0	4	4	3	2
The Low Affinity Fe2+ Transporter (FeT) Family	0	0	0	0	0	1	0	0	0	0	0
The Ferroportin (FP) Family	0	0	0	0	0	0	0	0	0	0	1
The Lysosomal Protein Import (LPI) Family	0	0	0	0	0	0	0	0	0	0	5
The Mg2+ Transporter-E (MgtE) Family	0	0	0	0	0	0	0	0	0	0	0
The Oxidase-dependent Fe2+ Transporter (OFeT) Family	0	0	0	0	0	2	0	0	0	0	0
The Peroxisomal Protein Importer (PPI) Family	0	0	0	0	0	0	3	0	0	6	6
The Peptide Uptake Permease (PUP) Family	0	0	0	4	0	0	0	0	0	0	0

Supplementary Table 10. Predicted apicoplast-targeted proteins in the *P. vivax* genome. Although earlier *in silico* reconstructions of *P. falciparum* apicoplast metabolism are broadly consistent with the present analysis, the number of hypothetical proteins predicted to target the apicoplast is substantially lower in this analysis than the initial estimate (159 in *P. vivax* compared to ~340 in *P. falciparum*). This is in part due to assignation of function to some of the former hypotheticals (*e.g.*, ribosomal proteins, chaperones, iron sulfur cluster assembly proteins, and RNA modifying enzymes), and also because of improved ability to discern false positives. Recent completed genome projects for related plastid-bearing algae (*Cyanidioschyzon merolae* [69] and *Thallasiosira pseudonanna* [70]) as well as apicomplexans with and without apicoplasts (*Theileria parva* and *T. annulata* [71] and, *Cryptosporidium parvum* [72] respectively) now provide ortholog data that allow refinement and improvement of prediction of apicoplast targeting. In addition, the elucidation of exportome motifs [73] has helped substantially to discern genuine apicoplast proteins from other proteins trafficked via the endomembrane system. Of particular interest is the confirmation of a glyoxalase pathway in the apicoplast, with glyoxalase I and glyoxalase II enzymes being targeted there. The enzymes are probably required for the detoxification of methylglyoxal, a spontaneous breakdown product of the dihydroxyacetone phosphate that must be imported into the apicoplast for the synthesis of isoprenoids. Since glyoxalases require glutathione as a coenzyme, the apicoplast likely synthesises or imports glutathione and its essential accessory enzymes.

The table is appended as a separate file to the supplementary information.

Supplementary Table 11. A list of the ~150 hypothetical genes that appear to be unique in *P. vivax*.

Identifier	Product name
PVX_000000	hypothetical protein
PVX_000010	hypothetical protein
PVX_000025	hypothetical protein
PVX_000740	hypothetical protein
PVX_001105	hypothetical protein
PVX_001110	hypothetical protein
PVX_001620	hypothetical protein
PVX_001630	hypothetical protein
PVX_001640	hypothetical protein
PVX_001655	hypothetical protein
PVX_001660	hypothetical protein
PVX_001783	hypothetical protein
PVX_002512	hypothetical protein
PVX_002560	hypothetical protein
PVX_003525	hypothetical protein
PVX_003530	hypothetical protein
PVX_003710	hypothetical protein
PVX_003875	hypothetical protein
PVX_004505	hypothetical protein
PVX_004537	hypothetical protein
PVX_005040	hypothetical protein
PVX_005055	hypothetical protein
PVX_005065	hypothetical protein
PVX_005565	hypothetical protein
PVX_005570	hypothetical protein
PVX_005575	hypothetical protein
PVX_020170	hypothetical protein
PVX_027190	hypothetical protein
PVX_034690	hypothetical protein
PVX_035690	hypothetical protein
PVX_036690	hypothetical protein
PVX_037690	hypothetical protein
PVX_038190	hypothetical protein
PVX_060190	hypothetical protein
PVX_061690	hypothetical protein
PVX_065690	hypothetical protein
PVX_074190	hypothetical protein
PVX_081730	hypothetical protein
PVX_081845	hypothetical protein
PVX_082870	hypothetical protein
PVX_082930	hypothetical protein
PVX_084320	hypothetical protein
PVX_084405	hypothetical protein
PVX_084530	hypothetical protein
PVX_084860	hypothetical protein
PVX_086135	hypothetical protein
PVX_086865	hypothetical protein
PVX_086870	hypothetical protein
PVX_086885	hypothetical protein
PVX_086905	hypothetical protein
PVX_087665	hypothetical protein
PVX_087890	hypothetical protein
PVX_088795	hypothetical protein
PVX_088845	hypothetical protein
PVX_089005	hypothetical protein
PVX_089440	hypothetical protein
PVX_089545	hypothetical protein
PVX_089770	hypothetical protein
PVX_089780	hypothetical protein
PVX_090235	hypothetical protein
PVX_090295	hypothetical protein
PVX_090320	hypothetical protein
PVX_090835	hypothetical protein
PVX_091520	hypothetical protein
PVX_092475	hypothetical protein
PVX_092665	hypothetical protein
PVX_093700	hypothetical protein
PVX_093730	hypothetical protein
PVX_094230	hypothetical protein
PVX_094235	hypothetical protein
PVX_094250	hypothetical protein
PVX_094260	hypothetical protein
PVX_094275	hypothetical protein
PVX_094960	hypothetical protein
PVX_096010	hypothetical protein
PVX_096030	hypothetical protein
PVX_096035	hypothetical protein
PVX_096040	hypothetical protein
PVX_096045	hypothetical protein
PVX_096050	hypothetical protein
PVX_096055	hypothetical protein
PVX_096965	hypothetical protein
PVX_097010	hypothetical protein
PVX_097542	hypothetical protein
PVX_097550	hypothetical protein
PVX_097565	hypothetical protein
PVX_097580	hypothetical protein
PVX_097730	hypothetical protein
PVX_098005	hypothetical protein
PVX_098030	hypothetical protein
PVX_099170	hypothetical protein
PVX_099610	hypothetical protein

PVX_099730	hypothetical protein
PVX_099755	hypothetical protein
PVX_099900	hypothetical protein
PVX_100010	hypothetical protein
PVX_100620	hypothetical protein
PVX_101375	hypothetical protein
PVX_101380	hypothetical protein
PVX_101500	hypothetical protein
PVX_101530	hypothetical protein
PVX_101545	hypothetical protein
PVX_101595	hypothetical protein
PVX_101605	hypothetical protein
PVX_102130	hypothetical protein
PVX_102645	hypothetical protein
PVX_105705	hypothetical protein
PVX_106215	hypothetical protein
PVX_106730	hypothetical protein
PVX_107230	hypothetical protein
PVX_107735	hypothetical protein
PVX_107740	hypothetical protein
PVX_108255	hypothetical protein
PVX_108765	hypothetical protein
PVX_110815	hypothetical protein
PVX_110830	hypothetical protein
PVX_110845	hypothetical protein
PVX_110875	hypothetical protein
PVX_111505	hypothetical protein
PVX_112100	hypothetical protein
PVX_112635	hypothetical protein
PVX_112640	hypothetical protein
PVX_112650	hypothetical protein
PVX_113960	hypothetical protein
PVX_114165	hypothetical protein
PVX_114765	hypothetical protein
PVX_115990	hypothetical protein
PVX_117070	hypothetical protein
PVX_117710	hypothetical protein
PVX_117720	hypothetical protein
PVX_118075	hypothetical protein
PVX_118215	hypothetical protein
PVX_121350	hypothetical protein
PVX_121860	hypothetical protein
PVX_121875	hypothetical protein
PVX_121895	hypothetical protein
PVX_121905	hypothetical protein
PVX_121940	hypothetical protein
PVX_121955	hypothetical protein
PVX_122925	hypothetical protein
PVX_123015	hypothetical protein
PVX_124700	hypothetical protein
PVX_124725	hypothetical protein
PVX_125726	hypothetical protein
PVX_125735	hypothetical protein
PVX_183275	hypothetical protein
PVX_231290	hypothetical protein

Supplementary Table 12. Search for genetic determinants of the hypnozoite stage in *P. vivax*. Sequences associated with dormancy phenotypes in a range of organisms were used to search the *P. vivax* predicted proteome.

Genes associated with dormancy in other species	*P. vivax* homologs	Other data
RTOA (Ratio-A) from *Dictyostelium discoideum*, involved in directing cell differentiation to prestalk or prespore development [74].	PVX_096975 Hypothetical protein	Significant hit to PKH_020060 kir-like protein (score 464, *E* value e-131, identity 242/499).
ADP-ribosylation factor (Arfs). 21-kDa proteins of the Ras superfamily of GTP-binding proteins associated with dormancy phenotype in yeast.	PVX_110990 ADP-ribosylation factor-like protein PVX_080550 small GTPase Rab1, putative PVX_080610 small GTPase Rab1A, putative PVX_081430 small GTPase Rab5c, putative PVX_082950 small GTPase rab11b, putative PVX_084835 ABC transporter, putative PVX_088180 small GTPase Rab18, putative PVX_089930 small GTP-binding protein sar1, putative PVX_089010 translation initiation factor IF-2, putative PVX_091320 hypothetical protein, conserved PVX_092850 small GTPase Rab6, putative PVX_091515 GTP-binding nuclear protein Ran, putative PVX_099440 ADP-ribosylation factor-like protein PVX_098605 small GTPase Rab7, putative PVX_101235 hypothetical protein, conserved PVX_113515 translation initiation factor IF-2, putative PVX_117930 hypothetical protein, conserved PVX_118390 ADP-ribosylation-like factor, putative PVX_124195 small GTPase Rab2, putative PVX_122840 small GTPase rab11, putative PVX_122635 ADP-ribosylation factor, putative PVX_122360 small GTPase Rab5b, putative PVX_001905 ADP-ribosylation factor, putative PVX_002970 small GTPase rab5, putative	PFAM domain: PF00025 PFAM domains: PF00025 & PF00071 PF00025 & PF00071 PF00025 & PF00071
Ras Associated with dormancy in yeast [75].	PVX_113975 hypothetical protein, conserved PVX_118390 ADP-ribosylation-like factor, putative PVX_124195 small GTPase Rab2, putative PVX_123540 GTP-binding protein, putative	PFAM domain: PF00071 PF00025 & PF00071

	PVX_122840 small GTPase rab11, putative PVX_122360 small GTPase Rab5b, putative PVX_001905 ADP-ribosylation factor, putative PVX_002970 small GTPase rab5, putative	 PF00025 & PF00071 PF00025 & PF00071
Gcs1 Required for re-entry of yeast cells into the cell cycle after stationary-phase [76].	PVX_079935 Arf GTP-ase activating protein, putative PVX_093515 GTP-ase activating protein, putative PVX_101125 GTP-ase activating protein for Arf protein	PFAM domain: PF01412

Supplementary Table 13. Comparison of invasion genes in *P. falciparum*, *P. vivax* and *P. y. yoelii*. (Note: Experimental verification of gene sequences and structures are necessary since there may be errors in assembly of the subtelomeric ends of *P. vivax* and *P. knowlesi* chromosomes.)

A. Comparison of 10 *P. vivax* RBP genes with six *P. falciparum* Rh/NBP genes, three putative *P. knowlesi* RBP genes and one of the 14 *P. y. yoelii* Py235 genes. * Two genes previously identified from the Belem isolate. ** Fragments of a single RBP gene. **pI**: isoelectric point; ESTs: expressed sequence tags identified from current EST databases indicate gene transcription; NA: not available; ST: subtelomeric; CT: centromeric; NK: not known. Arrows in the orientation column indicate that the genes either point towards the telomere (T) or centromere (C).

Species	Name	Identifier	MW	pI	ESTs	Chr.	Location	Orientation
P. vivax (strain Salvador 1)	PvRBP1 (Belem)*	B42771/Q00798	303,297	5.57	-	7	ST	T <- C
	PvRBP1a	PVX_098585	324,366	4.84	2			
	PvRBP1b	PVX_098582	333,325	5.34	0			
	PvRBP1 (partial)	PVX_125738	90,852	5.03	0	NK	NK	NK
	PvRBP2a	PVX_121920	286,633	6.83	1	14	ST	T -> C
	PvRBP2b	PVX_094255	326,909	5.65	2	8	ST	T -> C
	PvRBP2 (Belem)*	Q00799	331,402	5.44	-	5	ST	T -> C
	PvRBP2c	PVX_090325	326,397	5.37	3			
	PvRBP2 (partial)	PVX_090330	77,367	8.35	0			
	PvRBP2d (pseudogene)	PVX_101585	328,821	5.33	0	14	ST	T -> C
	PvRBP2 (partial)	PVX_101590	74,092	9.11	0			
	PvRBP3 (pseudogene)	PVX_101495	321,167	5.62	1			
P. falciparum (clone 3D7)	PfRH1	PFC0110W	358,502	7.88	many	4	ST	T -> C
	PfRH2a	PF13_0198	370,401	5.29	many	13	CT	-><-
	PfRH2b	MAL13P1.178	382,847	5.01	7			
	PfRH3	PFL2520w	328,916	5.91	2	12	ST	T <- C
	PfRH4	PFD1150c	203,361	6.95	2	4	ST	T <- C
	PfRH5	PFD1145c	62,990	8.56	2			
P. knowlesi (strain H)	PkRBP1	PKH_070003	344,069	6.25	NA	7	ST	T -> C
	PkRBP3a (partial**)	PKH_146970	187,520	5.91	NA	14	ST	T <- C
	PkRBP3a (partial**)	PKH_146980	79,210	4.49	NA			
	PkRBP3b (partial)	PKH_000020	43,800	6.00	NA	NK	NK	NK
P. y. yoelii	Py235_E8	AAB41263	325,607	6.04	many	-	ST	-

B. Percent identity at the protein level for invasion-related genes. Numbers in red font indicate values for percent identity; numbers in blue font indicate values for percent identity and percent similarity. Low sequence similarity among *P. vivax* RBPs (<50%) is consistent with their apparent lack of antigenic cross-reactivity, as observed among the RH/NBP paralogs of *P. falciparum*. In addition to extensive paralog diversity, considerable allelic variation is evident between the *rbp2* genes of different isolates.

Identity Scores (%)

	PvRBP1a PVX098585	PvRBP1 Q00798	PvRBP1b PVX09852	PvRBP1 PVX125738	PvRBP2a PVX121920	PvRBP2b PVX094255	PvRBP2c PVX090325	PvRBP2 PVX090330	PvRBP2 Q00799	PvRBP2d PVX101585	PvRBP2 PVX101590	PvRBP3 PVX101495	PkRBP3b PKH000020	PkRBP3a _PKH146970	PkRBP3b PKH146980	PfRh1_PfNBP1 PFD0110w	PfRh2a PF13_0198	PfRH2b MAL131P1.176
PvRBP1a_PVX098585	**100**	99	29	7	14	17	16	3	16	17	3	16	14	7	5	11	12	12
PvRBP1_Q00798	100	**100**	29	6	14	17	16	3	16	17	3	16	14	7	5	11	12	12
PvRBP1b_PVX09852	48	48	**100**	17	14	16	14	2	14	15	3	15	14	6	3	10	13	13
PvRBP1_PVX125738	12	12	17	**100**	2	4	4	0	4	4	0	5	3	3	17	2	4	4
PvRBP2a_PVX121920	30	30	29	5	**100**	29	26	5	26	26	6	16	22	7	3	8	14	13
PvRBP2b_PVX094255	35	35	33	9	47	**100**	29	4	29	29	5	18	25	8	5	10	16	15
PvRBP2c_PVX090325	34	34	32	9	43	49	**100**	8	83	48	7	16	34	8	5	9	15	15
PvRBP2_PVX090330	6	6	7	0	10	9	13	**100**	8	9	35	4	6	6	0	2	3	3
PvRBP2_Q00799	34	34	31	9	42	48	87	12	**100**	48	8	16	34	8	5	9	15	14
PvRBP2d_PVX101585	34	34	31	9	43	48	65	13	65	**100**	12	17	36	8	5	10	16	16
PvRBP2_PVX101590	7	7	7	0	10	9	12	53	12	16	**100**	4	7	6	0	2	3	3
PvRBP3_PVX101495	33	33	32	10	31	35	34	7	34	34	7	**100**	16	23	15	10	17	17
PkRBP3b_PKH000020	31	31	31	7	40	44	52	10	51	53	11	32	**100**	8	5	10	15	14
PkRBP3a_PKH146970	17	17	15	9	16	17	18	13	17	17	12	31	16	**100**	5	5	8	7
PkRBP3b_PKH146980	9	9	7	35	5	10	10	0	10	10	0	18	9	11	**100**	3	5	5
PfRh1_PfNBP1_PFD0110w	28	28	25	7	23	27	27	6	27	28	6	26	26	14	7	**100**	9	9
PfRh2a_PF13_0198	28	28	28	8	28	32	31	7	30	31	7	33	29	17	9	25	**100**	87
PfRH2b_MAL131P1.176	27	27	26	8	27	31	30	7	29	30	6	32	28	16	9	24	89	**100**
PfRh3_PFL2520w	33	33	31	9	31	35	33	8	34	33	8	34	31	16	9	27	31	30
PfRh4_PFD1150c	16	16	17	8	14	17	17	11	17	17	9	17	17	18	8	17	16	15
PfRh5_PFD1145c	4	4	4	0	5	5	5	19	4	4	18	4	4	7	0	4	4	4

Similarity Scores (%)

Supplementary Table 14. MSP gene families in *Plasmodium* species. The number of loci is indicated for each family, followed by the gene identifier. * Three gene members identified from previous studies [77] have been reclassified. ** Two *P. falciparum* MSP3 and MSP6 genes, in addition to four contiguous *P. falciparum* genes (PF10_0347, PF10_0348, PF10_0352 and PF10_0355) in a locus on chromosome 10, share structural features and motifs linking them as a family that are also related to the *P. vivax* MSP3 family.

Gene family	*P. vivax*	*P. knowlesi*	*P. falciparum*	*P. y. yoelii*
MSP1	1: PVX_099980	1: PKH_072850	1: PF11_0344	1: PY05748
MSP2	0	0	1: PFB0300c	0
MSP3	11: MSP3.1 (γ*): PVX_097670 MSP3.2: PVX_097675 MSP3.3 (β*): PVX_097680 MSP3.4: PVX_097685 MSP3.5: PVX_097690 MSP3.6: PVX_097695 MSP3.7: PVX_097700 MSP3.8: PVX_097705 MSP3.9: PVX_097710 MSP3.10 (α*): PVX_097720 MSP3.11: PVX_097725	2: PKH_103020 PKH_145630	1: PF10_0345**	0
MSP4	1: PVX_003755	1: PKH_041300	1: PFB0310c	1: PY05967
MSP5	1: PVX_003770	1: PKH_041310	1: PFB0305c	0
MSP6	0	0	1: PF10_0346**	0
MSP7	11: PVX_082700 PVX_082695 PVX_082690 PVX_082685 PVX_082680 PVX_082675 PVX_082670 PVX_082665 PVX_082655 PVX_082650 PVX_082645	4: PKH_121820 PKH_121830 PKH_121840 PKH_121860	6: PF13_0197 PF13_0196 MAL13P1.175 (fragment) MAL13P1.174 PF13_0193 MAL13P1.173	3: PY02149 PY02148 PY02147
MSP8	1: PVX_097625	1: PKH_103110	1: PFE0120c	1: PY06415
MSP9	1: PVX_124060	1: PKH_144540	1: PFL1385c	1: PY02883
MSP10	1: PVX_114145	1: PKH_112880	1: PFF0995c	1: PY04226
MAP1	1: PVX_099975	1: PKH_072840	0	0

Supplementary Table 15. Analysis of Salvador I *vir* gene expression. Total RNA from *P. vivax* Salvador I (Sal I) and three different human patients were used in RT-PCR analysis of members from the different subfamilies A-L. + indicates a positive amplification from at least one member of a subfamily; primers (F, forward; R, reverse) for subfamilies G, K, L are shown; * primers and RT-PCR conditions for subfamilies A-F have been described elsewhere [48,49].

vir gene sub-family	Sal I cDNA	Patient 1 cDNA	Patient 2 cDNA	Patient 3 cDNA	Primer sequences
A	+	+	+	+	*
B	+	+	+	+	*
C	+	+	+	+	*
D	+	+	+	+	*
E	+	+	+	+	*
G	+	+	+	+	F 5'-CTTCARYRCCYRTRGGAATTR-3' R 5'-TGAATAYTCRAARGMRCYATA-3'
K	+	+	+	+	F 5'-TKTTWAAYKWYTGGBTAAAT-3' R 5'-TCRBHRMAACATDHAKYWGA-3'
L	+	+	+	+	F 5'-AYTGGYTHYWYGATVAAVTA-3' R 5'-AAKHTSHTYRWRDWTTTYWARATA-3’

Supplementary Table 16. Gene families in four species of *Plasmodium*, with specific emphasis on *P. vivax* and *P. knowlesi*. * Estimated number of complete genes; ** functional homologs to *P. falciparum var* gene.

Family name	**Other name**	***P. vivax* copy no.**	***P. knowlesi* copy no.**	**Homologs in *P. y. yoelii* and *P. falciparum***	**Putative function or location**
Pv-fam-a	PvTRAG	36	21	*P. falciparum*, *P. y. yoelii*	Tryptophan-rich antigen
Pv-fam-b	-	6	4	None	Hypothetical
Pv-fam-c	-	7	0	None	Hypothetical
Pv-fam-d	HYPB	16	2	None	Hypothetical
Pv-fam-e	RAD	44	14	*P. falciparum*	Hypothetical
Pv-fam-g	-	3	3	*P. falciparum*	Hypothetical
Pv-fam-h	HYP16	4	3	*P. falciparum*	Hypothetical
Pv-fam-i	HYP11	6	3	*P. falciparum*	Hypothetical
SICAVAR	-	0	107*	None**	Antigenic variation, immune evasion
Pk-fam-a	-	0	15	None	Hypothetical
Pk-fam-b	-	0	10	None	Hypothetical
Pk-fam-c	-	0	6	None	Hypothetical
Pk-fam-d	-	0	6	None	Hypothetical
Pk-fam-e	-	0	4	None	Hypothetical
MSP3	-	11	2	*P. falciparum*	Merozoite surface protein
MSP7	-	11	4	*P. falciparum*, *P. y. yoelii*	Merozoite surface protein
PIR	-	346* *vir*	68* *kir*	*P. y. yoelii*	Immune evasion
PST-A	-	11*	5	*P. falciparum*, *P. y. yoelii*	Alpha beta hydrolase
ETRAMP	-	9	9	*P. falciparum*, *P. y. yoelii*	Parasitophorous vacuole membrane
RBL	Rh/NBP	10*	3	*P. falciparum*, *P. y. yoelii*	Host cell selection, antigenic variation
DBL	EBL	1	3	*P. falciparum*, *P. y. yoelii*	Host cell selection, antigenic variation
CLAG	RhopH-1	3	2	*P. falciparum*, *P. y. yoelii*	High MW rhoptry antigen complex
PvSTP1	Pf-fam-b	11*	0	*P. falciparum*	Hypothetical
PHIST	Pf-fam-h	20	37	*P. falciparum*	Hypothetical, exported
SERA	-	13	7	*P. falciparum*, *P. y. yoelii*	Cysteine protease

Supplementary Table 17. Codon usage analysis of *P. vivax* and *P. falciparum* genes. T3s, C3s, A3s, G3s: proportion of codons with a T, C, A or G at the third position; CBI: Codon bias index, measures the extent to which a gene uses a subset of optimal codons (in a gene with extreme codon bias, CBI will equal 1.0; in a gene with random codon usage, CBI will equal 0.0, although it is possible for the number of optimal codons to be less than expected by random change, which results in a negative value for CBI); N_c: a measure of overall codon bias (the reported value of N_c is between 20 [codon usage completely biased and only one codon is effectively used for each amino acid] to 61 [codon usage completely random]); GC3s: the fraction of codons that are synonymous at the third codon position, and which have either a guanine or cytosine at that third codon position; GC: the G+C content of the genes. Approximately 3,800 orthologs between *P. vivax* and *P. falciparum* and ~460 *P. vivax* genes belonging to subtelomeric families were analyzed. Based upon the orthologs, *P. vivax* has a more random usage of codons (N_c =54) while the usage trend in *P. falciparum* is more biased (Nc=37). The results for the GC3 show that the fraction of codons that are synonymous at the third codon position are dramatically reduced in *P. falciparum* compared to *P. vivax*, and as also evidenced by the T3s, C3s, G3s and A3s indices. Interestingly, analysis of *P. vivax* protein families (located in A+T-rich regions) showed the codon usage to resemble that found in *P. falciparum* genes. The Nc value is 47.21, lower than the Nc value of *P. vivax* orthologs (54.24), indicating that the codon usage seems to be less random, and the frequency of synonymous codons ending with either G or C is closer to the value for *P. falciparum* orthologs.

Parameter	*P. vivax* orthologs	*P. falciparum* orthologs	*P. vivax* subtelomeric gene families
T3s	0.2583	0.5867	0.4751
C3s	0.3957	0.1003	0.1748
A3s	0.2943	0.6268	0.5285
G3s	0.3861	0.1283	0.1764
CBI	0.030	-0.248	-0.115
Nc	54.24	37.47	47.21
GC3s	0.577	0.147	0.252
GC	0.471	0.230	0.341

Supplementary References

1. Collins, W.E., Contacos, P.G., Krotoski, W.A. & Howard, W.A. Transmission of four Central American strains of Plasmodium vivax from monkey to man. *J Parasitol* **58**, 332-5 (1972).
2. Gardner, M.J. et al. Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum. *Science* **282**, 1126-32 (1998).
3. Venter, J.C. et al. The sequence of the human genome. *Science* **291**, 1304-51 (2001).
4. Myers, E.W. et al. A whole-genome assembly of Drosophila. *Science* **287**, 2196-204 (2000).
5. Gajer, P., Schatz, M. & Salzberg, S.L. Automated correction of genome sequence errors. *Nucleic Acids Res* **32**, 562-9 (2004).
6. Delcher, A.L., Phillippy, A., Carlton, J. & Salzberg, S.L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478-83 (2002).
7. Carlton, J.M., Galinski, M.R., Barnwell, J.W. & Dame, J.B. Karyotype and synteny among the chromosomes of all four species of human malaria parasite. *Mol Biochem Parasitol* **101**, 23-32 (1999).
8. Huang, X., Adams, M.D., Zhou, H. & Kerlavage, A.R. A tool for analyzing and annotating genomic sequences. *Genomics* **46**, 37-45 (1997).
9. Haas, B.J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654-66 (2003).
10. Tatusov, R.L. et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**, 22-8 (2001).
11. Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783-95 (2004).
12. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-80 (2001).
13. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-64 (1997).
14. Wootton, J.C. & Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* **266**, 554-71 (1996).
15. Gardner, M.J. et al. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498-511 (2002).
16. Wingender, E., Dietze, P., Karas, H. & Knuppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24**, 238-41 (1996).
17. Kummerfeld, S.K. & Teichmann, S.A. DBD: a transcription factor prediction database. *Nucleic Acids Res* **34**, D74-81 (2006).
18. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-84 (2002).
19. Coulson, R.M. & Ouzounis, C.A. The phylogenetic diversity of eukaryotic transcription. *Nucleic Acids Res* **31**, 653-60 (2003).
20. Elemento, O., Slonim, N. & Tavazoie, S. A universal framework for regulatory element discovery across all genomes and data types. *Mol Cell* **28**, 337-50 (2007).
21. Bozdech, Z. et al. The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum. *PLoS Biol* **1**, E5 (2003).
22. Le Roch, K.G. et al. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301**, 1503-8 (2003).

23. Young, J.A. et al. The Plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol Biochem Parasitol* **143**, 67-79 (2005).
24. Zhou, Y. et al. In silico gene function prediction using ontology-based pattern identification. *Bioinformatics* **21**, 1237-45 (2005).
25. Pavesi, G., Mauri, G. & Pesole, G. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* **17 Suppl 1**, S207-14 (2001).
26. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-7 (2000).
27. Levy, S. & Hannenhalli, S. Identification of transcription factor binding sites in the human genome sequence. *Mamm Genome* **13**, 510-4 (2002).
28. Hughes, J.D., Estep, P.W., Tavazoie, S. & Church, G.M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* **296**, 1205-14 (2000).
29. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-80 (1999).
30. Crabtree, J., Angiuoli, S.V., Wortman, J.R. & White, O. Sybil: Methods and Software for Multiple Genome Comparison and Visualization. in *Gene Function Analysis* (ed. Ochs, M.) 93-108 (Humana Press, New Jersey, 2007).
31. Li, L., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-89 (2003).
32. Kooij, T.W. et al. A Plasmodium whole-genome synteny map: indels and synteny breakpoints as foci for species-specific genes. *PLoS Pathog* **1**, e44 (2005).
33. Yang, Z., Nielsen, R. & Hasegawa, M. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol* **15**, 1600-11 (1998).
34. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* **13**, 555-556 (1997).
35. Yang, Z., Nielsen, R. & Hasegawa, M. Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution* **15**, 1600-11 (1998).
36. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418-26 (1986).
37. Yang, Z. & Bielawski, J.P. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution* **15**, 496-503 (2000).
38. Pinney, J.W., Shirley, M.W., McConkey, G.A. & Westhead, D.R. metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. *Nucleic Acids Res* **33**, 1399-409 (2005).
39. Karp, P.D., Paley, S. & Romero, P. The Pathway Tools software. *Bioinformatics* **18 Suppl 1**, S225-32 (2002).
40. Ren, Q., Kang, K.H. & Paulsen, I.T. TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res* **32**, D284-8 (2004).
41. Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A. & Durbin, R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **26**, 320-2 (1998).
42. Finn, R.D. et al. Pfam: clans, web tools and services. *Nucleic Acids Res* **34**, D247-51 (2006).
43. Haft, D.H., Selengut, J.D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res* **31**, 371-3 (2003).
44. Haas, B.J. et al. Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. *BMC Biol* **3**, 7 (2005).

45. Jaccard, P. The Distribution of the Flora in the Alpine Zone. *The New Phytologist* **11**, 37-50 (1912).
46. Price, A.L., Jones, N.C. & Pevzner, P.A. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**, i351-i358 (2005).
47. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28-36 (1994).
48. del Portillo, H.A. et al. A superfamily of variant genes encoded in the subtelomeric region of Plasmodium vivax. *Nature* **410**, 839-42 (2001).
49. Merino, E.F. et al. Multi-character population study of the vir subtelomeric multigene superfamily of Plasmodium vivax, a major human malaria parasite. *Mol Biochem Parasitol* **149**, 10-16 (2006).
50. Swofford, D.L. *PAUP*: phylogenetic analysis using parsimony and other methods.*, (Sinauer Press, New York, 1998).
51. Gilson, P.R. et al. Identification and stoichiometry of glycosylphosphatidylinositol-anchored membrane proteins of the human malaria parasite Plasmodium falciparum. *Mol Cell Proteomics* **5**, 1286-99 (2006).
52. Foth, B.J. et al. Dissecting apicoplast targeting in the malaria parasite Plasmodium falciparum. *Science* **299**, 705-8 (2003).
53. Toyoshima, C. & Nomura, H. Structural changes in the calcium pump accompanying the dissociation of calcium. *Nature* **418**, 605-11 (2002).
54. Zhang, Z. et al. Electron transfer by domain movement in cytochrome bc1. *Nature* **392**, 677-84 (1998).
55. Iwata, S. et al. Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex. *Science* **281**, 64-71 (1998).
56. Galinski, M.R. et al. Plasmodium vivax merozoite surface proteins-3beta and-3gamma share structural similarities with P. vivax merozoite surface protein-3alpha and define a new gene family. *Mol Biochem Parasitol* **115**, 41-53 (2001).
57. Galinski, M.R. et al. Plasmodium vivax merozoite surface protein-3 contains coiled-coil motifs in an alanine-rich central domain. *Mol Biochem Parasitol* **101**, 131-47 (1999).
58. del Portillo, H.A., Longacre, S., Khouri, E. & David, P.H. Primary structure of the merozoite surface antigen 1 of Plasmodium vivax reveals sequences conserved between different Plasmodium species. *Proc Natl Acad Sci U S A* **88**, 4030-4 (1991).
59. Benjamin, P.A. et al. Antigenic and sequence diversity at the C-terminus of the merozoite surface protein-1 from rodent malaria isolates, and the binding of protective monoclonal antibodies. *Mol Biochem Parasitol* **104**, 147-56 (1999).
60. Herrera, S., Corradin, G. & Arevalo-Herrera, M. An update on the search for a Plasmodium vivax vaccine. *Trends Parasitol* **23**, 122-8 (2007).
61. He, X.L., Grigg, M.E., Boothroyd, J.C. & Garcia, K.C. Structure of the immunodominant surface antigen from the Toxoplasma gondii SRS superfamily. *Nat Struct Biol* **9**, 606-11 (2002).
62. Sanders, P.R. et al. Distinct protein classes including novel merozoite surface antigens in Raft-like membranes of Plasmodium falciparum. *J Biol Chem* **280**, 40169-76 (2005).
63. DePristo, M.A., Zilversmit, M.M. & Hartl, D.L. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. *Gene* **378**, 19-30 (2006).
64. Kaiser, K., Matuschewski, K., Camargo, N., Ross, J. & Kappe, S.H. Differential transcriptome profiling identifies Plasmodium genes encoding pre-erythrocytic stage-specific proteins. *Mol Microbiol* **51**, 1221-32 (2004).

65. Spielmann, T., Fergusen, D.J. & Beck, H.P. etramps, a new Plasmodium falciparum gene family coding for developmentally regulated and highly charged membrane proteins located at the parasite-host cell interface. *Mol Biol Cell* **14**, 1529-44 (2003).
66. Imwong, M. et al. Contrasting genetic structure in Plasmodium vivax populations from Asia and South America. *Int J Parasitol* (2007).
67. Imwong, M. et al. Microsatellite variation, repeat array length, and population history of Plasmodium vivax. *Mol Biol Evol* **23**, 1016-8 (2006).
68. Ralph, S.A., Foth, B.J., Hall, N. & McFadden, G.I. Evolutionary pressures on apicoplast transit peptides. *Mol Biol Evol* **21**, 2183-94 (2004).
69. Matsuzaki, M. et al. Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. *Nature* **428**, 653-7 (2004).
70. Armbrust, E.V. et al. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science* **306**, 79-86 (2004).
71. Gardner, M.J. et al. Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes. *Science* **309**, 134-7 (2005).
72. Abrahamsen, M.S. et al. Complete genome sequence of the apicomplexan, Cryptosporidium parvum. *Science* **304**, 441-5 (2004).
73. Marti, M., Good, R.T., Rug, M., Knuepfer, E. & Cowman, A.F. Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* **306**, 1930-3 (2004).
74. Wood, S.A. et al. RtoA links initial cell type choice to the cell cycle in Dictyostelium. *Development* **122**, 3677-85 (1996).
75. Jazwinski, S.M. The RAS genes: a homeostatic device in Saccharomyces cerevisiae longevity. *Neurobiol Aging* **20**, 471-8 (1999).
76. Ireland, L.S. et al. A member of a novel family of yeast 'zn-finger' proteins mediates the transition from stationary phase to cell proliferation. *Embo J* **13**, 3812-21 (1994).
77. Bruce, M.C., Galinski, M.R., Barnwell, J.W., Snounou, G. & Day, K.P. Polymorphism at the merozoite surface protein-3alpha locus of Plasmodium vivax: global and local diversity. *Am J Trop Med Hyg* **61**, 518-25 (1999).