

Supplementary Information for:

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

Shin-Ichiro Tachibana, Steven A. Sullivan, Satoru Kawai, Shota Nakamura, Hyunjae R. Kim, Naohisa Goto, Nobuko Arisue, Nirianne M. Q. Palacpac, Hajime Honma, Masanori Yagi, Takahiro Tougan, Yuko Katakai, Osamu Kaneko, Toshihiro Mita, Kiyoshi Kita, Yasuhiro Yasutomi, Patrick L. Sutton, Rimma Shakhbatyan, Toshihiro Horii, Teruo Yasunaga, John W. Barnwell, Ananias A. Escalante, Jane M. Carlton, and Kazuyuki Tanabe

Corresponding authors:

Kazuyuki Tanabe, PhD

Department of Molecular Protozoology

Research Institute for Microbial Diseases

Osaka University

3-1, Yamada-Oka, Suita, 565-0871

Japan

Phone: +81-6-6879-8279

Fax: +81-6-6879-8281

E-mail: kztanabe@biken.osaka-u.ac.jp

Jane M. Carlton, PhD

Professor and Faculty Director of Genomic Sequencing

Center for Genomics and Systems Biology

Department of Biology

New York University

12 Waverly Place

New York, NY 10003

USA

Phone: +1-212-992-6981

Fax: +1-212-995-4266

Email: jane.carlton@nyu.edu

Author affiliations are provided in the manuscript.

List of supplementary information:

- Supplementary Table 1.** Genome sequence data for three strains of *P. cynomolgi*: B, Berok and Cambodian.
- Supplementary Table 2.** Comparison of genome assembly and annotation characteristics between *P. vivax*, *P. cynomolgi* and *P. knowlesi*.
- *Supplementary Table 3.** List of orthologs between the genomes of *P. cynomolgi*, *P. vivax* and *P. knowlesi*.
- Supplementary Table 4.** Examples of gene duplications in *P. cynomolgi*, *P. vivax* and *P. knowlesi*.
- *Supplementary Table 5.** List of multigene families in *P. cynomolgi*, *P. vivax* and *P. knowlesi*.
- Supplementary Table 6.** Copy number variations and orthologous relationships in the *dbl* and *rbl* genes of *P. vivax*, *P. knowlesi* and three strains of *P. cynomolgi*.
- *Supplementary Table 7.** List of *P. cynomolgi* *cyir* genes, *P. vivax* *vir* genes, *P. knowlesi* *kir* genes, and *P. knowlesi* *SICAvar* genes and their homologs in *P. cynomolgi*, *P. vivax* and *P. knowlesi*.
- Supplementary Table 8.** Candidate genes for hypnozoite formation.
- Supplementary Table 9.** Genome-wide comparative analysis between three strains of *P. cynomolgi* and *P. vivax* Salvador I.
- *Supplementary Table 10.** List of polymorphic microsatellite loci identified between *P. cynomolgi* strains B and Berok.
- *Supplementary Table 11.** Ds-Dn within 4,605 orthologs of *P. cynomolgi* strains B and Berok.
- *Supplementary Table 12.** Ds-Dn between 4,605 orthologs of *P. cynomolgi* strains B and Berok, and *P. vivax* Salvador I.
- Supplementary Table 13.** Comparison of the *P. cynomolgi* B genome sequence with 23 *P. cynomolgi* protein-coding genes obtained by Sanger sequencing.
- Supplementary Table 14.** List of primer sequences.
- Supplementary Figure 1.** A maximum likelihood phylogenetic tree of 192 protein genes from seven *Plasmodium* species.
- Supplementary Figure 2.** A maximum likelihood phylogenetic tree of DBP from *P. vivax* and related monkey malaria parasites.
- Supplementary Figure 3.** A maximum likelihood phylogenetic tree of RBP/NBP from *P. vivax*, *P. cynomolgi* and *P. knowlesi*.
- Supplementary Figure 4.** A region in a *P. cynomolgi* CYIR protein with high sequence similarity to primate and monkey CD99.
- Supplementary Figure 5.** Copy number variation in the *P. cynomolgi* genome.
- Supplementary Note.** Additional methods and references.

* indicates supplementary Excel files

Supplementary Table 1. Genome sequence data for three strains of *P. cynomolgi*: B, Berok and Cambodian.

Strain	GS FLX Titanium reads (avg. read length; total length)	Illumina GAll reads (avg. read length; total length)	Sanger reads (avg. read length; total length)	Scaffolds (total length)	Singleton contigs (length)	Genome coverage*
B	i. <u>GS FLX only</u> 424,140 (253 nt; 107,388,987 nt) ii. <u>GS FLX Titanium</u> 1,694,412 (304 nt; 514,413,326 nt)	49,764,484 (75 nt; 3,732,336,300 nt)	5,568 (1,226 nt; 6,824,982 nt)	14 (22.7 Mb)	1,649 (3.5 Mb)	161
Berok	i. <u>Single end</u> : 1,141,791 (338 nt; 385,925,358 nt) ii. <u>Paired end</u> : 2,000,376 (154 nt; 308,057,904 nt)	-	-	-	-	26
Cambodian	1,633,472 (287 nt; 468,806,464 nt)	-	-	-	-	17

* Assuming *P. cynomolgi* predicted genome size of ~27 Mb

Supplementary Table 2. Comparison of genome assembly and annotation characteristics between *P. vivax*, *P. cynomolgi* and *P. knowlesi*.

Chr #	<i>P. cynomolgi</i> (B strain)					<i>P. vivax</i> (Salvador I) ^a					<i>P. knowlesi</i> (H strain) ^a				
	Contig no.	Length (bp)	Sequence gap (bp)	No. of genes	GC (%)	Contig no.	Length (bp)	Sequence gap (bp)	No. of genes	GC (%)	Contig no.	Length (bp)	Sequence gap (bp)	No. of genes	GC (%)
Chr01	1	828,170	33,817	166	44.4	1	830,022	4,000	168	47.1	1	838,594	19,200	179	39.9
Chr02	1	725,503	28,467	156	43.0	1	755,035	3,000	155	44.9	1	726,886	7,802	159	39.3
Chr03	1	1,037,329	79,518	210	43.2	1	1,011,127	19,000	210	44.8	1	973,297	36,100	202	40.1
Chr04	1	801,051	26,441	186	43.8	1	876,652	3,300	207	45.1	1	785,142	18,101	182	40.1
Chr05	1	1,296,396	61,024	287	42.5	1	1,370,936	4	311	44.3	1	1,324,984	70,301	287	39.3
Chr06	1	1,030,703	53,027	230	43.9	1	1,033,388	3,001	244	45.9	1	1,053,092	61,304	235	40.0
Chr07	1	1,492,599	49,712	342	43.2	1	1,497,819	100	353	45.6	1	1,526,735	37,202	350	39.2
Chr08	1	1,717,921	71,065	375	42.7	1	1,678,596	201	374	45.4	1	1,770,351	112,700	387	39.0
Chr09	1	1,985,783	55,262	436	43.0	1	1,923,364	0	432	46.1	1	2,147,124	61,200	460	39.0
Chr10	1	1,459,515	75,419	329	42.3	1	1,419,739	2,800	317	45.0	1	1,486,039	23,500	326	38.9
Chr11	1	2,076,872	52,909	454	42.4	1	2,067,354	4,000	459	45.1	1	2,372,884	82,100	491	38.7
Chr12	1	3,118,530	120,793	693	41.7	1	3,004,884	6,000	690	44.6	1	3,128,370	27,806	707	38.0
Chr13	1	2,094,899	32,653	447	42.7	1	2,031,768	0	445	45.7	1	2,200,295	87,600	465	38.5
Chr14	1	3,063,064	85,114	679	40.9	1	3,120,417	3,000	693	43.0	1	3,159,096	134,602	692	37.8
Unassigned	1,649 ^b	3,453,009	17,847	732	26.8	2,547 ^b	4,323,804	5,369	374	28.3	67	236,903	22	75	36.7
Total	1,663	26,181,344	843,068	5,722	40.4	2,561	26,944,905	53,775	5,432	42.3	81	23,729,792	779,540	5,197	38.8

^a Data are from PlasmoDB build 8.0 (<http://plasmodb.org>).

^b Contigs >1,000 bp in *P. cynomolgi* and >500 bp in *P. vivax*.

Supplementary Table 4. Examples of gene duplications in *P. cynomolgi*, *P. vivax* and *P. knowlesi*. Homologous genes (putative paralogs) were searched against the *P. cynomolgi* genome or PlasmoDB using BLASTP (<1e-16 with low complexity filtering).

#	Products	<i>P. cynomolgi</i>	<i>P. vivax</i>	<i>P. knowlesi</i>	<i>P. falciparum</i>	<i>P. berghei</i>	<i>P. chabaudi</i>	<i>P. yoelii</i>	Gene ID hit (BLASTP search)
1	sexual stage surface protein Pvs28	PCYB_062530	No ortholog	No ortholog	No ortholog	No ortholog	No ortholog	No ortholog	PCYB_007100 (1.0e-94) PCYB_062510 (1.0e-52) PCYB_062520 (5.0e-22)
2	knob-associated His-rich protein	No ortholog	PVX_081835	No ortholog	No ortholog	No ortholog	No ortholog	No ortholog	PVX_081835 (5.7e-157) PVX_003520 (1.5e-44) PVX_003525 (1.5e-18)
3	PV1H14010_P	No ortholog	PVX_119215	No ortholog	No ortholog	No ortholog	No ortholog	No ortholog	PVX_106730 (8.4e-22) PVX_121350 (3.3e-19)
4	glutathione synthetase	No ortholog	No ortholog	PKH_011060	No ortholog	No ortholog	No ortholog	No ortholog	PKH_102080 (0.0)
5	replication factor c subunit	No ortholog	No ortholog	PKH_052270	No ortholog	No ortholog	No ortholog	No ortholog	PKH_123880 (4.4e-179) PKH_145890 (4.4e-179) PKH_124240 (3.5e-51) PKH_040210 (8.5e-50)
6	MAEBL	No ortholog	No ortholog	PKH_070002	No ortholog	No ortholog	No ortholog	No ortholog	PKH_000510 (0.0)
7	replication factor c subunit	No ortholog	No ortholog	PKH_123880	No ortholog	No ortholog	No ortholog	No ortholog	PKH_145890 (2.1e-179) PKH_052270 (4.4e-179) PKH_124240 (3.5e-51) PKH_040210 (8.5e-50)
8	glutathione peroxidase	No ortholog	No ortholog	PKH_125860	No ortholog	No ortholog	No ortholog	No ortholog	PKH_131090 (1.9e-98)
9	pf47	No ortholog	No ortholog	PKH_142580	No ortholog	No ortholog	No ortholog	No ortholog	PKH_120710 (1.9e-139)

Supplementary Table 6. Copy number variations and orthologous relationships in the *dbl* and *rbl* genes of *P. vivax*, *P. knowlesi* and three strains of *P. cynomolgi*. Orthologous genes inferred from chromosomal location are in the same row. Pseudogenes are preceded by Ψ .

Gene	Chrom.	<i>P. vivax</i> Salvador I	<i>P. cynomolgi</i> B	<i>P. cynomolgi</i> Berok	<i>P. cynomolgi</i> Cambodian	<i>P. knowlesi</i> H
DBP	3		DBP2 (PCYB_033090)	PcyBer_DBP2(JQ422036)	PcyC_DBP2 (AB617789)	
	6	DBP (PVX_110810)	DBP1 (PCYB_063270)	PcyBer_DBP1(JQ422035)	PcyC_DBP1 (AB617788)	DBPa (PKH_062300)
	13	-	-	-	-	DBPg (PKH_134580)
	unknown	-	-	-	-	DBPb (PKH_000490)
RBP/NBP-1	7	PvRBP1a (PVX_098585)	PcyRBP1 (PCYB_071060)	PcyBer_RBP1a (JQ422037)	PcyC_RBP1a (JQ422044)	Ψ PkNBP1 (AY151130)
	7	PvRBP1b (PVX_098582)	-	PcyBer_RBP1b (JQ422038)	-	-
	unknown	PvRBP1 (partial) (PVX_125738)	-	-	-	-
RBP/NBP-2	5	PvRBP2c (PVX_090325)	PcyRBP2c (PCYB_053840)	PcyBer_RBP2c (JQ422040)	PcyC_RBP2c (JQ422047)	-
	5	PvRBP2 (partial) (PVX_090330)	PcyRBP2d (truncated) (PCYB_053850)	PcyBer_RBP2d (truncated) (JQ422041)	PcyC_RBP2d (truncated) (JQ422048)	-
	7	-	PcyRBP2e (PCYB_071010)	PcyBer_RBP2e (JQ422042)	PcyC_RBP2e (JQ422049)	PkNBPXb (EU867792)
	8	PvRBP2b (PVX_094255)	PcyRBP2b (PCYB_081060)	PcyBer_RBP2b (JQ422039)	PcyC_RBP2b (JQ422046)	-
	14	PvRBP2a (PVX_121920)	PcyRBP2a (fragment) (PCYB_141090)	-	PcyC_RBP2a (fragment) (JQ422045)	-
	14	Ψ PvRBP2d (PVX_101585)	-	-	-	-
	14	PvRBP2 (partial) (PVX_101590)	-	-	-	-
	unknown		PcyRBP2f (fragment) (PCYB_002010)	-	-	-
	RBP/NBP-3	14	Ψ PvRBP3 (PVX_101495)	PcyRBP3 (PCYB_147650)	PcyBer_RBP3 (JQ422043)	PcyC_RBP3 (JQ422050)

Supplementary Table 8. Candidate genes for hypnozoite formation. Ran (PCYB_092380), also identified in a previous screen for dormancy-related candidates in *P. vivax*¹, is involved in diverse cellular processes including the regulation of nucleocytoplasmic transport². We speculate that transport of yet-unidentified molecule(s) through the nuclear pore may trigger expression of a gene(s) associated with hypnozoite formation.

A. Three genes with sporozoite-specific ApiAP2 motifs in *P. vivax* and *P. cynomolgi*-unique genes.

	<i>P. vivax</i>	<i>P. cynomolgi</i>	Product
1	PVX_093700	PCYB_012620	Hypothetical protein
2	PVX_101530	PCYB_147720	Hypothetical protein (including unread regions)
3	PVX_101545	PCYB_147750	Hypothetical protein

B. Nine genes with sporozoite-specific ApiAP2 motifs in *P. vivax* and *P. cynomolgi* homologs of dormancy genes in GenBank.

	<i>P. vivax</i>	<i>P. cynomolgi</i>	Product
1	PVX_089980	PCYB_053210	cdc2-related protein kinase 1
2	PVX_099825	PCYB_073530	serine/threonine protein phosphatase
3	PVX_119815	PCYB_083540	serine/threonine protein phosphatase
4	PVX_091515	PCYB_092380	GTP-binding nuclear protein Ran
5	PVX_114825	PCYB_112360	protein kinase Crk2
6	PVX_083360	PCYB_121440	sexual stage-specific protein kinase
7	PVX_118025	PCYB_126580	transcription factor IIIb subunit
8	PVX_117925	PCYB_126400	elongation factor 2
9	PVX_085535	PCYB_133830	RNA helicase

Supplementary Table 9. Genome-wide comparative analysis between three strains of *P. cynomolgi* and *P. vivax* Salvador I.

	Berok	Cambodian	Berok and Cambodian combined alignment	<i>P. cynomolgi</i> B vs. <i>P. vivax</i> Salvador I
No. reads aligned	1,619,199	1,023,588	2,286,713	1,218,895
Unique coverage (%)	1,602,765 ^a /3,142,167 (51%)	1,023,452/1,633,472 (62.7%)	2,2815,70/4,101,464 (55.6%)	1,182,905/2,134,338 (55.4%)
SNPs (with filtering)	49,655 (1/543 bp)	30,907 (1/873 bp)	178,732 (1/151 bp)	1,388,772 (1/19 bp)
Synonymous Site SNPs	28,826	13,473	101,407	807,033
Non-synonymous site SNPs	10,403	5,148	35,148	312,910
Intron SNPs	3,113	2,627	12,651	68,657
Intergenic SNPs	7,313	9,659	29,526	157,071
Ts/Tv ratio^b	1.983	1.865	2.068	1.689
dN/dS	ND	ND	ND	0.251

^a A smaller number of reads were used to calculate the unique coverage because during the merging process of single end with paired end reads, duplicate reads were automatically removed.

^b The transition/transversion ratio. The ratio of transitions to transversions in a particular set of SNP calls is a useful diagnostic of correctly called SNPs, since SNPs do not occur randomly and transitions are expected to occur twice as frequently as transversions.

Supplementary Table 13. Comparison of the *P. cynomolgi* B genome sequence with 23 *P. cynomolgi* protein-coding genes obtained by Sanger sequencing.

	Gene	GenBank identifier	Gene identifier ^a	Chr #	Gene length (bp)	Length covered by assembly (bp)	Length covered by assembly (%)	Identity (bp)	Identity (%)
1	P-type Ca ²⁺ -ATPase (serca)	This study	PCYB_021770	2	3558	3558	100	3553	99.86
2	Gametocyte surface protein	AB574620	PCYB_042090	4	8403	8403	100	8386	99.80
3	Serine-repeat antigen 1 (sera1)	AB576872 ^b	PCYB_042190	4	3069	3069	100	3066	99.90
4	Serine-repeat antigen 2 (sera2)	AB576872 ^b	PCYB_042200	4	3759	3616	96.2 ^c	3614	99.94
5	Serine-repeat antigen 3 (sera3)	AB576872 ^b	PCYB_042210	4	3132	3132	100	3121	99.65
6	Serine-repeat antigen 4 (sera4)	AB576872 ^b	PCYB_042220	4	3252	3252	100	3252	100
7	Serine-repeat antigen 5 (sera5)	AB576872 ^b	PCYB_042230	4	3222	3222	100	3205	99.47
8	Serine-repeat antigen 6 (sera6)	AB576872 ^b	PCYB_042240	4	3300	3300	100	3296	99.88
9	Serine-repeat antigen 7 (sera7)	AB576872 ^b	PCYB_042250	4	1678	1678	100	1678	100
10	Serine-repeat antigen 8 (sera8)	AB576872 ^b	PCYB_042270	4	3090	3090	100	3065	99.19
11	Serine-repeat antigen 9 (sera9)	AB576872 ^b	PCYB_042280	4	3243	3243	100	3242	99.97
12	Serine-repeat antigen 10	AB576872 ^b	PCYB_042290	4	3060	3060	100	3060	100
13	Serine-repeat antigen 11 (sera11)	AB576872 ^b	PCYB_042300	4	3243	3243	100	3242	99.97
14	Serine-repeat antigen 12	AB576872 ^b	PCYB_042310	4	2157	2157	100	2157	100
15	Pyruvate kinase	This study	PCYB_063070	6	2112	2112	100	2107	99.76
16	Merozoite surface protein 1	AB266188	PCYB_073770	7	5394	5394	100	5380	99.74
17	Phosphoglycerate kinase	This study	PCYB_073010	7	1251	1251	100	1251	100
18	Enolase	This study	PCYB_082570	8	1341	1341	100	1340	99.93
19	Malate dehydrogenase	This study	PCYB_113900	11	942	942	100	941	99.89
20	RNA binding protein	This study	PCYB_112120	11	1545	1545	100	1545	100
21	Phosphoenolpyruvate	This study	PCYB_122090	12	1800	1800	100	1800	100
22	Lactate dehydrogenase	This study	PCYB_123790	12	951	951	100	949	99.79
23	Glucose-6-phosphate isomerase	This study	PCYB_132290	13	1779	1779	100	1776	99.83
	Total				65281	65138	99.8	65026	99.83

^a Sequences were obtained using next generation sequencing platforms Roche 454 GS FLX and Illumina GA II.

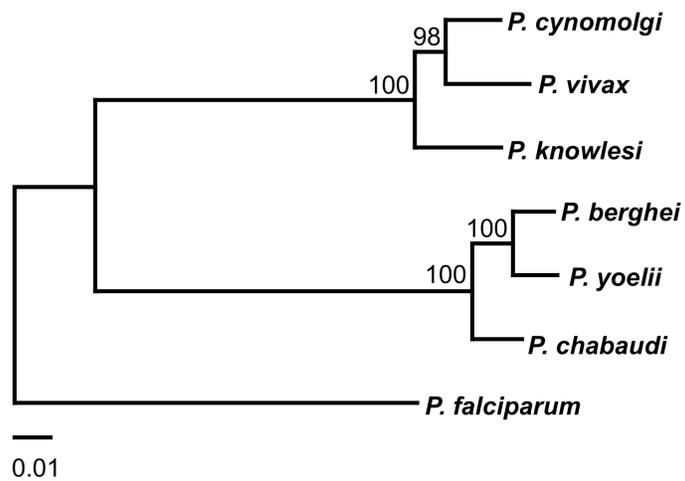
^b Sequences in GenBank were obtained from the Mulligan strain (ATCC30155). As described in the supplementary methods, the two strains have been shown to be identical due to a laboratory mix-up and are considered to be the same strain with different strain names.

^c The low sequence coverage was caused by insertions/deletions of low complexity repeat sequences.

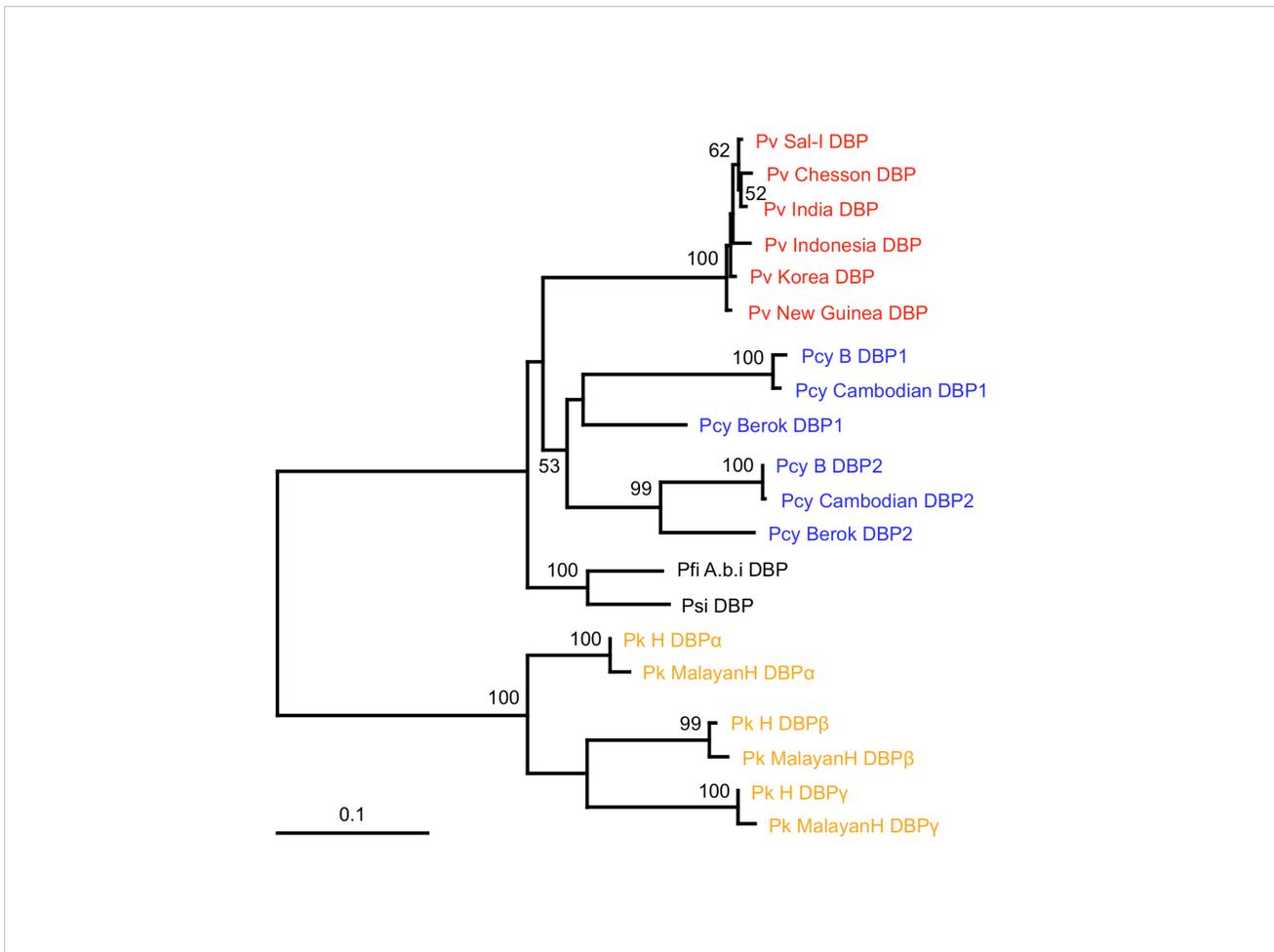
Supplementary Table 14. List of primer sequences.

Gene	Species/ application	Forward primer	Reverse primer
5'-region of <i>rbp3</i>	<i>P. cynomolgi</i>	PcyRBP3-F: 5'-ATTACTGTAAAAGTCAAAGTGAAGTTTCA-3'	PcyRBP3-R: 5'-AAAAAGGAACTAGGACTTCCATTAGGTTCA-3'
5'-region of <i>rbp3</i>	<i>P. vivax</i>	PvRBP3-F: 5'-TAAAAGTCAAAGTGAAGTTTCGACAGATCT-3'	PvRBP3-R: 5'-TTGGACACCATTCTTCTCTAATATGATCA -3'
<i>dbp</i>	<i>P. fieldi</i> & <i>P. simiovale</i>	conDBP-F1: 5'-ATATAYAGRYGGATTTCGAGAATGGGGA-3'	conDBP-R1: 5'-TAAAATCRTCCTTYTGWAAATCCTTTTC-3'
<i>dbp</i>	For sequence verification	conDBP-F2: 5'-AYAGRYGGATTTCGAGAATGGGGA-3'	conDBP-R2: 5'-TCRTCCTTYTGWAAATCCTTTTC-3'
<i>dbp</i>	For full length amplification	conDBP-F3: 5'-ACACTTTCTGTTCTGAATAATATWACCACA-3'	conDBP-R3: 5'-TGCTTGTAATCATTTCATGCATAAGCTTGA-3'

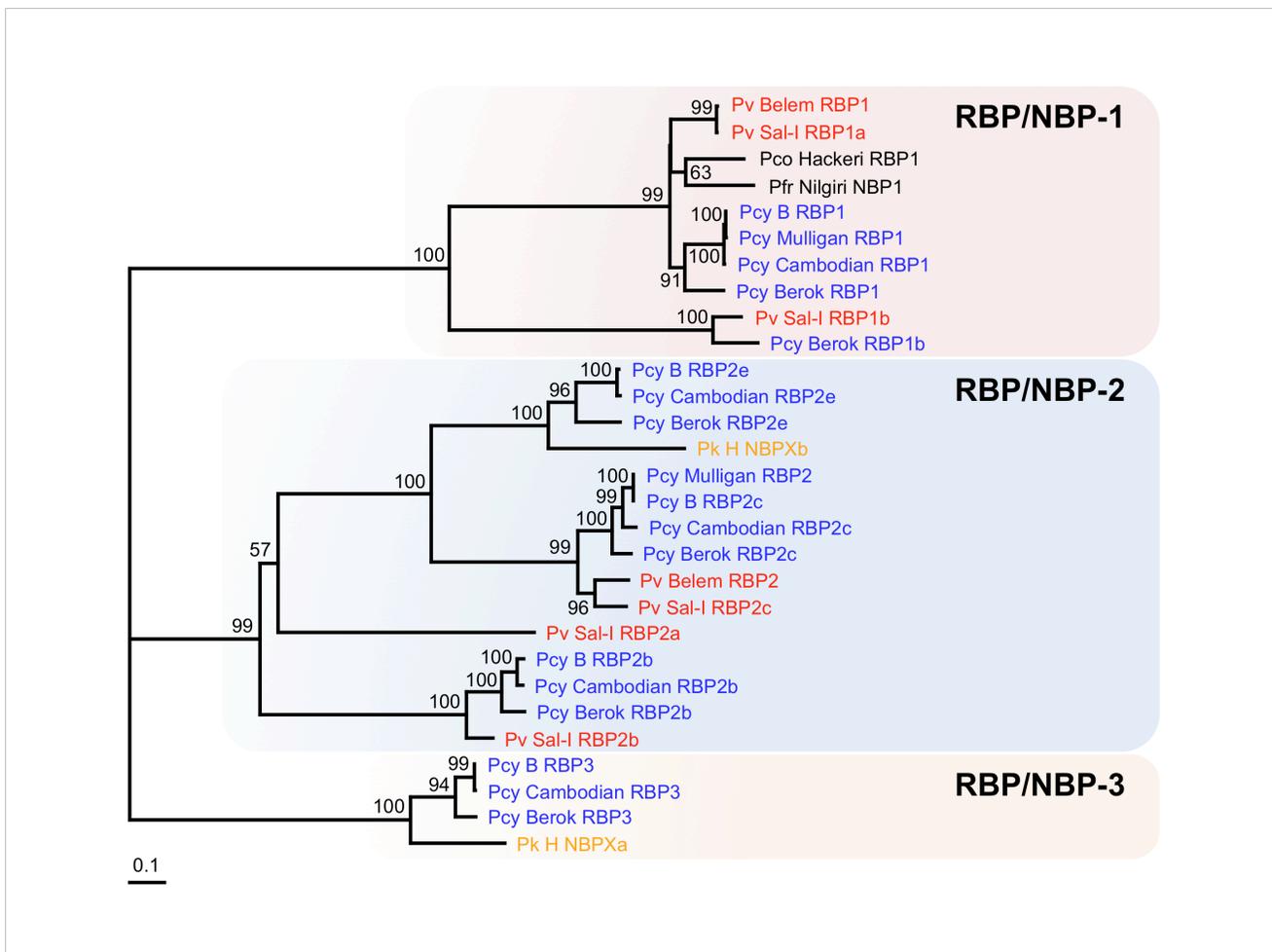
Supplementary Figure 1. A maximum likelihood phylogenetic tree of 192 protein genes from seven *Plasmodium* species. This tree was inferred using 49,617 amino acid sites under JTT + γ model (eight categories, $\alpha = 0.36$). Numbers shown along nodes represent bootstrap values.



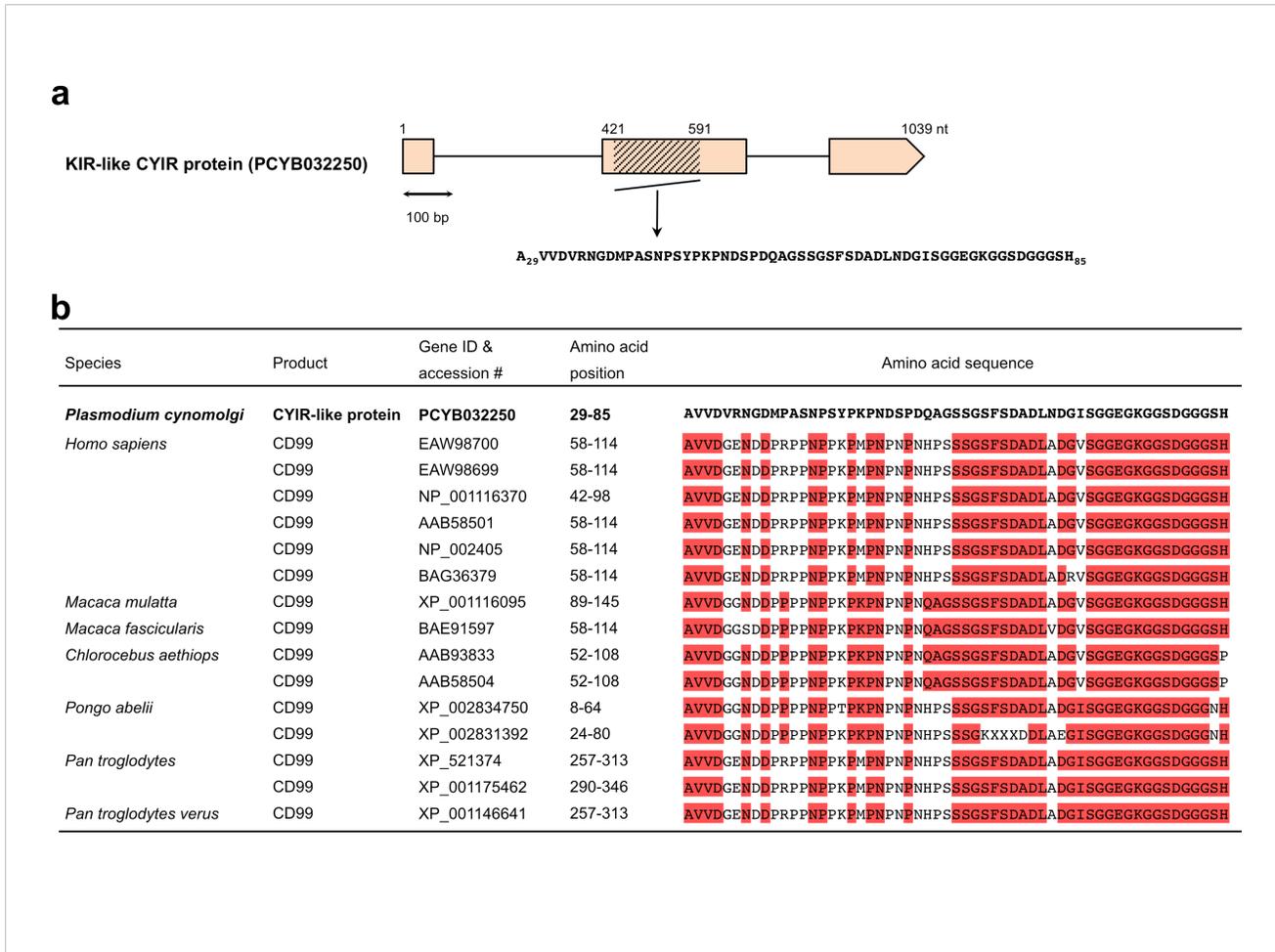
Supplementary Figure 2. A maximum likelihood phylogenetic tree of DBP from *P. vivax* and related monkey malaria parasites. The tree was inferred from 20 *dbp* genes using 715 amino acid sites under JTT + F + γ model (eight categories, $\alpha = 1.44$). Numbers shown along nodes represent bootstrap values. Abbreviations: Pv = *P. vivax*, Pcy = *P. cynomolgi*, Pfi = *P. fieldi*, Psi = *P. simiovale*, Pk = *P. knowlesi*. GenBank accession numbers are: Pv Sal-I DBP (PVX_110810), Pv Chesson DBP (EU395589), Pv India DBP (DQ156514), Pv Indonesia DBP (EU395591), Pv Korea DBP (DQ156523), Pv New Guinea DBP (DQ156519), Pcy B DBP1 (PCYB_063270), Pcy Cambodian DBP1 (AB617788), Pcy Berok DBP1 (JQ422035), Pcy B DBP2 (PCYB_033090), Pcy Cambodian DBP2 (AB617789), Pcy Berok DBP2 (JQ422036), Pfi A.b.i DBP (AB617790), Psi DBP (AB617791), Pk H DBP α (PKH_062300), Pk MalayanH DBP α (M90446), Pk H DBP β (PKH_000490), Pk MalayanH DBP β (M90694), Pk H DBP γ (PKH_134580), and Pk MalayanH DBP γ (M90695).



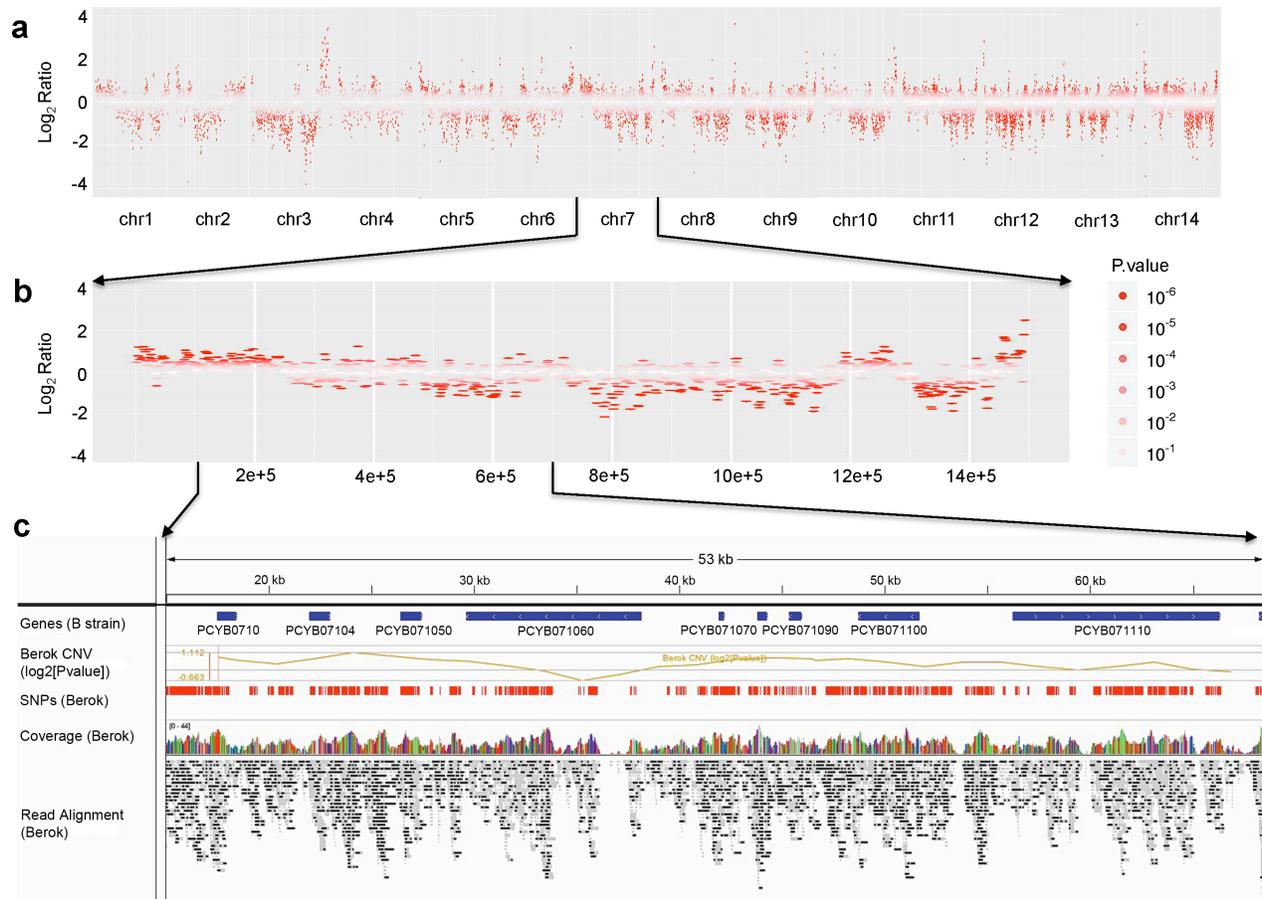
Supplementary Figure 3. A maximum likelihood phylogenetic tree of RBP/NBP from *P. vivax*, *P. cynomolgi* and *P. knowlesi*. The tree was inferred from 29 *rbp/nbp* genes using 1,506 amino acid sites under JTT - F + γ model (eight categories, $\alpha = 4.02672$). Numbers shown along nodes represent bootstrap values. Abbreviations: Pv = *P. vivax*, Pco = *P. coatneyi*, Pfr = *P. fragile*, Pcy = *P. cynomolgi*, Pk = *P. knowlesi*. GenBank Accession Numbers are: Pv Belem RBP1 (M88097), Pv Salvador I RBP1a (PVX_098585), Pco Hackeri RBP1 (DQ973816), Pfr Nilgiri RBP1 (DQ973815), Pcy B RBP1 (PCYB_071060), Pcy Mulligan RBP1 (DQ011582), Pcy Cambodian RBP1 (JQ422044), Pcy Berok RBP1 (DQ011581), Pv Salvador I RBP1b (PVX_098582), Pcy Berok RBP1b (JQ422038), Pcy B RBP2e (PCYB_071010), Pcy Cambodian RBP2e (JQ422049), Pcy Berok RBP2e (JQ422042), Pk H NBPXb (EU867792), Pcy Mulligan RBP2 (DQ011584), Pcy B RBP2c (PCYB_053840), Pcy Cambodian RBP2c (JQ422047), Pcy Berok RBP2c (DQ011583), Pv Belem RBP2 (AF184623), Pv Salvador I RBP2c (PVX_090325), Pv Salvador I RBP2a (PVX_121920), Pcy B RBP2b (PCYB_081060), Pcy Cambodian RBP2b (JQ422046), Pcy Berok RBP2b (JQ422039), Pv Salvador I RBP2b (PVX_094255), Pcy B RBP3 (PCYB_147650), Pcy Cambodian RBP3 (JQ422050), Pcy Berok RBP3 (JQ422043), Pk H NBPXa (EU867791).



Supplementary Figure 4. A region in a *P. cynomolgi* CYIR protein with high sequence similarity to primate and monkey CD99. (a) Gene structure of the CYIR protein (PCYB_032250) with boxes and lines indicating exons and introns, respectively. The 57 amino acid CD99 region is designated by diagonal lines. (b) Amino acid sequence region in the CYIR protein showing high sequence similarity to CD99 proteins from seven primate and monkey species. Residues identical to the CYIR protein are highlighted in red.



Supplementary Figure 5. Copy number variation in the *P. cynomolgi* genome. (a) A genome-wide \log_2 ratio plot of CNVs identified from a comparison of *P. cynomolgi* strain B with Berok. (b) A plot for chromosome 7. The red color gradient in (a) and (b) represents $\log_{10} p$ calculated on each of the ratios. (c) A detailed view of a ~53 kb region on chromosome 7 containing the *rbp1* gene in strain B (PCYB_071060) viewed using Integrative Genomics Viewer v2.0. SNPs between the B and Berok strains are indicated in red, and read alignment and coverage are shown in the bottom panel.



Supplementary Note

Parasite material. The *P. cynomolgi* B (*P. cynomolgi bastianellii*) strain was originally isolated from Malaya (Malaysia) in the 1950's³. Subsequent genetic testing of archived cryopreserved samples at CDC has shown the strain to be identical to the Mulligan strain isolated in Malaya in the 1930s⁴, indicating that there has been a mix-up and the two strains are now the same⁵ (J. W. Barnwell, personal communication). The Berok strain was isolated from a *Macaca nemestrina* monkey in the State of Perak in Malaysia in 1964⁶, and the Cambodian strain was isolated from a *Macaca irus* monkey captured in Cambodia in 1962⁷. DNA and frozen stabiliates of the strains are available through ATCC (<http://www.atcc.org/>). Individual vials of *P. cynomolgi* B strain (ATCC #30129), Cambodian strain (ATCC #30046), and Berok strain (CDC Repository of Primate Malaria Species) were thawed, inoculated into splenectomized *Macaca* monkeys, and at high parasitemia (9.4% strain B; 9.1% strain Cambodian; range of 3-10% in various monkeys for strain Berok) infected blood was withdrawn. For the *P. cynomolgi* B and Cambodian strains: whole blood was centrifuged to remove the buffy coat, filtrated twice through IMUGARD-III (Terumo, Japan) to remove leukocytes, and hemolyzed with 0.5 % saponin in HBS (HEPES-buffered saline). Erythrocyte-free parasites were sedimented by centrifugation at 15,000 rpm for 10 min at 4 °C. For the Berok strain, 120 mg of adenosine di-phosphate was added to the whole blood, then passaged through an acid washed glass bead (0.1mm diameter) column to remove platelets. Recovered blood was passaged through a Plasmodipur filter to remove leukocytes, and then through a column of Whatman CF11 cellulose fibers pre-wet with Hank's Basal Salt medium to remove residual host monocytes and lymphocytes. The parasites were matured before DNA extraction by *in vitro* culture in RPMI-1640 medium with 10% human AB serum. Infected red blood cells were subsequently concentrated by layering over a 48% Percoll cushion and centrifuging at 2200 rpm for 20 minutes, and the interface containing >90% iRBC was extracted for DNA purification. All parasite genomic DNA was purified with QIAamp DNA Blood Mini Kit (QIAGEN, Germany). *P. cynomolgi* cDNA was prepared from B strain-infected monkey blood using Takara PrimeScript High Fidelity RT-PCR Kit according to the manufacturer's instruction. We also used *P. vivax* cDNA previously prepared from infected patients⁸.

RBP sequencing. cDNA sequences of the 5'-region of *rbp3* from *P. cynomolgi* and *P. vivax* were obtained by PCR using the primers indicated in **Supplementary Table 14**.

Duffy-binding protein sequencing. The *dbp* genes of *P. fieldi* (A.b.i. strain; ATCC #30164) and *P. simiovale* (ATCC #30140) were amplified using primers conserved in *P. cynomolgi* (PCYB_063270 and PCYB_033090), *P. vivax* (PVX_110810) and *P. knowlesi* (PKH_062300, PKH_000490 and PKH_134580) using the primers indicated in **Supplementary Table 14**. *P. fieldi*, *P. simiovale* and DNA of the three parasite species as positive controls were used in 20 µl reaction mixtures containing 1 µl DNA, 0.2 µM of each primer, 1 U Takara LA-Taq DNA polymerase (Takara Bio Inc., Japan), 10× LA PCR Buffer II, 2.5 mM MgCl₂, and 0.4 mM dNTP, with PCR conditions of denaturation at 93 °C for 1 min, 35 cycles of amplification (93 °C for 20 s and 62 °C for 5 min), final elongation period of 72 °C for 10 min. PCR products were cloned into pCR XL TOPO (Invitrogen), and >12 clones sequenced (3130 Genetic Analyzer, Applied Biosystems, Foster City, CA, USA) with primers conDBP-f2:

and conDBP-r2. Full-length *P. fieldi* and *P. simiovale dbp* sequences were obtained by PCR-amplification and direct sequencing using primers conDBP-F3 and conDBP-R3 designed from the 5'- and 3'-UTR sequences of *P. vivax* and *P. cynomolgi dbp* genes, and with eleven primers designed to cover target regions in both directions.

SNP calling. We found multiple SNP candidates with low alignment quality located at the end of chromosomes, which in *Plasmodium* are known to be repetitive regions containing multicopy genes. We increased the thresholds (minimum root mean of the square mapping quality for SNPs = 25, minimum read depth = 4, 2 SNPs within 9 bp around a gap removed) to remove SNPs in these regions, and the boundaries for these regions as defined by the optimal parameters of the variation filter are shown below:

Chr	Start	End	No. bases
1	43,127	828,164	785,037
2	14,011	721,668	707,657
3	3,965	998,016	994,051
4	1,1016	46,0705	44,9689
5	2,542	1,209,953	1,207,411
6	100,822	1,030,566	929,744
7	10,033	1,492,473	1,482,440
8	20,164	1,717,905	1,697,741
9	326,963	1,984,637	1,657,674
10	544,450	1,459,499	915,049
11	12,872	2,064,127	2,051,255
12	23	2,576,113	2,576,090
13	80	2,0930,41	2,092,961
14	7,237	3,051,752	3,044,515
Total			20,591,314

References

1. Carlton, J.M. *et al.* Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* **455**, 757-63 (2008).
2. Yudin, D. & Fainzilber, M. Ran on tracks--cytoplasmic roles for a nuclear regulator. *J Cell Sci* **122**, 587-93 (2009).
3. Garnham, P.C. A new sub-species of *Plasmodium cynomolgi*. *Riv. Di Parassit.* **20**, 273-8 (1959).
4. Mulligan, H.W. Description of two species of monkey *Plasmodium* isolated from *Silenus irus*. *Arch. f. Protist.* **84**, 285-314 (1935).
5. Galinski, M.R. *et al.* The circumsporozoite gene of the *Plasmodium cynomolgi* complex. *Cell* **48**, 311-9

(1987).

6. Coatney, G.R., Collins, W.E., Warren, M. & P.G., C. *The Primate Malaria*, (U.S. Department of Health, Education and Welfare., Washington, D.C., 1971).
7. Bennett, G.F. & Warren, M. Transmission of a New Strain of *Plasmodium cynomolgi* to Man. *J Parasitol* **51**, 79-80 (1965).
8. Palacpac, N.M. *et al.* *Plasmodium vivax* serine repeat antigen (SERA) multigene family exhibits similar expression patterns in independent infections. *Mol Biochem Parasitol* **150**, 353-8 (2006).