

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 content of 40.5%), 14 positionally conserved centromeres and the presence of intrachromosomal telomeric sequences (ITs; GGGTT(T/C)A), which were discovered in the *P. knowlesi* genome⁹ but are absent in *P. vivax* (Fig. 1, Table 1 and Supplementary Table 2).

We annotated the *P. cynomolgi* strain B genome using a combination of *ab initio* gene prediction programs trained on high-quality data sets and sequence similarity searches against the annotated *P. vivax* and *P. knowlesi* genomes. Not unexpectedly for species from the same monkey malaria clade, gene synteny along the 14 chromosomes is highly conserved, although numerous microsyntenic breaks are present in regions containing multigene families (Fig. 2 and Table 2). This genome-wide view of synteny in six species of *Plasmodium* also 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*. The *P. cynomolgi* genome contains 5,722 genes, of which approximately half encode conserved hypothetical proteins of unknown function, as is the case in all the *Plasmodium* genomes sequenced to date. A maximum-likelihood phylogenetic tree constructed using 192 conserved ribosomal and translation- and transcription-related genes (Supplementary Fig. 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 duplicated homolog of *P. vivax* *Pvs28*—which encodes a sexual stage surface antigen that is a transmission-blocking vaccine candidate¹³—in *P. cynomolgi* (Supplementary Table 4). Genes common only to *P. cynomolgi* and *P. vivax* ($n = 214$) outnumber those that are restricted to *P. cynomolgi* and *P. knowlesi* ($n = 100$) or *P. vivax* and *P. knowlesi* ($n = 17$). Such figures establish the usefulness of *P. cynomolgi* as a model species for studying the more intractable *P. vivax*.

Notably, 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 numbers of the genes composing multigene families were generally greater in the *P. cynomolgi*–*P. vivax* lineage than in *P. knowlesi*, suggesting repeated gene duplication in the ancestral lineage of *P. cynomolgi* and *P. vivax* (or repeated gene deletion in the *P. knowlesi* lineage). Thus, the genomes of *P. cynomolgi*, *P. vivax* and *P. knowlesi* can largely be distinguished by variations in the copy number of multigene family members. 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*).

In malaria-causing parasites, invasion of host erythrocytes, mediated by specific interactions between parasite ligands and erythrocyte receptors, is a crucial component of the parasite lifecycle. Of great interest are the *eb1* and *rbl* gene families, which encode parasite ligands required for the recognition of host erythrocytes. The *eb1* genes encode erythrocyte binding-like (EBL) ligands such as the Duffy-binding proteins (DBPs) that bind to Duffy antigen receptor for chemokines (DARC) on human and monkey erythrocytes. The *rbl* 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 receptors¹⁴. We confirmed the presence of two *dbp* genes in *P. cynomolgi*¹⁵ (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 thus its infectivity of Old World monkey erythrocytes, after divergence from a common *P. vivax*–*P. cynomolgi* ancestor. This hypothesis is also supported by our identification of single-copy *dbp* genes in two other closely related Old World monkey malaria-causing parasites, *Plasmodium fieldi* and *Plasmodium simiovale*, which are incapable of infecting humans¹⁶. These two Old World monkey species lost one or more *dbp* genes during divergence that confer infectivity to humans, whereas *P. cynomolgi* and *P. knowlesi* retained *dbp* genes that allow invasion of human erythrocytes (Supplementary Fig. 2).

Figure 1 Architecture of the *P. cynomolgi* genome and associated genome-wide variation data. Data are shown for each of the 14 *P. cynomolgi* chromosomes. The six concentric rings, from outermost to innermost, represent (i) the location of 5,049 *P. cynomolgi* genes, excluding those on small contigs (cyan lines); (ii) 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 gray lines), 13 *msp7* (merozoite surface protein 7; gray lines), 25 *rad* (silver lines), 8 *etramp* (orange lines), 16 *Pf-fam-b* (light blue lines) and 7 *Pv-fam-d* (light green lines); (iii) plot of d_S/d_N for 4,605 orthologs depicting genome-wide polymorphism within *P. cynomolgi* strains B and Berok (black line) and 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); (iv) heatmap indicating SNP density of 3 *P. cynomolgi* strains plotted per 10-kb windows: red, 0–83 SNPs per 10 kb (regions of lowest SNP density); blue, 84–166 SNPs per 10 kb; green, 167–250 SNPs per 10 kb; purple, 251–333 SNPs per 10 kb; orange, 334–416 SNPs per 10 kb; yellow, 417–500 SNPs per 10 kb (regions of highest SNP density); (v) \log_2 ratio plot of CNVs identified from a comparison of *P. cynomolgi* strains B and Berok; and (vi) map of 182 polymorphic intergenic microsatellites (MS, black dots). The figure was generated using Circos software (see URLs).

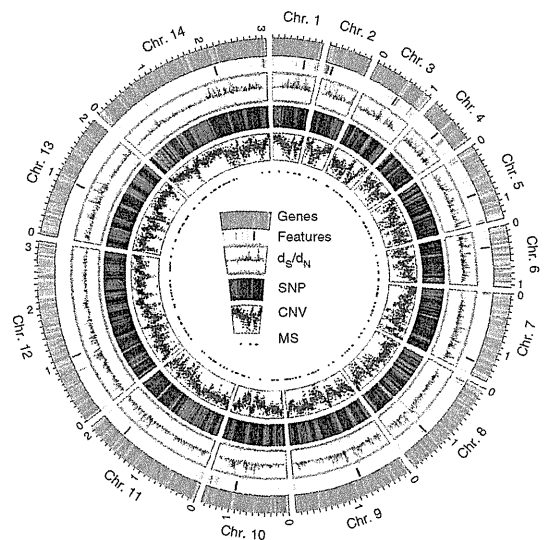


Table 1 Comparison of genome features between *P. cynomolgi*, *P. vivax* and *P. knowlesi*, three species of the monkey malaria clade

Feature	<i>P. cynomolgi</i>	<i>P. vivax</i> ¹²	<i>P. knowlesi</i> ⁹
Assembly			
Size (Mb)	26.2	26.9	23.7
Number of scaffolds ^a	14 (1,649)	14 (2,547)	14 (67)
Coverage (fold)	161	10	8
GC content (%)	40.4	42.3	38.8
Genes			
Number of genes	5,722	5,432	5,197
Mean gene length (bp)	2,240	2,164	2,180
Gene density (bp per gene) ^b	4,428.2	4,950.5	4,416.1
Percentage coding ^b	51.0	47.1	49.0
Structural RNAs			
Number of tRNA genes	43	44	41
Number of 5S rRNA genes	3	3	0 ^c
Number of 5.8S, 18S and 28S rRNA units	7	7	5
Nuclear genome			
Number of chromosomes	14	14	14
Number of centromeres	14	14	14
Isochore structure ^d	+	+	-
Mitochondrial genome			
Size (bp) ^e	5,986 (AB444123)	5,990 (AY598140)	5,958 (AB444108)
GC content (%)	30.3	30.5	30.5
Apicoplast genome			
Size (bp)	29,297 ^f	5,064 ^g	N/A
GC content (%)	13.0	17.1	N/A

N/A, not available.

^aSmall unassigned contigs indicated in parentheses. ^bSequence gaps excluded. ^cNot present in *P. knowlesi* assembly version 4.0. ^dRegions of the genome that differ in density and are separable by CsCl centrifugation; isochores correspond to domains differing in GC content.^eIdentified in other studies (GenBank accessions given in parentheses). ^fPartial sequence (~86% complete) generated during this project. ^gPartial sequence of reference genome only published¹²; actual size is ~35 kb.

We found multiple *rpb* genes, some truncated or present as pseudogenes, in the *P. cynomolgi* genome (Fig. 1 and Table 2). Phylogenetic analysis showed that *rbl* genes 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 Fig. 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 from the RBP/NBP-3 and RBP/NBP-1 groups, respectively. Thus, *rbl* gene family expansion seems to have occurred after speciation, indicating that the three species have multiple species-specific erythrocyte invasion mechanisms. Notably, we found an ortholog of *P. vivax rpb1b* in some strains of *P. cynomolgi* but not in others (Supplementary Table 6). To our knowledge, this

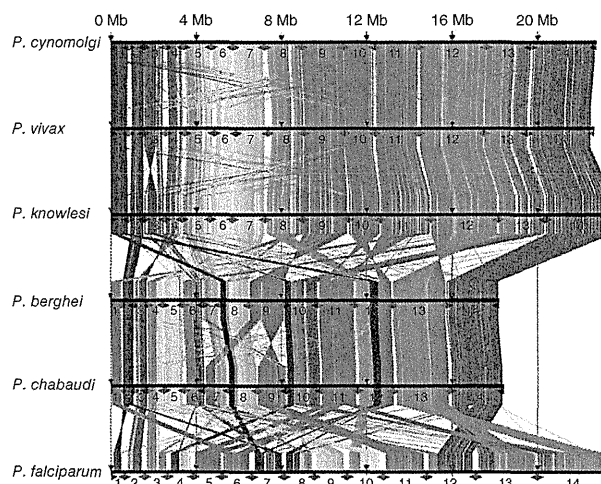
Figure 2 Genome synteny between six species of *Plasmodium* parasite. Protein-coding genes of *P. cynomolgi* are shown aligned with those of five other *Plasmodium* genomes: two species belonging to the monkey malaria clade, *P. vivax* and *P. knowlesi*; two species of rodent malaria, *P. berghei* and *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) with respect to genomic position of *P. cynomolgi*.

is the first example of a CNV for a *rpb* gene between strains of a single *Plasmodium* species, highlighting how repeated creation and destruction of *rbl* genes, a signature of adaptive evolution, may have enabled species of the monkey malaria clade to expand or switch between monkey and human hosts.

The largest gene family in *P. cynomolgi*, consisting of 256 *cyir* (cynomolgi-interspersed repeat) genes, is part of the *pir* (plasmodium-interspersed repeat) superfamily that includes *P. vivax vir* genes ($n = 319$) and *P. knowlesi kir* genes ($n = 70$) (Table 2). *Pir*-encoded proteins reside on the surface of infected erythrocytes and have an important role in immune evasion¹⁷. 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, notably, 11 *cyir* genes have sequence similarity to *P. knowlesi kir* genes (Supplementary Table 7) and occur more internally in the chromosomes, as do the *kir* genes in *P. knowlesi*. As with 'molecular mimicry' in *P. knowlesi* (mimicry of host sequences by pathogen sequences)⁹, one CYIR protein (encoded by PCYB_032250) has a region of 56 amino acids that is highly similar to the extracellular domain of primate CD99 (Supplementary Fig. 4), a molecule involved in the regulation of T-cell function. A new finding is that *P. cynomolgi* has two genes whose sequences are similar to *P. knowlesi SICAvir* genes (Supplementary Table 7) that are expressed on the surfaces of schizont-infected macaque erythrocytes and are involved in antigenic variation¹⁸.

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*¹⁹. In a search for candidate genes involved in the hypnozoite stage, we identified nine coding for 'dormancy-related' proteins that had the upstream ApiAP2 motifs²⁰ 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 of this elusive stage.

We sequenced *P. cynomolgi* strains Berok (from Malaysia) and Cambodian (from Cambodia) to 26 \times and 17 \times 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) at a frequency of 1 SNP per 151 bp, a polymorphism level somewhat



LETTERS

Table 2 Components of multigene families of *P. cynomolgi*, *P. vivax* and *P. knowlesi* differ in copy number

Family	Multigene family	Localization	Arrangement	<i>P. cynomolgi</i>	<i>P. vivax</i>	<i>P. knowlesi</i>	Putative function and other information
1	<i>pir</i> (<i>vir</i> -like)	Subtelomeric	Scattered and clustered	254	319 ^a	4	Immune evasion
2	<i>pir</i> (<i>kir</i> -like)	Subtelomeric and central	Scattered and clustered	11	2	66 ^a	Immune evasion
3	<i>SICAvar</i>	Subtelomeric and central	Scattered and clustered	2	1	242 ^a	Antigenic variation, immune evasion
4	<i>msp3</i>	Central	Clustered	12	12	3	Merozoite surface protein
5	<i>msp7</i>	Central	Clustered	13	13	5	Merozoite surface protein
6	<i>dbl</i> (<i>dbp/ebf</i>)	Subtelomeric	Scattered	2	1	3	Host cell recognition
7	<i>rbl</i> (<i>rpb/hbp/rh</i>)	Subtelomeric	Scattered	8 ^a	10 ^a	3 ^a	Host cell recognition
8	<i>Pv-fam-a</i> (<i>trag</i>)	Subtelomeric	Scattered and clustered	36	36	26 ^a	Tryptophan-rich
9	<i>Pv-fam-b</i>	Central	Clustered	3	6	1	Unknown
10	<i>Pv-fam-c</i>	Subtelomeric	Unknown ^b	1	7	0	Unknown
11	<i>Pv-fam-d</i> (<i>hypb</i>)	Subtelomeric	Scattered	18	16	2	Unknown
12	<i>Pv-fam-e</i> (<i>rad</i>)	Subtelomeric	Clustered	27	44	16	Unknown
13	<i>Pv-fam-g</i>	Central	Clustered	3	3	3	Unknown
14	<i>Pv-fam-h</i> (<i>hyp16</i>)	Central	Clustered	6	4	2	Unknown
15	<i>Pv-fam-i</i> (<i>hyp11</i>)	Subtelomeric	Scattered	6	6	5	Unknown
16	<i>Pk-fam-a</i>	Central	Scattered	0	0	12 ^a	Unknown
17	<i>Pk-fam-b</i>	Subtelomeric	Scattered	0	0	9	Unknown
18	<i>Pk-fam-c</i>	Subtelomeric	Scattered	0	0	6 ^a	Unknown
19	<i>Pk-fam-d</i>	Central	Scattered	0	0	3 ^a	Unknown
20	<i>Pk-fam-e</i>	Subtelomeric	Scattered	0	0	3 ^a	Unknown
21	<i>PST-A</i>	Subtelomeric and central	Scattered	9 ^a	11 ^a	7	β hydrolase
22	<i>ETRAMP</i>	Subtelomeric	Scattered	9	9	9	Parasitophorous vacuole membrane
23	<i>CLAG</i> (<i>RhopH-1</i>)	Subtelomeric	Scattered	2	3	2	High-molecular-weight rophtry antigen complex
24	<i>PvSTP1</i>	Subtelomeric	Unknown ^b	3	10 ^a	0	Unknown
25	<i>PHIST</i> (<i>Pf-fam-b</i>)	Subtelomeric	Scattered and clustered	21	20	15	Unknown
26	<i>SERA</i>	Central	Clustered	13 ^a	13 ^a	8 ^a	Cysteine protease

^aPseudogenes, truncated genes and gene fragments included. ^bGene arrangement could not be determined due to localization on unassigned contigs.

similar to that found when *P. falciparum* genomes are compared^{21,22}. We calculated the pairwise nucleotide diversity (π) as 5.41×10^{-3} across the genome, which varies little between the chromosomes. We assessed genome-wide CNVs between the *P. cynomolgi* B and Berok strains, using a robust statistical model in the CNV-seq program²³, by which we identified 1,570 CNVs (1 per 17 kb), including 1 containing the *rpb1b* gene on chromosome 7 (Supplementary Fig. 5). Finally, mining of the *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 studies of genetic diversity and population structure 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*²⁴.

We estimated the difference between the number of synonymous changes per synonymous site (d_S) and the number of nonsynonymous changes per nonsynonymous site (d_N) over 4,563 pairs of orthologs within *P. cynomolgi* strains B and Berok and 4,601 pairs of orthologs between these two *P. cynomolgi* strains and *P. vivax* Salvador I, using a simple Nei-Gojobori model²⁵. We found 63 genes with $d_N > d_S$ within the two *P. cynomolgi* strains and 3,265 genes with $d_S > d_N$ (Supplementary Table 11). Genes with relatively high d_N/d_S ratios include those encoding transmembrane proteins, such as antigens and transporters, among which is a transmission-blocking target antigen, Pcy230 (encoded by PCYB_042090). Notably, the *P. vivax* ortholog (PVX_003905) does not show evidence for positive selection²⁶, 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 83 genes under possible accelerated evolution but 3,739 genes under possible purifying selection (Supplementary Table 12). This conservative

estimate indicates that at least 81% of loci have diverged under strong constraint, compared with 1.8% of loci under less constraint or positive selection (Fig. 1), indicating that, overall, the genome of *P. cynomolgi* is highly conserved in single-locus genes compared to *P. vivax* and emphasizing the value of *P. cynomolgi* as a biomedical and evolutionary model for studying *P. vivax*.

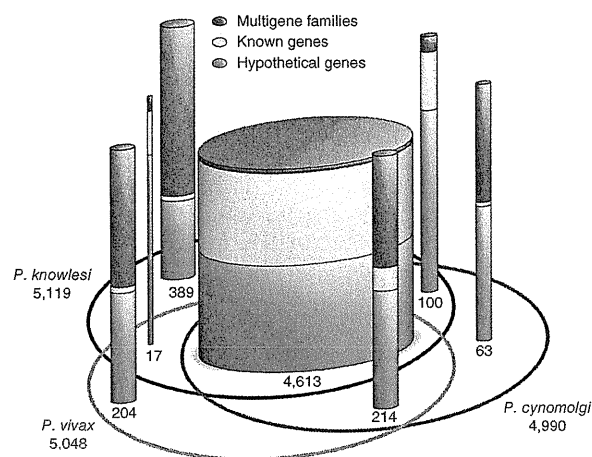


Figure 3 Comparison of the genes of *P. cynomolgi*, *P. vivax* and *P. knowlesi*. The Venn ellipses represent the three genomes, with the total number of genes assigned to the chromosomes indicated under the species name. Cylinders depict orthologous and non-orthologous genes between the three genomes, with the number of genes in each indicated and represented graphically by cylinder relative width. In each cylinder, genes are divided into three categories whose thickness is represented by colored bands proportional to category percentage.



Our generation of the first *P. cynomolgi* genome sequences is a critical step in the development of a robust model system for the intractable and neglected *P. vivax* species²⁷. Comparative genome analysis of *P. vivax* and the Old World monkey malaria-causing parasites *P. cynomolgi* and *P. knowlesi* presented here provides the foundation for further insights into traits such as host specificity that will enhance prospects for the eventual elimination of vivax-caused malaria and global malaria eradication.

URLs. PlasmoDB, <http://plasmodb.org/>; Circos, [## METHODS](http://circos.ca/MicroSatellite Identification tool (MISA), http://pgrc.ipk-gatersleben.de/misa/; dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/snp_viewBatch.cgi?sbid=1056645.</p>
</div>
<div data-bbox=)

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence data for the *P. cynomolgi* B, Cambodian and Berok strains have been deposited in the DNA Data Bank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL) and the GenBank databases under the following accessions: B strain sequence reads DRA000196, genome assembly BAEJ01000001–BAEJ01003341 and annotation DF157093–DF158755; Cambodian strain sequence reads DRA000197; and Berok strain sequence reads SRA047950. SNP calls have been submitted to dbSNP (NYU_CGSB_BIO; 1056645) and may also be downloaded from the dbSNP website (see URLs). Sequences of the *dbp* genes from *P. cynomolgi* (Cambodian strain), *P. fieldi* (A.b.i. strain) and *P. simiovale* (AB617788–AB617791) and the *P. cynomolgi* Berok strain (JQ422035–JQ422036) and *rbp* gene sequences from the *P. cynomolgi* Berok and Cambodian strains (JQ422037–JQ422050) have been deposited. A partial apicoplast genome of the *P. cynomolgi* Berok strain has been deposited (JQ522954). The *P. cynomolgi* B reference genome is also available through PlasmoDB (see URLs).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank H. Sawai for suggestions on genome analysis, D. Fisher for help with genome-wide evolutionary analyses and the 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 US National Institutes of Health (NIH)). Genome and phylogenetic analyses used the Genome Information Research Center in the Research Institute of Microbial Diseases at Osaka University. This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan (18073013, 18GS03140013, 20390120 and 22406012) to K.T., an NIH grant (R01 GM080586) to A.A.E. and a Burroughs Wellcome Fund grant (1007398) and an NIH International Centers of Excellence for Malaria Research grant (U19 AI089676-01) to J.M.C. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

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.

Published online at <http://www.nature.com/doi/10.1038/ng.2375>.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA) license. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/>.

- Mendis, K., Sina, B.J., Marchesini, P. & Carter, R. The neglected burden of *Plasmodium vivax* malaria. *Am. J. Trop. Med. Hyg.* **64**, 97–106 (2001).
- Mueller, I. *et al.* Key gaps in the knowledge of *Plasmodium vivax*, a neglected human malaria parasite. *Lancet Infect. Dis.* **9**, 555–566 (2009).
- Baird, J.K. Resistance to chloroquine unhinges vivax malaria therapeutics. *Antimicrob. Agents Chemother.* **55**, 1827–1830 (2011).
- Rayner, J.C., Liu, W., Peeters, M., Sharp, P.M. & Hahn, B.H. A plethora of *Plasmodium* species in wild apes: a source of human infection? *Trends Parasitol.* **27**, 222–229 (2011).
- Cornejo, O.E. & Escalante, A.A. The origin and age of *Plasmodium vivax*. *Trends Parasitol.* **22**, 558–563 (2006).
- Escalante, A.A. *et al.* A monkey's tale: the origin of *Plasmodium vivax* as a human malaria parasite. *Proc. Natl. Acad. Sci. USA* **102**, 1980–1985 (2005).
- Mu, J. *et al.* Host switch leads to emergence of *Plasmodium vivax* malaria in humans. *Mol. Biol. Evol.* **22**, 1686–1693 (2005).
- Singh, B. *et al.* A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. *Lancet* **363**, 1017–1024 (2004).
- Pain, A. *et al.* The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* **455**, 799–803 (2008).
- Eyles, D.E., Coatney, G.R. & Getz, M.E. Vivax-type malaria parasite of macaques transmissible to man. *Science* **131**, 1812–1813 (1960).
- Gibbs, R.A. *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222–234 (2007).
- Carlton, J.M. *et al.* Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* **455**, 757–763 (2008).
- Saxena, A.K., Wu, Y. & Garboczi, D.N. *Plasmodium* p25 and p28 surface proteins: potential transmission-blocking vaccines. *Eukaryot. Cell* **6**, 1260–1265 (2007).
- Iyer, J., Gruner, A.C., Renia, L., Snounou, G. & Preiser, P.R. Invasion of host cells by malaria parasites: a tale of two protein families. *Mol. Microbiol.* **65**, 231–249 (2007).
- Okenu, D.M., Malhotra, P., Lalitha, P.V., Chitnis, C.E. & Chauhan, V.S. Cloning and sequence analysis of a gene encoding an erythrocyte binding protein from *Plasmodium cynomolgi*. *Mol. Biochem. Parasitol.* **89**, 301–306 (1997).
- Coatney, G.R., Collins, W.E., Warren, M. & Contacos, P.G. *The Primate Malariae* (US Department of Health, Education and Welfare, Washington, DC, 1971).
- Cunningham, D., Lawton, J., Jarra, W., Preiser, P. & Langhorne, J. The *pir* multigene family of *Plasmodium*: antigenic variation and beyond. *Mol. Biochem. Parasitol.* **170**, 65–73 (2010).
- al-Khedery, B., Barnwell, J.W. & Galinski, M.R. Antigenic variation in malaria: a 3' genomic alteration associated with the expression of a *P. knowlesi* variant antigen. *Mol. Cell* **3**, 131–141 (1999).
- Krotoski, W.A. The hypnozoite and malarial relapse. *Prog. Clin. Parasitol.* **1**, 1–19 (1989).
- Campbell, T.L., De Silva, E.K., Olszewski, K.L., 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.* **6**, e1001165 (2010).
- Mu, J. *et al.* Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nat. Genet.* **39**, 126–130 (2007).
- Volkman, S.K. *et al.* A genome-wide map of diversity in *Plasmodium falciparum*. *Nat. Genet.* **39**, 113–119 (2007).
- Xie, C. & Tammi, M.T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**, 80 (2009).
- Lee, K.S. *et al.* *Plasmodium knowlesi*: reservoir hosts and tracking the emergence in humans and macaques. *PLoS Pathog.* **7**, e1002015 (2011).
- Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
- Doi, M. *et al.* Worldwide sequence conservation of transmission-blocking vaccine candidate Pvs230 in *Plasmodium vivax*. *Vaccine* **29**, 4308–4315 (2011).
- Carlton, J.M., Sina, B.J. & Adams, J.H. Why is *Plasmodium vivax* a neglected tropical disease? *PLoS Negl. Trop. Dis.* **5**, e1160 (2011).





ONLINE METHODS

Parasite material. Details of the origin of the *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 the Roche 454 GS FLX (Titanium) and Illumina/Solexa Genome Analyzer IIx platforms to 161× coverage. In addition, 2,784 clones (6.8 Mb) of a ~40-kb insert fosmid library in pCC1FOS (EpiCentre Biotechnologies) was 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 of >500 bp were mapped to the *P. vivax* Salvador I reference assembly¹² (PlasmoDB; see URLs). *P. cynomolgi* contigs were iteratively arrayed through alignment 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 ($e < 1 \times 10^{-15}$; identity > 90%; coverage > 200 bp) implemented in GenomeMatcher software version 1.65 (ref. 28). Additional linkages (699 regions) were made using PCR across the intervening sequence gaps with primers designed from neighboring contigs. The 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 ~22.73 Mb, encompassing ~80% of the predicted *P. cynomolgi* genome. A total of 1,651 contigs (>1 kb) with a total length of 3.45 Mb was identified as unassigned subtelomeric sequences by searching against the *P. vivax* genome using blastn. Additionally, 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). 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**). The *P. cynomolgi* Berok and Cambodian strains were sequenced to 26× and 17× coverage, respectively, using the Roche 454 GS FLX 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 Table 14** and **Supplementary Note**).

Prediction and annotation of genes. Gene prediction for the 14 supercontigs and 1,651 unassigned contigs was performed using the MAKER genome annotation pipeline²⁹ with *ab initio* gene prediction programs trained on proteins and ESTs from PlasmoDB Build 7.1. For gene annotation, blastn ($e < 1 \times 10^{-15}$; identity > 70%; coverage > 100 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 were 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* were determined on synteny. A unique identifier, PCYB_#####, was assigned to *P. cynomolgi* genes, where the first two of the six numbers indicate chromosome number. Paralogs of genes that seemed to be specific to either *P. cynomolgi*, *P. vivax* or *P. knowlesi* were searched using blastp with default parameters, using a cutoff *e* value of 1×10^{-16} .

Multiple genome sequence alignment. Predicted proteins of *P. cynomolgi* B strain were concatenated and aligned with those from the 14 chromosomes of 5 other *Plasmodium* genomes: *P. vivax*, *P. knowlesi*, *P. falciparum*, *P. berghei* and *P. chabaudi*, using Murasaki software version 1.68.6 (ref. 30).

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 cutoff *e* value of 1×10^{-12} and compositional adjustment (no adjustment) against the nonredundant protein sequence data set of the *M. mulatta* proteome in NCBI.

Phylogenetic analyses. Genes were aligned using ClustalW version 2.0.10 (ref. 31) with manual corrections, and unambiguously aligned sites were selected for phylogenetic analyses. Maximum-likelihood phylogenetic trees were constructed using PROML programs in PHYLIP version 3.69 (ref. 32) under the 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 γ distribution, with eight categories for approximating the site-rate distribution. CODEML programs in PAML 4.4 (ref. 33) were used for estimating the γ shape parameter, α 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* genome¹² to search the *P. cynomolgi* B sequence. For the orthologs identified in both species, ~1 kb of sequence 5' to the coding sequence was searched for four specific ApiAP2 motifs²⁰—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 upstream of the coding sequence for the presence of the four ApiAP2 motifs. Data for *P. vivax*, *P. knowlesi*, *P. falciparum*, *P. berghei*, *Plasmodium chabaudi* and *Plasmodium 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 (ref. 34), 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 the reads with the highest mapping quality were retained. We used a statistical method³⁵ 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 (π) 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. Letting π be the genetic distance between allele *i* and allele *j*, then the nucleotide diversity within the population is

$$\pi = \sum_{i,j} P_i P_j f_{ij}$$

where P_i and P_j are the overall allele frequencies of *i* and *j*, respectively. Mean π was calculated by averaging over sites, weighting each by $\frac{\pi-1}{\sum_{i=1}^{\pi} i}$, where n is the number of aligned sites. Average d_N/d_S ratios were

estimated using the modified Nei-Gojobori/Jukes-Cantor method in MEGA 4 (ref. 36).

CNV-seq²³ was used to identify potential CNVs 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 were used as the reference genome, and the *P. cynomolgi* Berok strain was used as a test genome; the sequence coverage of the Cambodian strain was considered too low for analysis. The test reads were mapped to the reference genome, and CNVs were detected by computing the number of reads for each test strain in a sliding window. The validity of the observed ratios was 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 nucleotides) 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 the Burrows-Wheeler Aligner (BWA)³⁷ and allowing for gaps. Using the Phred-scaled probability of the base being misaligned by SAMtools³⁵, 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 4,563 orthologs between *P. cynomolgi* strains B and Berok and 4,601 orthologs between these strains and *P. vivax* Salvador I used MUSCLE³⁸ 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 model²⁵. To detect values that could not be explained by chance, we estimated the standard error by a bootstrap procedure with 200 pseudoreplicates for each gene. The expected value for d_S/d_N is 0 if a given pair of sequences is diverging without obvious effects on fitness. In the case of the comparison within *P. cynomolgi*, values with a difference of ± 2 s.e.m. from 0 were considered indicative of an excess of synonymous ($d_S/d_N > 0$) or nonsynonymous ($d_S/d_N < 0$) changes. In the case of the comparison between *P. cynomolgi* and *P. vivax*, we used a more stringent criterion of ± 3 s.e.m. from 0.

28. Ohtsubo, Y., Ikeda-Ohtsubo, W., Nagata, Y. & Tsuda, M. GenomeMatcher: a graphical user interface for DNA sequence comparison. *BMC Bioinformatics* **9**, 376 (2008).
29. Cantarel, B.L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
30. Popendorf, K., Tsuyoshi, H., Osana, Y. & Sakakibara, Y. Murasaki: a fast, parallelizable algorithm to find anchors from multiple genomes. *PLoS ONE* **5**, e12651 (2010).
31. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
32. Felsenstein, J. *PHYLIP, Phylogeny Inference Package*, 3.6a3 edn (University of Washington, Seattle, 2005).
33. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
34. Ning, Z., Cox, A.J. & Mullikin, J.C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
35. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
36. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007).
37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
38. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).



RESEARCH ARTICLE

Open Access

Novel type of linear mitochondrial genomes with dual flip-flop inversion system in apicomplexan parasites, *Babesia microti* and *Babesia rodhaini*

Kenji Hikosaka^{1,2}, Naotoshi Tsuji³, Yoh-ichi Watanabe², Hiroe Kishine⁴, Toshihiro Horii⁵, Ikuo Igarashi⁶, Kiyoshi Kita^{2*} and Kazuyuki Tanabe^{1,5*}

Abstract

Background: Mitochondrial (mt) genomes vary considerably in size, structure and gene content. The mt genomes of the phylum Apicomplexa, which includes important human pathogens such as the malaria parasite *Plasmodium*, also show marked diversity of structure. *Plasmodium* has a concatenated linear mt genome of the smallest size (6-kb); *Babesia* and *Theileria* have a linear monomeric mt genome (6.5-kb to 8.2-kb) with terminal inverted repeats; *Eimeria*, which is distantly related to *Plasmodium* and *Babesia/Theileria*, possesses a mt genome (6.2-kb) with a concatemeric form similar to that of *Plasmodium*; *Cryptosporidium*, the earliest branching lineage within the phylum Apicomplexa, has no mt genome. We are interested in the evolutionary origin of linear mt genomes of *Babesia/Theileria*, and have investigated mt genome structures in members of archaeopiroplasmid, a lineage branched off earlier from *Babesia/Theileria*.

Results: The complete mt genomes of archaeopiroplasmid parasites, *Babesia microti* and *Babesia rodhaini*, were sequenced. The mt genomes of *B. microti* (11.1-kb) and *B. rodhaini* (6.9-kb) possess two pairs of unique inverted repeats, IR-A and IR-B. Flip-flop inversions between two IR-As and between two IR-Bs appear to generate four distinct genome structures that are present at an equi-molar ratio. An individual parasite contained multiple mt genome structures, with 20 copies and 2 – 3 copies per haploid nuclear genome in *B. microti* and *B. rodhaini*, respectively.

Conclusion: We found a novel linear monomeric mt genome structure of *B. microti* and *B. rodhaini* equipped with dual flip-flop inversion system, by which four distinct genome structures are readily generated. To our knowledge, this study is the first to report the presence of two pairs of distinct IR sequences within a monomeric linear mt genome. The present finding provides insight into further understanding of evolution of mt genome structure.

Keywords: Mitochondrial genome, *Babesia/Theileria*, Piroplasma, Apicomplexa, Flip-flop inversion

Background

Mitochondria, organelles essential for energy transduction, are present in almost all eukaryotes and have their own genome. Like nuclear genomes, mitochondrial (mt) genomes vary considerably in size, structure, and gene content [1]. There are two major mt genome forms: circular and linear. Circular forms are present in animal mt genomes with sizes ranging from 15 kb to 20 kb and

gene arrangements in the genomes are remarkably stable [2]. Some animal circular mt genomes are composed of more than two chromosomes or minicircles (e.g., the sucking louse *Pediculus humanus*, [3]). Circular forms are also found in higher-plant mt genomes. A higher-plant mt genome is characterized by a multipartite structure, which contains several subgenomic circular molecules with various organizational features, with sizes ranging from 200 kb to 2400 kb [4].

Linear mt genomes are found in diverse, unrelated organisms and have terminal inverted repeat (TIR) on both ends [5]. In some organisms, the mt genomes are

* Correspondence: kitak@m.u-tokyo.ac.jp; kztanabe@biken.osaka-u.ac.jp
²Department of Biomedical Chemistry, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
Full list of author information is available at the end of the article

divided into multiple chromosomes (e.g., the colorless green alga *Polytomella parva*, [6]), or several hundred chromosomes (e.g., the ichthyosporean *Amoebidium parasiticum*, [7]). In the phylum Apicomplexa, which includes important pathogens such as the causative agents of malaria (*Plasmodium*), coccidiosis (*Eimeria*), and piroplasmiasis (*Babesia* and *Theileria*), the mt genome structure is also diverse. Monomeric linear mt genomes with TIR on both ends are found in the *Babesia* and *Theileria* genera [8]. The *Babesia/Theileria* mt genomes are from 6.6 kb to 8.2 kb in size and encode only three protein coding genes (cytochrome *c* oxidase subunit I [*cox1*] and III [*cox3*] and cytochrome *b* [*cob*]) in addition to 24 fragmented small subunit (SSU) and large subunit (LSU) ribosomal RNA (rRNA) sequences [8,9]. *Plasmodium*, closely related to *Babesia/Theileria* [10], has the minuscule 6-kb tandemly repeated linear or concatenated mt genome, which encodes the same three protein coding genes as *Babesia/Theileria* [11,12]. The gene arrangements and transcriptional direction are however different from *Babesia/Theileria*. Furthermore, SSU and LSU rRNA genes of *Plasmodium* are highly fragmented with 27 pieces [9,13] and the pattern of fragmentation differs from *Babesia/Theileria* [8,9]. *Eimeria*, which is distantly related to *Babesia/Theileria* and *Plasmodium*, possess a concatemeric form and contains the same three protein-coding genes and 20 rRNA gene fragments as *Plasmodium* [9,14,15]. A recent phylogenetic study suggests that a concatenated form appears to be the ancestral mt genome structure in the phylum Apicomplexa, with the monomeric linear form of *Babesia/Theileria* having evolved in the lineage [14]. We are interested in the evolution of linear mt genomes of *Babesia/Theileria*, and investigated mt genome structure of the rodent piroplasms, *Babesia microti* and *Babesia rodhaini*, which belong to archaeopiroplasmids group, a lineage which branched off earlier from *Babesia/Theileria* [16]. Results revealed that both *B. microti* and *B. rodhaini* have a monomeric linear mt genome, in which two pairs of unique IR sequences are present, and that flip-flop inversions in each pair of the IRs appear to generate four distinct mt genome structures.

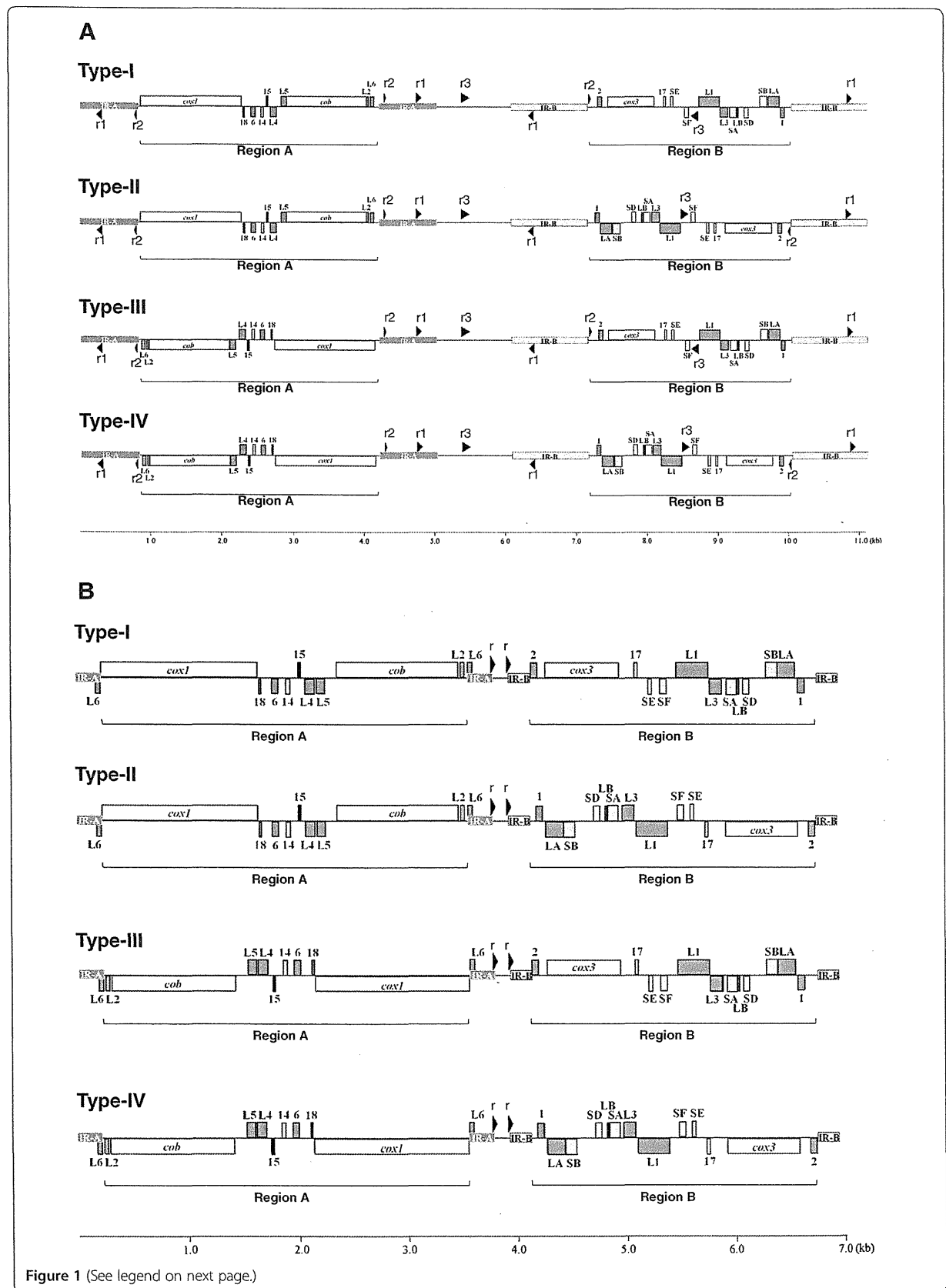
Results and discussion

Mitochondrial genome organization

We obtained the complete mt genome sequences of *B. microti* (Munich strain) and *B. rodhaini* (Australian strain). The *B. microti* mt genome was a monomeric linear molecule of 11.1 kb (Figures 1-A and 2-A) and contained three protein-coding genes, *cox1*, *cob* and *cox3*, and seven SSU and 12 LSU rRNA gene fragments (Figure 1-A). Unexpectedly, the *B. microti* mt genome possessed two pairs of long inverted repeats, inverted repeats A (IR-A) and B (IR-B). The nucleotide sequences

of IR-A (817 bp) and IR-B (1082 bp) were dissimilar to each other. A 3.4-kb region between two IR-As contained *cox1*, RNA18, RNA6, RNA14, RNA15, LSU4, LSU5, *cob*, LSU2 and LSU6, and a 2.9-kb region between two IR-Bs contained RNA2, *cox3*, RNA17, SSUE, SSUE, LSU1, LSU3, SSUA, LSUB, SSUD, SSUB, LSUA and RNA1. The predicted secondary structures for fragments comprising the *B. microti* SSU and LSU rRNA, except for the LSU1-LSU6 whose secondary structures have been predicted in *Theileria parva* and *Babesia gibsoni* [8,17], are shown in Additional file 1: Figure S1-A. RNA15 is a transcript of unknown function in *Plasmodium falciparum* and *T. parva* [9]. An intervening region (1080 bp) between IR-A and IR-B does not appear to contain any gene or gene fragment. Searches for other repeat sequences identified additional three short repeats with lengths of 23, 65 and 103 bp in the *B. microti* mt genome (Figure 1-A and Additional file 2: Table S1).

Interestingly, the *B. microti* mt genome displayed four distinct genome structures, types I, II, III and IV (Figure 1-A). The four genome structures can be generated by two inversions: one is an inversion of a region containing the 3.4-kb region between the IR-As (here termed Region A) and the other an inversion of a region containing the 2.9-kb region between the IR-Bs (Region B). Southern blot hybridization with probe *cox1* (Bm-1) against undigested *B. microti* genomic DNA produced a clear signal at 11.1 kb (lane 1 in Figure 2-A). Hybridization against DNA digested with *DraI* gave two bands at 10.0 kb and 7.4 kb (lane 2). Hybridization against *Eco065I*-digested DNA yielded two bands at 8.9 kb and 2.8 kb (lane 3). These results are consistent with the hypothesis that the four genome structures, types I, II, III and IV are generated by dual 'flip-flop' inversions of Region A and Region B. Thus, *DraI* digestion produced 10-kb and 1.1-kb fragments (types I and III in Figure 1-A), and additionally produced 7.4-kb and 3.7-kb fragments (types II and IV, if Region B was inverted); and *Eco065I* digestion produced 8.9-kb, 2.2-kb (types I and II in Figure 1-A) and 8.3-kb and 2.8-kb fragments (types III and IV, if Region A was inverted). Hybridization with a *cox3* probe (Bm-2) revealed a clear signal at 11.1 kb against undigested genomic DNA (lane 4), two bands at 10.0 kb and 3.7 kb against DNA digested with *DraI* (lane 5), and two bands at 8.9 kb and 8.3 kb against DNA with *Eco065I* (lane 6). Another *B. microti* probe from an intervening region (Bm-3) gave a band at 11.1 kb against undigested genomic DNA (lane 7), two bands at 10.0 and 7.4 kb against *DraI*-treated DNA (lane 8), and two bands at 8.9 kb and 8.3 kb against *Eco065I*-digested DNA (lane 9). These results obtained with Bm-2 and Bm-3 are consistent with the idea of dual 'flip-flop' inversions.



(See figure on previous page.)

Figure 1 Four distinct mitochondrial (mt) genome structures in *Babesia microti* (A) and *Babesia rodhaini* (B). These mt genomes possess two pairs of inverted repeats, IR-A and IR-B. Genes shown above bold line are transcribed from left to right and those below from right to left. Light and dark gray blocks indicate fragments of small subunit (SSU) and large subunit (LSU) rRNA genes, respectively. Abbreviations: *cox1*, cytochrome c oxidase subunit 1 gene; *cox3*, cytochrome c oxidase subunit 3 gene; *cob*, cytochrome b gene. Black arrowheads, r1, r2 and r3 in the *B. microti* mt genome and r in the *B. rodhaini* mt genome, indicate short direct or inverted repeat sequences (see Additional file 2: Table S1 and Additional file 1: Figure S4).

The *B. rodhaini* mt genome (6.9 kb) also possessed two pairs of long inverted repeats, IR-A (220 bp) and IR-B (184 bp) and four distinct genome structures (types I, II, III and IV) (Figure 1-B). IR-A contained LSU6. A 3.3-kb region between two IR-As (Region A) contained *cox1*, RNA18, RNA6, RNA14, RNA15, LSU4, LSU5, *cob* and LSU2. A 2.6-kb region between two IR-Bs (Region B) contained RNA2, *cox3*, RNA17, SSUE, SSUF, LSU1, LSU3, SSUA, LSUB, SSUD, SSUB, LSUA and RNA1. The transcriptional direction of SSUE and LSU5 of the *B. rodhaini* mt genome is different from that of the *B. microti* mt genome. The predicted secondary structures for fragments comprising the *B. rodhaini* SSU and LSU rRNA, except for the LSU1-LSU6, are shown in Additional file 1: Figure S1-B. RNA15 seems to be functionally important since its nucleotide sequence is highly conserved among *B. microti*, *B. rodhaini*, *T. parva* and *P. falciparum* (Additional file 1: Figure S2). An intervening region (152 bp) between IR-A and IR-B does not appear to contain any gene and gene fragment. In addition to the two IRs, a pair of short direct repeat was identified (Figure 1-B and Additional file 2: Table S1).

As in the *B. microti* mt genome, the four genome structures of *B. rodhaini* can also be generated by dual flip-flop inversions of Region A and Region B. Thus, hybridization with Br-1 against undigested DNA produced a clear signal at 6.9 kb (lane 1 in Figure 2-B), as expected from the genome sequence. Hybridization against *Hind*III-digested DNA gave two bands at 6.6 and 3.5 kb (lane 2), the former being expected in types III and IV (Figure 1-B), and the latter expected in types I and II (Figure 1-B). Hybridization against *Xho*I-digested DNA yielded two bands at 6.0 and 4.8 kb, (lane 3), the former being expected in types I and III, and the latter expected in types II and IV. Hybridization with Br-2 yielded a band at 6.9 kb against undigested DNA (lane 4), two bands at 6.6 and 3.4 kb against *Hind*III-digested DNA (lane 5), and two bands at 6.0 and 2.0 kb against *Xho*I-digested DNA (lane 6). The intervening region probe (Br-3) gave a band at 6.9 kb against undigested DNA (lane 7), two bands at 6.6 and 3.4 kb against *Hind*III-digested DNA (lane 8), and two bands at 6.0 and 4.8 kb against *Xho*I-digested DNA (lane 9). These signals were consistent with the four genome structures (Figure 2-B).

All monomeric linear mt genomes characterized to date have a single pair of TIR on both ends [5]. Thus, to

our knowledge this is the first study to show two pairs of distinct IR sequences. In both *B. microti* and *B. rodhaini*, dual flip-flop inversions of Region A and Region B appear to generate four mt genome structures. We postulate that the dual flip-flop inversions are mediated through recombination in palindromes of IR-A and IR-B (Figure 3). Recombination between a pair of IR-As (and IR-Bs) produces an isomeric form characterized by a flip-flop of Region A (and Region B), thus generating four distinct mt genome structures.

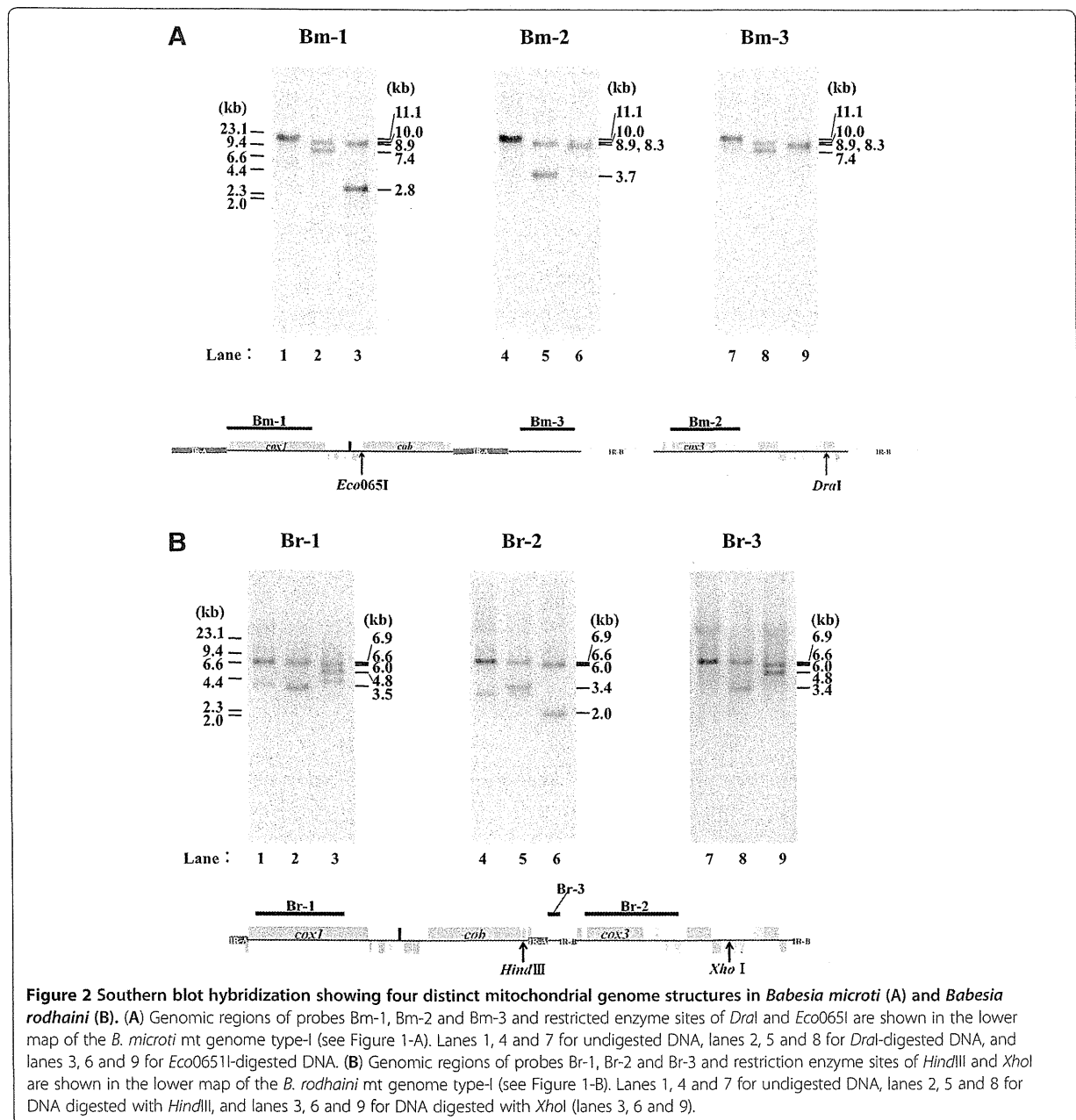
Flip-flop inversion of nuclear or organelle genomes have been found in some organisms. In the bacterium *Staphylococcus aureus*, large-scale inversion of its chromosome switches on or off different phenotypes, including the expression of dozens of genes [18]. The mt activity of the *Plasmodium* genus, which is closely related to the *Babesia* and *Theileria* genera, has been reported to be different between mosquito stages and vertebrate stages [19]. Therefore, it is likely that flip-flop inversions of the *B. microti* and *B. rodhaini* mt genomes may switch on or off expression of mt genes and gene fragments in their lifecycles.

Transcription

RT-PCR using three separate primer sets targeting about 500-bp sequences of *cox1*, *cox3* and *cob* of *B. microti* gave the expected transcript size using cDNA but not RNA (Additional file 1: Figure S3). Similarly, expected PCR sized fragments were obtained using primers specific to *B. rodhaini* for *cox1*, *cox3* and *cob* (Additional file 1: Figure S3). Results confirm the transcription of the three protein-coding genes. We were unable to perform additional transcription analysis, such as a northern blotting, for 19 SSU and LSU gene fragments, due to extreme difficulties in obtaining an adequate amount of parasites from infected mice. Two reports of Kairo et al. [17] on the transcription of five LSU rRNA gene fragments (LSU1-LSU5) in *T. parva* and Hikosaka et al. [8] on the transcription of LSU6 in *B. gibsoni*, however, suggest that at least these six fragments are transcribed in *B. microti* and *B. rodhaini*.

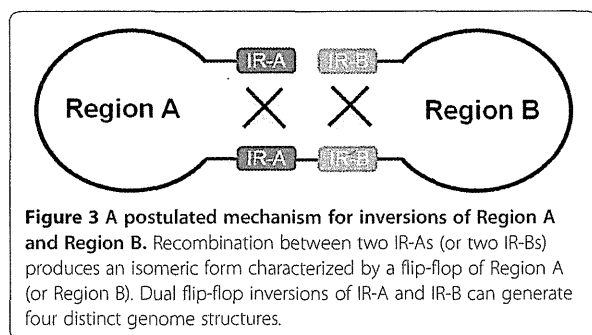
Estimations of the molar ratio of the four mt genome structures and mt genome copy number

We estimated the molar ratio of the four genome structures of *B. microti*. The intensity ratios of the two signals



produced by hybridization of the Bm-1 probe with *DraI*-digested DNA (lane 2 in Figure 2-A) and *Eco065I*-digested DNA (lane 3 in Figure 2-A) were 0.7 and 0.8 (Additional file 2: Table S2), respectively. Likewise, hybridization of the Bm-2 probe with *DraI*-digested DNA (lane 5 in Figure 2-A), and the Bm-3 probe with *DraI*-digested DNA (lane 8 in Figure 2-A) produced same signal intensities of 0.8 each. These suggest that a molar ratio of types-I, II, III and IV of the *B. microti* mt genome is approximately 1:1:1:1. We infer that the four

distinct genome structures is generated from one parasite because the four genome structures was confirmed by Southern blot analysis for genomic DNA extracted from parasites derived from a single parasite cloned by limiting dilutions (data not shown). In addition, copy number analysis using Southern hybridization estimated about 20 copies of the mt genome per haploid nuclear genome. Taken together, these findings suggest that one parasite possesses four types of mt genome structure in *B. microti*.



In *B. rodhaini*, the molar ratio of the four mt genome structures (types I, II, III and IV) was also approximately 1:1:1:1. Thus, intensity ratio of the two signals produced in each case by hybridization of the Br-1 probe with DNA digested by *Hind*III or *Xho*I (lanes 2 and 3 in Figure 2-B), and of the Br-2 probe (lanes 5 and 6 in Figure 2-B), and of the Br-3 probe (lanes 8 and 9 in Figure 2-B) ranged from 0.9 to 1.2 (Additional file 2: Table S2). The equivalent molar ratio was confirmed by Southern blot analysis using DNA of *B. rodhaini* cloned by limiting dilutions (data not shown). Copy number of the *B. rodhaini* mt genome was estimated to be two to three copies per haploid nuclear genome. This suggests that one parasite has one to three types of mt genome structures, and four distinct mt genome structures may be generated during cell proliferation.

Phylogeny

Both *B. microti* and *B. rodhaini* belong to an ancestral group of *Babesia/Theileria*, represented by the Archaeopiroplasmid group, according to phylogenetic analysis using 18S rRNA gene sequence [16]. The maximum likelihood (ML) tree inferred from concatenated COX1 and COB amino acid sequences (Figure 4) revealed a monophyletic relationship between Babesids, Theilerids and Archaeopiroplasmids with a bootstrap proportion (BP) value of 100%, clearly separated from a clade of *Plasmodium* species. *B. microti* and *B. rodhaini* were grouped into a clade (BP value 100%), which was located at the branch leading to the common ancestor of Babesids and Theilerids. These results indicate that the monomeric linear mt genomes found in the group of Babesids, Theilerids and Archaeopiroplasmids were generated specifically in this lineage.

Conclusions

We found a novel linear monomeric mt genome structure in the rodent piroplasmids, *Babesia microti* and *Babesia rodhaini*, equipped with dual flip-flop inversion system, by which four distinct genome structures are readily generated. Such a unique linear mt genome structure has not been known in not only apicomplexan

parasites but also in other organisms. The present findings would provide insight into further understanding of evolution of mt genome structure.

Methods

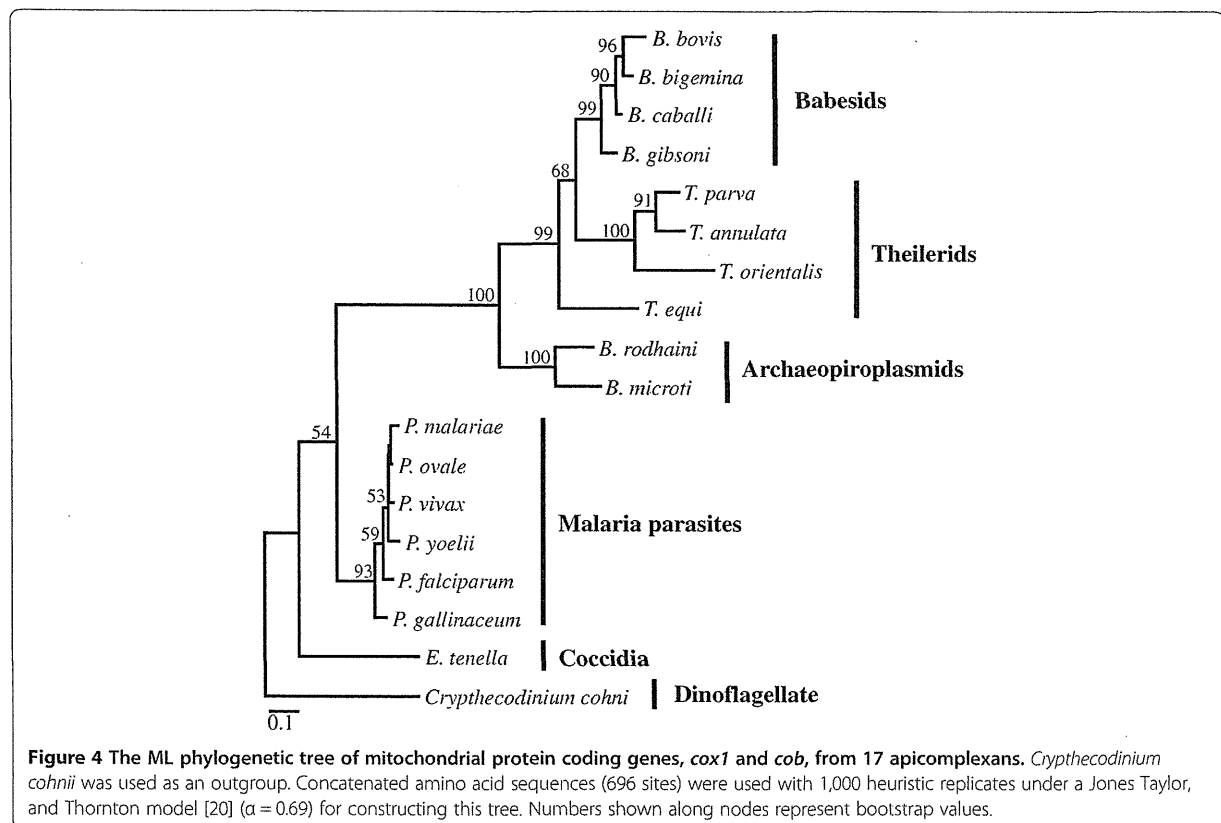
Animals

Four-week-old female ICR mice were purchased from Japan SLC (Shizuoka, Japan) and housed in microisolator cages within a modified pathogen-free barrier facility at the Animal Resource Center for Infectious Diseases, Research Institute for Microbial Diseases, Osaka University. All animals had free access to food and water *ad libitum*, and all of the experimental procedures followed our institutional guidelines.

DNA sequencing

B. microti (Munich strain) and *B. rodhaini* (Australian strain), were maintained by routine passage through mice. Infected blood was collected by cardiac puncture. Leukocytes were removed using Plasmodipur filters (EuroProxima, Arnhem, the Netherlands) [21]. Parasite genomic DNA was extracted using a QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The mt genomes of *B. microti* and *B. rodhaini* were directly sequenced using specific primers (Additional file 2: Table S3-A). Primers were designed by aligning reported mt genome sequences of *Plasmodium falciparum* (DDB)/EMBL/GenBank accession # M76611), *Plasmodium mexicanum* (EF079653), *Plasmodium yoelii* (MALPY00209), *Babesia bovis* (AB499088), *Theileria annulata* (NW_001091933), *Theileria equi* (AB499091) and *Theileria parva* (AB499089). Mt DNA was amplified in a 20 μ l reaction mixture containing 0.2 μ M each of forward and reverse primers, 400 μ M each of dNTP, 1 unit of LA-Taq (Takara Bio, Shiga, Japan), 2 μ l of 10 \times PCR buffer, 2.5 mM of MgCl₂, and 1 μ l of genomic DNA. PCR conditions were as follows: initial denaturation at 94°C for 1 min, and amplification for 40 cycles of 94°C for 30 s, 55-68°C (depending on primers used) for 30 s, and 72°C for 1-6 min, depending on amplification size (1 min per kb), followed by a final extension at 72°C for 10 min.

Sequences of telomeric regions of the mt genomes of *B. microti* and *B. rodhaini* were determined by using the terminal deoxynucleotidyl transferase (TdT) tailing method [22] with minor modifications. Briefly, following initial denaturation of genomic DNA (150 ng) for 5 min at 95°C, the 3'-ends was tailed with cytosine for 30 min at 37°C in a reaction mixture containing 200 μ M dCTP, 1 U of TdT (Takara Bio), 20 mM Tris-HCl (pH 8.4), 50 mM KCl, and 1.5 mM MgCl₂, and then heat-inactivated at 65°C for 10 min. The first PCR was done in a 50- μ l reaction mixture containing 2 μ l of the tailed DNA fragments, 1.25 units of AmpliTaq DNA



Polymerase (Applied Biosystems, Life Technologies, Carlsbad, CA), 2.5 mM MgCl₂, 200 μM dNTPs, 0.4 μM of a mt genome-specific primer (Additional file 2: Table S3-B) and a selective anchor primer (5'-CTACTACTAC TAGGCCACGCGTCTAGTACGGGGGGGGGGGGGGGG GG-3'). The PCR was performed by initial denaturation at 95°C for 2 min, and 40 cycles of 94°C for 30 s, 62°C for 3 min, followed by an extension step at 72°C for 10 min. For each sample, 1 μl of the first PCR products was used for the nested PCR amplification in a 50-μl reaction mixture as mentioned above, containing a nested primer (Additional file 2: Table S3-B) and a universal amplification primer (5'-CTACTACTACTAGGCCACGCGTCTAGTACTAGTAC-3'). The second PCR was performed by initial denaturation at 95°C for 2 min, and 25 cycles of 94°C for 30 s, 62°C for 2 min, followed by an extension step at 72°C for 10 min. PCR products were purified using QIAquick PCR purification kit (Qiagen), and sequenced directly from two independent PCR products, using the BigDye[®] Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Life Technologies) in an ABI 3130 Genetic Analyzer (Applied Biosystems, Life Technologies). Sequencing primers were designed to cover target regions in both directions. The sequences obtained in this study have been deposited in DDBJ/EMBL/GenBank with the

following accession numbers, AB624353 – AB624356 (*B. microti* mt genome structures type-I to type-IV) and AB624357 – AB624360 (*B. rodhaini* mt genome structures type-I to type-IV).

Gene annotation

Nucleotide sequences of the mt genomes from *B. microti* and *B. rodhaini* and their deduced amino acid sequences were aligned with reported sequences from *P. falciparum* (M76611), *B. bovis* (AB499088), *T. annulata* (NW_001091933), *T. equi* (AB499091) and *T. parva* (AB499089) by Clustal W [23] with manual correction. Protein-coding regions were inferred using previously annotated sequences from *T. parva* and *B. bovis*.

To identify putative rRNA genes, mt DNA sequences or annotated rRNA gene fragments from *B. bovis* (EU075182) and *T. parva* (Z23263) were used as a query under suggested algorithm parameters [24] in NCBI BLAST 2.2 [25]. In silico analysis was also performed with Probalign beta version 1.2 [26] and SSEARCH 3.5 [27] using known rRNA gene fragments and suggested advanced search options [24,26]. Newly identified candidate rRNA genes were, likewise, used as input sequences. Information from sequence alignments using CLUSTAL W [23] and putative base-pairings between fragments

proposed for *T. parva* mt rRNA fragments [9,17] were used to determine the termini of candidate rRNA genes.

Search for repeat sequences

Repeat sequences were searched using the REPFIND program (<http://zlab.bu.edu/repfind/>) [28] with cut-off of >20 nucleotides and a *P*-value < 0.0001. Inverted repeat sequences were searched using a 'self against self' BLASTN search [29] with cut-off of >20 nucleotides. Additional searches for repeats and inverted repeats were performed using GENETYX soft ware (Version 8; SDC, Tokyo, Japan).

Southern blot hybridization

Genomic DNA of *B. microti*, either undigested or digested with *Dra*I or *Eco*065I, and that of *B. rodhaini*, either undigested or digested with *Hind*III or *Xho*I, were electrophoresed on 0.8% (w/v) agarose gels in TAE (40 mM Tris-acetate, 1 mM EDTA) and then transferred to a positively charged nylon membrane (Amersham Hybond-N+, GE Healthcare, Little Chalfont, England). Specifically amplified PCR products from *B. microti* and *B. rodhaini* genomic DNA (Additional file 2: Table S3-C) were labeled with digoxigenin-dUTP using the DIG High Prime DNA Labeling and Detection Starter Kit II (Roche Diagnostics, Rotkreuz, Switzerland). The DIG-labeled DNA probes were incubated with the nylon membrane, and blots were washed twice with 2 × SSC, 0.1% SDS and twice with 0.5 × SSC, 0.1% SDS, at 65°C for 15 min. Hybridization signals were detected using the Detection Starter Kit II. Chemiluminescence signals were quantitated using LAS-4000mini (GE Healthcare BioSciences AB, Uppsala, Sweden).

RNA preparation and analysis

Transcription of *cox1*, *cox3* and *cob* in *B. microti* and *B. rodhaini* was analyzed by RT-PCR. Total RNA was extracted with RNeasy Mini Kit (Qiagen). Residual DNA in the RNA preparation was removed by DNase I treatment. cDNA synthesis and DNA amplification were carried out using specific primers (Additional file 2: Table S2-D) with PrimeScript[®] High Fidelity RT-PCR Kit (Takara Bio). RNA extracts that were not treated with reverse transcriptase gave no PCR products.

Copy number estimation

Copy numbers of mt genomes of *B. microti* and *B. rodhaini* were estimated using dot blot hybridization. Briefly, DNA fragments of the mt genome and of the *B. microti* and *B. rodhaini* beta-tubulin genes (nuclear genome) were amplified by PCR using specific primers (Additional file 2: Table S3-C), and DNA amount was measured. Serial dilutions of control PCR products of known DNA amounts were dot-blotted onto a nylon

membrane, following heat denaturation (99°C, 10 min). Genomic DNA were electrophoresed on agarose gels and then transferred to a nylon membrane. A PCR product specifically amplified from target regions of the mt genome and the beta-tubulin gene was labeled as described. Chemiluminescence signals were quantitated using LAS-4000mini.

Phylogenetic analysis

Concatenated amino acid sequences of COX1 and COB (696 sites) from 17 apicomplexan parasites (Additional file 2: Table S4) were used for phylogenetic analysis. A free living dinoflagellate, *Cryptocodinium cohnii* [30,31], was included as an outgroup. COX3 were not used for phylogenetic analysis, due to very high divergence in *Babesia/Theileria* species [8]. We constructed the ML phylogenetic tree by the PROML program in PHYLIP version 3.68 [32]. CODEML program in PAML version 4.2 [33] was used to estimate the Γ shape parameter value α . Bootstrap analysis was done by applying PROML to 100 re-sampled datasets produced by SEQBOOT program in PHYLIP. BP values were calculated for internal branches of the ML-tree using CONSENSE in PHYLIP.

Additional files

Additional file 1: This file contains Supplemental Figures S1-S4.

Additional file 2: This file contains Supplemental Tables S1-S4.

Abbreviations

mt genome: Mitochondrial genome; kb: Kilobase; IR: Inverted repeats; TIR: Terminal inverted repeat; *cox1*: Cytochrome *c* oxidase subunit I gene; *cox3*: Cytochrome *c* oxidase subunit III gene; *cob*: Cytochrome *b* gene; LSU: Large subunit; rRNA: Ribosomal RNA; SSU: Small subunit; bp: Base pairs; ML: Maximum likelihood; BP: Bootstrap proportion; min: Minute; TdT: Terminal deoxynucleotidyl transferase; dCTP: Deoxycytidine 5'-triphosphate.

Competing interests

The authors have declared that no competing interests exist.

Authors' contributions

KH carried out the molecular experiments. KT coordinated all experiments. KH, KK and KT contributed to the research design. KH and KT drafted the manuscript. NT and II provided parasite samples. YW analyzed data. HK and TH provided critical comments about this study. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by Grant-in-Aids for Scientific Research from Japan Society for Promotion of Sciences (18GS03140013 and 20390120) and Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists. We thank E.O. Balogun for language corrections.

Author details

¹Laboratory of Malariaology, Research Institute for Microbial Diseases, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan. ²Department of Biomedical Chemistry, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. ³Laboratory of Parasitic Diseases, National Institute of Animal Health, National Agriculture and Food

Research Organization, Tsukuba, Ibaraki, Japan. ⁴Department of Molecular Biology, Research Institute for Microbial Diseases, Osaka University, Suita, Osaka, Japan. ⁵Department of Molecular Protozoology, Research Institute for Microbial Diseases, Osaka University, Suita, Osaka, Japan. ⁶National Research Center for Protozoan Diseases, Obihiro University of Agriculture and Veterinary Medicine, Obihiro, Hokkaido, Japan.

Received: 24 April 2012 Accepted: 29 October 2012
Published: 14 November 2012

References

1. Gray MW, Lang BF, Burger G: Mitochondria of protists. *Annu Rev Genet* 2004, **38**:477–524.
2. Boore JL: Animal mitochondrial genomes. *Nucleic Acids Res* 1999, **27**(8):1767–1780.
3. Shao R, Kirkness EF, Barker SC: The single mitochondrial chromosome typical of animals has evolved into 18 minichromosomes in the human body louse, *Pediculus humanus*. *Genome Res* 2009, **19**(5):904–912.
4. Palmer JD, Soltis D, Soltis P: Large size and complex structure of mitochondrial DNA in two nonflowering land plants. *Curr Opin Genet Evol* 1992, **21**(2):125–129.
5. Nosek J, Tomaska L: Mitochondrial genome diversity: evolution of the molecular architecture and replication strategy. *Curr Opin Genet Evol* 2003, **44**(2):73–84.
6. Fan J, Lee RW: Mitochondrial genome of the colorless green alga *Polytomella parva*: two linear DNA molecules with homologous inverted repeat termini. *Mol Biol Evol* 2002, **19**(7):999–1007.
7. Burger G, Forget L, Zhu Y, Gray MW, Lang BF: Unique mitochondrial genome architecture in unicellular relatives of animals. *Proc Natl Acad Sci USA* 2003, **100**(3):892–897.
8. Hikosaka K, Watanabe Y, Tsuji N, Kita K, Kishine H, Arisue N, Palacpac NM, Kawazu S, Sawai H, Horii T, Igarashi I, Kita K, Tanabe K: Divergence of the mitochondrial genome structure in the apicomplexan parasites, *Babesia* and *Theileria*. *Mol Biol Evol* 2010, **27**(5):1107–1116.
9. Feagin JE, Hairrel MI, Lee JC, Coe KJ, Sands BH, Cannone JJ, Tami G, Schnare MN, Gutell RR: The fragmented mitochondrial ribosomal RNAs of *Plasmodium falciparum*. *PLoS One* 2012, **7**(6):e38320.
10. Lau AO: An overview of the *Babesia*, *Plasmodium* and *Theileria* genomes: a comparative perspective. *Mol Biochem Parasitol* 2009, **164**(1):1–8.
11. Feagin JE: Mitochondrial genome diversity in parasites. *Int J Parasitol* 2000, **30**(4):371–390.
12. Hikosaka K, Watanabe YI, Kobayashi F, Waki S, Kita K, Tanabe K: Highly conserved gene arrangement of the mitochondrial genomes of 23 *Plasmodium* species. *Parasitol Int* 2011, **60**(2):175–180.
13. Feagin JE, Mericle BL, Werner E, Morris M: Identification of additional rRNA fragments encoded by the *Plasmodium falciparum* 6kb element. *Nucleic Acids Res* 1997, **25**(2):438–446.
14. Hikosaka K, Nakai Y, Watanabe YI, Tachibana SI, Arisue N, Palacpac NM, Toyama T, Honma H, Horii T, Kita K, Tanabe K: Concatenated mitochondrial DNA of the coccidian parasite *Eimeria tenella*. *Mitochondrion* 2011, **11**(2):273–278.
15. Lin RQ, Qiu LL, Liu GH, Wu XY, Weng YB, Xie WQ, Hou J, Pan H, Yuan ZG, Zou FC, Hu M, Zhu XQ: Characterization of the complete mitochondrial genomes of five *Eimeria* species from domestic chickens. *Gene* 2011, **480**(1–2):28–33.
16. Criado-Fornelio A, Martinez-Marcos A, Buling-Sarana A, Barba-Carretero JC: Molecular studies on *Babesia*, *Theileria* and *Hepatozoon* in southern Europe. Part II. Phylogenetic analysis and evolutionary history. *Vet Parasitol* 2003, **114**(3):173–194.
17. Kairo A, Fairlamb AH, Gobright E, Nene V: A 7.1kb linear DNA molecule of *Theileria parva* has scrambled rDNA sequences and open reading frames for mitochondrially encoded proteins. *EMBO J* 1994, **13**(4):898–905.
18. Cui L, Neoh H, Iwamoto A, Hiramatsu K: Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria. *Proc Natl Acad Sci USA* 2012, **109**(25):E1647–E1656.
19. Hino A, Hirai M, Tanaka TQ, Watanabe Y, Matsuoka H, Kita K: Critical roles of the mitochondrial complex II in oocyst formation of rodent malaria parasite *Plasmodium berghei*. *J Biochem* 2012, **152**(3):259–268.
20. Jones DT, Taylor WR, Thornton JM: The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 1992, **8**(3):275–282.
21. Janse CJ, Camargo A, Del Portillo HA, Herrera S, Kumlien S, Mons B, Thomas A, Waters AP: Removal of leucocytes from *Plasmodium vivax*-infected blood. *Ann Trop Med Parasitol* 1994, **88**(2):213–216.
22. Bah A, Bachand F, Clair E, Autexier C, Wellinger RJ: Humanized telomeres and an attempt to express a functional human telomerase in yeast. *Nucleic Acids Res* 2004, **32**(6):1917–1927.
23. 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**(22):4673–4680.
24. Freyhult EK, Bollback JP, Gardner PP: Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res* 2007, **17**(1):117–125.
25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, **215**(3):403–410.
26. Roshan U, Chikkagoudar S, Livesay DR: Searching for evolutionary distant RNA homologs within genomic sequences using partition function posterior probabilities. *BMC Bioinforma* 2008, **9**:61.
27. Pearson WR: Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* 1991, **11**(3):635.
28. Betley JN, Frith MC, Graber JH, Choo S, Deshler JO: A ubiquitous and conserved signal for RNA localization in chordates. *Curr Biol* 2002, **12**(20):1756.
29. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, **25**(17):3389–3402.
30. Norman JE, Gray MW: A complex organization of the gene encoding cytochrome oxidase subunit 1 in the mitochondrial genome of the dinoflagellate, *Cryptocodinium cohnii*: homologous recombination generates two different *cox1* open reading frames. *J Mol Evol* 2001, **53**(4–5):351–363.
31. Zhang H, Lin S: Mitochondrial cytochrome b mRNA editing in dinoflagellates: possible ecological and evolutionary associations?. *J Eukaryot Microbiol* 2005, **52**(6):538–545.
32. Felsenstein J, Churchill GA: A hidden markov model approach to variation among sites in rate of evolution. *Mol Biol Evol* 1996, **13**(1):93–104.
33. Yang Z: PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 1997, **13**(5):555–556.

doi:10.1186/1471-2164-13-622

Cite this article as: Hikosaka et al.: Novel type of linear mitochondrial genomes with dual flip-flop inversion system in apicomplexan parasites, *Babesia microti* and *Babesia rodhaini*. *BMC Genomics* 2012 **13**:622.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



Pharmacophore identification of ascofuranone, potent inhibitor of cyanide-insensitive alternative oxidase of *Trypanosoma brucei*

Received October 15, 2012; accepted November 6, 2012; published online November 23, 2012

Hiroyuki Saimoto¹, Yasutoshi Kido^{2,*},
Yasushi Haga¹, Kimitoshi Sakamoto^{2,†} and
Kiyoshi Kita^{2,‡}

¹Department of Chemistry and Biotechnology, Graduate School of Engineering, Tottori University, Tottori 680-8552, Japan; and
²Department of Biomedical Chemistry, Graduate School of Medicine, The University of Tokyo, Tokyo 113-0033, Japan

*Yasutoshi Kido, Department of Biomedical Chemistry, Graduate School of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. Tel: +81-3-5841-3526, Fax: +81-3-5841-3444, email: yasutoshikido@gmail.com

‡Kiyoshi Kita, Department of Biomedical Chemistry, Graduate School of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. Tel: +81-3-5841-3526; Fax: +81-3-5841-3444; email: kitak@m.u-tokyo.ac.jp

†Present address: Kimitoshi Sakamoto, Faculty of Agriculture and Life Science, Hirosaki University, Hirosaki 036-8561, Japan.

Trypanosoma brucei is a parasite that causes human African trypanosomiasis (HAT). The parasites depend on the cyanide-insensitive trypanosome alternative oxidase (TAO) for their vital aerobic respiration. Ascofuranone (AF), a potent and specific sub-nanomolar inhibitor of the TAO quinol oxidase, is a potential novel drug with selectivity for HAT, because mammalian hosts lack the enzyme. To elucidate not only the inhibition mechanism but also the inhibitor–enzyme interaction, AF derivatives were designed and synthesized, and the structure–activity relationship was evaluated. Here we identified the pharmacophore of AF that interacts with TAO. The detailed inhibitory profiles indicated that the 1-formyl and 6-hydroxyl groups, which might contribute to intramolecular hydrogen bonding and/or serve as hydrogen-bonding donors, were responsible for direct interaction with the enzyme.

Keywords: alternative oxidase/ascofuranone/neglected tropical disease/trypanosome/structure–activity relationship.

Abbreviations: AF, ascofuranone; AOX, alternative oxidase; HAT, human African trypanosomiasis; IC₅₀, 50% inhibitory concentration; rAOX, recombinant alternative oxidase; rTAO, recombinant trypanosome alternative oxidase; SAR, structure–activity relationship; SHAM, salicyl hydroxamic acid; TAO, trypanosome alternative oxidase.

Trypanosoma brucei, which causes human African trypanosomiasis (HAT), depends on a cyanide-insensitive respiratory pathway for survival (1). Such

cyanide-insensitive oxygen consumption has been recognized in plants since 1920s (2). Intensive biochemical studies revealed that a mitochondrial membrane enzyme, designated alternative oxidase (AOX), is responsible for cyanide-insensitive respiration (1–3). To date, AOX has been detected not only in higher plants and protozoa including trypanosomes but also in algae, yeast, slime moulds, free-living amoebae, eubacteria and nematodes (4, 5).

The cyanide-insensitive ubiquinol oxidase activity of AOX catalyzes the four-electron reduction of dioxygen to water and has been characterized as a salicyl hydroxamic acid (SHAM)-inhibited activity (6). Although the activity inhibited by SHAM has been assigned to the ubiquinol oxidase moiety of AOX, inhibition of the activity requires a relatively high concentration (micromolar) of SHAM (6, 7). We previously reported that ascofuranone (AF), isolated from pathogenic fungus *Ascochyta viciae*, specifically inhibits the ubiquinol oxidase activity of trypanosome mitochondrial AOX (TAO) at sub-nanomolar levels, revealing a novel class of potent AOX inhibitors (8). Furthermore, since TAO is essential for trypanosomal survival and is absent from mammalian hosts, AF represents a lead compound for development of selective anti-HAT drugs (1, 9, 10). Indeed, we have demonstrated the chemotherapeutic efficacy of AF both *in vitro* and *in vivo* (8, 11, 12). However, the structural requirements for AF inhibition of AOX remain poorly defined.

In this study we have designed and synthesized derivatives of AF. Biochemical evaluation of the derivatives elucidates the relationship between the chemical structures and inhibition of TAO. The structure–activity relationship (SAR) study for the inhibition of TAO identified the profile of the AF pharmacophore. Our data shed light upon the potential for drug development for HAT, a neglected tropical disease, while also clarifying the molecular interaction between AF and TAO.

Materials and Methods

Production of recombinant TAO and recombinant *Sauromatum guttatum* AOX

Recombinant TAO (rTAO) was produced using a haeme synthesis-deficient strain (FN102) harbouring a TAO-encoding plasmid, as described in a previous study (13). The strain FN102/pTbAO was grown aerobically at 30°C and induced to express rTAO. Inner membranes from the induced strain were used to assay enzymatic activity in the presence of a series of AF derivatives. The 50% inhibitory concentration (IC₅₀), which is a molar concentration needed to halve the control enzymatic activity, was used to evaluate the inhibitory activity. Recombinant *S. guttatum* AOX was also produced using FN102 harbouring a *S. guttatum* AOX-encoding plasmid as rTAO was expressed.

Ubiquinol oxidase assay

Ubiquinol oxidase activity was measured by recording the absorbance change of ubiquinol-1 ($\epsilon_{278} = 15,000 \text{ M}^{-1} \text{ cm}^{-1}$) at 278 nm (SHIMADZU spectrophotometer UV-3000). Reactions (1 ml each) were constituted from 0.35 μg of FN102/pTbAO membrane fraction in 50 mM Tris-HCl (pH 7.3). Following pre-incubation (2 min at 25°C), reactions were initiated by the addition of ubiquinol-1 to a final concentration of 150 μM . The sigmoidal curve of the inhibition was observed using a broad dilution series of all AF derivatives ranged from none of inhibition to complete inhibition. Following the depiction of the sigmoidal curve, a sequential analysis of the inhibition was performed using at least five different concentrations of all AF derivatives around IC_{50} values. The measurement of ubiquinol oxidase activity of recombinant *S. gutatum* AOX was identical to that of rTAO. The presence of dithiothreitol in the ubiquinol oxidase assay of rTAO does not affect the activity at all despite other AOX's activity occasionally regulated by a disulphide/sulphydryl system via conserved cysteine residue.

Synthesis of test compounds

A series of AF derivatives synthesized and tested in this study are numbered from 1 to 30. The synthetic procedures are described in the Supplementary data.

Results and Discussions

The chemical structure of AF (Fig. 1A) consists of three moieties: aromatic ring, linker and furanone ring. The functional groups at those moieties were chemically modified and synthesized. The inhibitory activity of all AF derivatives was evaluated for ubiquinol oxidase activity of rTAO by IC_{50} . At nanomolar concentrations, AF inhibits not only rTAO but also recombinant AOXs (rAOXs) derived from other organisms such as plants and pathogenic microorganisms, although AOXs have diverse and unique physiological functions in each organism (14, 15). Table I summarizes the IC_{50} s for rAOXs from various organisms, suggesting that AF is a potent and universal inhibitor of various AOXs (13, 15).

For rTAO tested under our experimental conditions, the IC_{50} of AF was 0.13 nM. For comparison, the IC_{50} s of SHAM and ascochlorin (Fig. 1B; another metabolite from the fungus *A. viciae*) were 4 μM and 1.5 nM, respectively. This difference between AF and ascochlorin suggested that the structure and asymmetric carbon at the furanone ring might be involved in the binding to TAO as judged from the structural relation between the two compounds. However, a significant difference was not observed in the derivatives in which the furanone ring was substituted, as demonstrated by the IC_{50} s for colletochlorin B (0.20 nM) and compound 1 (0.40 nM) (Table II). The other furanone ring-substituted derivatives (2 and 3) showed potencies (1.2 and 1.4 nM, respectively) similar to that of ascochlorin. These data indicated that the furanone ring

was not essential for the strong inhibition of AF. The linker structure might interact with the enzyme because the linker of ascochlorin was different from the geranyl chain of AF.

To confirm the function of the geranyl chain, we synthesized and investigated a series of compounds that replaced the linear alkyl chain while maintaining the furanone ring as in AF (as shown in Table III). The IC_{50} s of compounds 4 (with two branched methyl groups), 5 (with a $\text{C}\alpha$ -double bond) and 6 (with a linear alkyl chain) were 0.30, 1.0 and 0.5 nM, respectively. Thus, the IC_{50} of each of these derivatives was several times as high as that of AF. Although the geranyl structure of AF was obviously favourable for inhibition, the linker structure was likely to be flexible to inhibit TAO strongly at a nanomolar level. This slight decrease of the inhibitory activity suggested that the linker structure in the proximity of the aromatic ring may affect the conformational range available to these compounds.

In addition to the structure, we examined the effect of the total linker length in the inhibitory potency, as shown in Table IV. Six compounds were prepared varying the side chain length at the 5-position from propenyl (C_3) to dodec-1-enyl (C_{12}) (7a–7f). Table IV shows that the optimum length for potent inhibition was observed at (7d; C_9) and (7e; C_{10}). The compound with the shortest linker (7a; C_3) exhibited a remarkable decrease in potency (IC_{50} 1,000 times higher than that of AF), suggesting that hydrophobicity is required for inhibition, consistent with the fact that TAO is a membrane protein. In contrast, the IC_{50} values of compound (7f; C_{12}) with a longer chain above the optimum length seemed to be lower; we hypothesize that this compound might be trapped in the hydrophobic lipid bilayer of the membrane. A similar tendency is generally observed for other hydrophobic inhibitors (16).

Since the geranyl linker and the furanone ring were not essential for potent inhibition, the effect of the steric tolerance for this portion was examined (8, 9 and 10) to elucidate the interaction between AF and TAO (Table V). The size of this portion did not affect the inhibitory activity, suggesting that the chemical moiety that is attached to the C9 (or beyond) is not associated with the enzyme. In other words, those chemical structures might be located at the surface or outside of TAO in the AF-TAO complex. Such expected steric interaction between AF and TAO was supported by the fact that the inhibitory activity of 13 was much more potent than that of 11, despite the presence of a hydroxyl group at the end of the

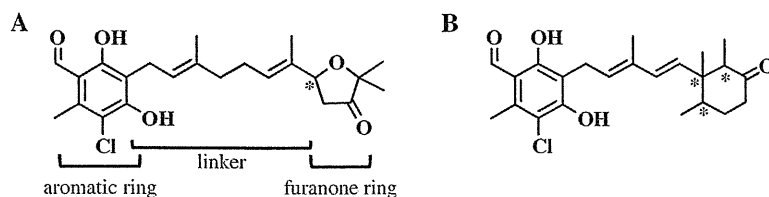


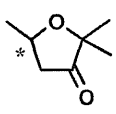
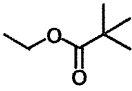


Fig. 1 Chemical structure of inhibitors of AOX. (A) Ascofuranone. (B) Ascochlorin. Asterisk represents asymmetric carbon.

Table I. IC₅₀ values of ascofuranone for ubiquinol oxidase activity derived from various organisms.

Species	IC ₅₀ (nM)
<i>T. brucei</i> ^a	0.13 ± 0.04
<i>S. gutatum</i> ^b	1.4 ± 0.3
<i>Antonospora locustae</i> ^c	>91% inhibition at 10 nM
<i>Trachipleistophora hominis</i> ^c	>93% inhibition at 10 nM

An index 50% inhibitory concentration (IC₅₀), which is a molar concentration needed to halve the control enzymatic activity, was used to evaluate the inhibitory activity of AF for the ubiquinol oxidase activity of rAOXs from various organisms. ^aFrom Kido *et al.* (13). ^bFrom this study. ^cFrom Williams *et al.* (15).

Table II. TAO inhibition by furanone ring substituted derivatives.

Compound	R	IC ₅₀ (nM)
Ascofuranone		0.13 ± 0.04
Coltochlorin B	—CH ₃	0.20
1		0.40
2		1.2
3		1.4

linker, which had a detrimental effect on the potency (6 versus 12).

Since we have recently reported that AF showed mixed-type inhibition for the TAO against ubiquinol (13), we speculated that the aromatic ring and the geranyl moiety of AF would be responsible for the direct interaction with TAO. To investigate the role of functional groups on the aromatic ring, AF derivatives substituted at 1-, 2-, 3- and 4-positions were synthesized and examined. The functional group at the 1-position in AF is a formyl group, and an intramolecular hydrogen bond between the 1- and 6-positions of the aromatic ring (=O of formyl group and —OH of 6-position) was expected in AF. Table VI shows the inhibitory effect of 1-position substitutions of AF on TAO. The acetyl-substituted derivatives (all containing the geranyl linker) (14, 15 and 16) showed an inhibitory activity similar to that of AF, whereas the inhibitory activity was significantly decreased in the

Table III. TAO inhibition by the derivatives with various alkyl chain.

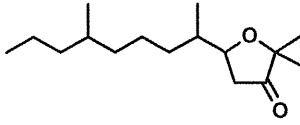
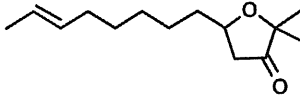
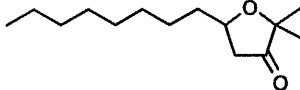
Compound	R	IC ₅₀ (nM)
4		0.30
5		1.0
6		0.50

Table IV. TAO inhibition by the derivatives varying the side chain length.

Compound	R	IC ₅₀ (nM)
7a	CH ₃	100
7b	CH ₂ CH ₂ CH ₃	3.8
7c	CH ₂ CH ₂ CH ₂ CH ₂ CH ₃	0.70
7d	CH ₂ CH ₂ CH ₂ CH ₂ CH ₂ CH ₂ CH ₃	0.38
7e	CH ₂ CH ₂ CH ₂ CH ₂ CH ₂ CH ₂ CH ₂ CH ₃	0.38
7f	CH ₂ CH ₂ CH ₂ CH ₂ CH ₂ CH ₂ CH ₂ CH ₂ CH ₂ CH ₃	0.45

hydroxyimino and methoxycarbonyl substitutions (17 and 18) (to 28 and 45 nM, respectively). These slightly larger substituted groups presumably hampered the access of inhibitors to the binding site due to their steric congestion.

The chloro-substituted compound 19 drastically decreased inhibitory activity (to a micromolar level), although the linker also was substituted to a linear alkyl chain. In contrast, the extent of the decrease in activity caused by the nitro-substituted and cyano-substituted derivatives (20 and 21) was moderate. Comparison of 1-position-substituted derivatives shows that the intramolecular hydrogen bond or the hydrogen-bonding donor activity of =O of the formyl group was crucial for strong inhibition (Tables VI and VII). The electronic property of the formyl group, or

Table V. TAO inhibition by the derivatives varying the bulkiness of the furanone ring portion.

Compound	Structure	IC ₅₀ (nM)
8		0.50
9		0.30
10		0.32
11		6.0
12		40
13		4.2

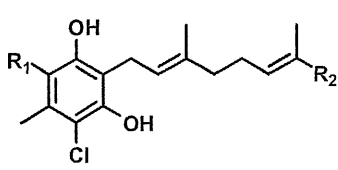
of the electron-withdrawing group, is not likely to be responsible for potent inhibition because the nitro-substituted derivative (**25**), which causes a stronger electron-withdrawing effect, exhibited a slightly decreased inhibitory activity. Surprisingly, the presence of the geranyl or C α -carbonyl group in proximity to the aromatic ring appeared to compensate for the detrimental effect of the acetyl-substituted and the chloro-substituted derivatives (**15** versus **22** and **19** versus **23**), suggesting that the role of 1-formyl and/or 6-hydroxyl groups depends on the linker structure.

Next, 2- and 3-position derivatives were examined. Removal of 2-methyl (**25** and **26**) or 3-chloro (**27**) did not alter inhibitory activity, whereas removal of both moieties (**28**) decreased inhibitory activity (Table VIII). These four compounds all included a geranyl linker.

The effects of 2- and 3-position substituents suggested that the major contribution of these functional groups for inhibition was not an electronic state affected by the 2-methyl and 3-chloro moieties, but a whole conformational preference, because the 2- and 3-position substituents did not diminish inhibitory activity when the compounds retained a geranyl moiety as a linker. In short, the 2- and/or 3-functional groups may affect the recognition of the geranyl moiety by the enzyme.

Finally, Table IX examines the role of AF's 4-hydroxyl group on TAO inhibition. The 4-methoxy substitution (**29**) resulted in low inhibitory activity, whereas the 4-hydrogen substitution (**30**) resulted in a complete loss of inhibition. This result suggests that the lone pair of electrons on the 4-hydroxyl's oxygen plays a role in the interaction of AF with the enzyme.

Table VI. TAO inhibition by 1-position substituted derivatives with a geranyl linker.




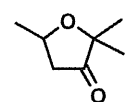

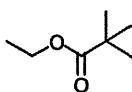


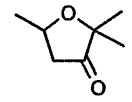

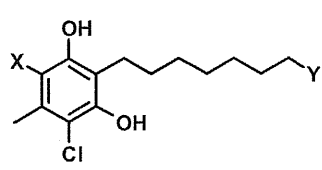
Compound	R1	R2	IC ₅₀ (nM)
14			2.5
15			0.70
16		-CH ₃	0.15
17			28
18		-CH ₃	45

Table VII. TAO inhibition by 1-position substituted derivatives with a linear alkyl linker.





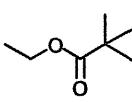

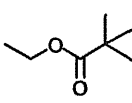
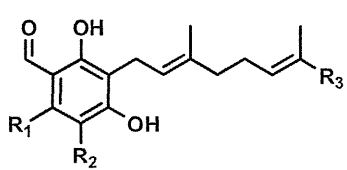
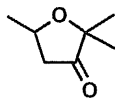
Compound	X	Y	IC ₅₀ (nM)
19	Cl-		10000
20	O ₂ N-		0.45
21	NC-		6.0
22			100

Table VIII. TAO inhibition by 2, 3-position substituted derivatives.



Compound	R1	R2	R3	IC ₅₀ (nM)
25	-H	-Cl		0.45
26	-H	-Cl	-CH ₃	0.23
27	-CH ₃	-H	-CH ₃	0.50
28	-H	-H	-CH ₃	38

Conclusion

This study revealed the structure–activity profiles as follows (Fig. 2). (i) The furanone ring is not essential for potent inhibition. (ii) One isoprene in proximity to the aromatic ring is recognized by the enzyme. (iii) Intramolecular hydrogen bonding and/or the hydrogen-bonding donor activity of the 1-formyl and 6-hydroxyl groups is responsible for direct interaction with the enzyme. (iv) The 2-methyl and/or 3-chloro groups contribute to the enzyme-bound conformation of the molecule. Notably, we have successfully designed and synthesized compound **24** (Table X), which exhibited more potent inhibitory activity than AF (0.06 nM). These AF derivatives and the present systematic structure–activity profiles provide important information about the reaction mechanism of TAO and AOXs. These data are expected to clarify the molecular mechanism of AF inhibition when considered in the context of the 3D structure of TAO (17, 18).

SHAM and propyl gallate have been known as specific inhibitors of AOX since 1971 and 1980, respectively (6, 19). Inhibitors have played a crucial role in the characterization of AOX, which has been recognized as the source of cyanide-insensitive, SHAM-sensitive oxygen consumption for years. The IC₅₀ values of the two inhibitors against rTAO under our experimental conditions were 4 μM and 200 nM, respectively. Another class of AOX inhibitor, aurachin C, was reported in 1995 (20). Historical data did not identify any other inhibitors with inhibition higher than that of the lead compounds; however, some components of the inhibitors, such as the hydroxyl acetamide group of SHAM, were revealed to be responsible for inhibition (7). Accordingly, the intramolecular hydrogen bond and/or the hydrogen-bonding donor activity on the hydroxyl acetamide group might be analogous to that of AF.

For drug development, structural modification often is required to improve the pharmacological activity