

Published in final edited form as:

*Nat Genet.* 2012 September ; 44(9): 1051–1055. doi:10.1038/ng.2375.

## ***Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade**

Shin-Ichiro Tachibana<sup>1,†</sup>, Steven A. Sullivan<sup>2</sup>, Satoru Kawai<sup>3</sup>, Shota Nakamura<sup>4</sup>, Hyunjae R. Kim<sup>2</sup>, Naohisa Goto<sup>4</sup>, Nobuko Arisue<sup>5</sup>, Nirianne M. Q. Palacpac<sup>5</sup>, Hajime Honma<sup>1,5</sup>, Masanori Yagi<sup>5</sup>, Takahiro Tougan<sup>5</sup>, Yuko Katakai<sup>6</sup>, Osamu Kaneko<sup>7</sup>, Toshihiro Mita<sup>8</sup>, Kiyoshi Kita<sup>9</sup>, Yasuhiro Yasutomi<sup>10</sup>, Patrick L. Sutton<sup>2</sup>, Rimma Shakhbatyan<sup>2</sup>, Toshihiro Horii<sup>5</sup>, Teruo Yasunaga<sup>4</sup>, John W. Barnwell<sup>11</sup>, Ananias A. Escalante<sup>12</sup>, Jane M. Carlton<sup>2,13</sup>, and Kazuyuki Tanabe<sup>1,5,13</sup>

<sup>1</sup>Laboratory of Malariology, Research Institute for Microbial Diseases, Osaka University, Suita, Osaka 565-0871, Japan

<sup>2</sup>Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY 10003, United States of America

<sup>3</sup>Laboratory of Tropical Medicine and Parasitology, Institute of International Education and Research, Dokkyo Medical University, Shimotsuga, Tochigi 321-0293, Japan

<sup>4</sup>Genome Information Research Center, Research Institute for Microbial Diseases, Osaka University, Suita, Osaka 565-0871, Japan

<sup>5</sup>Department of Molecular Protozoology, Research Institute for Microbial Diseases, Osaka University, Suita, Osaka 565-0871, Japan

<sup>6</sup>The Corporation for Production and Research of Laboratory Primates, Tsukuba, Ibaraki 305-0843, Japan

<sup>7</sup>Department of Protozoology, Institute of Tropical Medicine (NEKKEN) and GCOE, Nagasaki University, Nagasaki 852-8523, Japan

Correspondence should be addressed to K. Tanabe (kztanabe@biken.osaka-u.ac.jp) or J. Carlton (jane.carlton@nyu.edu).

<sup>13</sup>These authors jointly directed this work

<sup>†</sup>Present address: Career-Path Promotion Unit for Young Life Scientists, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

### URLs

PlasmoDB: <http://plasmodb.org>

Circos: <http://circos.ca/>

MISA: <http://pgrc.ipk-gatersleben.de/misa/>

**Accession codes.** The *P. cynomolgi* B, Cambodian and Berok strains have been deposited in DDBJ/EMBL/GenBank databases with the following accession numbers: B strain sequence reads DRA000196, genome assembly BAEJ01000001 to BAEJ01003341, and annotation DF157093-DF158755; Cambodian strain sequence reads DRA000197; Berok strain sequence reads SRA047950.1. SNP calls have been submitted to dbSNP (handle: NYU\_CGSB\_BIO; batch id: 1056645). Sequences of the *dbp* genes from *P. cynomolgi* (Cambodian strain), *P. fieldi* (A.b.i. strain), and *P. simiovale* (AB617788-AB617791), and *P. cynomolgi* Berok strain (JQ422035-JQ422036), and *rbp* gene sequences from *P. cynomolgi* Berok and Cambodian (JQ422037-JQ422050), were deposited. The complete apicoplast genome of *P. cynomolgi* Berok was also deposited (JQ522954). The *P. cynomolgi* B reference genome is also available through PlasmoDB (see URLs).

### AUTHOR CONTRIBUTIONS

K.T., J.M.C., A.A.E., and J.W.B. conceived and conducted the study. S.K., Y.K., Y.Y., S.I.T. and J.W.B. provided *P. cynomolgi* material. S.N., N.G., T.Y. and H.R.K. constructed a computing system for data processing, and S.I.T., H.H., P.L.S., S.A.S. and H.R.K. performed scaffolding of contigs and manual annotation of the predicted genes. S.N. performed sequence correction of supercontigs and gene prediction. S.I.T., S.N., N.G., N.A., M.Y., O.K., K.T., H.R.K., R.S., S.A.S. and J.M.C. analyzed data. S.I.T. N.M.Q.P., T.T., T.M., K.K., J.M.C., T.H., A.A.E., J.W.B. and K.T. wrote the manuscript.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

<sup>8</sup>Department of Molecular and Cellular Parasitology, Juntendo University, School of Medicine, Bunkyo-ku, Tokyo, 113-8421, Japan

<sup>9</sup>Department of Biomedical Chemistry, Graduate School of Medicine, The University of Tokyo, Hongo, Tokyo 113-0033, Japan

<sup>10</sup>Tsukuba Primate Research Center, National Institute of Biomedical Innovation, Tsukuba, Ibaraki 305-0843, Japan

<sup>11</sup>Division of Parasitic Diseases and Malaria, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, GA 30329, United States of America

<sup>12</sup>Center for Evolutionary Medicine and Informatics, The Biodesign Institute, Arizona State University, PO Box 874501, Tempe AZ 85287-4501, United States of America

## Abstract

*Plasmodium cynomolgi*, a malaria parasite of Asian Old World monkeys, is the sister taxon of *Plasmodium vivax*, the most prevalent human malaria species outside Africa. Since *P. cynomolgi* shares many phenotypic, biologic and genetic characteristics of *P. vivax*, we generated draft genome sequences of three *P. cynomolgi* strains and performed comparative genomic analysis between them and *P. vivax*, as well as a third previously sequenced simian parasite, *Plasmodium knowlesi*. Here we show that genomes of the monkey malaria clade can be characterized by CNVs in multigene families involved in evasion of the human immune system and invasion of host erythrocytes. We identify genome-wide SNPs, microsatellites, and CNVs in the *P. cynomolgi* genome, providing a map of genetic variation for mapping parasite traits and studying parasite populations. The *P. cynomolgi* genome is a critical step in developing a model system for *P. vivax* research, and to counteract the neglect of *P. vivax*.

---

Human malaria is transmitted by anopheline mosquitoes and caused by four species in the genus *Plasmodium*. Of these, *Plasmodium vivax* is the major malaria agent outside Africa, annually causing 80 to 100 million cases<sup>1</sup>. Often mistakenly regarded as benign and self-limiting, *P. vivax* treatment and control present challenges distinct from those of the more virulent *P. falciparum*. Biological traits including a dormant (hypnozoite) liver stage responsible for recurrent infections (relapses), early infective sexual stages (gametocytes), and transmission from low parasite densities in the blood<sup>2</sup>, coupled with emerging antimalarial drug resistance<sup>3</sup>, render *P. vivax* resilient to modern control strategies. Recent evidence indicates that *P. falciparum* derives from parasites of great apes in Africa<sup>4</sup>, while *P. vivax* is more closely related to parasites of Asian Old World monkeys (OWM)<sup>5-7</sup>, though not itself infective to OWM.

*P. vivax* cannot be cultured *in vitro*, and the small New World monkeys capable of hosting it are rare and do not provide an ideal model system. *P. knowlesi*, an Asian OWM parasite recently recognized as a significant zoonosis for humans<sup>8</sup>, offers a sequenced genome<sup>9</sup>, but the species is distantly related to *P. vivax* and phenotypically dissimilar. In contrast, *P. cynomolgi*, a simian parasite which can infect humans experimentally<sup>10</sup>, is the closest living relative (*i.e.*, a sister taxon) to *P. vivax* and possesses most of its genetic, phenotypic, and biologic characteristics: importantly, periodic relapses caused by dormant hypnozoites, early infectious gametocyte formation, and invasion of Duffy blood group-positive reticulocytes. *P. cynomolgi* thus offers a robust model for *P. vivax* in a readily-available laboratory host, the Rhesus monkey, whose genome was recently sequenced<sup>11</sup>. Here we report draft genome sequences of three *P. cynomolgi* strains, and comparative genomic analyses of *P. cynomolgi*, *P. vivax*<sup>12</sup> and *P. knowlesi*<sup>9</sup>, three members of the monkey malaria clade.

We sequenced the genome of *P. cynomolgi* 'B' strain, isolated from a monkey in Malaysia and grown in splenectomized monkeys (Online Methods). A combination of Sanger, Roche 454 and Illumina chemistries was deployed to generate a high quality reference assembly at 161-fold coverage, consisting of 14 supercontigs (corresponding to the 14 parasite chromosomes) and ~1,649 unassigned contigs, comprising a total length of ~26.2 Mb (Supplementary Table 1). Comparing genomic features of *P. cynomolgi*, *P. knowlesi* and *P. vivax* reveals many similarities, including GC content (mean GC 40.5%), fourteen conserved centromeres, and presence of intrachromosomal telomeric sequences (ITSs; GGGTT[T/C]A) discovered in the *P. knowlesi* genome<sup>9</sup> but absent in *P. vivax* (Table 1, Supplementary Table 2 and Fig. 1).

We annotated the *P. cynomolgi* strain B genome using a combination of *ab initio* gene prediction programs trained on high quality datasets and sequence similarity searches against the annotated *P. vivax* and *P. knowlesi* genomes. Unsurprisingly for species from the same monkey malaria clade, gene synteny along the 14 chromosomes is highly conserved, though numerous microsyntenic breaks occur in regions containing multigene families (Fig. 2 and Table 2). The *P. cynomolgi* genome contains 5,722 genes, of which about half encode conserved hypothetical proteins of unknown function, as is the case in all *Plasmodium* genomes sequenced to date. A maximum likelihood (ML) phylogenetic tree using 192 conserved ribosomal and translation/transcription-related genes (Supplementary Figure 1) confirms the close relationship of *P. cynomolgi* to *P. vivax* compared to five other *Plasmodium* species. Approximately 90% of genes (4,613) have reciprocal best match orthologs in all three species (Fig. 3), enabling refinement of the existing *P. vivax* and *P. knowlesi* annotations (Supplementary Table 3). The high degree of gene orthology enabled us to identify specific examples of gene duplication (an important vehicle for genome evolution), including a *P. cynomolgi* duplicated homolog of *P. vivax* Pvs28, a sexual stage surface antigen that is a transmission-blocking vaccine candidate<sup>13</sup> (Supplementary Table 4). Genes common only to *P. cynomolgi* and *P. vivax* (n = 214) outnumber those exclusive to *P. cynomolgi* and *P. knowlesi* (n = 100) or *P. vivax* and *P. knowlesi* (n = 17). Such figures establish the utility of *P. cynomolgi* as a model species for studying the more intractable *P. vivax*.

Remarkably, most of the genes specific to a particular species belong to multigene families (excluding hypothetical genes; Table 2 and Supplementary Table 5). This suggests repeated lineage-specific gene duplication and/or gene deletion in multigene families within the three monkey malaria clade species. Moreover, copy number of multigene families was generally greater in the *P. cynomolgi*/*P. vivax* lineage than in *P. knowlesi*, suggesting repeated gene duplication in the ancestral lineage of *P. cynomolgi*/*P. vivax* (or repeated gene deletion in the *P. knowlesi* lineage). Thus the genomes of *P. cynomolgi*, *P. vivax* and *P. knowlesi* can almost be completely characterized by variations in the copy number of multigene families. Examples of such families include those that encode proteins involved in evasion of the human immune system (*vir*, *kir* and *SICAvar*), and invasion of host red blood cells (*dbp* and *rbp*), described below.

In malaria parasites, invasion of host erythrocytes, mediated by specific interactions between parasite ligands and erythrocyte receptors, is a crucial component of the parasite life-cycle. Of great interest are the *ebf* and *rbf* families, which encode parasite ligands required for the recognition of host erythrocytes. The *ebf* genes encode erythrocyte binding-like (EBL) ligands such as the Duffy binding proteins (DBP) that bind to Duffy Antigen/Receptor for Chemokines (DARC) on human and monkey erythrocytes. The *rbf* genes encode the reticulocyte binding-like (RBL) protein family, including reticulocyte binding proteins (RBPs) in *P. cynomolgi* and *P. vivax*, and normocyte binding proteins (NBPs) in *P. knowlesi*, which bind to unknown erythrocyte receptors<sup>14</sup>. We confirmed the presence of

two *dbp* genes in *P. cynomolgi*<sup>15</sup> (Supplementary Table 6), in contrast to the one *dbp* and three *dbp* genes identified in *P. vivax* and *P. knowlesi*, respectively. This raises an intriguing hypothesis -- that *P. vivax* lost one *dbp* gene, and hence its infectivity to OWM erythrocytes, after divergence from a common *P. vivax/P. cynomolgi* ancestor. This hypothesis is also supported by our identification of only single copies of *dbp* genes in two other closely related OWM malaria parasites *P. fieldi* and *P. simiovale*, which are incapable of infecting humans<sup>16</sup>. These latter OWM species lost one or more *dbp* genes during divergence that conferred infectivity to humans, while *P. cynomolgi* and *P. knowlesi* retained *dbp* genes that allow invasion of human erythrocytes (Supplementary Figure 2).

We found multiple *rbc* genes, some truncated or present as pseudogenes, in the *P. cynomolgi* genome (Fig. 1 and Table 2). Phylogenetic analysis shows that *rbc*s from *P. cynomolgi*, *P. vivax* and *P. knowlesi* can be classified into three distinct groups, RBP/NBP-1, RBP/NBP-2 and RBP/NBP-3 (Supplementary Figure 3), and suggests that these groups existed before the three species diverged. All three groups of RBP/NBP are represented in *P. cynomolgi*, whereas *P. vivax* and *P. knowlesi* lack functional genes of the RBP/NBP-3 and RBP/NBP-1 groups, respectively. Thus *rbc* gene family expansion appears to have occurred after speciation, indicating that the three species have multiple species-specific erythrocyte invasion mechanisms. Intriguingly, we found an ortholog of *P. vivax rbc1b* in some strains of *P. cynomolgi*, but not in others (Supplementary Table 6). This is the first example of an *rbc* copy number variant between strains of a single *Plasmodium* species that we are aware of, and highlights how repeated birth and death of *rbc* genes, a signature of adaptive evolution, may have enabled species of the monkey malaria clade to expand or switch hosts between monkeys and humans.

The largest gene family in *P. cynomolgi*, consisting of 256 *cyir* (*c*ynomolgi *i*nterspersed *r*epeat) genes, is part of the *pir* (*p*lasmodium *i*nterspersed *r*epeat) superfamily that include *P. vivax vir*s (n = 319) and *P. knowlesi kir*s (n = 70; Table 2). *Pir* proteins reside on the surface of infected erythrocytes and play an important role in immune evasion<sup>17</sup>. Most *cyir* genes have sequence similarity to *P. vivax vir* genes (n = 254; Supplementary Table 7) and are found in subtelomeric regions (Fig. 1), but interestingly 11 *cyirs* have sequence similarity to *P. knowlesi kir*s (Supplementary Table 7) and occur internally as the *kirs* do in *P. knowlesi*. As with the ‘molecular mimicry’ in *P. knowlesi*<sup>9</sup>, one CYIR protein (PCYB\_032250) has a 56-amino acid region highly similar to the extracellular domain of primate CD99 (Supplementary Figure 4), a molecule involved in the regulation of T-cell function. A novel finding is that *P. cynomolgi* has two genes whose sequences are similar to *P. knowlesi SICAv* genes (Supplementary Table 7) that are expressed on the surfaces of schizont-infected macaque erythrocytes and are involved in antigenic variation<sup>18</sup>.

The ability to form a dormant hypnozoite stage is common to both *P. cynomolgi* and *P. vivax*, and was first shown in laboratory infections of monkeys by mosquito-transmitted *P. cynomolgi*<sup>19</sup>. In a search for candidate hypnozoite genes, we identified nine coding for ‘dormancy-related’ proteins and possessing ApiAP2 motifs<sup>20</sup> necessary for stage-specific transcriptional regulation at the sporozoite (pre-hypnozoite) stage (Supplementary Table 8). The candidates include kinases that are involved in cell-cycle transition; hypnozoite formation may be regulated by phosphorylation of proteins specifically expressed at the pre-hypnozoite stage. Our list of *P. cynomolgi* candidate genes represents an informed starting point for experimental studies on this elusive stage.

We sequenced *P. cynomolgi* strains Berok (from Malaysia) and Cambodian (from Cambodia) to 26× and 17 coverage, respectively, to characterize *P. cynomolgi* genome-wide diversity through analysis of SNPs, CNVs and microsatellites. A comparison of the three *P. cynomolgi* strains identified 178,732 SNPs (Supplementary Table 9), a frequency of one

SNP per 151 bp, a polymorphism level somewhat similar to that found when *P. falciparum* genomes are compared<sup>21,22</sup>. We calculated the pairwise nucleotide diversity ( $\pi$ ) as  $5.41 \times 10^{-3}$  across the genome, which varies little between the chromosomes. We assessed genome-wide copy number variants (CNVs) between *P. cynomolgi* B and Berok strains, using a robust statistical model in the program CNV-seq<sup>23</sup>, identifying 1,570 CNVs (1 per 17 kb), including one containing the *rbp1b* gene on chromosome 7 described above (Supplementary Figure 5). Finally, mining of *P. cynomolgi* B and Berok strains identified 182 polymorphic intergenic microsatellites (Supplementary Table 10), the first set of genetic markers developed for this species. These provide a toolkit for genetic diversity and population structure studies of laboratory stocks or natural infections of *P. cynomolgi*, many of which have recently been isolated from screening hundreds of wild monkeys for the zoonosis *P. knowlesi*<sup>24</sup>.

We estimated the difference between the number of synonymous changes per synonymous site (Ds) and the number of non-synonymous changes per non-synonymous site (Dn) over 4,605 pairs of orthologs within *P. cynomolgi* strain B and Berok, and between these two *P. cynomolgi* genomes and *P. vivax* Salvador I, using a simple Nei-Gojobori model<sup>25</sup>. We found 196 genes with  $Dn > Ds$  within the two *P. cynomolgi* strains, and 1,742 genes with  $Ds > Dn$  (Supplementary Table 11). Genes with relatively high  $Dn > Ds$  include transmembrane proteins such as antigens and transporters, among which is a transmission blocking target antigen Pcyn230 (PCYB\_042090). Interestingly, the *P. vivax* ortholog (PVX\_003905) does not show evidence for positive selection<sup>26</sup>, suggesting species-specific positive selection. We explored the degree to which evolution of orthologs has been constrained between *P. cynomolgi* and *P. vivax*, and found 154 genes under possible accelerated evolution but 1,613 genes under possible purifying selection (Supplementary Table 12). This conservative estimate indicates that at least 35% of loci have diverged under strong constraint, compared with 3.3% of loci under less constraint or positive selection (Figure 1), indicating that overall the genome of *P. cynomolgi* is highly conserved in single locus genes when compared to *P. vivax*, and emphasizing the value of *P. cynomolgi* as a biomedical and evolutionary model for studying *P. vivax*.

Our generation of the first *P. cynomolgi* genome sequences is a critical component in the development of a robust model system for the intractable and neglected *P. vivax* species<sup>27</sup>. Comparative genome analysis of *P. vivax* and the OWM malaria parasites *P. cynomolgi* and *P. knowlesi* presented here provides the foundation for further insights into traits such as host specificity that will enhance the prospects for the eventual elimination of vivax malaria and global malaria eradication.

## ONLINE METHODS

### Parasite material

Details of the origin of *P. cynomolgi* B, Berok and Cambodian strains, their growth in macaques, and isolation of parasite material are given in the Supplementary Note.

### Genome sequencing and assembly

*P. cynomolgi* B strain was sequenced using Roche/454 GS FLX (Titanium) and Illumina/Solexa GAI platforms to  $161 \times$  coverage. In addition, 2,784 clones (6.8 Mb) of a ~40 kb insert fosmid library in pCC1FOS (Epidentre Biotechnologies, USA) were sequenced by the Sanger method. A draft assembly of strain B was constructed using a combination of automated assembly and manual gap closure. We first generated *de novo* contigs by assembling Roche 454 reads using GS De Novo Assembler version 2.0 with default parameters. Contigs > 500-bp were mapped to the *P. vivax* Salvador I reference assembly<sup>12</sup>

(PlasmoDB; see URLs). *P. cynomolgi* contigs were iteratively arrayed by aligning to *P. vivax*-assembled sequences with manual corrections. A total of 1,264 aligned contigs were validated by mapping paired-end reads from fosmid clones using BLASTN (e-value:  $<1e^{-15}$ , identity:  $>90\%$ , coverage:  $>200\text{bp}$ ) implemented in the software GenomeMatcher version 1.65<sup>28</sup>. Additional linkages (699 regions) were made using PCR across the intervening sequence gaps with primers designed from neighboring contigs. Length of sequence gaps was estimated from insert lengths of the fosmid paired-end reads, the size of PCR products and homologous sequences of the *P. vivax* genome. Supercontigs were then manually constructed from the aligned contigs. Eventually we obtained 14 supercontigs corresponding to the 14 chromosomes of the parasite, with a total length of  $\sim 22.73$  Mb, encompassing  $\sim 80\%$  of the predicted *P. cynomolgi* genome. A total of 1,651 contigs ( $>1$  kb) with a total length of 3.45 Mb were identified as unassigned subtelomeric sequences by searching against the *P. vivax* genome using BLASTN. Additionally, in order to improve sequence accuracy, we constructed a mapping assembly of Illumina paired-end reads and the 14 supercontigs and unassigned contigs as reference sequences using CLC Genomics Workbench version 3.0 with default settings (CLC Bio, Denmark). Comparison of the draft *P. cynomolgi* B sequence with 23 *P. cynomolgi* protein-coding genes (64 kb) obtained by Sanger sequencing showed 99.8 % sequence identity (Supplementary Table 13). *P. cynomolgi* Berok and Cambodian strains were sequenced to  $26 \times$  and  $17 \times$  coverage respectively, using the Roche/454 GS FLX (Titanium) platform, with single-end and 3 kb paired-end libraries made for the former and a single-end library only made for the latter. For phylogenetic analyses of specific genes, sequences were independently verified by Sanger sequencing (Supplementary Note and Supplementary Table 14).

### Prediction and annotation of genes

Gene prediction of the 14 supercontigs and 1,651 unassigned contigs was performed using the MAKER genome annotation pipeline<sup>29</sup> using *ab initio* gene prediction programs trained on protein and ESTs from PlasmoDB Build 7.1. For gene annotation, BLASTN (e-value:  $<1e^{-15}$ , identity:  $>70\%$ , coverage:  $>100\text{bp}$ ) searches of *P. vivax* (PvivaxAnnotatedTranscripts\_PlasmoDB-7.1.fasta) and *P. knowlesi* (PknowlesiAnnotatedTranscripts\_PlasmoDB-7.1.fasta) predicted proteomes were run and the best-hits identified. All predicted genes were manually inspected at least twice for gene structure and functional annotation, and orthologous relationships between *P. cynomolgi*, *P. vivax* and *P. knowlesi* determined based upon synteny. A unique identifier was assigned to *P. cynomolgi* genes as follows: PCYB\_#####, where the first two of the six numbers indicate chromosome number. Paralogs of genes that appeared to be specific to either *P. cynomolgi*, *P. vivax* or *P. knowlesi* were searched using BLASTP with default parameters using cut-off e-value of  $<1e^{-16}$ .

### Multiple genome sequence alignment

Predicted proteins of *P. cynomolgi* B strain were concatenated and aligned with those in the 14 chromosomes of five other *Plasmodium* genomes, *P. vivax*, *P. knowlesi*, *P. falciparum*, *P. berghei* and *P. chabaudi*, using the software Murasaki version 1.68.6<sup>30</sup>.

### Search for sequence showing high similarity to host proteins

Eleven *P. cynomolgi* CYIR proteins (with sequence similarity to *P. knowlesi* KIR) were subjected to BLASTP search for regions having high similarity to host *Macacca mulatta* CD99 protein, with cut-off e-value of  $<1e^{-12}$  and compositional adjustment (*i.e.*, no adjustment) against the non-redundant protein sequence dataset of the *M. mulatta* proteome in NCBI.

## Phylogenetic analyses

Genes were aligned using Clustal W version 2.0.10<sup>31</sup> with manual corrections, and unambiguously aligned sites were selected for phylogenetic analyses. Maximum Likelihood (ML) phylogenetic trees were constructed using PROML programs in PHYLIP version 3.69<sup>32</sup> under Jones-Taylor-Thornton (JTT) amino acid substitution model. To take the evolutionary rate heterogeneity across sites into consideration, the R (Hidden Markov Model rates) option was set for discrete G distribution with 8 categories for approximating the site-rate distribution. CODEML programs in PAML 4.4<sup>33</sup> were used for estimating the G shape parameter, a values. For bootstrap analyses, SEQBOOT and CONSENSE programs in PHYLIP were applied.

## Candidate genes for hypnozoite formation

We undertook two approaches. First, genes unique to *P. vivax* and *P. cynomolgi* (hypnozoite-forming parasites) and not found in other non-hypnozoite-forming *Plasmodium* species were identified. We used the 147 unique genes identified in the *P. vivax* genome<sup>12</sup> to search the *P. cynomolgi* B sequence. The orthologs identified in both species were then searched ~1 kb 5' upstream for four specific ApiAP2 motifs<sup>20</sup>, PF14\_0633: GCATGC, PF13\_0235\_D1: GCCCCG, PFF0670w\_D1: TAAGCC, and PFD0985w\_D2: TGTTAC, which are involved in sporozoite stage-specific regulation and expression (corresponding to the pre-hypnozoite stage). Second, 'dormancy related' proteins were retrieved from GenBank and used to search for *P. vivax* homologs. Candidate genes (n= 128) and orthologs of *P. cynomolgi* and five other parasite species were searched in the region ~1 kb 5' upstream for the presence of the four ApiAP2 motifs. Data of *P. vivax*, *P. knowlesi*, *P. falciparum*, *P. berghei*, *P. chabaudi* and *P. yoelii* were retrieved from PlasmoDB Build 7.1.

## Genome-wide screen for polymorphisms

For SNP identification, alignment of Roche 454 data from strains B, Berok and Cambodian was performed using SSAHA2<sup>34</sup>, with 0.1 mismatch rate and only unique matches reported. Potential duplicate reads generated during PCR amplification were removed, so that when multiple reads mapped at identical coordinates, only reads with the highest mapping quality were retained. We used a statistical method by Li *et al.*<sup>35</sup> implemented in SAMtools version 0.1.18 to call SNPs simultaneously in the case of duplicate runs of the same strain. SNPs with high read depth (>100) were filtered out, as were SNPs in poor alignment regions at the ends of chromosomes (Supplementary Note).

Nucleotide diversity ( $\pi$ ) was calculated as follows: for each site being compared, we calculated allele frequency by counting the two alleles and measured the proportion of nucleotide differences. Let  $\pi$  be the genetic distance between allele i and allele j, then the

nucleotide diversity within the population is  $\pi = \sum_{i,j} P_i P_j \pi_{ij}$  where  $P_i$  and  $P_j$  are the overall allele frequencies i and j respectively. Mean  $\pi$  was calculated by averaging over sites,

weighting each by  $\sum_{i=1}^{n-1} \frac{1}{i}$  where n is the number of aligned sites. Average dN/dS ratios were estimated using the modified Nei-Gojobori/Jukes-Cantor method in MEGA 4<sup>36</sup>.

CNV-seq<sup>23</sup> was used to identify potential copy number variants in *P. cynomolgi*. Briefly, this method is based on a statistical model that allows confidence assessment of observed copy number ratios from next generation sequencing data. Roche 454 sequences from *P. cynomolgi* strain B assembly was used as the reference genome, and the *P. cynomolgi* Berok strain used as a test genome; the Cambodian strain sequence coverage was considered too low for analysis. The test reads were mapped to the reference genome, and CNVs detected

by computing the number of reads for each test strain in a sliding window. The validity of the observed ratios were assessed by the computation of a probability of a random occurrence, given no copy number variation.

Polymorphic microsatellites (defined as repeat units of 1–6 nc) between *P. cynomolgi* strains B and Berok were identified by aligning contigs from a *de novo* assembly of Berok (generated using Roche GS Assembler version 2.6, with 40 bp minimum overlap, 90% identity) to the B strain using BWA<sup>37</sup> and allowing for gaps. Utilizing the Phred-scaled probability of the base being misaligned by SAMtools<sup>35</sup>, indel candidates were called from the alignment. In-house Python scripts were used to then cross-reference with the microsatellites found in the reference strain B assembly identified by MISA (see URLs). All homopolymer microsatellites were discarded to account for potential sequence errors introduced by 454 sequencing.

Selective constraint analysis of 4,605 orthologs between *P. cynomolgi* strains B and Berok and *P. vivax* Salvador I, utilized MUSCLE<sup>38</sup> alignments with stringent removal of gaps and missing data (*P. cynomolgi* Berok orthologs were identified through a reciprocal best hit BLAST search against strain B genes). Analyses were conducted using the Nei-Gojobori model<sup>25</sup>. In order to detect values that could not be explained by chance, we estimated the standard error (SE) by a bootstrap procedure with 200 pseudo-replicates for each gene. The expected value for Ds-Dn is 0 if a given pair of sequences is diverging without obvious effects on fitness. In the case of the intra-*P. cynomolgi* comparison, values with a difference of  $\pm 2$  SE from 0 were considered indicative of an excess of synonymous changes (Ds-Dn>0) or non-synonymous changes (Ds-Dn<0). In the case of the inter-*P. cynomolgi*-*P. vivax* comparison, we used a more stringent criterion of  $\pm 3$  SE from 0.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

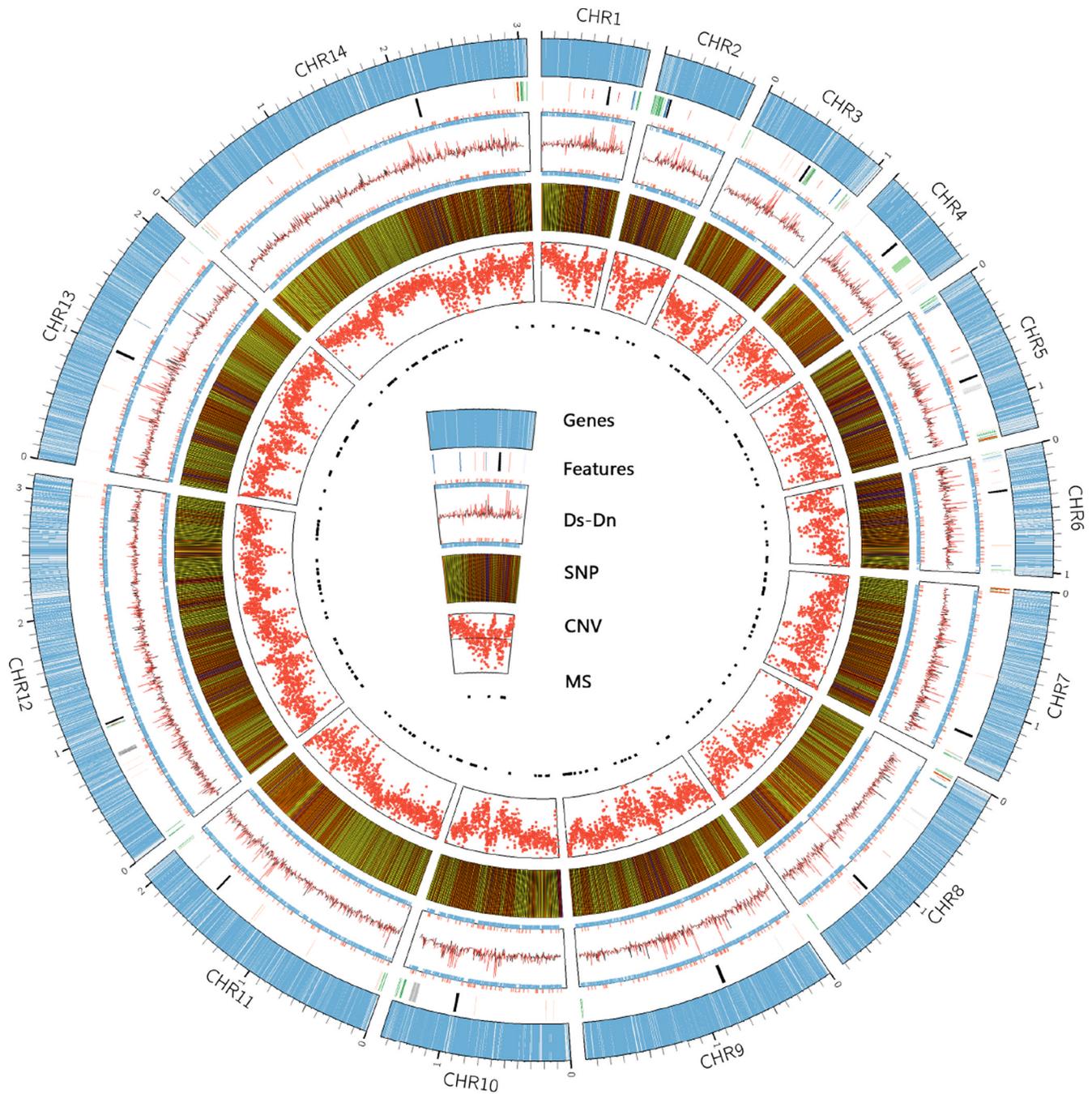
We thank Hiromi Sawai for suggestions of genome analysis, David Fisher for help with genome-wide evolutionary analyses, and NYU Langone Medical Center Genome Technology Core for access to Roche 454 sequencing equipment funded by grant S10 RR026950 to J.M.C. from the National Institutes of Health (NIH). Genome and phylogenetic analyses used the Genome Information Research Center in the Research Institute of Microbial Diseases, Osaka University. This work was supported by the Ministry of Education, Culture, Sports, Science and Technology grants (18073013, 18GS03140013, 20390120, 22406012 and 22406012) to K.T., NIH grant R01 GM07079393-01A2, Burroughs Wellcome Fund grant 1007398 and NIH International Centers of Excellence for Malaria Research grant U19 AI089676-01 to J.M.C., and NIH grant R01 GM080586 to AAE.

## REFERENCES

1. Mendis K, Sina BJ, Marchesini P, Carter R. The neglected burden of *Plasmodium vivax* malaria. *Am J Trop Med Hyg.* 2001; 64:97–106. [PubMed: 11425182]
2. Mueller I, et al. Key gaps in the knowledge of *Plasmodium vivax*, a neglected human malaria parasite. *Lancet Infect Dis.* 2009; 9:555–566. [PubMed: 19695492]
3. Baird JK. Resistance to chloroquine unhinges *vivax* malaria therapeutics. *Antimicrob Agents Chemother.* 2011; 55:1827–1830. [PubMed: 21383088]
4. Rayner JC, Liu W, Peeters M, Sharp PM, Hahn BH. A plethora of *Plasmodium* species in wild apes: a source of human infection? *Trends Parasitol.* 2011; 27:222–229. [PubMed: 21354860]
5. Cornejo OE, Escalante AA. The origin and age of *Plasmodium vivax*. *Trends Parasitol.* 2006; 22:558–563. [PubMed: 17035086]
6. Escalante AA, et al. A monkey's tale: the origin of *Plasmodium vivax* as a human malaria parasite. *Proc Natl Acad Sci U S A.* 2005; 102:1980–1985. [PubMed: 15684081]

7. Mu J, et al. Host switch leads to emergence of *Plasmodium vivax* malaria in humans. *Mol Biol Evol.* 2005; 22:1686–1693. [PubMed: 15858201]
8. Singh B, et al. A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. *Lancet.* 2004; 363:1017–1024. [PubMed: 15051281]
9. Pain A, et al. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature.* 2008; 455:799–803. [PubMed: 18843368]
10. Eyles DE, Coatney GR, Getz ME. Vivax-type malaria parasite of macaques transmissible to man. *Science.* 1960; 131:1812–1813. [PubMed: 13821129]
11. Gibbs RA, et al. Evolutionary and biomedical insights from the rhesus macaque genome. *Science.* 2007; 316:222–234. [PubMed: 17431167]
12. Carlton JM, et al. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature.* 2008; 455:757–763. [PubMed: 18843361]
13. Saxena AK, Wu Y, Garboczi DN. *Plasmodium* p25 and p28 surface proteins: potential transmission-blocking vaccines. *Eukaryot Cell.* 2007; 6:1260–1265. [PubMed: 17557884]
14. Iyer J, Gruner AC, Renia L, Snounou G, Preiser PR. Invasion of host cells by malaria parasites: a tale of two protein families. *Mol Microbiol.* 2007; 65:231–249. [PubMed: 17630968]
15. Okenu DM, Malhotra P, Lalitha PV, Chitnis CE, Chauhan VS. Cloning and sequence analysis of a gene encoding an erythrocyte binding protein from *Plasmodium cynomolgi*. *Mol Biochem Parasitol.* 1997; 89:301–306. [PubMed: 9364974]
16. Coatney, GR.; Collins, WE.; Warren, M.; PG, C. *The Primate Malariae*. Washington, DC: U. S. Department of Health, Education and Welfare; 1971.
17. Cunningham D, Lawton J, Jarra W, Preiser P, Langhorne J. The *pir* multigene family of *Plasmodium*: antigenic variation and beyond. *Mol Biochem Parasitol.* 2010; 170:65–73. [PubMed: 20045030]
18. al-Khedery B, Barnwell JW, Galinski MR. Antigenic variation in malaria: a 3' genomic alteration associated with the expression of a *P. knowlesi* variant antigen. *Mol Cell.* 1999; 3:131–141. [PubMed: 10078196]
19. Krotoski WA. The hypnozoite and malarial relapse. *Prog Clin Parasitol.* 1989; 1:1–19. [PubMed: 2491691]
20. Campbell TL, De Silva EK, Olszewski KL, Elemento O, Llinas M. Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog.* 2010; 6:e1001165. [PubMed: 21060817]
21. Mu J, et al. Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nat Genet.* 2007; 39:126–130. [PubMed: 17159981]
22. Volkman SK, et al. A genome-wide map of diversity in *Plasmodium falciparum*. *Nat Genet.* 2007; 39:113–119. [PubMed: 17159979]
23. Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics.* 2009; 10:80. [PubMed: 19267900]
24. Lee KS, et al. *Plasmodium knowlesi*: reservoir hosts and tracking the emergence in humans and macaques. *PLoS Pathog.* 2011; 7:e1002015. [PubMed: 21490952]
25. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 1986; 3:418–426. [PubMed: 3444411]
26. Doi M, et al. Worldwide sequence conservation of transmission-blocking vaccine candidate Pvs230 in *Plasmodium vivax*. *Vaccine.* 2011; 29:4308–4315. [PubMed: 21514344]
27. Carlton JM, Sina BJ, Adams JH. Why is *Plasmodium vivax* a neglected tropical disease? *PLoS Negl Trop Dis.* 2011; 5:e11160. [PubMed: 21738804]
28. Ohtsubo Y, Ikeda-Ohtsubo W, Nagata Y, Tsuda M. GenomeMatcher: a graphical user interface for DNA sequence comparison. *BMC Bioinformatics.* 2008; 9:376. [PubMed: 18793444]
29. Cantarel BL, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 2008; 18:188–196. [PubMed: 18025269]
30. Popendorf K, Tsuyoshi H, Osana Y, Sakakibara Y, Murasaki: a fast, parallelizable algorithm to find anchors from multiple genomes. *PLoS One.* 2010; 5:e12651. [PubMed: 20885980]

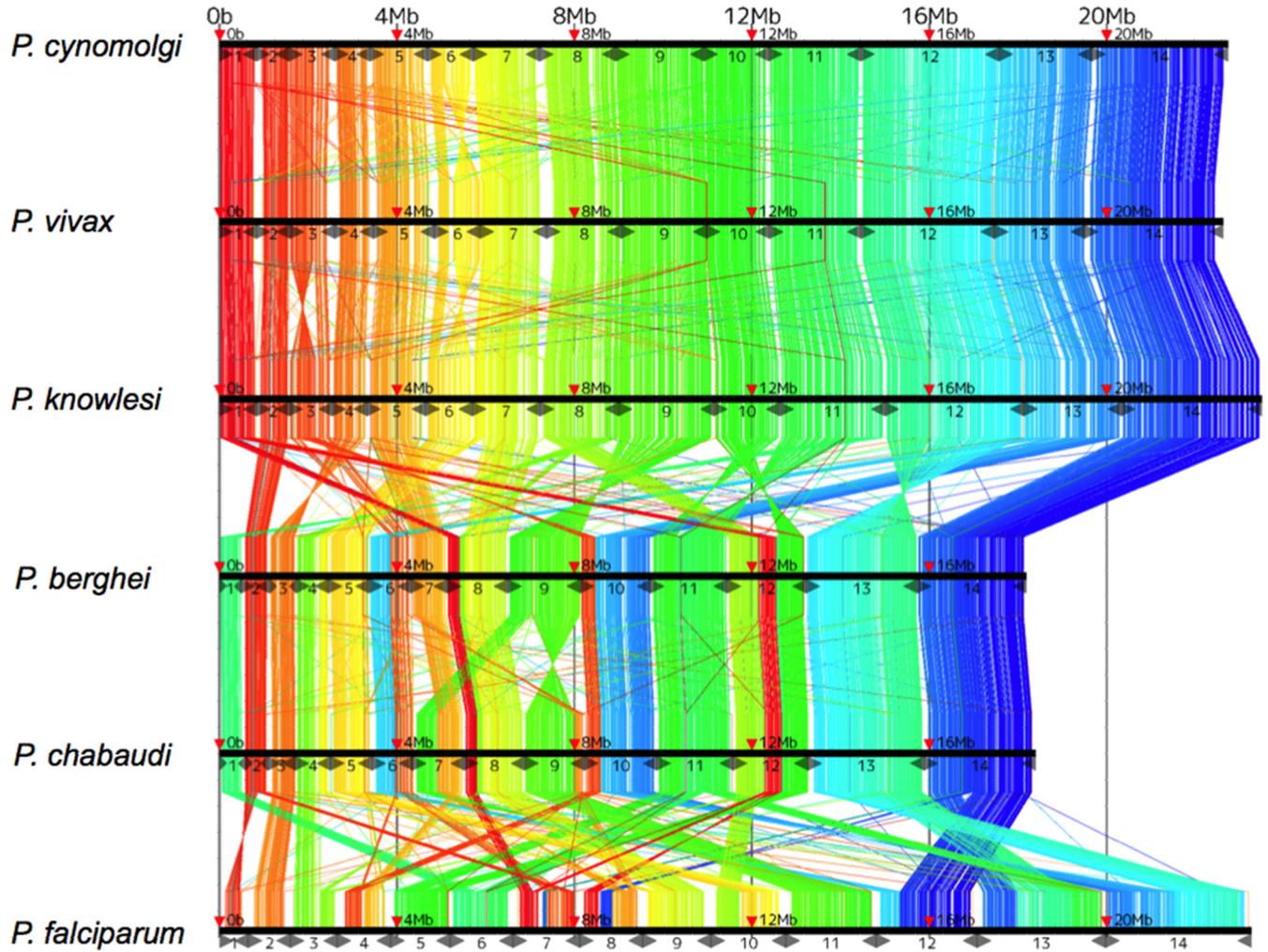
31. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994; 22:4673–4680. [PubMed: 7984417]
32. Felsenstein, J. PHYLIP, Phylogeny Inference Package. 3.6a3 edn. Seattle: University of Washington; 2005.
33. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 2007; 24:1586–1591. [PubMed: 17483113]
34. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. *Genome Res.* 2001; 11:1725–1729. [PubMed: 11591649]
35. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* 2011; 27:2987–2993. [PubMed: 21903627]
36. Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4. 0. *Mol Biol Evol.* 2007; 24:1596–1599. [PubMed: 17488738]
37. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009; 25:1754–1760. [PubMed: 19451168]
38. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004; 32:1792–1797. [PubMed: 15034147]



**Figure 1.**

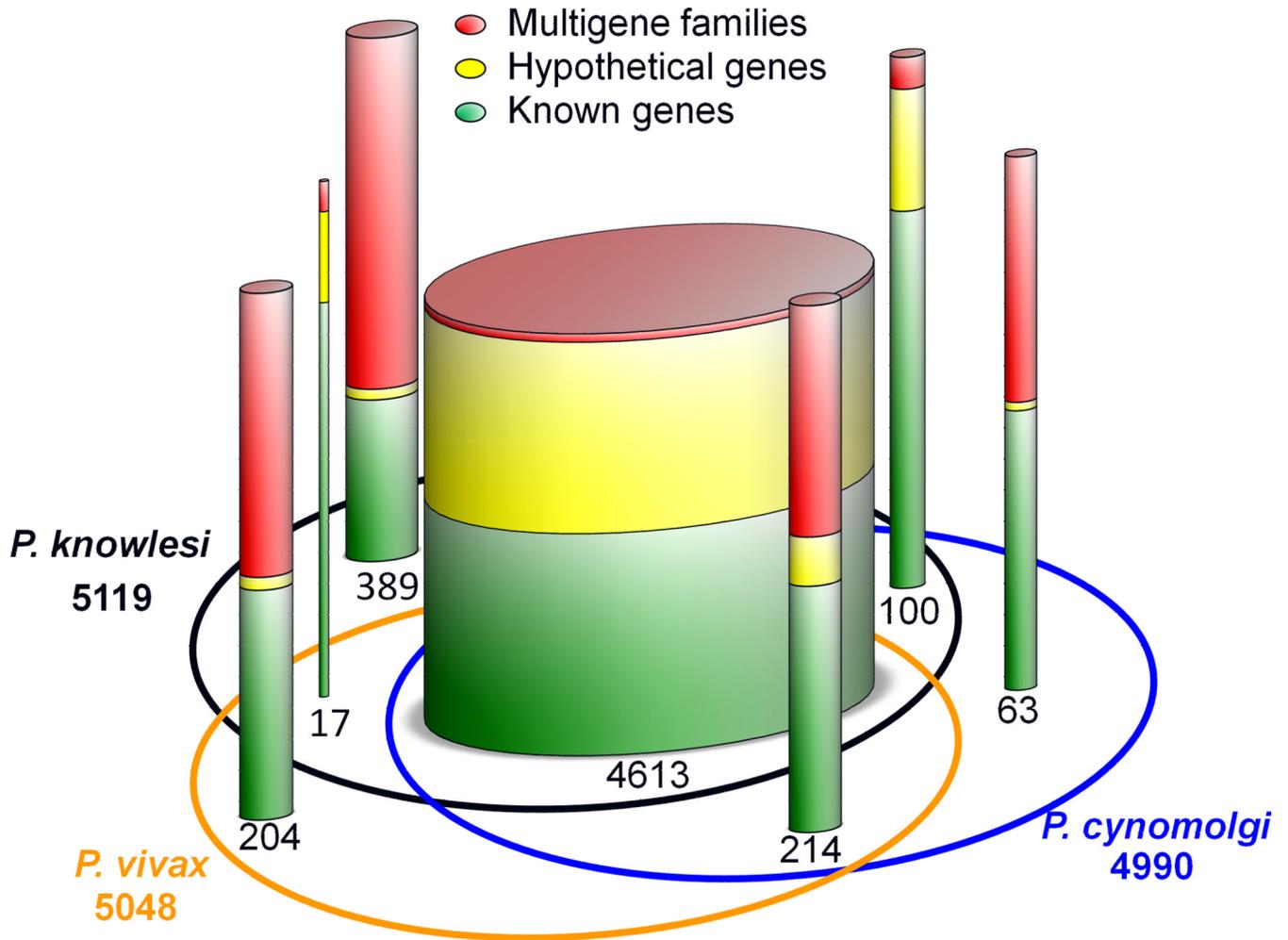
Architecture of the *P. cynomolgi* genome and associated genome-wide variation data. Each of the 14 *P. cynomolgi* chromosomes is indicated, and one chromosome slice is shown annotated in the center. The six co-centric rings represent: (1) Outer ring: localization of 5,049 *P. cynomolgi* genes excluding those on small contigs (cyan lines); (2) Second ring: genome features including 14 centromeres (thick black lines), 43 telomeric sequence repeats (short red lines), 43 tRNA genes (red lines), 10 rRNAs (dark blue lines), and several gene family members including: 53 *cyir* (dark green lines), 8 *RBP* (brown lines), 13 *SERA* (serine-rich antigen; pink lines), 25 *TRAG* (tryptophan-rich antigen; purple lines), 12 *MSP3* (merozoite surface protein 3; light grey lines), 13 *MSP7* (merozoite surface protein 7; grey

lines), 25 *RAD* (silver lines), 8 *etrapm* (orange lines), 16 *Pf-fam-b* (light blue lines), 7 *Pv-fam-d* (light green lines lines); (3) Third ring: plot of Ds-Dn for 4,605 orthologs depicting genome-wide (i) polymorphism within *P. cynomolgi* strains B and Berok (black line); and (ii) divergence between *P. cynomolgi* strains B and Berok, and *P. vivax* Salvador I (red line); a track above the plot indicates *P. cynomolgi* genes under positive selection (red) and purifying selection (blue), and a track below the plot indicates *P. cynomolgi/P. vivax* orthologs under positive selection (red) and purifying selection (blue); (4) Fourth ring: heat map indicating SNP density of three *P. cynomolgi* strains plotted per 10 kb window: red, 0–83 SNPs/10 kb (regions of lowest SNP density); blue, 84–166 SNPs/10 kb; green, 166–250 SNPs/10 kb; purple, 251–333 SNPs/10 kb; orange, 334–416 SNPs/10 kb; yellow, 417–500 SNPs/10 kb (regions of highest SNP density); (5) Fifth ring:  $\log_2$  ratio plot of CNVs identified from a comparison of *P. cynomolgi* strain B strain with Berok; and (6) Inner ring: map of 182 polymorphic intergenic microsatellites (MS; black dots). Figure was generated using Circos software (see URLs).



**Figure 2.**

Genome synteny between six species of *Plasmodium* parasite. Protein coding genes of *P. cynomolgi* are shown aligned with those of five other *Plasmodium* genomes: two species belonging to the monkey malaria clade *P. vivax* and *P. knowlesi*, two species of rodent malaria *P. berghei* and *P. chabaudi*, and *P. falciparum*. Highly conserved protein coding regions between the genomes are colored in order from red (5'-end of chromosome 1) to blue (3'-end of chromosome 14) after the genomic position of *P. cynomolgi*. A scale in Mb is shown on top of each genome alignment. This genome-wide view of synteny identified two apparent errors in existing public sequence databases: an inversion in chromosome 3 of *P. knowlesi*, and an inversion in chromosome 6 of *P. vivax*.



**Figure 3.**

A comparison of the genes of *P. cynomolgi*, *P. vivax* and *P. knowlesi*. The three ellipses represent the three genomes, with total number of genes assigned to chromosomes indicated under the species name. The Venn diagram delineates orthologous and non-orthologous genes between the three genomes, with the number of genes in each indicated and represented graphically by a cylinder of proportional width. In each cylinder, genes are divided into three categories (putatively known function, hypothetical, and members of multigene families) represented by colored bands proportional to their percentage.

**Table 1**

Comparison of genome features between *P. cynomolgi*, *P. vivax* and *P. knowlesi*, three species of the monkey malaria clade.

Feature	<i>P. cynomolgi</i>	<i>P. vivax</i> <sup>12</sup>	<i>P. knowlesi</i> <sup>9</sup>
<b>Assembly</b>			
Size (Mb)	26.2	26.9	23.7
No. scaffolds <sup>a</sup>	14 (1,649)	14 (2,547)	14 (67)
Coverage (fold)	161	10	8
G+C content (%)	40.4	42.3	38.8
<b>Genes</b>			
No. genes	5,722	5,432	5,197
Mean gene length	2,240	2,164	2,180
Gene density (bp per gene) <sup>b</sup>	4,428.2	4,950.5	4,416.1
Percentage coding <sup>b</sup>	51.0	47.1	49.0
<b>Structural RNAs</b>			
No. tRNA genes	43	44	41
No. 5S rRNA genes	3	3	0 <sup>d</sup>
No. 5.8S/18S/28S rRNA units	7	7	5
<b>Nuclear genome</b>			
No. chromosomes	14	14	14
No. centromeres	14	14	14
Isochore structure <sup>c</sup>	Yes	Yes	No
<b>Mitochondrial genome</b>			
Size (bp) <sup>e</sup>	5,986 (AB444123)	5,990 (AY598140)	5,958 (AB444108)
G+C content (%)	30.3	30.5	30.5
<b>Apicoplast genome</b>			
Size (bp)	29,297 <sup>f</sup>	5,064 <sup>g</sup>	Not available
G+C content (%)	13.0	17.1	Not available

<sup>a</sup>Small unassigned contigs indicated in parentheses

<sup>b</sup>Sequence gaps excluded

<sup>c</sup>Regions of the genome that differ in their density and are separable by CsCl centrifugation; isochores correspond to domains differing in their GC content

<sup>d</sup>Not present in *P. knowlesi* assembly version 4.0

<sup>e</sup>Identified in other studies, see Accession Numbers

<sup>f</sup>Partial sequence (~86% complete) identified during this project

<sup>g</sup>Partial sequence of reference genome only published<sup>12</sup>; actual size is ~35 kb

Table 2

Multigene families of *P. cynomolgi*, *P. vivax* and *P. knowlesi* differ in their copy number.

#	Multigene family	Localization	Arrangement	<i>P. cynomolgi</i>	<i>P. vivax</i>	<i>P. knowlesi</i>	Putative function & other information
1	pir (vir-like)	subtelomeric	scattered/clustered	254	319 <sup>a</sup>	4	Immune evasion
2	pir (kir-like)	subtelomeric/central	scattered/clustered	11	2	66 <sup>b</sup>	Immune evasion
3	SICAvar	subtelomeric/central	scattered/clustered	2	1	242 <sup>a</sup>	Antigenic variation, immune evasion
4	msp3	central	clustered	12	12	3	Merozoite surface protein
5	msp7	central	clustered	13	13	5	Merozoite surface protein
6	DBL (DBP/EBL)	subtelomeric	scattered	2	1	3	Host cell recognition
7	RBL (RBP/NBP/Rh)	subtelomeric	scattered	8 <sup>a</sup>	10 <sup>a</sup>	3 <sup>a</sup>	Host cell recognition
8	Pv-fam-a (PvTRAG)	subtelomeric	scattered/clustered	36	36	26 <sup>a</sup>	Tryptophan-rich
9	Pv-fam-b	central	clustered	3	6	1	Unknown
10	Pv-fam-c	subtelomeric	unknown <sup>b</sup>	1	7	0	Unknown
11	Pv-fam-d (HYPB)	subtelomeric	scattered	18	16	2	Unknown
12	Pv-fam-e (RAD)	subtelomeric	clustered	27	44	16	Unknown
13	Pv-fam-g	central	clustered	3	3	3	Unknown
14	Pv-fam-h (HYPI6)	central	clustered	6	4	2	Unknown
15	Pv-fam-i (HYPI1)	subtelomeric	scattered	6	6	5	Unknown
16	Pk-fam-a	central	scattered	0	0	12 <sup>a</sup>	Unknown
17	Pk-fam-b	subtelomeric	scattered	0	0	9	Unknown
18	Pk-fam-c	subtelomeric	scattered	0	0	6 <sup>a</sup>	Unknown
19	Pk-fam-d	central	scattered	0	0	3 <sup>a</sup>	Unknown
20	Pk-fam-e	subtelomeric	scattered	0	0	3 <sup>a</sup>	Unknown
21	PST-A	subtelomeric/central	scattered	9 <sup>a</sup>	11 <sup>a</sup>	7	Alpha beta hydrolase
22	ETRAPM	subtelomeric	scattered	9	9	9	Parasitophorous vacuole membrane
23	CLAG (RhopH-1)	subtelomeric	scattered	2	3	2	High MW rhoptry antigen complex
24	PvSTPI	subtelomeric	unknown <sup>b</sup>	3	10 <sup>a</sup>	0	Unknown

#	Multigene family	Localization	Arrangement	<i>P. cynomolgi</i>	<i>P. vivax</i>	<i>P. knowlesi</i>	Putative function & other information
25	PHIST (PF-fam-b)	subtelomeric	scattered/clustered	21	20	15	Unknown
26	SERA	central	clustered	13 <sup>a</sup>	13 <sup>a</sup>	8 <sup>a</sup>	Cysteine protease

<sup>a</sup>Pseudogenes, truncated genes and gene fragments included.

<sup>b</sup>Gene arrangement could not be determined due to localization on unassigned contigs.