





## RESEARCH ARTICLE

# An improved *Plasmodium cynomolgi* genome assembly reveals an unexpected methyltransferase gene expansion [version 1; referees: 2 approved]

Erica M Pasini<sup>1</sup>, Ulrike Böhme<sup>2</sup>, Gavin G. Rutledge <sup>2</sup>,  
Annemarie Voorberg-Van der Wel<sup>1</sup>, Mandy Sanders<sup>2</sup>, Matt Berriman<sup>2</sup>,  
Clemens HM Kocken<sup>1</sup>, Thomas Dan Otto <sup>2</sup>

<sup>1</sup>Biomedical Primate Research Centre, Rijswijk, Lange Kleiweg 161, 2288GJ Rijswijk, Netherlands

<sup>2</sup>The Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

**v1** First published: 16 Jun 2017, 2:42 (doi: [10.12688/wellcomeopenres.11864.1](https://doi.org/10.12688/wellcomeopenres.11864.1))  
Latest published: 16 Jun 2017, 2:42 (doi: [10.12688/wellcomeopenres.11864.1](https://doi.org/10.12688/wellcomeopenres.11864.1))

## Abstract

**Background:** *Plasmodium cynomolgi*, a non-human primate malaria parasite species, has been an important model parasite since its discovery in 1907. Similarities in the biology of *P. cynomolgi* to the closely related, but less tractable, human malaria parasite *P. vivax* make it the model parasite of choice for liver biology and vaccine studies pertinent to *P. vivax* malaria. Molecular and genome-scale studies of *P. cynomolgi* have relied on the current reference genome sequence, which remains highly fragmented with 1,649 unassigned scaffolds and little representation of the subtelomeres.



**Methods:** Using long-read sequence data (Pacific Biosciences SMRT technology), we assembled and annotated a new reference genome sequence, PcyM, sourced from an Indian rhesus monkey. We compare the newly assembled genome sequence with those of several other *Plasmodium* species, including a re-annotated *P. coatneyi* assembly.

**Results:** The new PcyM genome assembly is of significantly higher quality than the existing reference, comprising only 56 pieces, no gaps and an improved average gene length. Detailed manual curation has ensured a comprehensive annotation of the genome with 6,632 genes, nearly 1,000 more than previously attributed to *P. cynomolgi*. The new assembly also has an improved representation of the subtelomeric regions, which account for nearly 40% of the sequence. Within the subtelomeres, we identified more than 1300 *Plasmodium* interspersed repeat (*pir*) genes, as well as a striking expansion of 36 methyltransferase pseudogenes that originated from a single copy on chromosome 9.

**Conclusions:** The manually curated PcyM reference genome sequence is an important new resource for the malaria research community. The high quality and contiguity of the data have enabled the discovery of a novel expansion of methyltransferase in the subtelomeres, and illustrates the new comparative genomics capabilities that are being unlocked by complete reference genomes.

## Open Peer Review

Referee Status:  

	Invited Referees	
	1	2
<b>version 1</b> published 16 Jun 2017	 report	 report
1 <b>Aaron Jex</b> , Walter and Eliza Hall Institute of Medical Research, Australia		
2 <b>Richárd Bártfai</b> , Radboud University, Netherlands		

## Discuss this article

Comments (0)

**Corresponding author:** Thomas Dan Otto ([tdo@sanger.ac.uk](mailto:tdo@sanger.ac.uk))

**Author roles:** **Pasini EM:** Data Curation, Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; **Böhme U:** Data Curation, Visualization, Writing – Review & Editing; **Rutledge GG:** Formal Analysis, Writing – Review & Editing; **Voorberg-Van der Wel A:** Resources, Writing – Review & Editing; **Sanders M:** Project Administration, Writing – Review & Editing; **Berriman M:** Conceptualization, Writing – Review & Editing; **Kocken CH:** Conceptualization, Resources, Writing – Review & Editing; **Otto TD:** Formal Analysis, Methodology, Resources, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** None of the authors declared competing interest.

**How to cite this article:** Pasini EM, Böhme U, Rutledge GG *et al.* **An improved *Plasmodium cynomolgi* genome assembly reveals an unexpected methyltransferase gene expansion [version 1; referees: 2 approved]** Wellcome Open Research 2017, 2:42 (doi: [10.12688/wellcomeopenres.11864.1](https://doi.org/10.12688/wellcomeopenres.11864.1))

**Copyright:** © 2017 Pasini EM *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Grant information:** This work was supported by the Wellcome Trust (098051), EVIMalaR (contract number 242095) and Gates Foundation Project OPP1023583. GGR is supported by the Medical Research Council (MR/J004111/1).

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**First published:** 16 Jun 2017, 2:42 (doi: [10.12688/wellcomeopenres.11864.1](https://doi.org/10.12688/wellcomeopenres.11864.1))

## Introduction

*Plasmodium cynomolgi*, a non-human primate malaria parasite first mentioned by Mayer in 1907<sup>1</sup> and established as a separate species from *P. inui* by Mulligan in 1935<sup>2</sup>, has been used as a model parasite species since its discovery. First used to establish the level of susceptibility of Malaysian Anophelines to non-human primate malaria<sup>3</sup>, *P. cynomolgi* forms hypnozoites (a dormant liver stage), similar to those of human-infective *P. vivax* and *P. ovale* species. Other shared characteristics between *P. cynomolgi* and *P. vivax* include erythrocyte morphology (e.g. Schüffner's stippling), amoeboidity and the tertian periodicity of intraerythrocytic asexual development (48h life-cycle). *P. cynomolgi* is thus regarded as a powerful model for *P. vivax* and potentially *P. ovale* human malaria. The use of *P. cynomolgi* as a model organism is further reinforced by it being readily infective to and transmitted by a large number of mosquito species<sup>4–7</sup>, and by having a wide range of natural<sup>8–10</sup> and experimental hosts<sup>3,11</sup>.

A particular strength of the *P. cynomolgi* system is access to chronic infections and to the developing and dormant liver stages in a parasite similar to *P. vivax*. An *in vivo-vitro* shuttle system for the study of *P. cynomolgi* liver stages<sup>12</sup> is being exploited to better understand hypnozoite biology using molecular tools and genome-scale approaches, which rely on the availability of a complete and well annotated *P. cynomolgi* reference genome sequence. However, the current *P. cynomolgi* B reference is very fragmented<sup>13</sup>, and lacks large parts of the subtelomeric regions, thought to harbour genes involved in host-parasite interactions. Other closely related malaria parasite species have been sequenced, including *P. coatneyi*<sup>14</sup> which is closely related to *P. knowlesi*, and *P. simiovale* that was sequenced but never systematically assembled<sup>15</sup>.

In this paper, we describe the improved genome sequence assembly of the *P. cynomolgi* M strain and compare it the genomes of five other *Plasmodium* species (*P. vivax*, *P. falciparum*, *P. knowlesi*, *P. coatneyi*, *P. simiovale*) that infect humans or monkeys, to uncover similarities and differences that may inform future studies aimed at harnessing *P. cynomolgi* as a model for *P. vivax* human malaria.

## Methods

### Samples

DNA was obtained from a blood stage infection of an Indian rhesus macaque donor with *P. cynomolgi* M strain stocks originally provided by Dr. Bill Collins from the Center for Disease Control, Atlanta. After PlasmodiPur filtration, parasites were matured *in vitro* overnight. Parasites were purified over a 15.1% (w/v) Nycodenz gradient and DNA was isolated using the Gentra Puregene Blood kit (Qiagen) and processed according to the manufacturers' instructions. The material was handled carefully in order to ensure the integrity of the DNA was maintained.

### Ethical approval

Ethical approval for the donor infection was provided under DEC750 following Dutch and European legislation in terms of animal experimentation. Prior to the start of the experiment, ethical approval for the donor monkey infection was provided by the local independent ethical committee, complying with Dutch law (BPRC Dier Experimenten Commissie, DEC; agreement number

DEC# 750). The monkey was healthy as assessed by a veterinarian and as determined by clinical and hematological parameters measured before the start of the experiment. The experiment was performed according to Dutch and European laws. The Council of the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC International) has awarded BPRC full accreditation. Thus, BPRC is fully compliant with the international demands on animal studies and welfare as set forth by the European Council Directive 2010/63/EU, and Convention ETS 123, including the revised Appendix A as well as the 'Standard for humane care and use of Laboratory Animals by Foreign institutions' identification number A5539-01, provided by the Department of Health and Human Services of the United States of America's National Institutes of Health (NIH) and Dutch implementing legislation.

The donor monkey (*Macaca mulatta*, male, age 5 years, Indian origin) used in this study was captive-bred and socially housed. Animal housing was according to international guidelines for non-human primate care and use. Besides the standard feeding regime, and drinking water ad libitum via an automatic watering system, the animal followed an environmental enrichment program in which, next to permanent and rotating non-food enrichment, an item of food-enrichment was daily offered to the macaque. Monitoring of parasitemia was done by thigh pricks each time followed by a reward. The intravenous injection and large blood collection were performed under ketamine sedation, and all efforts were made to minimize any suffering of the animal. The monkey was daily monitored for health and discomfort. Immediately after taking blood from the monkey, the monkey was cured from malaria by intramuscular injection of chloroquine (7.5 mg/kg, on 3 consecutive days) and the absence of parasites was verified two weeks after treatment by microscopy of Giemsa stained slides of thigh prick blood of the monkey.

### Sequencing, assembly and annotation of *P. cynomolgi*

Genomic DNA was sheared into 250–350 base-pair fragments by focused ultrasonication (Covaris Adaptive Focused Acoustics technology (AFA Inc., Woburn, USA), and amplification-free Illumina libraries were prepared<sup>16</sup>. Paired 76-base reads were generated on the Illumina GAII platform according to the manufacturer's standard sequencing protocol.

We also generated a SMRTbell template library using the Pacific Biosciences issued protocol (20 kb Template Preparation Using BluePippin Size-Selection System). Five SMRT cells were sequenced on the PacBio RS II platform using P5 polymerase and the chemistry version 3 (C3/P5).

Raw sequence data were deposited in the European Nucleotide Archive under accession number [ERP000298](https://www.ebi.ac.uk/ena/browser/view/ERP000298).

Sequence data from the SMRT cells were assembled with HGAP<sup>17</sup> (version 2.3.0), assuming an assembly size of 30 Mb. The resulting draft assembly was further improved using the IPA script (<https://github.com/ThomasDOtto/IPA>), version 1.0.1. This script performs the following steps:

- 1) deletes small contigs,

- 2) identifies overlapping contigs with low Illumina coverage,
- 3) orders contigs against the *P. vivax* P01 reference using ABACAS2<sup>18</sup> (version 1),
- 4) corrects errors with Illumina reads using iCORN2<sup>19</sup> (version 0.95),
- 5) circularizes the two plasmid genomes with Circlator<sup>20</sup> (version 0.12.0); and
- 6) renames the chromosomes and contigs.

Draft genome annotation was transferred from *P. vivax* P01 using RATT<sup>21</sup> (version 1), and supplemented with the output of the Augustus<sup>22</sup> gene finder, trained on *P. vivax* P01 as described in<sup>23</sup>. This was followed by manual curation of the gene models in Artemis<sup>24</sup> (version from January 2015).

### Re-annotation of *P. coatneyi*

The published *P. coatneyi* genome assembly<sup>14</sup> (accession numbers CP016239 to CP016252 from NCBI) contains several large open reading frames that appear to correspond to coding sequences, especially in the subtelomeric regions. Using the reference genomes of *P. vivax* P01 and *P. knowlesi*, we re-annotated *P. coatneyi* using Companion<sup>25</sup> (version 1.0.1). Default settings were used, with the exception of a cut-off of 0.2 for the “Augustus” parameter.

### Analysis of *P. simiovale*

Short reads of *P. simiovale* were obtained from the SRA<sup>15</sup> (accession number SRR826495). The reads were assembled with MaSuRCA<sup>26</sup> (version 2.1.0), improved with PAGIT<sup>27</sup> (version 1) and annotated with Companion<sup>25</sup> (version 1.0.1), reference *P. vivax* P01 and default settings.

### OrthoMCL

To identify orthologues, genes from the following eleven genome sequences were clustered using OrthoMCL<sup>28</sup> (version 1.4): the present *P. cynomolgi* M, *P. vivax* P01<sup>29</sup>, *P. falciparum* 3D7<sup>30</sup>, *P. reichenowi* CDC<sup>31</sup>, the re-annotated *P. coatneyi*, the rodent malaria parasites (*P. yoelii*, *P. chabaudi* and *P. berghei*<sup>32</sup>), *P. knowlesi*<sup>33</sup>, *P. malariae* and *P. ovale curtisi*<sup>34</sup>. We used the May 2016 version of the genome annotations, taken from GeneDB<sup>35</sup>. The amino acid sequences were compared using a BLASTp all-against-all, with an E-value cut-off of 1e-6. OrthoMCL version 1.4 was used, and a PERL script ascribed the gene functions to each gene ID.

### MSP analysis

All the genes annotated as ‘merozoite surface protein’ from *P. falciparum*, *P. reichenowi* CDC, *P. ovale curtisi*, *P. malariae*, *P. cynomolgi* M, *P. vivax* P01, *P. coatneyi* and *P. knowlesi* were selected and compared with a BLASTp (E-value 1e-6 -F F). The results were visualized with Gephi<sup>36</sup> (version 0.9.1). Genes that clustered together in that analysis were aligned with mafft<sup>37</sup> (version 7.205, parameter --auto). The alignment was trimmed with GBLOCKS<sup>38</sup> (version 0.91b) in Seaview<sup>39</sup> (version 4.6.1) and the tree was built with raxML<sup>40</sup> (version 8.0.24) using the PROTGAMMAGTR model and a bootstrap of 100. Visualization was done in FigTree<sup>41</sup> (version 1.4.2).

### Methyltransferases

Genes with the product ‘methyltransferase’ were all selected as nucleotide sequences. A selection of these genes, based on sequence

similarity, was aligned with mafft. The phylogenetic tree was generated as the MSP tree, using the PROTGAMMAGTR model. Potential transposons were analysed with <http://www.girinst.org><sup>42</sup> (using the RepbaseSubmitter section).

### PIR analysis

The amino acid sequences of the *Plasmodium* interspersed repeat (*pir*) genes were extracted from five genomes (PcyM, *P. vivax* P01, *P. coatneyi*, *P. ovale curtisi* and *P. knowlesi*). First, low complexity sequences were trimmed with seg<sup>43</sup>. Next, proteins smaller than 250aa were excluded. A BLASTp all-against-all comparison was run (E-value 1e-6, -F F, allowing for up to 4500 hits). The results were visualized in Gephi<sup>36</sup>, clustered with the force field and the Reingold-Watermann algorithm. We also clustered the *pir* genes from the same BLAST with TribeMCL<sup>44</sup>, using an inflation coefficient of 1.5.

## Results and discussion

### Improved genome assembly and annotation

The existing *P. cynomolgi* reference (B-strain, referred henceforth as PcyB) is highly fragmented, with 1,649 unassigned scaffolds. We generated a new reference genome sequence (*P. cynomolgi* M strain – PcyM) using high-depth (>100x) Pacific Bioscience long-read sequence data and further improved it with Illumina sequencing reads. The new PcyM assembly is significantly larger than the PcyB assembly (31 versus 26 Mb) (see Table 1), more contiguous (N90 of 370kb versus 3.9kb), and has no sequencing gaps (0 versus 1943 gaps). The unassigned scaffolds have been reduced from 1,649 in PcyB to just 40 in the new PcyM assembly (see Figure 1).

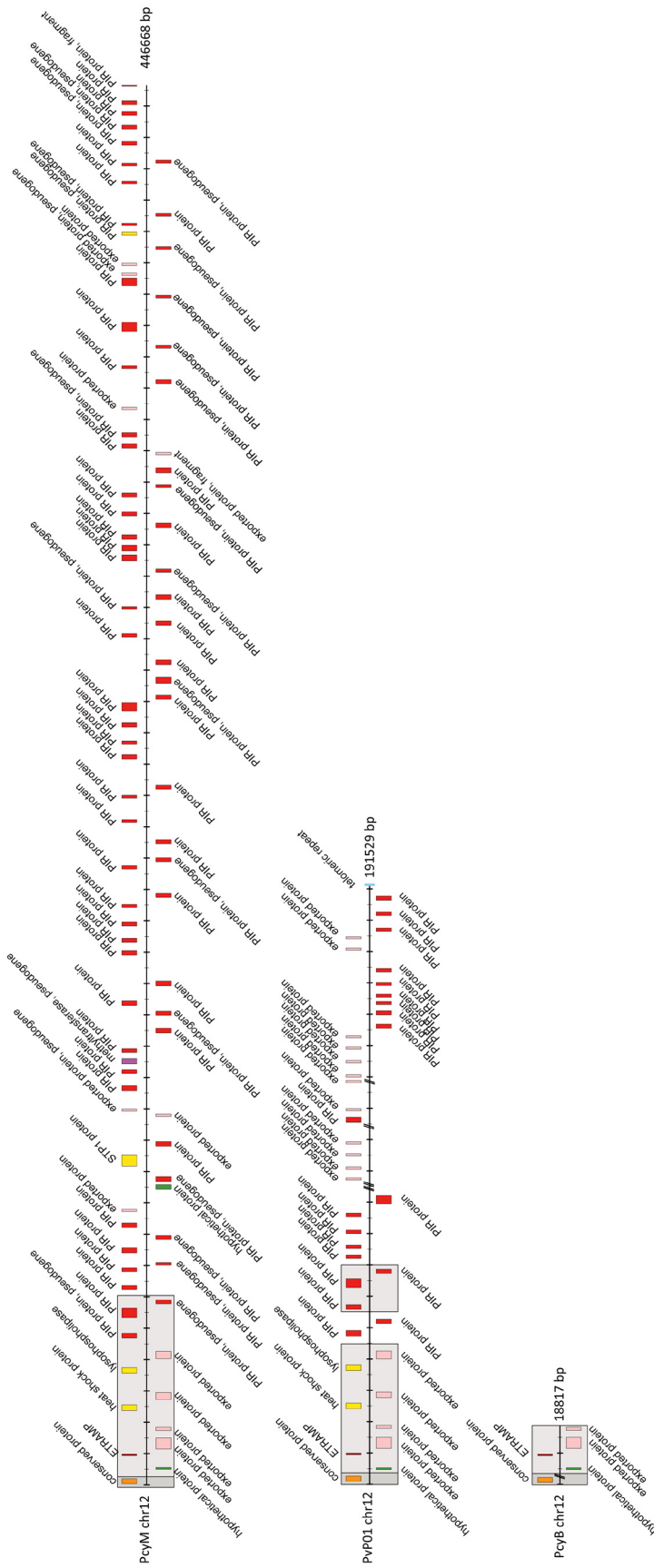
**Table 1. Comparison of *P. cynomolgi* M, *P. cynomolgi* B and *P. vivax* P01 genome features.**

Genome features	PcyM	PcyB <sup>a</sup>	PvP01 <sup>b</sup>
<b>Nuclear genome</b>			
Assembly size (Mb)	30.6	26.2	29.0
Coverage (fold)	>150	161	212
G + C content (%)	37.3	40.4	39.8
No. contigs assigned to chrom.	14	14	14
No. unassigned contigs	40	1,649	226
# Sequencing Gaps	0	1943	560
No. genes <sup>c</sup>	6,632	5,722	6,642
Average gene length (bp) <sup>d</sup>	758	622	741
No. <i>pir</i> genes	1,373	265	1,212
<b>Mitochondrial genome<sup>c</sup></b>			
Assembly size (bp)	6,017	5,986	5,989
G + C content (%)	30.3	30.3	30.5
<b>Apicoplast genome</b>			
Assembly size (kb)	34.5	29.3	29.6
G + C content (%)	14.2	13.0	13.3
No. genes	30	23	30

<sup>a,b</sup>: Published sequences

<sup>c</sup>: Including pseudogenes and partial genes, excluding non-coding RNA genes.

<sup>d</sup>: Based on 1-1 orthologous



**Figure 1. Organization of subtelomeric regions of chromosomes 12 of *P. cynomolgi* M, *P. vivax* P01 and *P. cynomolgi* B.** The order and orientation of the genes in the subtelomeric region of chromosomes 12 (right hand side) of *P. cynomolgi* M (PcM), *P. vivax* P01 (PVP01) and *P. cynomolgi* B (PcB) are shown. Exons are shown in coloured boxes. // lines in PVP01 represent gaps. The dark shaded/grey areas mark the start of the conserved, syntenic regions to other *Plasmodium* species, e.g. *P. falciparum*. The lighter shaded/grey areas mark the syntenic regions between PcM, PVP01 and PcB.

These improvements in contiguity and reduction of gaps had a large impact on the quality of the gene models. Overall, genes in PcyM are similar in size to their orthologues in *P. vivax* P01, while those in PcyB are around 20% shorter. In terms of annotation, 966 new genes were found in the PcyM assembly compared to PcyB, with most of these genes being found in the subtelomeres (see Table 2). The new genes, however, also include 119 genes that are 1-1 orthologous to genes in *P. vivax*. Due to the manual curation, 12% more genes have been assigned a gene function in the new assembly. These systematic improvements make the PcyM genome sequence a better reference for the community to use when studying the biology of *P. cynomolgi* and relapsing malaria parasites in general.

The genome sequences were obtained from samples that were originally described as being two different strains, Mulligan (M strain) and Bastianelli (B-strain). However, a genome-wide comparison of the gene repertoires reveals that 67% of the 1:1 orthologues are identical, which is much more than the number of identical genes observed (32%) between two *P. vivax* isolates (P01 versus C01). This is in line with the findings in the original publication describing

the PcyB genome assembly<sup>13</sup>, suggesting that the two strains are likely derived from the same isolate. This was further confirmed by a recent study that analysed the diversity of several *P. cynomolgi* isolates<sup>45</sup>. Although the authors proposed to call the isolate M/B, we will use the Mulligan nomenclature for continuity.

### OrthoMCL clustering

To look for conserved orthologues between species, an OrthoMCL<sup>28</sup> clustering of genes from eleven genome assemblies was performed (see Methods and Supplementary Table 1). We used the clustering to look further into genes potentially involved in the formation and development of the dormant hypnozoite stage. There are 103 gene clusters (see Figure 2) that are common to the relapsing parasites, but absent in *P. knowlesi* and *P. coatneyi*. Of these, 73 gene clusters are uniquely shared between *P. vivax* P01, PcyM and *P. ovale curtisi* GH01. The remaining 30 clusters are either shared with various combinations of the other nine parasite species (see Supplementary Table 1) or only with *P. malariae* (20 out of the 30 clusters).

The 73 clusters unique to the relapsing parasites include three tryptophan rich protein clusters where the orthology is 1:1:1 with the exception of one cluster in which *P. vivax* presents an expansion to four genes; two PHIST proteins (before named RAD and Pv-fam-e) clusters containing 1:1:1 orthologues; 11 clusters featuring 1:1:1 orthologues annotated as 'Plasmodium exported proteins'; three clusters of 1:1:1 hypothetical protein orthologues; one cluster annotated as MSP-7 or MSP-7-like and 56 *pir* gene clusters showing different degrees of expansion in the three relapsing species. While their specificity is interesting, clusters corresponding to multigene

**Table 2. Number of gene members of different (subtelomeric) multigene families in the genomes of *P. cynomolgi* B, *P. cynomolgi* M, *P. vivax* P01.**

Subtelomeric genes*				other (previous) names
	PcyM	PcyB**	PvP01**	
<b>Gene family</b>				
PIR protein	1373	265	1212	vir-like, kir-like
tryptophan-rich protein	39	36	40	Pv-fam-a, TRAG, tryptophan-rich antigen
methyltransferase, pseudogene	36	26***	0	
lysophospholipase	8	9	10	PST-A protein
STP1 protein	51	3	10	PvSTP1
early transcribed membrane protein (ETRAMP)	9	9	9	
Plasmodium exported protein (PHIST), unknown function	54	48	84	Phist protein (Pf-fam-b), RAD protein (Pv-fam-e)
reticulocyte binding protein	6	8	9**	reticulocyte-binding protein, RBP
exported protein****	276	175	447	

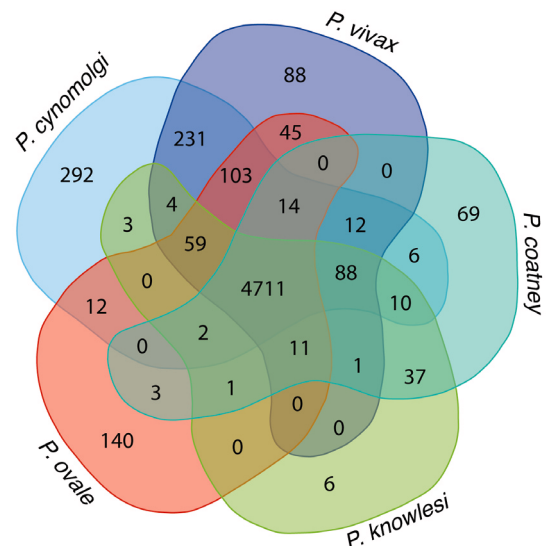
Key:

\*Numbers including pseudogenes and partial genes

\*\*Published sequence

\*\*\*annotated as hypothetical protein

\*\*\*\*ExportPred



**Figure 2. Orthologous classes of five genomes.** Shared orthologous clusters produced using OrthoMCL version 1.4 with default parameters. The high number of shared clusters between species confirms the number of shared genes between the species. The 242 clusters between *P. vivax* and *P. cynomolgi* emphasise that they are closer related. The 103 clusters shared between *P. ovale*, *P. vivax* and *P. cynomolgi* might give insight into genes associated to the hypnozoite stages.

families are probably less likely to have a direct function in dormancy. The hypothetical protein clusters (PcyM\_0326800, PcyM\_0423700 and PcyM\_0904700), however, being specific to the three relapsing *Plasmodium* species, are intriguing, as is the MSP-like protein cluster.

**Paralogous expansion of the merozoite surface protein (MSP) family.** Although the specific function of the different merozoite surface proteins (MSPs) remains elusive, MSP-1 and MSP-3 are currently under evaluation as vaccine candidates. The OrthoMCL clustering shows that MSP-1, MSP-1 paralog, MSP-4, MSP-5, MSP-9 and MSP-10 are highly conserved and present across different *Plasmodium* species. MSP-2 and MSP-6 are present only in *P. falciparum* and *P. reichenowi* (see [Figure 3A](#)). In contrast, MSP-3 and MSP-7/7-like are highly expanded. MSP-3 is expanded in *P. vivax*, *P. malariae*, *P. ovale* and *P. cynomolgi* (see [Figure 3B](#)). Interestingly, while in *P. malariae* and to *P. ovale*, MSP-3 paralogs seem to be species-specific, in *P. cynomolgi*, *P. vivax*, *P. coatneyi* and *P. knowlesi* many of the paralogs seem to predate speciation, indicating that MSP-3 duplicated in the common ancestor of the latter four species. These findings of MSP-3 expansions are in line with the finding of multi-allelic diversification reported previously<sup>46</sup>, but also confirm the expansion in *P. malariae* and *P. ovale*. In addition to the pre-speciation expansion in *P. cynomolgi*, a species-specific expansion of MSP-3 (see area indicated with “\*” in [Figure 3B](#)) genes suggests ongoing evolutionary pressure on these genes.

We also observed an expansion of MSP-7/7-like genes. In the OrthoMCL clustering, the genes were distributed in nine different clusters: 108, 4913, 5404, 5550, 5065, 6376 and 5765–5767 ([Supplementary Table 1](#)). A phylogenetic tree of the MSP-7/7-like proteins revealed a complex evolutionary relationship (see [Figure 3C](#)), splitting the tree into three major clades. Across the tree we find paralogous expansions of different ages, some of which predate speciation. A particularly striking branch comprises only genes from the three hypnozoite-forming species. As a result of the large amount of genome sequences now available for different *Plasmodium* species, a complex pattern now emerges in the MSP7/7-like tree, suggesting that the different MSP7 proteins likely have different functions.

#### Improved sub-telomeres reveals insights into subtelomeric gene families

The new high-quality PcyM assembly has an improved representation of the subtelomeric regions of the genome, which now encompass nearly 40% of the genome sequence. Manual curation of the gene annotation enabled the complete set of subtelomeric genes to be resolved (see [Table 2](#)). In *P. vivax*, genes encoding the exported protein family ‘PHIST’, and exported proteins in general (as predicted by ExportPred<sup>47</sup>), have paralogously expanded compared to *P. cynomolgi* (84 vs 54). It is tempting to speculate about the reason for the higher number of exported proteins in *P. vivax*. One hypothesis is that it could be due to differences in the blood cells of humans compared to primates; while another could be that they are involved in the regulation of genes involved in host parasite interaction. In *P. falciparum*, it was suggested that PHISTb regulates *var*

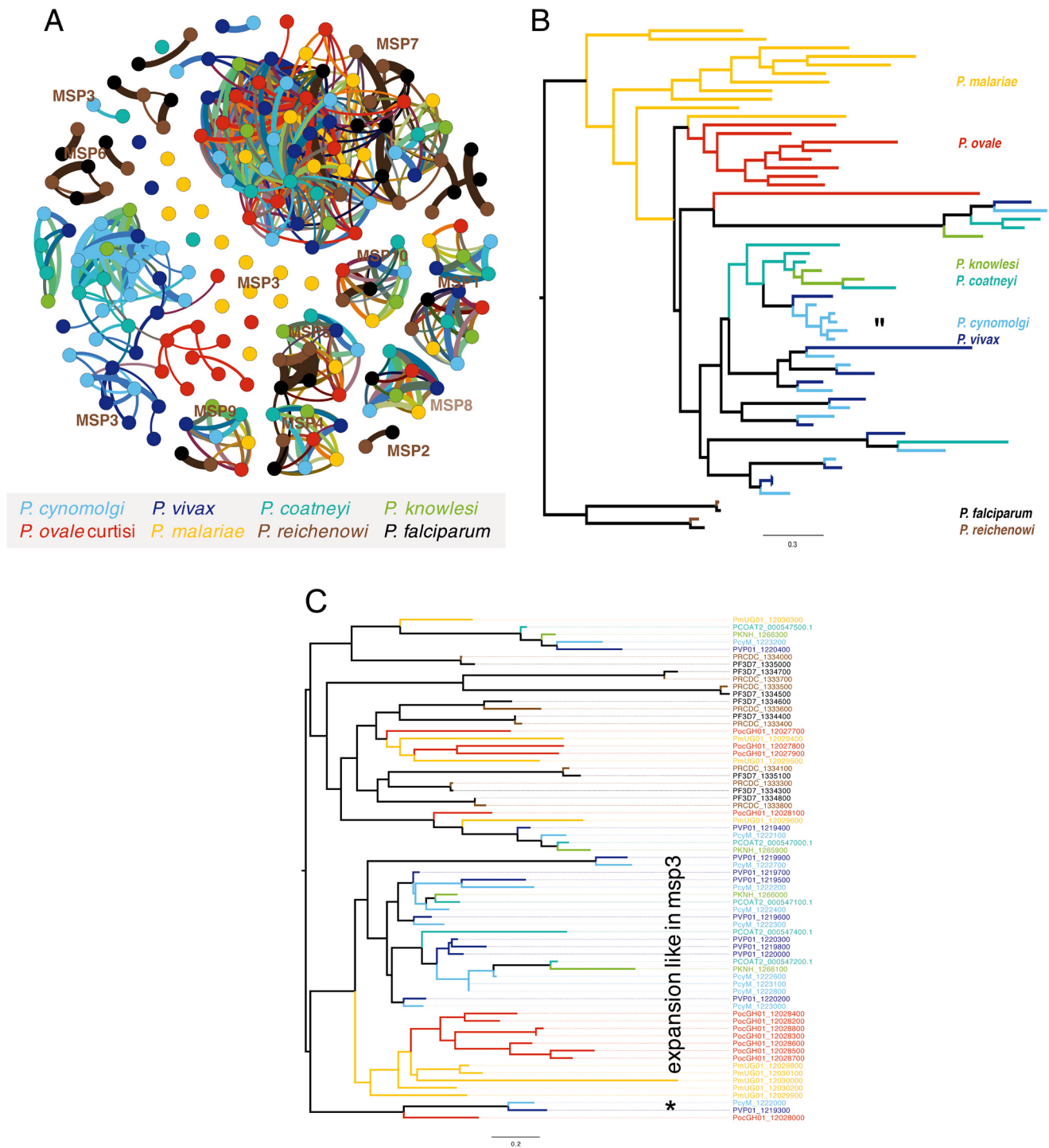
genes<sup>48</sup>. In *P. cynomolgi*, we observed an expansion of the STP1 family (51 genes). STP1 proteins are common in *P. malariae* and *P. ovale curtisi* (166 and 70 genes, respectively), but are contracted in number in *P. vivax* (10 genes). One may also speculate that the expansion of PHIST and exported proteins in *P. vivax* compensates for the lack of STP1 proteins.

The largest multigene family in *P. cynomolgi* comprises *pir* genes. The *pir* superfamily occurs in all *Plasmodium* species<sup>49</sup>, but their function remains poorly understood. Recent studies suggest a possible role in the regulation of the establishment of chronic infections<sup>50</sup> and they have been found expressed in liver stage infections of rodent parasites<sup>51</sup>. An extensive repertoire of 1373 *pir* genes was identified in the PcyM assembly, compared to 263 in PcyB. This updated number puts the *P. cynomolgi* *pir* gene repertoire at a similar size to that of *P. vivax* (1,216), while *P. ovale curtisi* has an even larger repertoire (1,949). Conversely, *P. knowlesi*, has only 70 *pir* genes present. Interestingly, the re-annotated *P. coatneyi* genome that clusters closely to *P. knowlesi* has 827 *pir* genes (see [Figure 4B](#)). In the published annotation it has just 256 *pir* genes.

As previously reported<sup>29,52</sup>, the *pir* genes can be grouped based on sequence similarity. We observe that the diversity of the *pir* repertoire is dramatically reduced in *P. coatneyi* and *P. knowlesi*. Most of the *pir* genes form the same cluster (cluster 0; [Figure 4A](#)). However, that cluster splits into two groups in the gene-gene network due to the different lengths of the *pir* genes in *P. coatneyi* and *P. knowlesi* (see [Figure 4B](#)). One hypothesis for the loss of other *pir* types might be the occurrence of sicaVAR genes in *P. knowlesi* and *P. coatneyi*<sup>33</sup>. The reduction of the *pir* repertoire is an interesting parallel to the *Laverania*, where the amount of *rif* genes (analogous to *pir* genes) is reduced but a new gene family evolved, the *var* genes. Additionally, in the *Laverania* the number of *rif* genes drops further when the parasite is in the human compared to the primate<sup>31</sup>.

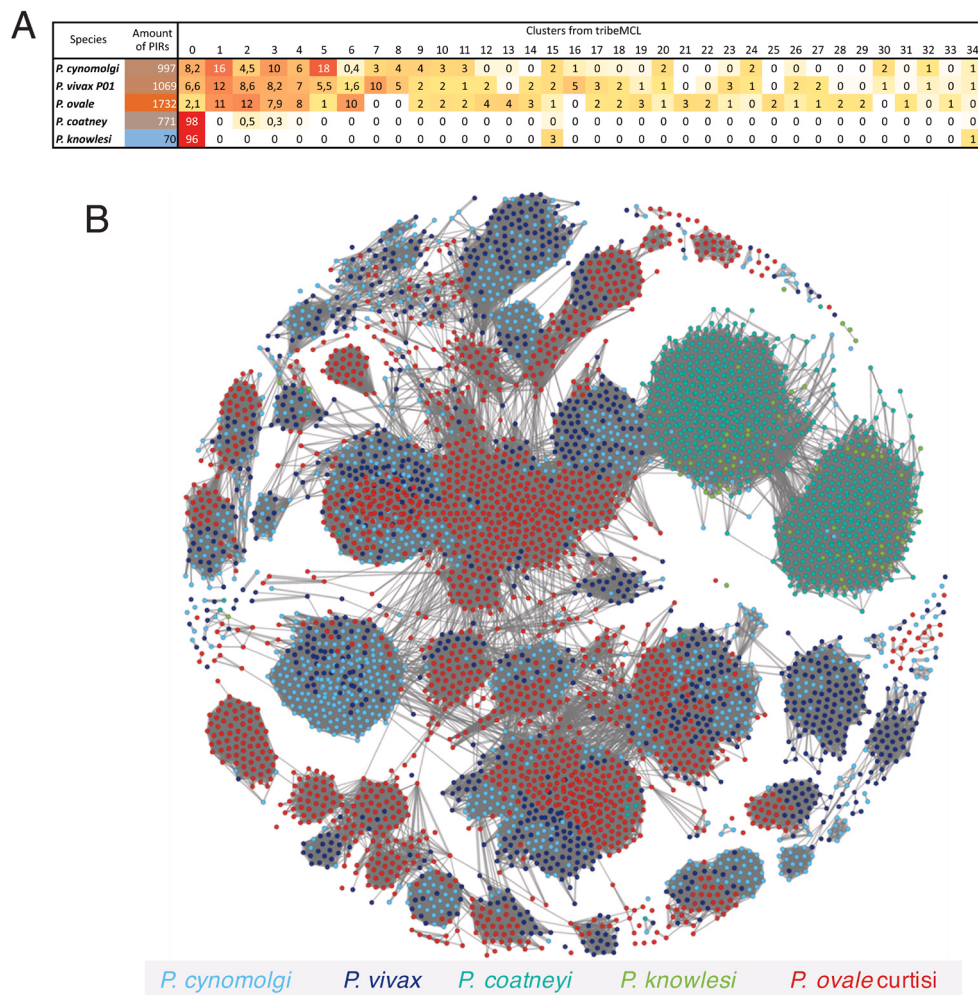
As for the other clusters, it seems that the underlying structure of the *pir* genes predates the speciation of *P. ovale*, *P. vivax* and *P. cynomolgi*. Depending on the type of *pir*, the amount can fluctuate, as can be seen by the large variance in number of genes per cluster. Some clusters are specific to *P. ovale* and some others contain just the two human malaria parasites, *P. vivax* and *P. ovale*. Interestingly, several *pir* genes have 1:1 orthologues across the different species ([Supplementary Table 1](#), see [Figure 4B](#)). As those genes seem to be conserved across evolutionary time, it is unlikely that they are extracellular (where they would be under immune pressure), rather they must have more conserved core functions.

**Expansion of methyltransferases.** While paralogous expansions of *pir* genes and genes encoding MSP genes have been described in other *Plasmodium* species, *P. cynomolgi* exhibits an unexpected expansion of 36 methyltransferase pseudogenes. These pseudogenes are found in the subtelomeres, and were annotated as encoding 26 hypothetical proteins in the PcyB assembly. The role of pseudogenes in *Plasmodium* is little understood, but in several malaria parasite species conserved pseudogenes are found in the subtelomeres. In the OrthoMCL clustering, all 36 methyltransferase



**Figure 3. Analysis of expansion merozoite surface proteins.** (A) BLAST-based graph of all the merozoite surface proteins (MSP), (cut-off 20% global similarity). The different MSP types form clusters, apart from MSP3 which seem to be more diverse. (B) Maximum likelihood tree (PROTGAMMAJTTF model, bootstrap at all branches in 100) of MSP3 and 2 laverania MSP6, shows species complex specific expansions in some species. The expansion was pre-speciation of *P. cynomolgi* and *P. vivax*. We also observe a MSP3 expansion in *P. cynomolgi*. "\*" indicates an expansion of MSP3 in *P. cynomolgi*. (C) As in (B), but with MSP7 and MSP7. The tree is more complex, showing different types of MSP7. Some clades have a similar structure to MSP3, with specific expansions. "\*" highlights a cluster containing MSP7 from parasites that have the hypnozoite stage.





**Figure 4. Cluster analysis illustrating the relatedness between the PIR proteins in five genomes. (A)** Classification (generated with tribeMCL inflation parameter 1.5) highlights the different types of PIR clusters. *P. coatneyi* and *P. knowlesi*. **(B)** Network illustrating the relatedness of PIR between *P. ovale* (red), *P. cynomolgi* (light blue) and *P. vivax* (blue), and between *P. coatneyi* (light green) and *P. knowlesi* (green). *P. ovale*, *P. cynomolgi* and *P. vivax* clearly share the same general topology of PIR architecture. In contrast, *P. coatneyi* and *P. knowlesi* have a reduced number PIRs and also of PIR classes.

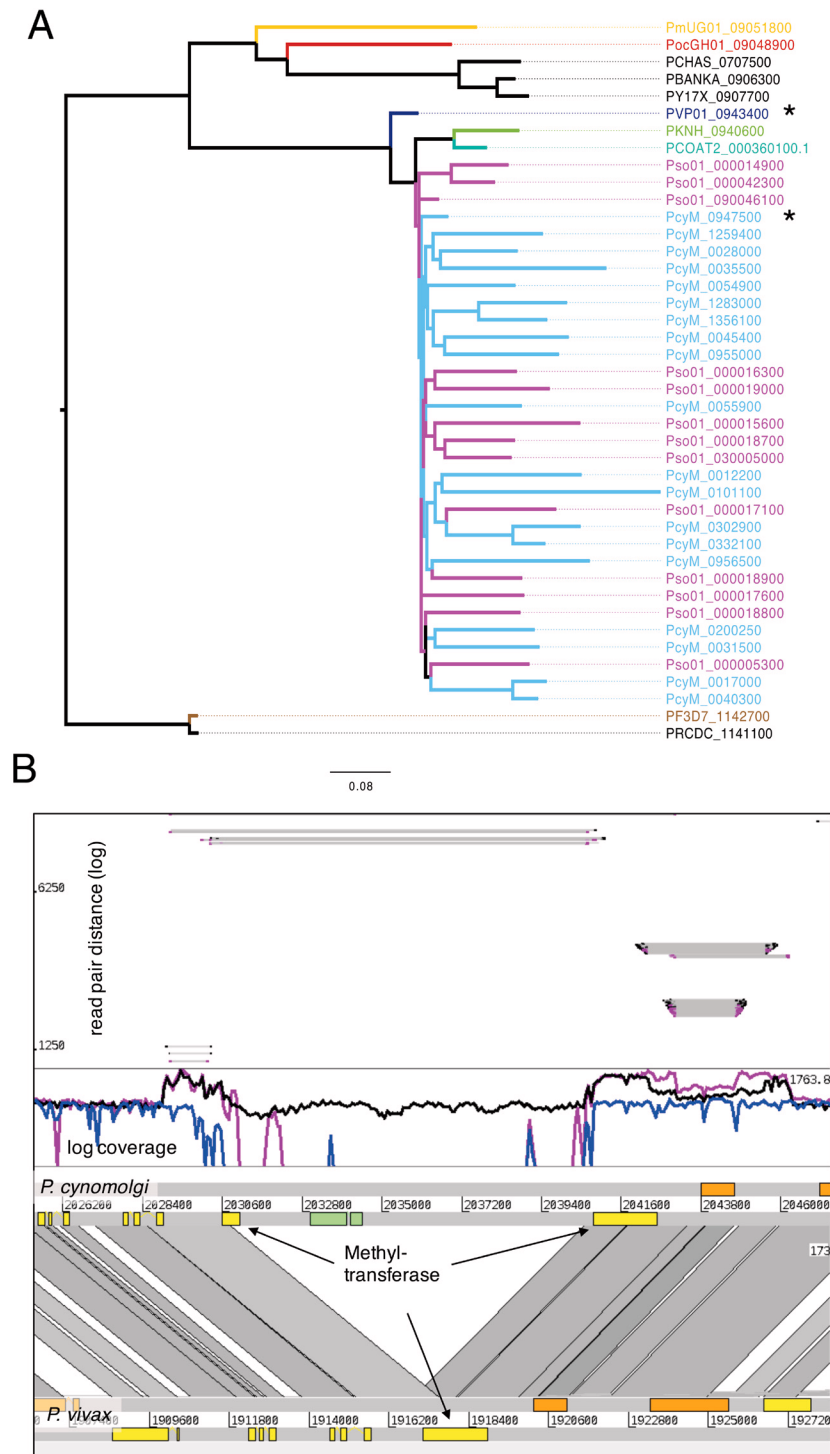
pseudogenes cluster with one full-length core gene (PcyM\_0947500, Figure 5A). This gene is found on chromosome 9 and has one conserved orthologue across all other *Plasmodium* species (cluster 51, Supplementary Table 1), and is found in many other species on OrthoMCL as cluster OG5\_129798. The 36 copies are spread evenly throughout the subtelomeres, without evidence of spatial clustering.

The methyltransferase pseudogenes contain motifs of the Caulimovirus, a virus often found integrated in to plant genomes, and of different retrotransposons families such as *aedes aegypti*, Gypsy, Helitron-5, CACTA-1, RTEK and CR1 (see Supplementary Table 2). While the Caulimovirus insert was mostly found to have occurred in an antisense orientation hinting towards a role in stability, the LTR and non-LTR insertions were found most often to have occurred

in a sense orientation<sup>53</sup>. The hits were mostly to low complexity regions, suggesting that recombination in the subtelomeres may be a result of mechanisms similar to those used by retro elements.

We also found evidence that this duplication of methyltransferases was also found in *P. simiovale*, a close outgroup to *P. cynomolgi*, *P. vivax*, and *P. knowlesi*. Fewer copies were observed in the *P. simiovale* assembly (13), but this may be due to the fragmentation of the assembly. Although they are generally less degenerate at their 5' ends, they are nevertheless pseudogenized.

To further understand the duplication, we mapped the reads of *P. cynomolgi*, *P. simiovale* and *P. vivax* P01 against the locus on chromosome 9 containing the ancestral methyltransferase in *P. cynomolgi* (see Figure 5B). Although the coverage is shown as log



**Figure 5. Expansion of Methyltransferase in *P. cynomolgi* and *P. simiovale*.** (A) Tree of methyltransferase in Plasmodium, including the expansion of those genes in *P. cynomolgi* (36) and *P. simiovale* (at least 15). The closest core genes are PVP01\_0943400 and PcyM\_0947500. (B) Comparative view of *P. cynomolgi* and *P. vivax* on the locus of methyltransferase (\*) of panel A. Interestingly, the locus in *P. cynomolgi* has an insertion with a subtelomeric gene that has a weak hit with a putative DNA translocase Ftsk domain. Coverage plot mapped from *P. cynomolgi* reads (black), *P. vivax* (blue) and *P. simiovale* (magenta) is shown in log scale on *P. cynomolgi*. The methyltransferases are duplicated more than 35 times. As the height is roughly similar between the two duplications, we expect around the same number of methyltransferases in *P. simiovale* than in *P. cynomolgi*. The insert of the green gene is found just in *P. cynomolgi*, due to the missing coverage. The upper panel shows the distance of read pairs; the insertion of the region probably occurred after the duplication of the gene into the subtelomeres, as all reads from the duplications are connected over the insertion. The next core gene is also duplicated.

scale, the coverage across the methyltransferase seems to be identical for *P. simiovale* and *P. cynomolgi*, but significantly lower for *P. vivax*. This leaves us to speculate that the number of methyltransferases is roughly the same in both *P. simiovale* and *P. cynomolgi*. Further, the coverage plot also reveals that the next core gene of unknown function, PcyM\_0947600, is also duplicated. In PcyM we find two further paralogous genes: PcyM\_0054800 PcyM\_0012100. Furthermore, it is more often duplicated in *P. simiovale*, as the coverage of that gene is high (Figure 5B). A search for structural similarity using I-TASSER<sup>54</sup> yielded no conclusive results.

A phylogenetic tree (see Figure 5A) shows the methyltransferase paralogs in *P. cynomolgi* and *P. simiovale* compared to the orthologues in the other species. The genes generally follow the species tree, but they are expanded in *P. cynomolgi* and *P. simiovale*. As *P. simiovale* is thought to be an outgroup to *P. cynomolgi* and *P. vivax*, we expect that *P. vivax* has lost the expansions.

We compared the location of the ancestral methyltransferase between PcyM and *P. vivax*. To our surprise, we found a potential open reading frame inserted between two methyltransferases in *P. cynomolgi*. A tBLASTn of that CDS against the Nucleotide NCBI database revealed no significant similarity to any other sequence, except for the subtelomeres of *P. vivax* and *P. cynomolgi*. A very weak hit (e-value of e-4) to a DNA translocase FtsK, is an interesting finding, in light of the potential LTR transposon-like sequences discussed previously, but is to be taken with caution. This particular open reading frame is absent in *P. simiovale* and seems that have occurred subsequent to the expansion and is not likely to be implicated in the expansion itself.

It remains speculative if the paralogs of the methyltransferase genes and the adjacent gene were functional in the ancestor. Hypothetical roles of the methyltransferase could involve any of the following: 1) the epigenetic control of differential *pir* gene expression in acute and chronic infections<sup>50</sup>, 2) the sequence may have a role in genome stability and recombination, or 3) this could be a selfish gene that was able to transpose.

## Conclusion

The availability of a new and improved *P. cynomolgi* reference genome sequence will enable in-depth studies of this widely used model parasite, including investigations into dormant stages and the selection of new drug targets and vaccine candidates. High quality genomics related studies will now be possible, including studies of previously missed core genes. In particular, the improved subtelomeres have enabled us to dissect the *pir* gene family

further, and have revealed a novel and unexpected expansion of methyltransferase genes.

## Data and software availability

The project number of the *P. cynomolgi* raw reads is deposited in the European Nucleotide Archive under accession number ERP000298. The submitted genome is under the project number PRJEB2243.

The chromosomes have the accession: LT841379-LT841394, and the scaffolds: FXLJ01000001-FXLJ01000040.

The annotation can be found at: <ftp://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/cynomolgi/M/Jan2017/>

The automated re-annotation of *P. coatneyi* and the draft assembly of *P. simiovale* can be found at: <ftp://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/coatneyi/ReAnnotation/> and <ftp://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/simiovale/May2017/>, respectively.

The IPA software is available on GitHub: <https://github.com/ThomasDOtto/IPA>. Version 1.0.1 was used for this work.

The software is also available on Zenodo: <https://doi.org/10.5281/zenodo.806818><sup>55</sup>

License: GNU General Public License v3.0

---

## Author contributions

AV infected the Rhesus monkey, isolated the parasites and extracted the DNA. MS organized the sample sequencing. GGR performed the *P. simiovale* analysis. UB and EP performed manual curation of the gene models. TDO perform the bioinformatics analysis. EP, CK, MB and TDO conceived the study. EP and TDO wrote the paper. All authors read, corrected and approved the manuscript.

## Competing interests

None of the authors declared competing interest.

## Grant information

This work was supported by the Wellcome Trust (098051), EVI-MalaR (contract number 242095) and Gates Foundation Project OPP1023583. GGR is supported by the Medical Research Council (MR/J004111/1).

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

## Supplementary material

**Supplementary Table 1: Annotated OrthoMCL of 11 species.**

[Click here to access the data.](#)

**Supplementary Table 2: Results of the search for motifs associated with transposons, <http://www.girinst.org>.**

[Click here to access the data.](#)

## References

1. Mayer M: **Über malaria beim Affen.** *Med Klin, Berl.* 1907; 579–580.
2. Mulligan HW: **Descriptions of two species of monkey *Plasmodium* isolated from *Silenus irus*.** *Arch Protistenkunde.* 1935; 84(2): 285–314.  
[Reference Source](#)
3. Garnham PC: **A new sub-species of *Plasmodium cynomolgi*.** *Rivista di Parassitologia.* 1959; 20(4): 273–278.  
[Reference Source](#)
4. Eyles DE: **The species of simian malaria: taxonomy, morphology, life cycle, and geographical distribution of the monkey species.** *J Parasitol.* 1963; 49(6): 866–887.  
[PubMed Abstract](#) | [Publisher Full Text](#)
5. Cheong WH, Coombs GL: **Transmission of *Plasmodium cynomolgi* (Perlis strain) to man.** *Se Asian J Trop Med Pub Hlth.* 1970; 302.
6. Bennet GF, Warren M, Cheong WH: **Biology of the simian malarials of Southeast Asia. II. The susceptibility of some Malaysian mosquitoes to infection with five strains of *Plasmodium cynomolgi*.** *J Parasitol.* 1966; 52(4): 625–631.  
[PubMed Abstract](#) | [Publisher Full Text](#)
7. Warren M, Wharton RH: **The vectors of simian malaria: identity, biology, and geographical distribution.** *J Parasitol.* 1963; 49(6): 892–904.  
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Prakash S, Chakrabarti SC: **The isolation and description of *Plasmodium cynomolgi* and *Plasmodium inui* from naturally occurring mixed infections in *Macaca radiata radiata* monkeys of the Nilgiris, Madras state, India.** *Ind J Malariol.* 1962; 303–311.
9. Eyles DE, Laing AB, Warren MW, *et al.*: **Malaria parasites of Malayan leaf monkeys of the genus *Presbytis*.** *Med J Malaya.* 1962; 85–86.
10. Dissanaïke AS: **Simian malaria parasites of Ceylon.** *Bull World Health Organ.* 1965; 32(4): 593–597.  
[PubMed Abstract](#) | [Free Full Text](#)
11. Wolfson F, Winter MW: **Studies of *Plasmodium cynomolgi* in the rhesus monkey, *Macaca mulatta*.** *Am J Hyg.* 1946; 44(2): 273–300.  
[PubMed Abstract](#) | [Publisher Full Text](#)
12. Zeeman AM, van Amsterdam SM, McNamara CW, *et al.*: **KAI407, a potent non-8-aminoquinoline compound that kills *Plasmodium cynomolgi* early dormant liver stage parasites *in vitro*.** *Antimicrob Agents Chemother.* 2014; 58(3): 1586–1595.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. Tachibana S, Sullivan SA, Kawai S, *et al.*: ***Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade.** *Nat Genet.* 2012; 44(9): 1051–1055.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Chien JT, Pakala SB, Geraldo JA, *et al.*: **High-Quality Genome Assembly and Annotation for *Plasmodium coatneyi*, Generated Using Single-Molecule Real-Time PacBio Technology.** *Genome Announc.* 2016; 4(5): pii: e00883-16.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Hester J, Chan ER, Menard D, *et al.*: ***De novo* assembly of a field isolate genome reveals novel *Plasmodium vivax* erythrocyte invasion genes.** *PLoS Negl Trop Dis.* 2013; 7(12): e2569.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Kozarewa I, Ning Z, Quail MA, *et al.*: **Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes.** *Nat Methods.* 2009; 6(4): 291–295.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Chin CS, Alexander DH, Marks P, *et al.*: **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.** *Nat Methods.* 2013; 10(6): 563–569.  
[PubMed Abstract](#) | [Publisher Full Text](#)
18. Assefa S, Keane TM, Otto TD, *et al.*: **ABACAS: algorithm-based automatic contiguation of assembled sequences.** *Bioinformatics.* 2009; 25(15): 1968–1969.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
19. Otto TD, Sanders M, Berriman M, *et al.*: **Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology.** *Bioinformatics.* 2010; 26(14): 1704–1707.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
20. Hunt M, Silva ND, Otto TD, *et al.*: **Circulator: automated circularization of genome assemblies using long sequencing reads.** *Genome Biol.* 2015; 16: 294.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
21. Otto TD, Dillon GP, Degraeve WS, *et al.*: **RATT: Rapid Annotation Transfer Tool.** *Nucleic Acids Res.* 2011; 39(9): e57.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. Stanke M, Keller O, Gunduz I, *et al.*: **AUGUSTUS: *ab initio* prediction of alternative transcripts.** *Nucleic Acids Res.* 2006; 34(Web Server issue): W435–439.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
23. Otto TD: **From sequence mapping to genome assemblies.** *Methods Mol Biol.* 2015; 1201: 19–50.  
[PubMed Abstract](#) | [Publisher Full Text](#)
24. Carver T, Berriman M, Tivey A, *et al.*: **Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database.** *Bioinformatics.* 2008; 24(23): 2672–2676.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
25. Steinbiss S, Silva-Franco F, Brunk B, *et al.*: **Companion: a web server for annotation and analysis of parasite genomes.** *Nucleic Acids Res.* 2016; 44(W1): W29–34.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
26. Zimin AV, Marçais G, Puiu D, *et al.*: **The MaSuRCA genome assembler.** *Bioinformatics.* 2013; 29(21): 2669–2677.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
27. Swain MT, Tsai IJ, Assefa SA, *et al.*: **A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs.** *Nat Protoc.* 2012; 7(7): 1260–1284.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
28. Li L, Stoeckert CJ Jr, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res.* 2003; 13(9): 2178–2189.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. Auburn S, Böhme U, Steinbiss S, *et al.*: **A new *Plasmodium vivax* reference sequence with improved assembly of the subtelomeres reveals an abundance of *pir* genes [version 1; referees: 2 approved].** *Wellcome Open Res.* 2016; 1: 4.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
30. Gardner MJ, Hall N, Fung E, *et al.*: **Genome sequence of the human malaria parasite *Plasmodium falciparum*.** *Nature.* 2002; 419(6906): 498–511.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
31. Otto TD, Rayner JC, Böhme U, *et al.*: **Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts.** *Nat Commun.* 2014; 5: 4754.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
32. Otto TD, Böhme U, Jackson AP, *et al.*: **A comprehensive evaluation of rodent malaria parasite genomes and gene expression.** *BMC Biol.* 2014; 12: 86.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
33. Pain A, Böhme U, Berry AE, *et al.*: **The genome of the simian and human malaria parasite *Plasmodium knowlesi*.** *Nature.* 2008; 455(7214): 799–803.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
34. Rutledge GG, Böhme U, Sanders M, *et al.*: ***Plasmodium malariae* and *P. ovale* genomes provide insights into malaria parasite evolution.** *Nature.* 2017; 542(7639): 101–104.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
35. Logan-Klumpler FJ, De Silva N, Boehme U, *et al.*: **GeneDB—an annotation database for pathogens.** *Nucleic Acids Res.* 2012; 40(Database issue): D98–108.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
36. Bastian M, Heymann S, Jacomy M: *In International AAAI Conference on Weblogs and Social Media.* 2009.
37. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Mol Biol Evol.* 2013; 30(4): 772–780.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
38. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.** *Syst Biol.* 2007; 56(4): 564–577.  
[PubMed Abstract](#) | [Publisher Full Text](#)
39. Gouy M, Guindon S, Gascuel O: **SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building.** *Mol Biol Evol.* 2010; 27(2): 221–224.  
[PubMed Abstract](#) | [Publisher Full Text](#)
40. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics.* 2014; 30(9): 1312–1313.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
41. **FigTree v.1.4.2.** 2014.  
[Reference Source](#)
42. Kohany O, Gentles AJ, Hankus L, *et al.*: **Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor.** *BMC Bioinformatics.* 2006; 7: 474.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
43. Wootton JC, Federhen S: **Analysis of compositionally biased regions in sequencing databases.** *Methods Enzymol.* Academic Press, 1996; 266: 554–571.  
[PubMed Abstract](#) | [Publisher Full Text](#)
44. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res.* 2002; 30(7): 1575–1584.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
45. Sutton PL, Luo Z, Divis PC, *et al.*: **Characterizing the genetic diversity of the monkey malaria parasite *Plasmodium cynomolgi*.** *Infect Genet Evol.* 2016; 40: 243–252.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
46. Rice BL, Acosta MM, Pacheco MA, *et al.*: **The origin and diversification of the merozoite surface protein 3 (*msp3*) multi-gene family in *Plasmodium vivax* and related parasites.** *Mol Phylogenet Evol.* 2014; 78: 172–184.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
47. Boddey JA, Carvalho TG, Hodder AN, *et al.*: **Role of plasmeprin V in export of**

- diverse protein families from the *Plasmodium falciparum* exportome. *Traffic*. 2013; 14(5): 532–550.  
[PubMed Abstract](#) | [Publisher Full Text](#)
48. Oberli A, Slater LM, Cutts E, *et al.*: **A *Plasmodium falciparum* PHIST protein binds the virulence factor PfEMP1 and comigrates to knobs on the host cell surface.** *FASEB J*. 2014; 28: 4420–4433.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
49. Janssen CS, Phillips RS, Turner CM, *et al.*: ***Plasmodium* interspersed repeats: the major multigene superfamily of malaria parasites.** *Nucleic Acids Res*. 2004; 32(19): 5712–5720.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
50. Brugat T, Reid AJ, Lin JW, *et al.*: **Antibody-independent mechanisms regulate the establishment of chronic *Plasmodium* infection.** *Nat Microbiol*. 2017; 2: 16276.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
51. Fougère A, Jackson AP, Bechti DP, *et al.*: **Variant Exported Blood-Stage Proteins Encoded by *Plasmodium* Multigene Families Are Expressed in Liver Stages Where They Are Exported into the Parasitophorous Vacuole.** *PLoS Pathog*. 2016; 12(11): e1005917.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
52. Lopez FJ, Bernabeu M, Fernandez-Becerra C, *et al.*: **A new computational approach redefines the subtelomeric *vir* superfamily of *Plasmodium vivax*.** *BMC Genomics*. 2013; 14: 8.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
53. van de Lagemaat LN, Medstrand P, Mager DL: **Multiple effects govern endogenous retrovirus survival patterns in human gene introns.** *Genome Biol*. 2006; 7(9): R86.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
54. Roy A, Kucukural A, Zhang Y: **I-TASSER: a unified platform for automated protein structure and function prediction.** *Nat Protoc*. 2010; 5(4): 725–738.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
55. Otto TD: **ThomasDOtto/IPA: Release which is in Zendo.** *Zenodo*. 2017.  
[Data Source](#)

# Open Peer Review

Current Referee Status:  

---

## Version 1

Referee Report 30 June 2017

doi:[10.21956/wellcomeopenres.12821.r23586](https://doi.org/10.21956/wellcomeopenres.12821.r23586)



**Richárd Bártfai**

Department of Molecular Biology, Radboud University, Nijmegen, Netherlands

In this study, Pasini and colleagues improved the assembly and annotation of the *P. cynomolgi* genome. They performed short-read Illumina and long-read PacBio sequencing of DNA isolated from the PcyM strain. Assembly of these sequences yielded a markedly improved reference genome with substantially less gaps and much better coverage of the subtelomeric regions. Furthermore, manual curation resulted in substantially improved gene models and better annotation of gene functions. Comparative genome analysis highlighted interesting dynamics in copy number variation of specific gene families (MSP, STP1, PIR), and in particular a peculiar expansion of methyltransferase pseudogenes.

This manuscript is well written and describe a well-executed assembly and annotation of the *P. cynomolgi* reference genome. This reference genome will be well appreciated in the field and will likely fuel further exploration of genome evolution, vaccine candidates and hypnozoite biology alike.

“Major” comments:

- It would be important to clarify what is the relevance for the use of the *P. vivax* reference genome during the assembly. I.e. are there contigs which are purely linked based on their assumed synteny to *P. vivax*. If so how many such “connections” are present? Also a more detailed description of the manual annotation would be appreciated. I.e. what sort of changes has been made and based on what kind of evidences?
- I think it would be important to clarify if the PcyB and M strains are indeed represent one and the same (in which case the PcyB/M name would be appropriate) or two closely related isolates. If it can be concluded with high confidence this information should be mention in the abstract as well.
- The potential function and origin of the methyltransferase could perhaps be better analyzed and discussed. After a quick domain search I realized that this methyltransferases also contain multiple ankyrin domains. More importantly homology search suggests that this is/was a nicotinamide N-methyltransferase and hence might also play a role in nicotinamide metabolism. Intriguingly some methyltransferases (e.g. SET8, Kishore, BMC Evol Bio, 2013<sup>1</sup>) and some members of the nicotinamide pathway (O’Hara, PLOS One, 2014<sup>2</sup>) in Plasmodia might be resulted from horizontal gene transfer. Therefore, it could be interesting to investigate if this could be the case for this particular gene as well.

Minor points:

- Supplementary table 1 is not very useful in its current form. It is rather cumbersome to select out the genes which are only present in certain species. Perhaps the authors could sort the table according to the clusters presented on figure 2.
- It is nice that the same color scheme is used throughout the manuscript, but the color of Pcy, Pc and Pk are rather similar and difficult to tell apart. In particular on Figure 4 this is problematic.
- Perhaps it would be more logic to discuss the improved subtelomeres earlier in the manuscript (i.e before the OrthoMCL clustering)
- Figure 3 B and C could be better labelled.
- The improved reference genome should be made available in GeneDB and PlasmoDB as well.
- Finally, the authors mention that genes specific to hypnozoite forming parasites mainly belong to variant multigene families and unlikely to be relevant for hypnozoite formation. Perhaps it would be worthwhile identifying genes which are not unique, but substantially different in these parasites (indels, unexpectedly high number of SNPs, etc).

### References

1. Kishore SP, Stiller JW, Deitsch KW: Horizontal gene transfer of epigenetic machinery and evolution of parasitism in the malaria parasite *Plasmodium falciparum* and other apicomplexans. *BMC Evol Biol.* 2013; **13**: 37 [PubMed Abstract](#) | [Publisher Full Text](#)
2. O'Hara JK, Kerwin LJ, Cobbold SA, Tai J, Bedell TA, Reider PJ, Llinás M: Targeting NAD<sup>+</sup> metabolism in the human malaria parasite *Plasmodium falciparum*. *PLoS One.* 2014; **9** (4): e94061 [PubMed Abstract](#) | [Publisher Full Text](#)

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Partly

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Referee Report 29 June 2017

doi:10.21956/wellcomeopenres.12821.r23884



**Aaron Jex**

Population Health and Immunity Division, Walter and Eliza Hall Institute of Medical Research, Melbourne, Vic, Australia

The current manuscript describes the resequencing and finishing to near chromosomal completion of the *Plasmodium cynomolgi* genome using high coverage PacBIO sequencing, the reannotation of the parasite's ~6,500 coding gene models and a detailed comparative analysis of its major gene families relative to related primate clade species of *Plasmodium*, including *P. vivax*, *P. ovale* and *P. knowlesi* as well as a draft assembly and annotation of *P. simiovale* and a reannotation of *P. coatneyi*. As the authors describe, *P. cynomolgi* is an important model for *P. vivax*, a relapsing human infectious species that now predominates as the primary cause of malaria in the Asia-Pacific and Americas. The authors correctly note that the current draft assembly the *P. cynomolgi* genome, which was produced several years ago, is highly fragmented, incomplete and has truncated / inadequate gene models. Though of an acceptable standard at the time, it is clear that this prior genome is no longer adequate to act as a reference of *P. cynomolgi* research. Based on this, the current manuscript is a timely contribution that will provide an excellent resource for the malaria research community. The paper itself is well written, the figures are very nicely conceived and presented and the methods used are appropriate and expertly applied. I have no hesitation in recommending this study for publication.

Minor comments:

1. 'Anophelines' shouldn't be capitalized (paragraph 1 of the introduction)
2. 'Reference gnomes' under 'Re-annotation of *P. coatneyi*' should be 'Reference genomes'
3. It would be interesting to know if any of the *P. cynomolgi* genes represented by the 103 ortholog clusters unique to *P. cynomolgi*, *P. ovale* and *P. vivax* are represented in the recent liver-stage transcriptome (particularly the hypnozoite transcriptome) published by Cubi *et al* (<https://www.ncbi.nlm.nih.gov/pubmed/28256794>). I wonder if this might be considered by the authors as a minor addition?
4. Could the authors please provide a more detailed explanation either in their methods section or in the results section for how they define the expanded methyltransferases they identify in *P. cynomolgi* as pseudogenes? I wonder if this will otherwise not be immediately clear to the reader.
5. Under the 'Paralogous expansion of ... MSP' section - 'while in *P. malaria* and to *P. ovale*' - I assume 'to' should be deleted here.
6. In Figure 3C, should this be 'expansion in msp3-like'? If not, could the authors explain what they mean by 'expansion like'?

**Is the work clearly and accurately presented and does it cite the current literature?**



Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

***Competing Interests:*** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---