PNAS
www.pnas.org

Supplementary Information for

Evolutionary history of human *Plasmodium vivax* revealed by genome-wide analysis of related ape parasites

Dorothy E. Loy[#], Lindsey J. Plenderleith[#], Sesh A. Sundararaman, Weimin Liu, Jakub Gruszczyk, Yi-Jun Chen, Stephanie Trimboli, Gerald H. Learn, Oscar A. MacLean, Alex L.K. Morgan, Yingying Li, Alexa N. Avitto, Jasmin Giles, Sébastien Calvignac-Spencer, Andreas Sachse, Fabian H. Leendertz, Sheri Speede, Ahidjo Ayouba, Martine Peeters, Julian C. Rayner, Wai-Hong Tham, Paul M. Sharp[*] and Beatrice H. Hahn[*]

[#]contributed equally
*co-senior authors

Corresponding author: Beatrice H. Hahn
Email: bhahn@pennmedicine.upenn.edu

This PDF file includes:

SI Materials and Methods

Ape samples. Whole blood samples (5-10 ml) were obtained from sanctuary chimpanzees (*Pan troglodytes*) cared for at the Sanaga-Yong (SY) Chimpanzee Rescue Center in Cameroon as previously described (1, 2), including from members of the central (*P. t. troglodytes*) and Nigeria-Cameroonian (*P. t. ellioti*) subspecies. These samples were obtained for veterinary purposes only or represented leftover specimens from yearly health examinations. None of the chimpanzees exhibited symptoms of malaria at the time of sampling. Blood was preserved in RNAlater (1:1 vol/vol) without further processing at room temperature until shipment to the United States and long term storage at -80°C. Blood from a wild-living habituated chimpanzee ("Sagu") in the Tai Forest (*P. t. verus*) was obtained during an emergency field immobilization for treatment of a respiratory condition as described (3) and immediately frozen in liquid nitrogen prior to shipment to Germany and storage at -80°C. A small amount of blood (frozen directly without preservation) was also available from a western gorilla (*Gorilla gorilla*) of unknown geographic origin (Gor3157), which was killed by hunters and confiscated by the anti-poaching program of the Cameroonian Ministry of Environment and Forestry (2, 4). Ape fecal samples were selected from an existing bank of chimpanzee, bonobo, and gorilla specimens previously shown to contain *P. vivax* parasite DNA (1, 5, 6). These specimens were collected non-invasively from non-habituated apes living in remote forest areas, with a two-letter-code indicating their field site of origin (Fig. S5). Additionally, dried blood spots were available from two chimpanzees and one gorilla housed at the Mfou National Park Wildlife Rescue Center in Cameroon, which were previously shown to be *P. vivax* positive (5). DNA was extracted from whole blood and dried blood spots using the QIAamp Blood DNA Mini Kit or the Puregene Core Blood Kit (Qiagen). Fecal DNA was extracted using the QIAamp Stool DNA mini kit (Qiagen). All specimens were subjected to host mitochondrial DNA analysis to confirm their species and subspecies origin (1, 5, 6). Sample collection was approved by the Ministry of Environment and

Forestry in Cameroon and by the Ministry of the Environments and Forests in Cote d'Ivoire. All samples were shipped in compliance with Convention on International Trade in Endangered Species of Wild Fauna and Flora regulations and country-specific import and export permits.

PCR screening for *Plasmodium* infection. Ape blood and fecal samples were screened for *Plasmodium* sequences by diagnostic PCR using both pan-*Plasmodium* and *P. vivax*-specific primer sets as previously described (1, 5, 6). Briefly, nested PCR was used to amplify a 956 bp fragment of the *cytochrome b* gene (pan-*Plasmodium* primers*)* and a 295 bp fragment of the *cytochrome oxidase I* gene (*P. vivax*-specific primers) of the parasite mitochondrial genome. Amplicons were subjected to Sanger or Illumina sequencing, and phylogenetic analysis was performed to determine the species of the amplified *Plasmodium* sequences. Because the blood sample from the chimpanzee Sagu was previously reported to be positive for *P. vivax* (7)*,* it was not rescreened for this study.

Selective amplification of ape *P. vivax* genomes. To generate *P. vivax* genome sequences from unprocessed ape blood, we used selective whole genome amplification (SWGA), which utilizes the highly processive phi29 DNA polymerase to preferentially amplify pathogen sequences from complex mixtures of target and host DNA (2, 8). Six sets of *P. vivax*-specific primers were used (Table S1), of which pvset1 and pvset6 (also termed pvset1920) have previously been reported (9). The remaining four sets (pvset2-5) were newly designed to increase overall *P. vivax* genome coverage. Using custom scripts (2), we initially identified sequence motifs (6-12 bp in length) that occurred frequently in the *P. vivax* SalI reference genome (10), but only infrequently in the human genome. These were filtered to remove primers that exhibited extreme melting temperatures, were predicted to form homodimers, and/or bound the SalI mitochondrial genome or its subtelomeric regions more than three times. The resulting primer sets, including pvset1 (5'-CGTTG*C*G-3', 5'-TTTTTTC*G*C-3', 5'-TCGTG*C*G-3', 5'-

CGTTTTTT*T*T-3', 5'-TTTTTTTC*G*T-3', 5'-CCGTT*C*G-3', 5'-CGTTTC*G*T-3', 5'-CGTTTC*G*C-3', 5'-CGTTTT*C*G-3', and 5'-TCGTTC*G*T-3') and pvset2 (5'-CGAAAAAA*A*A-3', 5'-CGCAA*C*G-3', 5'-GCGAAA*T*G-3', 5'-CGCAC*G*A-3', 5'-GCGAAAA*A*A-3', 5'-AACGAAAA*A*A-3', 5'-AACGAA*C*G-3', 5'-ACGAAA*C*G-3', 5'-CGAACG*A*A-3', and 5'-CGAAAC*G*G-3'), exhibited high coverage of the *P. vivax* genome and low coverage of the human genome (asterisks indicate the location of phosphorothioate bonds necessary to prevent degradation by the phi29 polymerase). Two additional sets, including pvset3 (5'-CTTCGAA*C*G-3', 5'-GCGAAAC*G*T-3', 5'-GGCGAAAAA*A*A-3', 5'-TCGCGAA*A*A-3', 5'-TTTCGCG*T*A-3', and 5'-TTTCGTG*C*G-3') and pvset4 (5'-CGAAGCGG*A*G-3', 5'-CTTCGAA*C*G-3', 5'-GCGAAAC*G*T-3', 5'-GGCGAAAAA*A*A-3', 5'-TCGCGAA*A*A-3', 5'-TTTCGCG*T*A-3', and 5'-TTTCGTG*C*G-3'), were generated using the program *swga* (11), which is designed to select primer sets that bind evenly across the reference genome. After noticing preferential amplification of AT-rich subtelomere regions in SWGA products, we designed two final primer sets using only GC-rich regions of the SalI reference sequence as the foreground genome (9), resulting in pvset5 (5'-AGCGAAAAAA*A*A-3', 5'-AGCGAAC*G*T-3', 5'-CAAACGGG*T*G-3', 5'-CGAACGA*A*T-3', 5'-CGAAGCGG*A*G-3', 5'-CGAATGGG*G*G-3', 5'-CGAGCGA*A*C-3', 5'-CGTTTGG*C*G-3', 5'-GCGGGAAAA*A*A-3', 5'-GCGTGTA*C*G-3', 5'-TACGACG*A*G-3', and 5'-TTCAGCG*C*G-3') and pvset6 (5'-AACGAAGC*G*A-3', 5'-ACGAAGCG*A*A-3', 5'-ACGACGA*A*G-3', 5'-ACGCGCA*A*C-3', 5'-CAACGCG*G*T-3', 5'-GACGAAA*C*G-3', 5'-GCGAAAAA*G*G-3', 5'-GCGAAGC*G*A-3', 5'-GCGGAAC*G*A-3', 5'-GCGTCGA*A*G-3', 5'-GGTTAGCG*G*C-3', and 5'-AACGAAT*C*G-3').

SWGA was performed as described (2, 9) by amplifying whole blood DNA (100-750 ng) in a 50 µl reaction with 1x phi29 Buffer (New England Biolabs), 1 mM dNTPs (Roche), 3.5 µM of SWGA primers (an equimolar mix of primers in the set), 1% BSA and 30 units of phi29 polymerase (New England Biolabs). SWGA conditions included a 1 h ramp-down step (35^{o}C to 30^{o}C), followed by an amplification step for 16 h at 30^{o}C, followed by a phi29 denaturation step

for 10 min at 65°C. SWGA products were diluted 1:1 in water, purified using AMPure Beads (Beckman Coulter), and stored at 4°C. To mitigate the stochastic nature of SWGA at low template concentrations (2), genomic DNA from each ape-derived sample was amplified on multiple independent occasions with different primer sets (Table S1). Because pretreatment with restriction enzymes that selectively degrade host DNA can improve SWGA efficiency (2), some DNA aliquots were digested with the methylation sensitive enzymes MspJI and FspEI (5 units each) for 2 hours at 37°C prior to SWGA amplification (Table S1). To obtain sufficient quantities of parasite genomic DNA for sequencing, ape-derived DNA samples were subjected to multiple rounds (up to 4) of successive SWGA amplification, some of which were performed with alternating primer sets to improve genome coverage (9).

Illumina and PacBio sequencing. To generate chimpanzee *P. vivax* draft genomes, we used SWGA amplicons from samples SY43 and SY56 for Illumina and PacBio sequencing. These chimpanzee blood samples were selected because they were PCR positive for *P. vivax cytb* and *cox1* regions, but lacked *Laverania* sequences. For Illumina sequencing, we pooled second and fourth round SWGA products to prepare short insert libraries using the KAPA HyperPlus kit (Roche) with an enzyme fragmentation time of 3 minutes. Fragmentation products were purified with AMPure Beads (Beckman Coulter), followed by dual-sided solid phase reversible immobilization (SPRI) to select for fragments 550 bp in length. The resulting libraries were sequenced on an Illumina MiSeq platform using V2 chemistry. For PacBio sequencing, second and third round SWGA products were pooled for library preparation. Briefly, amplification products (7.5-40 µg) were incubated with S1 nuclease (15 units) for 30 min at 37°C for DNA linearization, purified using AMPure beads, and passed through a 26 gauge blunt end needle to reduce DNA fragment size from >60,000 bp to ~15,000 bp. DNA was purified, eluted in 40 µl of water, and sent to the University of Delaware Sequencing Core where fragments of 7,000-

18,000 bp length were size-selected using BluePippin (Sage Biosciences) prior to SMRT Bell library preparation and PacBio SMRT Cell sequencing.

Assembly of PvSY43 and PvSY56 draft genomes. Illumina MiSeq reads from chimpanzee blood samples SY43 and SY56 were error corrected using SPAdes (12) and then mapped to the chimpanzee reference genome using smalt (http://www.sanger.ac.uk/science/tools/smalt-0). Unmapped reads were extracted and converted to fastq files using SAMtools (13) and BEDtools (14), respectively. PacBio reads from samples SY43 and SY56 that were longer than 1,500 bp and 1,000 bp, respectively, were filtered to remove chimpanzee reads using BLASR (15). The resulting non-chimpanzee reads were corrected using proovread (16).

The non-chimpanzee MiSeq reads were iteratively mapped (10 times) with correction to the human *P. vivax* PvP01 reference sequence in Geneious (version 9) to generate a preliminary chimpanzee *P. vivax* consensus sequence (17). Errors were identified based on low read support as described (2), the assembly was separated into contigs by splitting at assembly gaps, and contigs <100 bp were removed using Geneious (17), SAMtools (13), BEDtools (14) and custom scripts. The resulting contigs were then mapped to the PvP01 reference genome using ABACAS (18). Gaps were closed using FGAP (19) with proovread-corrected, non-chimpanzee PacBio reads. After initial gap closure, regions with low read support were again removed. Finally, gaps were filled using IMAGE (20) and GapFiller (21), followed by removal of likely duplications and inversions at the edges of gaps and a final error correction with iCORN (22).

The PvSY56 genome assembly was improved by *de novo* assembly of subtelomere and internal hypervariable regions. Orthologs of genes that bounded the subtelomeres and internal hypervariable regions in PvSalI (23) were identified in PvP01 and PvSY56, and the subtelomeres and internal hypervariable regions of these genomes were then defined as the sequence between the boundary gene and the nearest chromosome end, or the sequence

between and including the two internal boundary genes, respectively. For *de novo* assembly, corrected Illumina reads from sample SY56 were mapped using smalt (http://www.sanger.ac.uk/science/tools/smalt-0) to a version of PvP01 in which the subtelomeres and internal hypervariable regions were masked. Reads that did not map to the masked genome were extracted using SAMtools (13) and BEDtools (14). Since these contained a large number of *Pseudomonas* reads, we removed these by mapping to the *Pseudomonas yamanorum* genome (GenBank accession number: LT629793). The resulting reads were assembled using SPAdes (12). *De novo* contigs were retained if they were at least 2 kb in length and either could be mapped to PvP01 subtelomeres by ABACAS or had a blastn hit (e-value $<10^{-4}$) to any of four human *P. vivax* genome assemblies (PvSalI, PvP01, PvC01, PvT01) (24). These were ordered to the PvP01 genome using Companion (25). Ordered *de novo* contigs that provided a better assembly of the subtelomeres or internal hypervariable regions than the initial PvSY56 assembly were exchanged.

Genome annotation. Annotations were transferred from the human *P. vivax* PvP01 reference sequence to the PvSY56 and PvSY43 genome assemblies using RATT (26), with additional genes in PvSY56 subtelomeres predicted using Companion (25). All annotations on the chromosomes were visually inspected, and manually corrected where necessary. In roughly 10% of genes, we noted one or two base pair insertions or deletions in homopolymer tracts. Since these small indels were restricted to homopolymer regions but found throughout the genome, we reasoned that they represented sequencing errors. We thus manually corrected the annotation to maintain the reading frame. Additional *ad hoc* manual corrections were performed when alignments indicated an error and mapped reads supported either sequence correction or removal.

Partial ape *P. vivax* genome sequences. To generate genome sequences from additional ape *P. vivax* strains, we subjected diagnostic PCR positive chimpanzee (SY81, SY90, Sagu) and gorilla (Gor3157) blood DNA to SWGA analysis (Table S1). The resulting amplification products were prepared for sequencing using Nextera protocols (generating 100-150 bp fragments) and sequenced using the IIlumina MiSeq and/or MiniSeq platforms. Reads were mapped to the respective host reference genome (chimpanzee or gorilla) using smalt (http://www.sanger.ac.uk/science/tools/smalt-0) and unmapped reads were extracted using SAMtools (13) and BEDtools (14).

We also mined a publicly available read database from a blood sample of a *P. malariae*-infected sanctuary chimpanzee from Gabon (GA02), which we noted contained a substantial number of ape *P. vivax* reads (27). Sequencing reads (ERS434565) were mapped to concatenated reference sequences of the chimpanzee *P. reichenowi* strain CDC (28), the human *P. vivax* strain P01 (24), and the human *P. malariae* strain UG01 (27) using smalt (http://www.sanger.ac.uk/science/tools/smalt-0). Reads that mapped to *P. vivax* were used for this study.

Variant calling. For human *P. vivax*, we included genome data from previously sequenced field isolates from different geographic regions (SRA accession numbers SRP046091, SRS805942, SRP046182, SRP046094, SRP045997, SRP046126, SRP046031) that were classified as high-quality single infections by the authors and were not monkey-adapted (29). For ape *P. vivax*, we analyzed genome data from four chimpanzee-derived samples (PvSY81, PvSY90, PvSagu, PvGA02) and one gorilla-derived sample (PvGor3157), with uncorrected Illumina sequencing reads from PvSY43 and PvSY56 also included to aid validation of variants. Variants were called using the Genome Analysis Toolkit (GATK) version 4.0 (30). Sequencing reads were mapped to either PvSY56 (ape *P. vivax* samples) or PvP01 (human *P. vivax* samples) using bwa (http://bio-bwa.sourceforge.net), then duplicate reads were removed and base quality scores recalibrated.

'Known variants' for base quality score recalibration (BQSR) were generated for each reference genome by a bootstrap procedure of variant calling, hard filtering with GATK suggested generic filters, and BQSR using the filtered variants, which was repeated until variants called in subsequent iterations showed little difference. The de-duplicated, recalibrated bam files were used to generate genomic variant call format files for each sample using HaplotypeCaller, with the OverClippedReadFilter applied. Variants were called jointly in human *P. vivax* and ape *P. vivax* samples, excluding insertion-deletion polymorphisms. Single nucleotide polymorphisms (SNPs) were hard-filtered using a set of annotations and values selected to minimize the number of SNPs removed at fourfold degenerate sites, while minimizing the number of subtelomeric SNPs retained (QualByDepth <2.0, FisherStrand >50.0, RMSMappingQuality <45.0, StrandOddsRatio >2.5, GenotypeQuality_Mean <35.0, Quality <30.0, MappingQualityRankSumTest <-2.0 or >2.0, ReadPosRankSumTest <-6.0 or >10.0, BaseQualityRankSumTest <-6.0 or >10.0). Dustmasker (BLAST+ suite) (31) was used with default settings to identify low-complexity regions within each reference, which were then excluded in all samples. Since the reference genome for ape *P. vivax* was PvSY56, sites that were called as SNPs with PvSY56 reads were assumed to be errors in the PvSY56 assembly and excluded. The ability to call sites was assessed using CallableLoci (minimum mapping quality 20, minimum depth 5, OverClippedReadFilter applied). Uncallable sites and sites with heterozygous SNP calls in a given sample were excluded from that sample.

Orthologous gene alignments. Gene coding sequences were extracted from the chimpanzee PvSY43 and PvSY56 genome assemblies, the human *P. vivax* reference strains PvP01 and PvSalI, the *P. cynomolgi* reference strain PcyM (32) and the *P. knowlesi* reference strain PknH (33). Low-complexity regions were masked using segmasker (BLAST+ suite) (31). Groups of 1:1 orthologs were identified from the RATT annotation transfer for PvP01, PvSY43 and PvSY56, and from PlasmoDB ortholog groups. Genes in subtelomeric or internal hypervariable

regions, genes that were annotated as pseudogenes or had internal stop codons, and all *vir* and *phist* genes were excluded from the analysis. Orthologous gene sequences were aligned from the *P. vivax* genome assemblies using TranslatorX/MUSCLE (34) and genes with unusually high divergence were inspected and manually corrected when necessary. Orthologs from *P. cynomolgi* and *P. knowlesi* were added to these alignments where available, excluding the 2% and 3% of genes with the highest divergence, respectively. To add additional ape and human *P. vivax* strains to the alignment, gene sequences were generated from SNP calls by changing the reference sequence to the alternative allele at variant sites. Sites that were excluded from SNP calling were masked. Sequences of genes with >60% of sites callable and no internal stop codons were added to the alignments, following the alignment of the appropriate reference sequence. Ambiguous and masked sites as well as assembly gaps in any sequence were masked in all sequences.

Polymorphism analysis. Nucleotide diversity and divergence were calculated from gene alignments using the dist.dna function with the 'raw' model in the ape R package (35) and custom scripts. Synonymous (P_s) and nonsynonymous (P_n) polymorphisms were counted by determining the effect of each variant allele on the codon in the reference genome (PvSY56 or PvP01). If a site had multiple alternative alleles, each of these was counted as a separate SNP. Synonymous and nonsynonymous fixed differences were counted by comparing codons between *P. cynomolgi* strain M and PvSY56 or PvP01, after exclusion of sites that were polymorphic in ape *P. vivax* or human *P. vivax,* respectively, and assuming the mutation path with the smallest number of nonsynonymous changes. The neutrality index [$NI = (P_n/P_s)/(D_n/D_s)$] of each gene was calculated from the number of polymorphisms within ape and human *P. vivax* and the number of fixed differences (D_s and D_n) from the *P. cynomolgi* strain PcyM. Density distributions of $\log_2(NI)$ were generated in R for genes that had a defined $\log_2(NI)$ value in both ape and human *P. vivax.* McDonald-Kreitman tests were performed in R, using a two-tailed

Fisher exact test (fisher.test), followed by correction of the *p*-values for multiple testing (p.adjust, method=fdr).

Site frequency spectra were generated from SNP calls from the MalariaGen *P. vivax* Genome Variation project, using a subset of Southeast Asian field isolates (Cambodia, Indonesia, Laos, Myanmar, Malaysia, Papua New Guinea, Thailand and Vietnam) for which ≥ 80% of variant loci had been called. 173 samples met this inclusion criterion. Heterozygous calls were counted as one occurrence of the reference and one of the alternative allele. The number of allele calls varied between sites because of missing data and heterozygous SNPs. To standardize the number of calls per site without losing large amounts of data, we down-sampled to 168 calls per site, excluding sites with fewer calls (this threshold was chosen such that 95% of sites were retained). fourfold and zerofold degenerate sites were identified in the PvSalI reference sequence, as this was the reference utilized in the MalariaGen SNP data set (23). Alignments of orthologous genes from PvSalI, PvSY56 and PcyM were generated with TranslatorX/MUSCLE and used with the est-sfs (36) unfolder to identify the derived allele at each site and calculate the frequency spectrum.

Single genome amplification. To increase the number of geographically diverse ape parasite sequences, we subjected *P. vivax* positive ape blood and fecal samples to single genome amplification (SGA), which utilizes end-point-diluted template DNA and thus generates *Plasmodium* sequences devoid of PCR-induced sequence artifacts (1, 5). Five nuclear gene regions were targeted to complement existing genome data, including PVP01_1216000 (610bp), PVP01_1418300 (476 or 491bp), PVP01_1418500 (809 or 815bp), PVP01_1418600 (351bp), and PVP01_1418800 (576bp). These regions were amplified using first round primers Pv6000F1 (5'-ATGGAAAGGCAGGGCGACGC-3') and Pv6000R1 (5'-GCTGCACAGGTAGGAGATGTACT-3'), Pv8300F1 (5'-AACGTGGAGATGTAATTCCTGCC-3' and Pv8300R1 (5'-TTGTGTGCATTTTCGAGCAGGCTG-3'), Pv8500F1p (5'-

ATGGAGGACGAGACGGAGAAC-3') and Pv8500R1 (5'-CTGAAATAGATGTAGTTGTAGAAGG-3'), Pv8600F1 (5'-AAGAMAAACATTTTGGAAAACGCAG-3') and Pv8600R1 (5'-TCAAAACTCCATGGGGATGTTCTGC-3'), as well as Pv8800F1 (5'-TGTACGACTCGAT GAGTTACTTCC-3') and Pv8800R1 (5'-TCACAGGAAGACCGTCGAAAAC G-3'), respectively. Samples were multiplexed using 2.5 µl of end-point diluted DNA in a 25 µl reaction volume containing 0.5 µl dNTPs (10mM of each dNTP), 2.5 µmol of each first round primer, 2.5 µl PCR buffer, 0.1 µl BSA solution (50µg/ml), and 0.25 µl expand long template enzyme mix (Expand Long Template PCR System, Sigma). Cycling conditions included an initial denaturing step of 2 min at 94°C, followed by 15 cycles of denaturation (94°C, 10 sec), annealing (45°C, 30 sec), and elongation (68°C, 1 min), followed by 35 cycles of denaturation (94°C, 10 sec), annealing (48°C, 30 sec), and elongation (68 °C, 1 min; with 10-sec increments for each successive cycle), followed by a final elongation step of 10 minutes at 68°C. For second round PCR, 2 µl of the first round product were amplified using second round primers Pv6000F2 (5'-TAGAGGAGCAAGAGCGAGTGC-3') and Pv6000R2 (5'-TTCGACTCCTGCATTTGCCACTTG-3'), Pv8300F2 (5'-TTAACACGGAGGAGGCCACAGAATG-3') and Pv8300R2 (5'-CTCTCTCGTTTGTCTGCCTTCTTCC-3'), Pv8500F2p (5'-CAGAACTTGAAATGTCCAGGGAG-3') and Pv8500R2 (5'-AGCTGCCAGTTGTGCTTGTCTGCG-3'), Pv8600F2 (5'-GCAAAGGACATGACGCAAAGTG-3') and Pv8600R2 (5'-TTTCATCAAACGTGCATCTCTTGG-3'), as well as Pv8800F2 (5'-CTTATTTTGCTACGAAGATTTGGG-3') and Pv8800R2 (5'-GCAATATATCCGCCTCTCTCCTC-3'), respectively. Cycling conditions included an initial denaturation step of 2 min at 94°C, followed by 60 cycles of denaturation (94°C, 10 sec), annealing (52°C, 30 sec), and elongation (68°C, 1 min), followed by a final elongation step of 10 minutes at 68°C. Amplification products were sequenced directly without interim cloning. GenBank accession numbers for all newly derived ape *P. vivax* sequences are listed in Table S4.

We also used SGA to amplify regions of ape *P. vivax rbp* genes, which in human *P. vivax* contained inactivating mutations. Four fragments were targeted, including those containing the *rbp2d* (81bp) and *rbp2e* (120bp) frameshifts, and the *rbp2e* (61bp), and *rbp3* (52bp) stop mutations. These regions were amplified using first round primers anfsRBP2d_1F (5'-AATGATGCAAAGAATTTTATTTCGGAT-3' and anfsRBP2d_1R (5'-ACGCTTTCCTTTTCACTATCAATT-3'), anfsRBP2e_1F (5'- TGCAAGAAAACCATCTCGCT-3' and anfsRBP2e_1R (5'-TGCTCTCTTCATTTCTTCGTCA-3'), anstopRBP2e_1F (5'-ACAAAGCAAAAGGGCGAAGT-3') and anstopRBP2e_1R (5'- AGCGGATTCTTTGTGACTCCT-3'), as well as anstopRBP3_1F (5'- AATGAAGGGGAACTGAAAGGT-3') and anstopRBP3_1R (5'-TTTCTTTCGCCGCACTATGG-3'), respectively. PCRs were multiplexed using 2.5 µl of end-point diluted sample DNA in a 25 µl reaction volume, containing 0.5 µl dNTPs (10mM of each dNTP), 2.5 µmol of each first round primer (4 pairs), 2.5µl PCR buffer, 0.1 µl BSA solution (50µg/ml), and 0.25 µl expand long template enzyme mix (Expand Long Template PCR System, Sigma). Cycling conditions included an initial denaturing step of 2 minutes at 94°C, followed by 50 cycles of denaturation (94°C, 10 sec), annealing (48°C, 30 sec), and elongation (68°C, 30 sec), followed by a final elongation step of 10 minutes at 68°C. For second round PCR, 2 µl of the first round product were amplified using second round primers anfsRBP2d_2F (5'-AGATGATCTGAATAAACGTTTCACA-3') and anfsRBP2d_2R (5'-ACAAATTCGTCAACGTTAAGTGT-3'), anfsRBP2e_2F (5'-AGGACAACACATATGCAGTTACT-3') and anfsRBP2e_2R (5'-ACTTTTATGGTCACCGTAGATACA-3'), anstopRBP2e_2F (5'-ACACATGATATTGATGCACTCAAAGA-3') and anstopRBP2e_2R (5'-TCTTGATTTGTCTCACTATTCTCTGT-3'), as well as anstopRBP3_2F (5'-ACAATGTGTGTAAGAATATTGAGACCA-3') and anstopRBP3_2R (5'-TGGGACACATTTTCTATACAGGCT-3'), respectively. Cycling conditions included an initial denaturation step of 2 minutes at 94°C, followed by 50 cycles of denaturation (94°C, 10 sec),

annealing (52°C, 30 sec), and elongation (68°C, 30 sec), followed by a final elongation step of 10 minutes at 68°C. Amplification products were sequenced directly without interim cloning.

Phylogenetic analyses and *rbp* gene comparisons. To examine the evolutionary relationships of ape and human *P. vivax* strains, phylogenetic trees were constructed from (i) nuclear gene sequences generated by SGA from infected ape blood or fecal DNA, (ii) PvP01, PvSalI, PvSY56, PvSY43, *P. cynomolgi* and *P. knowlesi* genome assemblies, and (iii) SNP data by changing the reference sequence to the alternative allele at variant sites. Sequences were aligned with TranslatorX/MUSCLE and the alignments manually corrected, including truncation to the SGA amplicon where appropriate and removal of ambiguously aligned regions. Trees were generated using PhyML with a GTR+Γ4 model of nucleotide evolution (37), 10 random starts and best of NNI/SPR trees, and bootstrap values calculated from 100 replicates. For neighbor-joining trees, matrices of pairwise genetic distances were calculated from alignments of genes that were covered in all strains, and unrooted trees were generated from these matrices in R with ape 'nj'. Phylogenetic networks were generated from the same alignments in SplitsTree4 (38) using SplitDecomposition with uncorrected p-distances.

Ape (PvSY43 and PvSY56) and human (PvSalI, PvP01, PvC01, and PvT01 see (24)) *P. vivax rbp* genes were identified from genome annotations. For *rbp1a*, sequences generated from SNP data were also included. For other *Plasmodium* species, *rbp* genes were identified from annotations in the *P. malariae* strain PmUG01 (PmUG01_07014300, PmUG01_07014200, PmUG01_08058500, PmUG01_12081700, PmUG01_14085600) (27), the *P. knowlesi* strain H (PKNH_0700200, PKNH_1472300) (33), the *P. cynomolgi* strains B and Berok (PCYB_071060, PCYB_081060, PCYB_053840, PCYB_071010, PCYB_053850, PCYB_147650, JQ422038) (39), the *P. inui* strain San Antonio 1 (C922_04999, C922_01465) (PlasmoDB) and the *P. fragile* strain Nilgiri (AK88_00929, AK88_00936) (PlasmoDB). Blastn search using *P. vivax rbp* genes also identified two non-annotated pseudogenes (an ortholog of *rbp2a* in the *P. cynomolgi* strain

B, and *nbp1* in *P. knowlesi*) as well as pseudogenes annotated as multiple gene fragments (*rbp1a*, *rbp2a* and *rbp2e* orthologs in *P. inui*; *rbp1b* and *rbp3* orthologs in *P. fragile*). Sequences were aligned with TranslatorX/MUSCLE, and alignments manually corrected. A tree was constructed for the most conserved region of this alignment (corresponding to nucleotides 478-7,938 in PvP01 *rbp1a*) using PhyML with a GTR+Γ4 model of nucleotide evolution (37), 10 random starts and best of NNI/SPR trees, and bootstrap values calculated from 100 replicates.

For tests of selection, branch-site models were fitted to the data using codeml in the PAML package (40), using alignments excluding *P. ovale* spp. and *P. malariae*, which were considered too divergent. The fit of the null model (no selection) was compared with the fit of the model with selection along the branch leading to human *P. vivax* (foreground branch), by comparing twice the difference in log-likelihood (2ΔlnL) with a X^2 distribution with one degree of freedom (and a p-value threshold of 0.05).

Published human *P. vivax* reads (23, 29) were downloaded from the SRA database and mapped to *rbp2d* (PVP01_1471400 and PVX_101585), *rbp2e* (PVP01_0700500), *rbp3* (PVP01_1469400), *rbp2p1* (PVP01_0534400), and *rbp2p2* (PVX_101590) sequences using smalt (http://www.sanger.ac.uk/science/tools/smalt-0). Mapped reads were extracted, converted to fastq files using SAMtools (13) and BEDtools (14), and then imported into Geneious (version 9) (17) to again be mapped to the references. Because a truncation of *rbp2d* has been identified in some human *P. vivax* field isolates (41), including PvP01, reads were mapped to both the SalI and PvP01 alleles of *rbp2d*, and the correct allele for each human *P. vivax* sample was chosen by visual inspection of read mapping. A majority consensus sequence was called for all positions with ≥ threefold coverage. Alignments of genes with complete coding sequences were made to chimpanzee *P. vivax rbp* genes and the positions of frameshifts and pseudogenes were noted. After identification of the ancestral stop or frameshift mutation, consensus sequences from samples with fewer than threefold coverage in some regions of the genes were

inspected to verify that the ancestral mutation was indeed in all sequenced human *P. vivax* strains.

To examine the ancestral mutations in additional ape *P. vivax* strains, sequencing reads from SY81, SY90, Sagu, and GA02 were mapped to *rbp2d, rbp2e,* and *rbp3* from PvSY56, and a majority consensus sequence was called for all regions with ≥ threefold coverage.

Recombinant RBP expression and polyclonal rabbit antibody production. Chimpanzee *P. vivax rbp2d*, *rbp2e* and *rbp3* gene sequences were generated by aligning PvSY43 and PvSY56 sequencing reads to human *P. vivax* reference sequences. Deduced amino acid sequences spanning positions 100 to 1,000 of RBP2d (from PvSY43), RBP2e (from PvSY56), and RBP3 (from PvSY56) were codon optimized for expression in *Escherichia coli*, and the synthesized genes were purchased from Eurofins Genomics (RBP2d and RBP3) and GenScript (RBP2e). These were then used to generate shorter gene fragments ($RBP2d_{165-967}$, $RBP2e_{156-957}$, $RBP3_{149-968}$) predicted to express stable recombinant proteins, which included the respective binding domain based on homology to human *P. vivax* RBP2a and RBP2b (42). $RBP2d_{165-967}$, $RBP2e_{156-957}$ and $RBP3_{149-968}$ were cloned into the pET-32a(+) vector (Novagen), which expresses proteins with a hexa-histidine tag (C-terminus) but also contains a tobacco etch virus (TEV) protease cleavage site to allow removal of this tag. All clones were sequence confirmed.

Proteins were expressed using *E. coli* strain SHuffle T7 (New England Biolabs) and Terrific Broth (TB) supplemented with 100 µg/ml of carbenicillin. Flasks containing 1 liter of medium were incubated in a Multitron shaker (Infors HT) at 37°C at 200 rpm. At OD_{600} of ~1.0, isopropyl-(β)-D-thiogalactopyranoside (IPTG, Astral) was added to a final concentration of 1.0 mM and protein expression was allowed to continue for 20 hours at 16°C. Cells were harvested by centrifugation at 6,000 x g, resuspended in freezing buffer containing 50 mM TrisHCl pH 7.5, 500 mM NaCl and 10% (v/v) glycerol supplemented with cOmplete EDTA-free protease inhibitor

cocktail (Roche), flash-frozen in liquid nitrogen and stored at -80°C until further processing.

For protein purification, bacterial cell pellets were thawed on ice and resuspended in freezing buffer supplemented with 0.5 mg/ml of DNase and 1.0 mg/ml of lysozyme (Sigma-Aldrich). Cells were sonicated and centrifuged at 30,000 x g for 45 minutes at 4°C, and the resulting supernatant was added to a 5 ml HisTrap column (GE Healthcare). All unbound material was washed off using at least 10 column volumes of washing buffer (50 mM TrisHCl pH 7.5, 500 mM NaCl, 10% (v/v) glycerol, 20 mM imidazole). Bound protein was eluted from the column using the same buffer, but with the imidazole concentration increased to 300 mM. The eluted fractions were pooled and dialyzed in the presence of TEV protease in a buffer containing (i) 20 mM CAPS pH 10.0 and 100 mM NaCl for RBP2d$_{165-967}$, (ii) 20 mM TrisHCl pH 8.5 and 100 mM NaCl for RBP2e$_{156-957}$, and (iii) 20 mM TrisHCl pH 8.0 and 100 mM NaCl for RBP3$_{149-968}$. The resulting sample was added to a 5 ml Q-Sepharose HiTrap column (GE Healthcare), with unbound material washed off using at least 10 column volumes of the buffer. The protein was eluted using a gradient (0-50%) of (i) 20 mM CAPS pH 10.0 and 1.0 M NaCl for RBP2d$_{165-967}$, (ii) 20 mM TrisHCl pH 8.5 and 1.0 M NaCl for RBP2e$_{156-957}$, and (iii) 20 mM TrisHCl pH 8.0 and 1.0 M NaCl for RBP3$_{149-968}$. Collected fractions (2.5 ml) were analyzed on SDS PAGE and those containing protein were concentrated using Vivaspin 15 Turbo centrifugal concentrators with a molecular weight cut-off 5 kDa (Sartorius) and injected onto S200 Superdex 16/600 size exclusion column (GE Healthcare) preequilibrated with (i) 20 mM NaHEPES pH 7.5, 500 mM NaCl, 5 mM β-mercaptoethanol and 5% (v/v) glycerol for RBP2d$_{165-967}$, (ii) 20 mM TrisHCl pH 8.5, 300 mM NaCl and 10% (v/v) glycerol for RBP2e$_{156-957}$, and (iii) 20 mM TrisHCl pH 8.5, 300 mM NaCl and 10% (v/v) glycerol for RBP3$_{149-968}$. Fractions containing pure proteins (2 ml) were pooled, concentrated and flash-frozen in liquid nitrogen. RBP2e$_{156-957}$ and RBP3$_{149-968}$ yielded monodisperse peaks; however, RBP2d$_{165-967}$ was heavily aggregated despite using high salt and a reducing agent in the buffer. To examine protein folding, circular dichroism (CD) data (Fig. S8) were collected using a CD spectrometer Model 410 (Aviv

Biomedical) and analyzed as described (43). Expression and purification of RBP2a$_{160\text{-}1000}$ and RBP2b$_{161\text{-}969}$ has been described previously (43, 44).

Polyclonal rabbit antibodies were generated by the Walter and Eliza Hall Institute (WEHI) Antibody Facility. Rabbits were immunized 5 times with 150 µg of the respective recombinant protein. The first immunization was administered in Complete Freund's adjuvant and the remainder in Incomplete Freund's Adjuvant. Rabbit IgG fractions were purified from serum using Protein G sepharose. These studies were approved by the WEHI Institutional Animal Care and Use Committee.

RBP binding assays. 30-100 ml of whole blood was collected in 10 ml ACD blood collection tubes (BD Biosciences) from five chimpanzees (New Iberia Research Center, Lafayette, Louisiana), one gorilla (Lincoln Park Zoo, Chicago, Illinois), and one rhesus macaque (BioIVT, Westbury, New York). All ape blood samples represented leftover specimens obtained during routine health screenings and were approved by the respective Institutional Animal Care and Use Committees. The macaque blood was purchased. Blood was also obtained from healthy human volunteers at the University of Pennsylvania under IRB protocol #813699 with informed consent (kindly provided by R. Collman). Whole blood was centrifuged at 1,500 x g for 20 minutes (maximum acceleration and low brake speed). Buffy coats containing leukocytes were removed and red blood cells were resuspended in their respective plasma or in PBS. Resuspended red blood cells were passed through a SepaCell R-500 II filter (Fenwal) or an Acrodisc filter with leukosorb media (PALL) to remove remaining leukocytes. Reticulocytes were enriched by carefully layering 5.5 ml of red blood cells diluted in plasma (50% hematocrit) over 6 ml of a 65-75% (v/v) isotonic Percoll cushion. The percent Percoll was varied to achieve maximum reticulocyte enrichment for each sample after testing a small aliquot on a 70% (v/v) isotonic Percoll cushion. Blood layered over Percoll was centrifuged at 1,650 x g for 20 min (maximum acceleration and low brake speed) and the resulting cell band at the Percoll interface

was removed, pooled, washed three times in PBS, and stored in RPMI (Gibco). Although this protocol achieved reticulocyte enrichment for human blood samples (up to 60%), this was not the case for ape and monkey blood samples. Although different centrifugation speeds (1,200 – 2,100 x g), Percoll densities (65-75%), and PBS formulations (Gibco 1x and 10x DBPS, Ambion 10x PBS pH 7.4, and human tonicity (HT)PBS were tested, we were unable to achieve reticulocyte enrichment of more than 2% for the great majority of ape and monkey samples, as measured by thiazole orange staining by flow cytometry (BD Retic-Count; BD Biosciences). The exception was a blood sample from an anemic chimpanzee, which yielded 4.0% reticulocytes after enrichment. For each chimpanzee and macaque, we selected the fraction with the highest percentage of reticulocytes for subsequent binding studies. The single gorilla blood sample was of limited quantity and thus only subjected to leukocyte filtration with no reticulocyte enrichment performed. Human control samples were diluted until the percentage of enriched reticulocytes matched those of the ape and monkey samples (0.5 – 4% by flow cytometry).

For binding assays, 800,000 red blood cells were incubated with 5 µg of recombinant protein in 100 µl of PBS with 1% (w/v) BSA at room temperature for one hour after gentle mixing. Cells were then washed with 180 µl PBS containing 1% BSA, spun at 4,000 x g for 1 minute, resuspended in 180 µl PBS containing 1% BSA, and incubated with a polyclonal rabbit antibody (1 µg) raised against the respective protein. Following incubation at room temperature for one hour, cells were washed again, and then incubated with a secondary (Alexa Fluor 647 labeled) chicken anti-rabbit antibody (1 µg) (ThermoFisher). After a final wash, 100 µl of room temperature thiazole orange (TO) (Retic-Count Reagent; BD Biosciences) was added and incubated for 30 min. Supernatant was removed and cells were resuspended in 1.2 ml PBS immediately prior to analysis on an Accuri flow cytometer. Antibody only controls were run in parallel, omitting the protein in the first incubation but otherwise following the same protocol. Approximately 200,000 events were captured for each sample. Data were analyzed using FlowJo software, gating on erythrocytes, then single cells, then TO-positive (reticulocyte) or TO-

negative (normocyte) populations. Binding signal for each sample was obtained for both reticulocyte and normocyte populations by subtracting the florescence value of the corresponding antibody-only control from the sample value. RBP binding experiments were performed three times per sample on different days, with the average of these values reported.

SI Figures

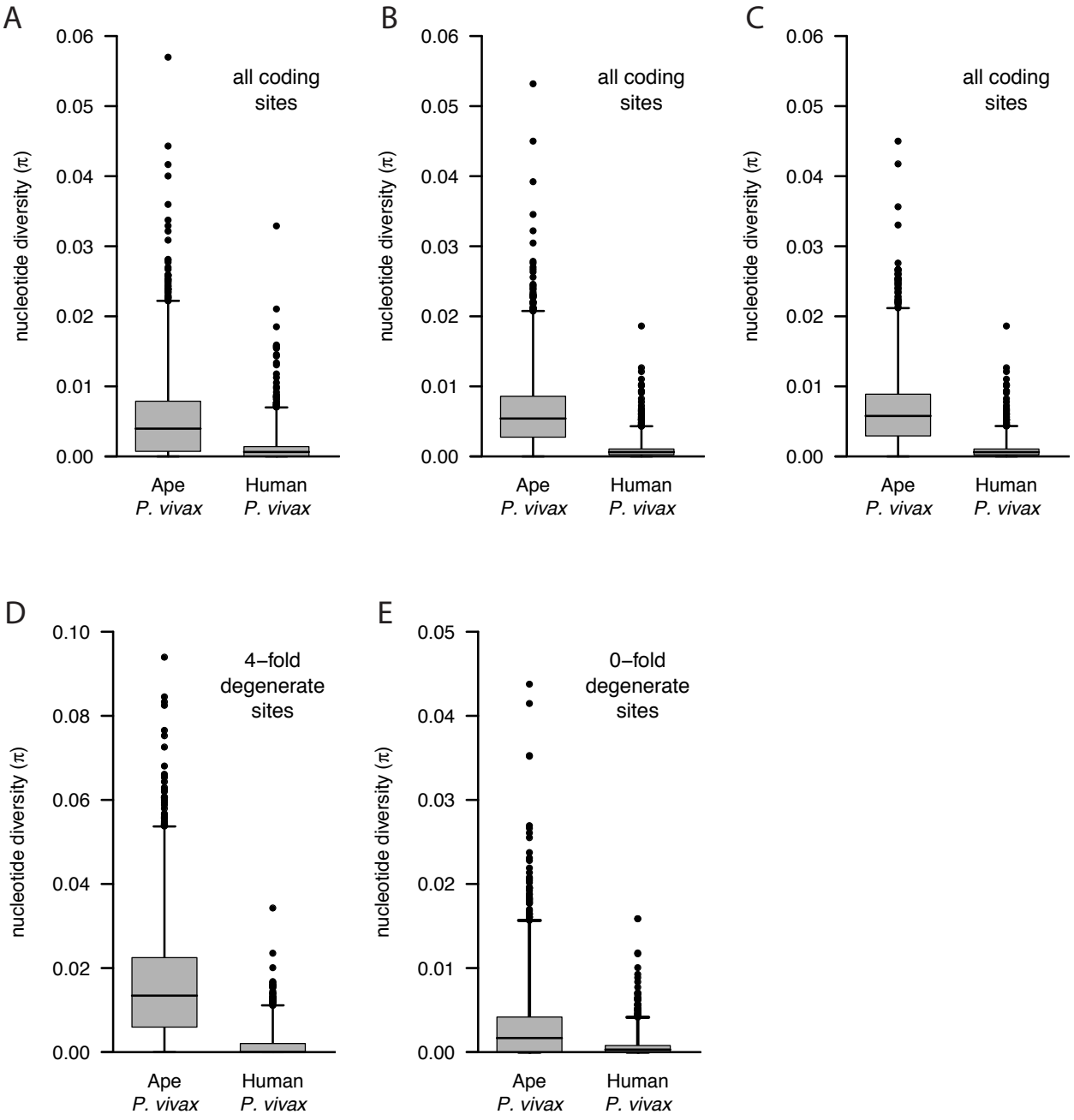

Fig. S1. Sequence diversity in ape and human *P. vivax*. (A) Nucleotide sequence diversity (π) calculated for all coding sites in 3,956 genes of two chimpanzee (PvSY43 and PvSY56) and two human (PvSalI and PvP01) *P. vivax* strains. (B) Nucleotide sequence diversity (π) for all coding sites in 4,260 genes of six chimpanzee and nine human *P. vivax* strains (Table S2). (C)

Nucleotide sequence diversity (π) for all coding sites in 3,914 genes of five chimpanzee and nine human *P. vivax* strains, after removal of the multiply infected PvSY43 sample. (D, E) Nucleotide sequence diversity calculated for chimpanzee and human *P. vivax* strains as in (B), but considering only fourfold degenerate sites (D, 3,836 genes) or zerofold degenerate sites (E, 4,260 genes). For all plots, the interquartile range is shown as a box, with the upper and lower 99th percentiles indicated by whiskers (outliers are shown as black dots). For each alignment set, genes with fewer than 35 aligned sites were excluded (2 genes in A, 3 genes in B and E, 3 genes in C, 427 genes in D) to avoid plotting spurious extreme values from very short sequences (these genes were retained in Table 2).

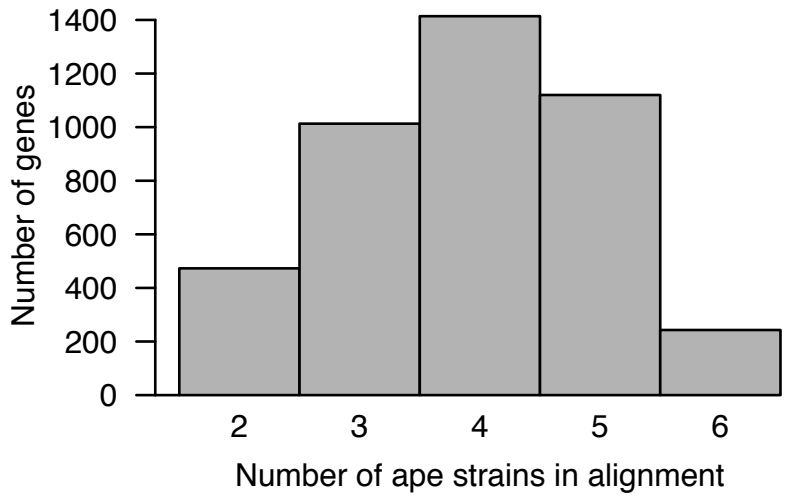

Fig. S2. Chimpanzee *P. vivax* core genes used for diversity analyses. The number of genes (y-axis) that could be compared for two or more chimpanzee *P. vivax* strains (x-axis) is indicated. PvSY56 and PvSY43 genes were included if they covered 90% or more of the length of the corresponding ortholog in the human PvP01 reference. Genes from the other chimpanzee parasites were included if at least 60% of sites in the corresponding PvSY56 reference gene were callable.

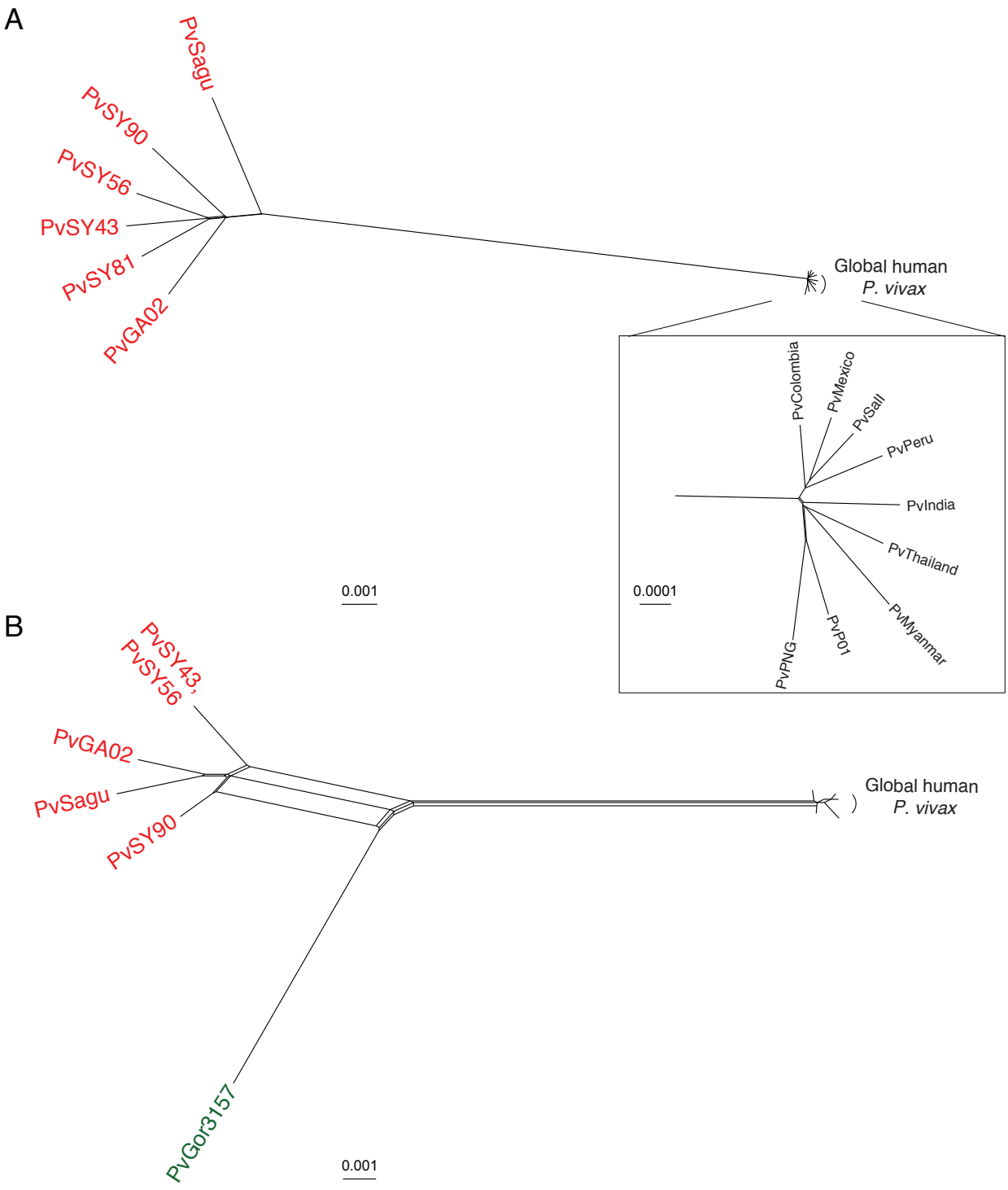

Fig. S3. Phylogenetic network analysis of ape and human *P. vivax* strains. (A) A phylogenetic network was constructed using split decomposition from pairwise distances in an

alignment of 241 nuclear genes. Nine human (black) and six chimpanzee (red) *P. vivax* strains are shown (the inset shows the human *P. vivax* strains in greater detail). The network supports the overall relationships of human and chimpanzee *P. vivax* depicted in Fig. 2A. (B) As in (A) but based on 6 nuclear genes with coverage in a gorilla *P. vivax* strain (green). The same human and chimpanzee *P. vivax* strains were also included, except for PvSY81, which did not cover these genes. The network suggests that some recombination has occurred between gorilla and chimpanzee parasites in these genes, but supports the overall topology shown in Fig. 2B. The scale bars represent substitutions per site.

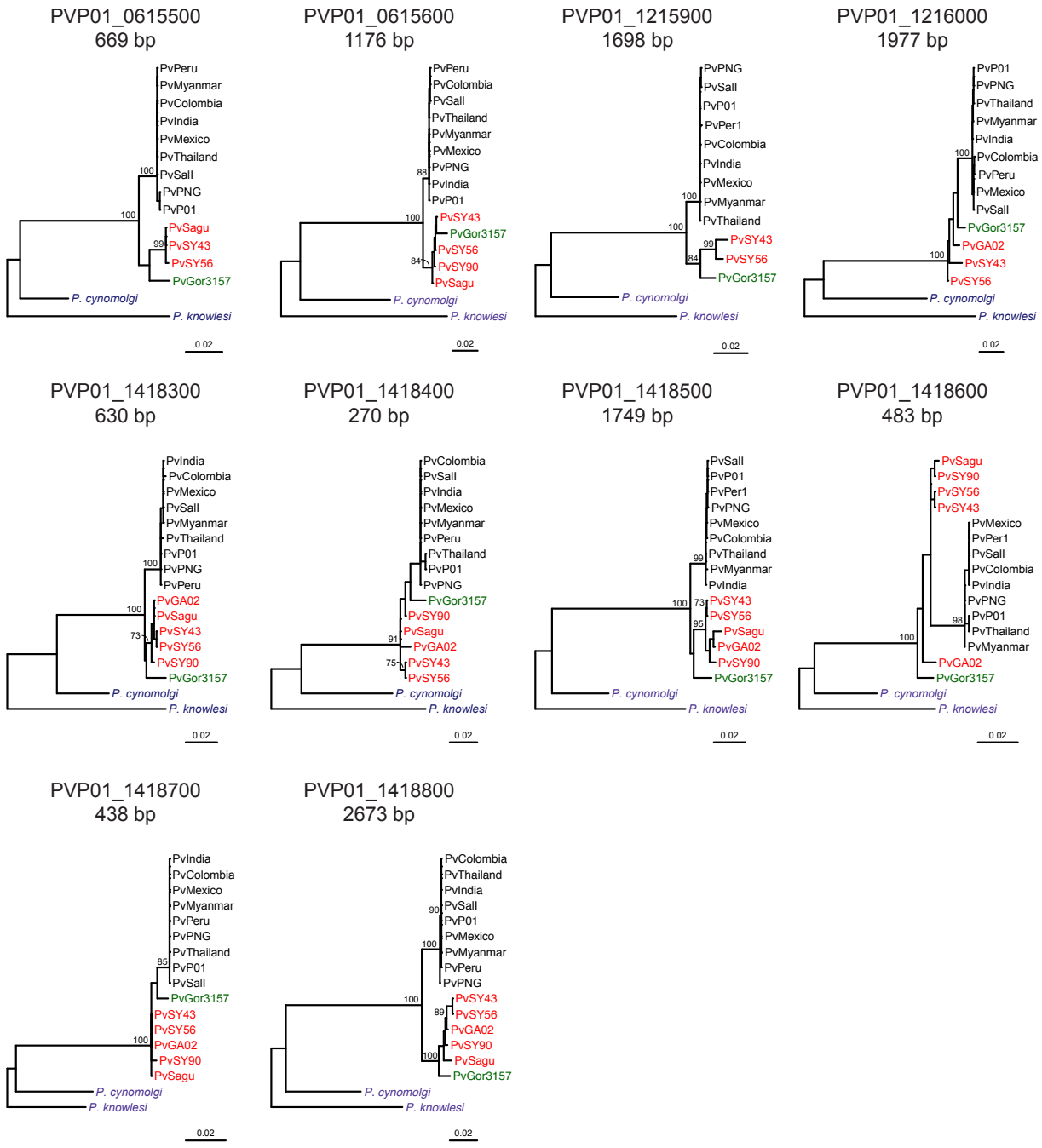

Fig. S4. Phylogenetic relationships of ape and human *P. vivax* using whole genome and SNP data. Maximum likelihood trees, rooted with *P. knowlesi*, are shown for 10 nuclear genes (gene names and lengths for the PvP01 reference are indicated). *P. vivax* sequences from humans, chimpanzees and gorillas are shown in black, red and green, respectively. The

monkey parasite species *P. cynomolgi* (strain M) and *P. knowlesi* (strain H) were included as outgroups (purple). Bootstrap values ≥70% are shown. The scale bars represent 0.02 substitutions per site.

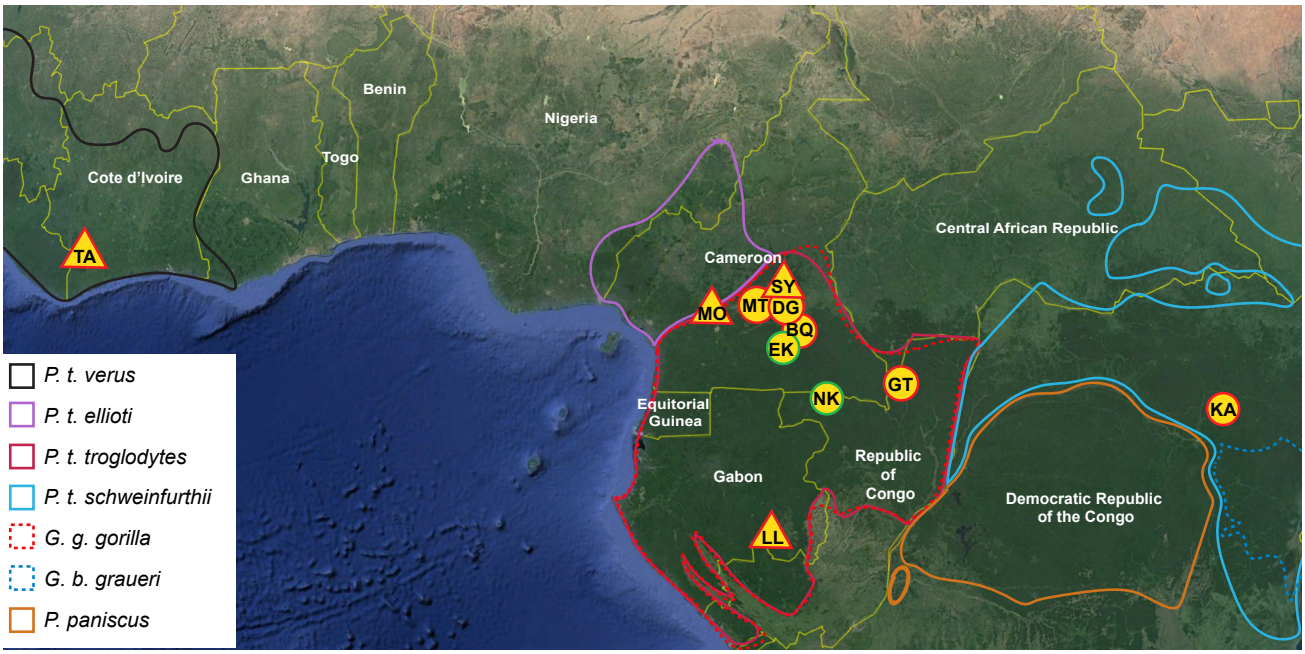

Fig. S5. Geographical origin of *P. vivax*-positive ape samples. Study sites are shown in relation to the natural ranges of chimpanzees (*Pan troglodytes verus,* black; *P. t. ellioti,* purple; *P. t. troglodytes,* red; *P. t. schweinfurthii,* blue), bonobos (*Pan paniscus,* orange), western lowland gorillas (*Gorilla gorilla gorilla,* dashed red) and eastern lowland gorillas (*Gorilla beringei graueri,* dashed blue) (45). Circles denote forest sites, while triangles indicate the location of sanctuaries, with colors denoting whether chimpanzee (red) or gorilla (green) *P. vivax* sequences were obtained (BQ, Belgique; DG, Diang; EK, E'kom; GT, Goualougo Triangle; KA, Kabuka; MO, Mfou National Park Wildlife Rescue Center; MT, Minta; NK, Ndongo; SY, Sanaga-Yong Chimpanzee Rescue Center; TA, Tai Forest). The chimpanzee infected with PvGA02 was previously reported (27) to have been sampled in Park of La Lékédi (LL).

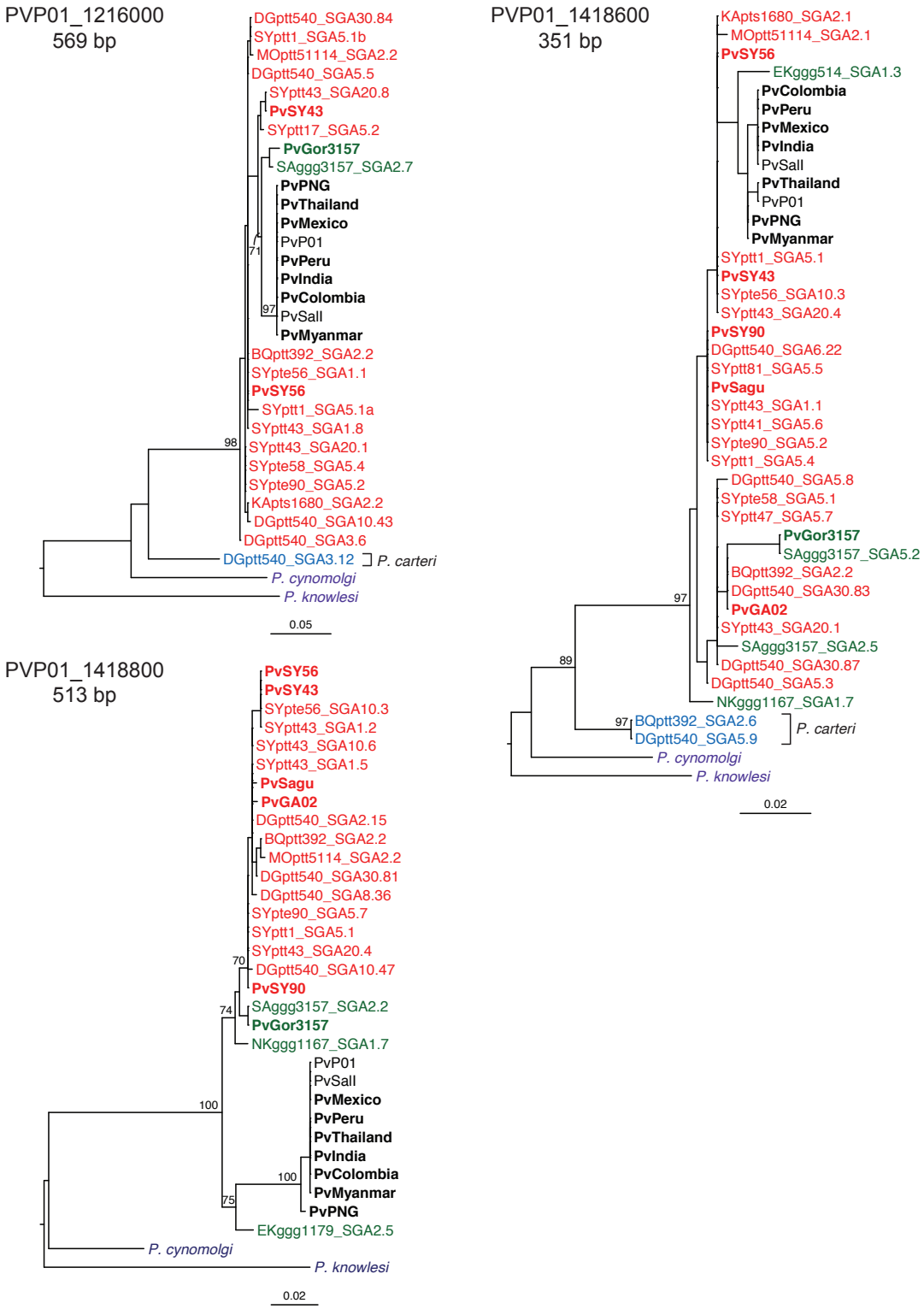

Fig. S6. Phylogenetic relationships of ape and human *P. vivax*. Maximum likelihood trees rooted with *P. knowlesi* are shown for fragments of three nuclear genes (PVP01_1216000,

PVP01_1418600, PVP01_1418800; the fragment size in PvP01 is indicated). *P. vivax* sequences from humans, chimpanzees and gorillas are shown in black, red and green, respectively. Sequences generated by SWGA or derived from published data (see Table S2) are shown in bold. SGA-amplified sequences from ape fecal and blood samples include a two-letter code to denote the field site (Fig. S5), lower case letters to indicate their species origin (ptt: *P. t. troglodytes*, red; pte: *P. t. ellioti*, red; pts: *P. t. schweinfurthii,* red; ggg: *G. g. gorilla*, green), the sample number, as well as the SGA dilution and well position (e.g., EKggg1179_SGA2.5 represents an SGA derived sequence amplified from a 1:2 dilution of fecal DNA from a western lowland gorilla sample 1179 and identified at position 5 of a plate of multiple PCR reactions). Sequences of *P. carteri*, a parasite species closely related to *P. vivax* that has thus far only been identified in wild-living chimpanzees, are shown in blue. GenBank accession numbers for SGA sequences are listed in Table S4. The monkey parasite species *P. cynomolgi* (strain M) and *P. knowlesi* (strain H) were included as outgroups (purple). Bootstrap values ≥70% are shown for clades containing at least two non-identical tips. The scale bars represent substitutions per site.

Fig. S7. **Ape *P. vivax* encode three intact *rbp* genes that are pseudogenized in human *P. vivax***. Alignments are shown for chimpanzee (red), gorilla (green) and human (black) *P.*

vivax rbp2d, rbp2e and *rbp3* partial gene sequences (numbers indicate the nucleotide position within the respective gene). 'Pv' denotes sequences from genome-wide analyses (Table S2), which were extracted from assembled genomes (PvSY43, PvSY56 and PvP01) or generated by mapping sequencing reads to the PvSY56 reference and calling bases covered by ≥3 reads (PvSY81, PvSY90, PvSagu, PvGA02). All gorilla *P. vivax* sequences were derived by single genome amplification (SGA) from the same multiply infected sample (SAggg3157) and represent individual parasites. Nucleotides that differ from the PvSY56 reference are highlighted, with dots indicating sequence identity. Blank spaces indicate insufficient read coverage. Inactivating mutations that cause frameshift and premature stop codons in all published human *P. vivax* strains are boxed.

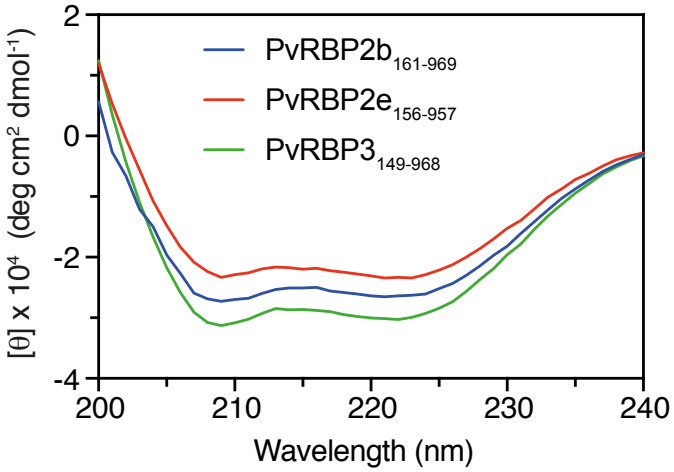

Fig. S8. Circular dichroism spectra of recombinant RBP proteins from chimpanzee *P. vivax* strains. Far UV circular dichroism spectra of two newly expressed chimpanzee *P. vivax* proteins PvRBP2e$_{156-957}$ (red) and PvRBP3$_{149-968}$ (green) are compared to the CD spectrum of the human *P. vivax* PvRBP2b$_{161-969}$ protein (blue) as previously reported (44). The mean residue molar ellipticity (θ, y-axis) is plotted relative to the wavelength (nm; x-axis). The three spectra superimpose, suggesting proper folding of the newly expressed proteins.

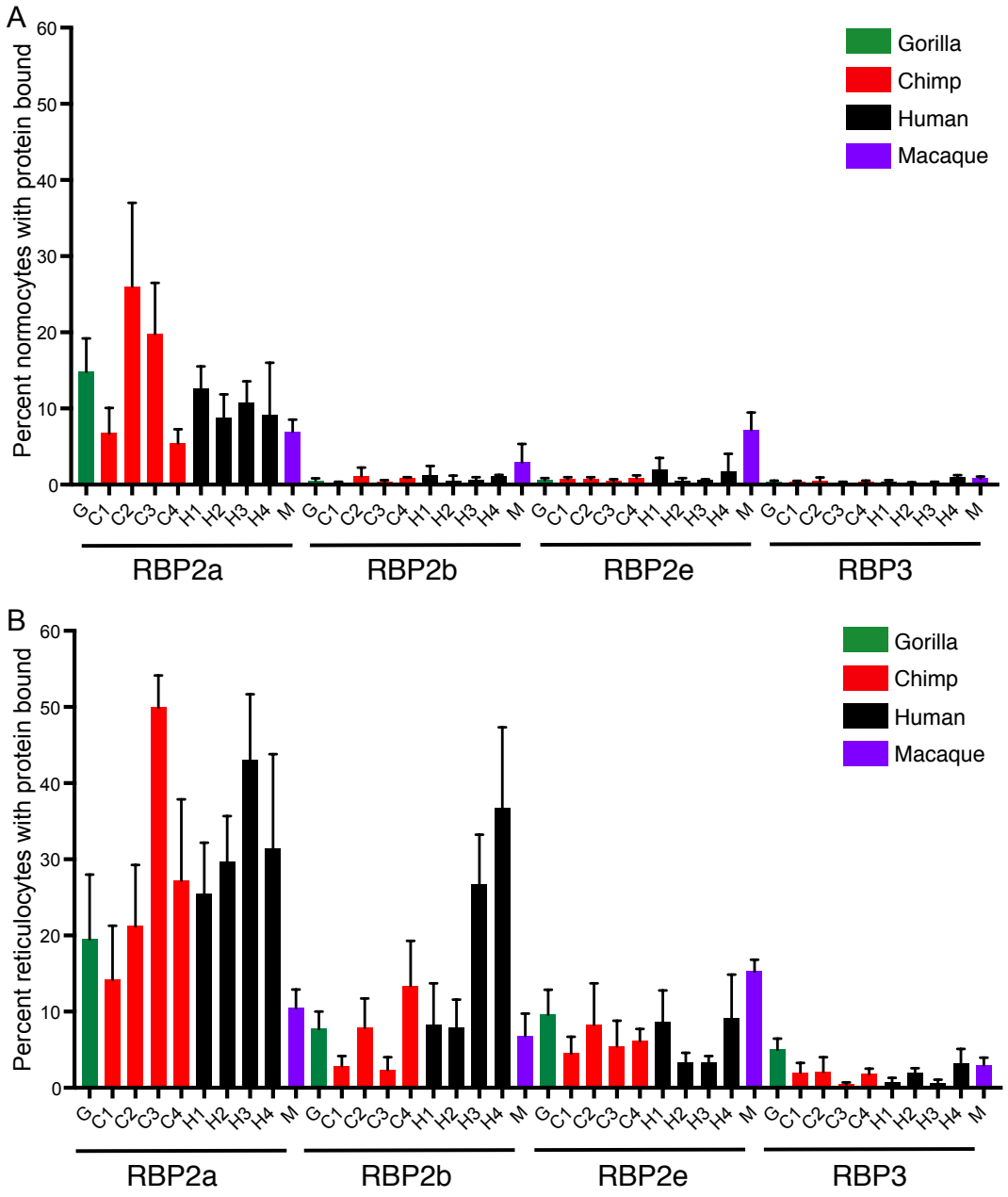

Fig. S9. Binding of chimpanzee and human *P. vivax* RBP proteins to red blood cells from different host species. The binding of human *P. vivax* RBP2a and RBP2b and chimpanzee *P. vivax* RBP2e and RBP3 proteins to normocytes (A) and reticulocytes (B) from one gorilla (G), four chimpanzees (C1-C4), four humans (H1-H4), and one macaque (M) is shown. Columns indicate the percentage of red blood cells (RBCs) that bound the respective RBP proteins after

subtracting background binding in the absence of protein. Error bars represent one standard deviation. All red blood cell preparations were tested three times in independent technical replicates.

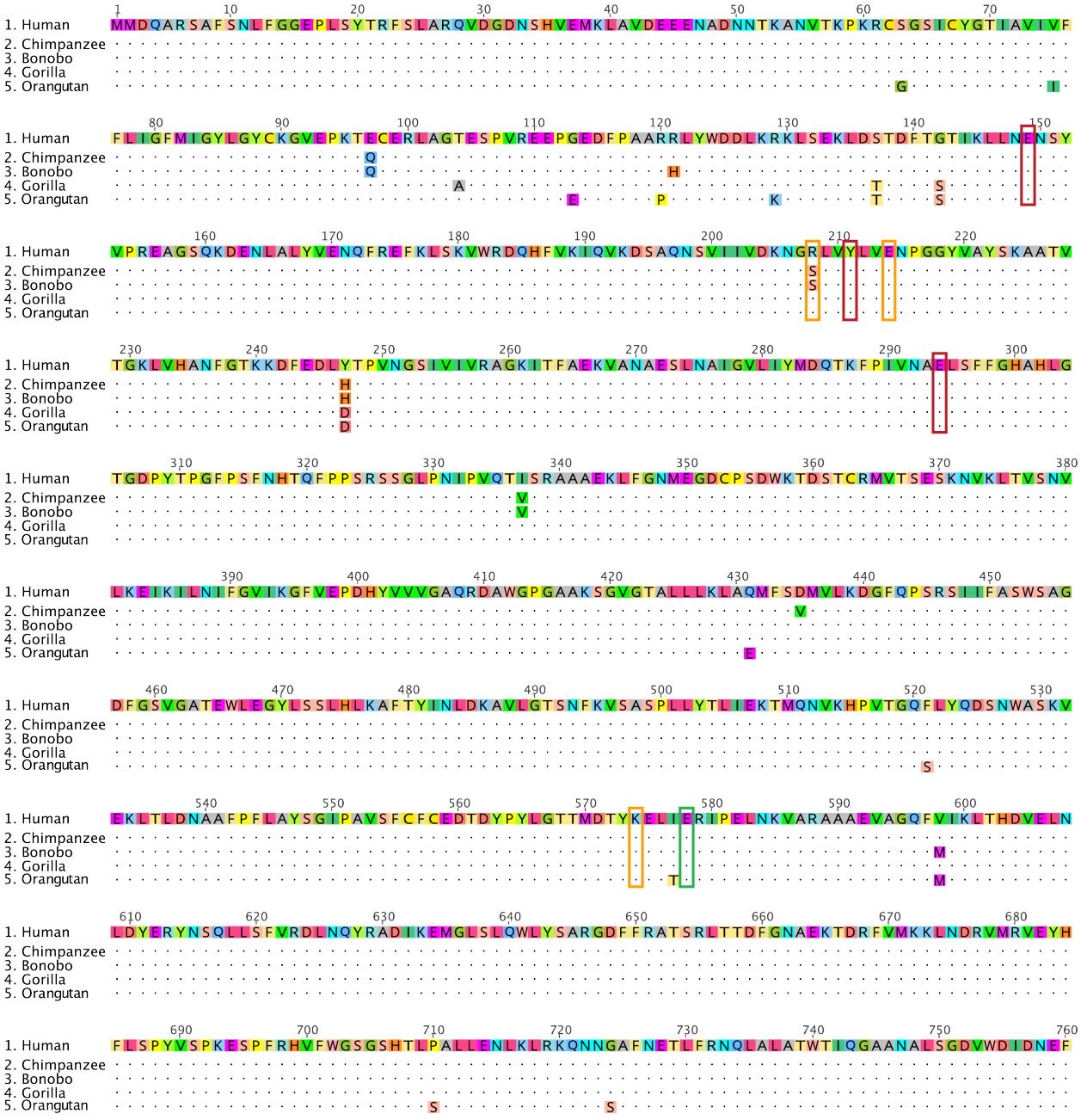

Fig S10. Alignment of human and ape transferrin receptor 1 (TfR1) protein sequences. An alignment of the deduced amino acid sequences of human (GRCh38, NC_000003.12), chimpanzee (*Pan troglodytes verus*, XM_003310191.3), bonobo (*Pan paniscus,* XM_003806407.2), gorilla (*Gorilla gorilla gorilla,* XM_004038250.2), and orangutan (*Pongo abelii,* NM_001131591.1) TfR1 sequences is shown. Residues that differ from the human

sequence are highlighted, with dots indicating sequence identity. Alanine mutagenesis of TfR1 residues that were identified to form stacking interactions or salt bridges with RBP2b are boxed (46), with border color indicating the degree to which the mutation abrogated invasion complex formation (no effect, green; moderate effect, orange; severe effect, red).

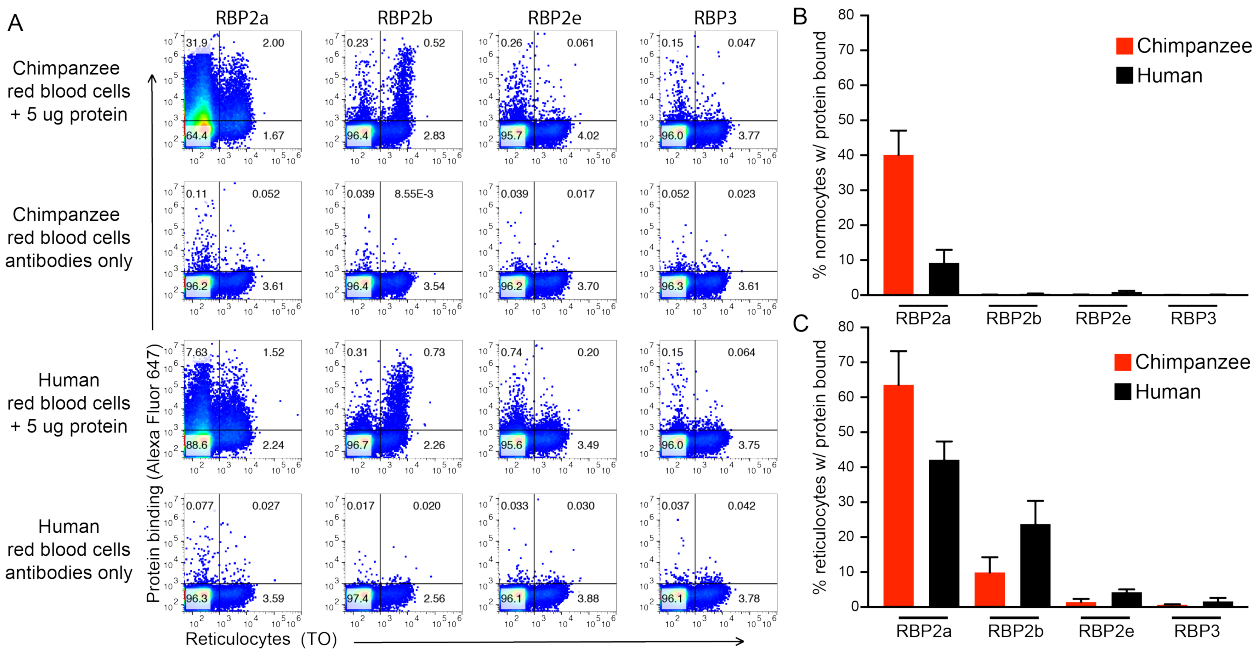

Fig. S11. Binding of chimpanzee and human *P. vivax* RBP proteins to reticulocyte-enriched chimpanzee and human red blood cells. (A) Dot plots are shown that depict the binding of human *P. vivax* RBP2a and RBP2b and chimpanzee *P. vivax* RBP2e and RBP3 proteins to chimpanzee (first row) and human (third row) red blood cells. Both the chimpanzee and the human blood sample contained a large fraction of reticulocytes (the human sample was diluted so that equivalent numbers of reticulocytes were tested in the binding assays). The x-axis depicts thiazole orange (TO) staining of reticulocytes; the y-axis indicates protein binding as detected using an RBP-specific polyclonal rabbit antibody, followed by a secondary chicken (Alexa Fluor 647 labeled) anti-rabbit antibody. The position of gates denoting normocytes versus reticulocytes as well as protein binding versus no protein binding are shown by vertical and horizontal lines, respectively. Numbers indicate the percentage of total cells within the respective gate. Antibody-only negative controls in which no protein was added are shown in the second (chimpanzee cells) and fourth (human cells) rows, respectively. (B, C) The percentage of chimpanzee (red) and human (black) normocytes (B) and reticulocytes (C) that

bound the respective RBP proteins are shown. Three independent replicates were performed and background signal from the antibody-only control was subtracted (see SI Appendix, Materials and Methods).

SI Tables

Table S1. Select whole genome amplification and sequencing of ape *P. vivax* genomes

Blood Sample*	Digest†	pvset1#	pvset2#	pvset3#	pvset4#	pvset5#	pvset6#	Illumina sequencing‡	PacBio sequencing§
SY43	+	+				+		+	+
SY56	+	+	+	+		+	+	+	+
SY81	+	+				+	+	+	
SY90	+	+				+	+	+	
Sagu		+		+		+	+	+	
Gor3157	+	+			+	+	+	+	

**P. vivax* genomes were amplified by SWGA from whole blood DNA of four sanctuary chimpanzees (SY43, SY56, SY81, SY90) housed at the Sanaga-Yong (SY) Chimpanzee Rescue Center, one wild-living habituated chimpanzee (Sagu) from the Tai forest, and one western lowland gorilla bushmeat sample (Gor3157) confiscated by the Cameroonian Ministry of Environment and Forestry.

#pvset, set of *P. vivax*-specific SWGA primers (see text for primer sequences).

†Aliquots of whole blood DNA were digested with the methylation dependent restriction enzymes MspJI and FspEI prior to SWGA to selectively degrade host DNA (2).

‡Illumina sequencing was performed on MiSeq or MiniSeq platforms.

§PacBio sequencing was performed on a PacBio RS II platform.

Table S2. Host species and geographic origin of ape and human *P. vivax* isolates

P. vivax strain	Host species*	Country	Genes analyzed	Reference
PvSY43	*P.t.t.*	Cameroon	3,974	This study
PvSY56	*P.t.e.*	Cameroon	4,263	This study
PvSY81	*P.t.t.*	Cameroon	695	This study
PvSY90	*P.t.e.*	Cameroon	2,220	This study
PvSagu	*P.t.v.*	Cote d'Ivoire	2,542	This study
PvGor3157	*G.g.g.*	Cameroon	10	This study
PvGA02[#]	*P.t.t.*	Gabon	3,005	27
PvColombia	human	Colombia	4,257	29
PvIndia	human	India	4,258	29
PvMexico	human	Mexico	4,258	29
PvMyanmar	human	Myanmar	4,258	29
PvPeru	human	Peru	4,259	29
PvPNG	human	Papua New Guinea	4,260	29
PvThailand	human	Thailand	4,254	29
PvSalI	human	El Salvador	4,242	10
PvP01	human	Indonesia	4,263	24

**P.t.t., Pan troglodytes troglodytes*; *P.t.e., P. t. ellioti; P.t.v., P. t. verus; G.g.g., Gorilla gorilla gorilla*;
[#]PvGA02 sequences were derived from a read database generated from the blood of a sanctuary chimpanzee (GA02) that was coinfected with ape *P. malariae* and *P. vivax* (27).

Table S3. Polymorphisms between chimpanzee and human *P. vivax* and *P. cynomolgi*

Parasites	Polymorphisms in *P. vivax*			Fixed differences from *P. cynomolgi*			
	NS*	S*	NS/S	NS†	S†	NS/S	NI‡
Chimpanzee *P. vivax*#	27,605	41,581	0.66	266,922	384,495	0.69	0.96
Human *P. vivax*#	8,906	6,728	1.32	277,933	401,852	0.69	1.91

*The number of nonsynonymous (NS) and synonymous (S) polymorphisms was calculated by counting the number of SNPs that changed (NS) or did not change (S) the amino acid sequence of PvSY56 (ape *P. vivax*) or PvP01 (human *P. vivax*).

†The number of nonsynonymous (NS) and synonymous (S) differences was calculated between *P. cynomolgi* strain M and PvSY56 (ape *P. vivax*) or PvP01 (human *P. vivax*); sites that were polymorphic in ape or human *P. vivax* were excluded from the respective comparisons.

‡NI, neutrality index.

#A common set of 3,913 genes was compared among six chimpanzee and nine human *P. vivax* strains as well as between these parasites and *P. cynomolgi* strain M.

.

Table S4. GenBank Accession numbers of SGA-derived ape *P. vivax* sequences

Sequence name*	Gene	Accession No.
DGptt540_SGA3.6	PVP01_1216000	MH443156
KApts1680_SGA2.2	PVP01_1216000	MH443160
BQptt392_SGA2.2	PVP01_1216000	MH443154
SYptt43_SGA20.1	PVP01_1216000	MH443166
SYpte58_SGA5.4	PVP01_1216000	MH443166
SYpte90_SGA5.2	PVP01_1216000	MH443166
DG540_SGA5.5	PVP01_1216000	MH443155
SYptt1_SGA5.1a	PVP01_1216000	MH443163
SYptt43_SGA1.8	PVP01_1216000	MH443165
SYpte56_SGA1.1	PVP01_1216000	MH443165
DGptt540_SGA30.84	PVP01_1216000	MH443159
SYptt1_SGA5.1b	PVP01_1216000	MH443159
MOptt51114_SGA2.2	PVP01_1216000	MH443161
DGptt540_SGA10.43	PVP01_1216000	MH443158
SAggg3157_SGA2.7	PVP01_1216000	MH443162
SYptt17_SGA5.2	PVP01_1216000	MH443164
SYptt43_SGA20.8	PVP01_1216000	MH443167
DGptt540_SGA3.12	PVP01_1216000	MH443157
DGptt540_SGA5.6	PVP01_1418300	MH443172
DGptt540_SGA3.11	PVP01_1418300	MH443170
DGptt540_SGA10.41	PVP01_1418300	MH443174
DGptt540_SGA6.21	PVP01_1418300	MH443173
SYptt43_SGA20.1	PVP01_1418300	MH443187
SYptt43_SGA20.4	PVP01_1418300	MH443188
SYptt1_SGA5.7	PVP01_1418300	MH443184
SYptt41_SGA5.6	PVP01_1418300	MH443186
SAggg3157_SGA2.4	PVP01_1418300	MH443179
SAggg3157_SGA5.7	PVP01_1418300	MH443181
SAggg3157_SGA2.5	PVP01_1418300	MH443180
SYpte58_SGA5.1	PVP01_1418300	MH443182
SYptt81_SGA5.5	PVP01_1418300	MH443189
DGptt540_SGA30.88	PVP01_1418300	MH443176
MOptt51114_SGA2.1	PVP01_1418300	MH443178
DGptt540_SGA15.58	PVP01_1418300	MH443175
SYpte90_SGA5.3	PVP01_1418300	MH443183
SYptt17_SGA5.2	PVP01_1418300	MH443185
GTptt314_SGA1.3	PVP01_1418300	MH443177
MTptt347_SGA1.3	PVP01_1418300	MH443177
BQptt392_SGA2.2	PVP01_1418300	MH443168
DGptt540_SGA5.3	PVP01_1418300	MH443171
BQptt392_SGA10.46	PVP01_1418300	MH443169
SYpte56_SGA10.2	PVP01_1418500	MH443190

SYptt43_SGA1.4	PVP01_1418500	MH443190
BQptt392_SGA2.2	PVP01_1418500	MH443191
DGptt540_SGA6.22	PVP01_1418500	MH443192
SYptt1_SGA5.5	PVP01_1418500	MH443193
MOptt51114_SGA2.1	PVP01_1418500	MH443195
MOptt51114_SGA2.6	PVP01_1418500	MH443194
SYptt43_SGA10.1	PVP01_1418500	MH443196
DGptt540_SGA5.7	PVP01_1418500	MH443197
SYptt43_SGA20.1	PVP01_1418500	MH443198
SYptt43_SGA1.6	PVP01_1418500	MH443199
SYptt1_SGA5.8	PVP01_1418500	MH443200
SYpte58_SGA5.1	PVP01_1418500	MH443201
SYptt43_SGA20.7	PVP01_1418500	MH443202
DGptt540_SGA3.9	PVP01_1418500	MH443203
SYptt81_SGA5.7	PVP01_1418500	MH443204
BQptt392_SGA2.6	PVP01_1418600	MH443205
DGptt540_SGA5.9	PVP01_1418600	MH443205
SYpte58_SGA5.1	PVP01_1418600	MH443206
SYptt43_SGA20.1	PVP01_1418600	MH443206
SYptt47_SGA5.7	PVP01_1418600	MH443206
DGptt540_SGA30.87	PVP01_1418600	MH443206
SAggg3157_SGA2.5	PVP01_1418600	MH443207
DGptt540_SGA5.8	PVP01_1418600	MH443208
BQptt392_SGA2.2	PVP01_1418600	MH443209
DG540_SGA30.83	PVP01_1418600	MH443209
DGptt540_SGA5.3	PVP01_1418600	MH443210
MOptt51114_SGA2.1	PVP01_1418600	MH443211
SYpttt1_SGA5.1	PVP01_1418600	MH443212
SYptt43_SGA20.4	PVP01_1418600	MH443212
KApts1680_SGA2.1	PVP01_1418600	MH443212
SYpte56_SGA10.3	PVP01_1418600	MH443212
SYptt43_SGA1.1	PVP01_1418600	MH443213
DGptt540_SGA6.22	PVP01_1418600	MH443213
SYptt1_SGA5.4	PVP01_1418600	MH443213
SYptt41_SGA5.6	PVP01_1418600	MH443213
SYptt81_SGA5.5	PVP01_1418600	MH443213
SYpte90_SGA5.2	PVP01_1418600	MH443213
SAggg3157_SGA5.2	PVP01_1418600	MH443214
EKggg514_SGA1.3	PVP01_1418600	MH443215
NKggg1167_SGA1.7	PVP01_1418600	MH443216
SYpte56_SGA10.3	PVP01_1418800	MH443217
SYptt43_SGA1.2	PVP01_1418800	MH443217
SYptt43_SGA20.4	PVP01_1418800	MH443218
SYptt1_SGA5.1	PVP01_1418800	MH443218
SYpte90_SGA5.7	PVP01_1418800	MH443218

DGptt540_SGA10.47	PVP01_1418800	MH443219
EKggg1179_SGA2.5	PVP01_1418800	MH443220
NKggg1167_SGA1.7	PVP01_1418800	MH443221
DGptt540_SGA2.15	PVP01_1418800	MH443222
SYptt43_SGA1.5	PVP01_1418800	MH443222
DGptt540_SGA30.81	PVP01_1418800	MH443223
MOptt51114_SGA2.2	PVP01_1418800	MH443224
BQptt392_SGA2.2	PVP01_1418800	MH443225
DGptt540_SGA8.36	PVP01_1418800	MH443226
SYptt43_SGA10.6	PVP01_1418800	MH443227
SAggg3157_SGA2.2	PVP01_1418800	MH443228

*Sequence names include a two-letter code indicating the field site of origin (Fig. S5), lower case letters denoting species and subspecies origin (ptt: *P. t. troglodytes*; pte: *P. t. ellioti*; pts: *P. t. schweinfurthii*; ggg: *G. g. gorilla*), and a sample number, followed by SGA dilution and well position (e.g., EKggg1179_SGA2.5 represents an SGA derived sequence amplified from a 1:2 dilution of fecal DNA from sample 1179 collected from a western lowland gorilla at field site EK, and identified at position 5 of a plate of multiple PCR reactions). SAggg3157 sequences are derived from a confiscated gorilla bushmeat sample of unknown origin.

SI References

1. Liu W, et al. (2010) Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. *Nature* 467:420–425.
2. Sundararaman SA, et al. (2016) Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. *Nat Commun* 7:11078.
3. Köndgen S, et al. (2011) *Pasteurella multocida* involved in respiratory disease of wild chimpanzees. *PLoS One* 6:e24236.
4. Liu W, et al. (2016) Multigenomic delineation of *Plasmodium* species of the *Laverania* subgenus infecting wild-living chimpanzees and gorillas. *Genome Biol Evol* 8:1929–1939.
5. Liu W, et al. (2014) African origin of the malaria parasite *Plasmodium vivax*. *Nat Commun* 5:3346.
6. Liu W, et al. (2017) Wild bonobos host geographically restricted malaria parasites including a putative new *Laverania* species. *Nat Commun* 8:1685.
7. Kaiser M, et al. (2010) Wild chimpanzees infected with 5 *Plasmodium* species. *Emerg Infect Dis* 16:1956–1959.
8. Leichty AR, Brisson D (2014) Selective whole genome amplification for resequencing target microbial species from complex natural samples. *Genetics* 198:473–481.
9. Cowell AN, et al. (2017) Selective whole-genome amplification is a robust method that enables scalable whole-genome sequencing of *Plasmodium vivax* from unprocessed clinical samples. *MBio* 8:e02257.
10. Carlton JM, et al. (2008) Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* 455:757–763.
11. Clarke EL, et al. (2017) *swga*: a primer design toolkit for selective whole genome amplification. *Bioinformatics* 33:2071–2077.
12. Bankevich A, et al. (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477.

13. Li H, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.

14. Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

15. Chaisson MJ, Tesler G (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory. *BMC Bioinformatics* 13:238.

16. Hackl T, Hedrich R, Schultz J, Förster F (2014) Proovread: Large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* 30:3004–3011.

17. Kearse M, et al. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28:1647–1649.

18. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M (2009) ABACAS: Algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* 25:1968–1969.

19. Piro VC, et al. (2014) FGAP: An automated gap closing tool. *BMC Res Notes* 7:371.

20. Tsai IJ, Otto TD, Berriman M (2010) Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol* 11:R41.

21. Nadalin F, Vezzi F, Policriti A (2012) GapFiller: A *de novo* assembly approach to fill the gap within paired reads. *BMC Bioinformatics* 13:S8.

22. Otto TD, Sanders M, Berriman M, Newbold C (2010) Iterative correction of reference nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* 26:1704–1707.

23. Pearson R, et al. (2016) Genomic analysis of local variation and recent evolution in *Plasmodium vivax*. *Nat Genet* 48:959–964.

24. Auburn S, et al. (2016) A new *Plasmodium vivax* reference sequence with improved assembly of the subtelomeres reveals an abundance of *pir* genes. *Wellcome Open Res*

1:4.

25. Steinbiss S, et al. (2016) Companion: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res* 44:W29–W34.
26. Otto TD, Dillon GP, Degrave WS, Berriman M (2011) RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res* 39:e57.
27. Rutledge GG, et al. (2017) *Plasmodium malariae* and *P. ovale* genomes provide insights into malaria parasite evolution. *Nature* 542:101–104.
28. Otto TD, et al. (2014) Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat Commun* 5:4754.
29. Hupalo DN, et al. (2016) Population genomics studies identify signatures of global dispersal and drug resistance in *Plasmodium vivax*. *Nat Genet* 48:953–958.
30. Auwera GA Van der, et al. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43:11.10.1-11.10.33.
31. Camacho C, et al. (2009) BLAST+: Architecture and applications. *BMC Bioinformatics* 10:421.
32. Pasini EM, et al. (2017) An improved *Plasmodium cynomolgi* genome assembly reveals an unexpected methyltransferase gene expansion. *Wellcome Open Res* 2:42.
33. Pain A, et al. (2008) The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* 455:799–803.
34. Abascal F, Zardoya R, Telford MJ (2010) TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res* 38:W7–W13.
35. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.
36. Keightley PD, Jackson BC (2018) Inferring the probability of the derived versus the ancestral allelic state at a polymorphic site. *Genetics* early online

doi.org/10.1534/genetics.118.301120

37. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Asessing the performance of PhyML 2.0. *Syst Biol* 59:307–321.

38. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23:254–267.

39. Tachibana SI, et al. (2012) *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. *Nat Genet* 44:1051–1055.

40. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.

41. Hietanen J, et al. (2016) Gene models, expression repertoire, and immune response of *Plasmodium vivax* reticulocyte binding proteins. *Infect Immun* 84:677–685.

42. van den Ent F, Löwe J (2006) RF cloning: a restriction-free method for inserting target genes into plasmids. *J Biochem Biophys Methods* 67:67–74.

43. Gruszczyk J, et al. (2016) Structurally conserved erythrocyte-binding domain in *Plasmodium* provides a versatile scaffold for alternate receptor engagement. *Proc Natl Acad Sci* 113:E191–E200.

44. Gruszczyk J, et al. (2018) Transferrin receptor 1 is a reticulocyte-specific receptor for *Plasmodium vivax*. *Science* 359:48–55.

45. Caldecott J, Miles L (2005) *World Atlas of Great Apes and their Conservation* (Univ of California Press).

46. Gruszczyk J, et al. (2018) Cryo-EM structure of an essential *Plasmodium vivax* invasion complex. *Nature* 559:135-139.