DATA NOTE

A new *Plasmodium vivax* reference sequence with improved assembly of the subtelomeres reveals an abundance of *pir* genes [version 1; peer review: 2 approved]

Sarah Auburn¹, Ulrike Böhme², Sascha Steinbiss², Hidayat Trimarsanto³, Jessica Hostetler²,⁴, Mandy J. Sanders², Qi Gao⁵, François Nosten⁶,⁷, Chris I. Newbold²,⁸, Matt Berriman², Ric N. Price¹,⁶, Thomas D. Otto²

¹Global and Tropical Health Division, Menzies School of Health Research and Charles Darwin University, Darwin, Australia
²Malaria Programme, Wellcome Trust Sanger Institute, Hinxton, UK
³Eijkman Institute for Molecular Biology, Jakarta, Indonesia
⁴Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, USA
⁵Jiangsu Institute of Parasitic Diseases, Key Laboratory of Parasitic Disease Control and Prevention (Ministry of Health), Jiangsu Provincial Key Laboratory of Parasite Molecular Biology, Jiangsu, China
⁶Centre for Tropical Medicine and Global Health, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK
⁷Shoklo Malaria Research Unit, Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Mae Sot, Thailand
⁸Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK

Abstract

*Plasmodium vivax* is now the predominant cause of malaria in the Asia-Pacific, South America and Horn of Africa. Laboratory studies of this species are constrained by the inability to maintain the parasite in continuous *ex vivo* culture, but genomic approaches provide an alternative and complementary avenue to investigate the parasite's biology and epidemiology. To date, molecular studies of *P. vivax* have relied on the Salvador-I reference genome sequence, derived from a monkey-adapted strain from South America. However, the Salvador-I reference remains highly fragmented with over 2500 unassembled scaffolds. Using high-depth Illumina sequence data, we assembled and annotated a new reference sequence, PvP01, sourced directly from a patient from Papua Indonesia. Draft assemblies of isolates from China (PvC01) and Thailand (PvT01) were also prepared for comparative purposes. The quality of the PvP01 assembly is improved greatly over Salvador-I, with fragmentation reduced to 226 scaffolds. Detailed manual curation has ensured highly comprehensive annotation, with functions attributed to 58% core genes in PvP01 versus 38% in Salvador-I. The assemblies of PvP01, PvC01 and PvT01 are larger than that of Salvador-I (28-30 versus 27 Mb), owing to improved assembly of the subtelomeres. An extensive repertoire of
over 1200 *Plasmodium* interspersed repeat (*pir*) genes were identified in PvP01 compared to 346 in Salvador-I, suggesting a vital role in parasite survival or development. The manually curated PvP01 reference and PvC01 and PvT01 draft assemblies are important new resources to study vivax malaria. PvP01 is maintained at GeneDB and ongoing curation will ensure continual improvements in assembly and annotation quality.

**Keywords**
Plasmodium vivax, genome, reference, subtelomere, pir, vir

This article is included in the Wellcome Sanger Institute gateway.

This article is included in the Mahidol Oxford Tropical Medicine Research Unit (MORU) gateway.

This article is included in the Wellcome Centre for Integrative Parasitology gateway.

**Corresponding author:** Thomas D. Otto (tdo@sanger.ac.uk)
**Competing interests:** No competing interests were disclosed.
**Grant information:** This work was supported by Wellcome Trust [098051], [099198], [091625].
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2016 Auburn S et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

**How to cite this article:** 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; peer review: 2 approved] Wellcome Open Research 2016, 1:4 https://doi.org/10.12688/wellcomeopenres.9876.1

**First published:** 15 Nov 2016, 1:4 https://doi.org/10.12688/wellcomeopenres.9876.1
Introduction

Infection with Plasmodium vivax is associated with significant direct and indirect morbidity that impacts on the poorest communities of malarious countries, with an estimated annual global cost of $1-2.7 billion\(^1\). Accumulating reports of drug-resistant infection and life-threatening disease underscore the urgency to reduce the burden of P. vivax and ensure its ultimate elimination\(^2\). Efforts to contain P. vivax are constrained by a limited understanding of the parasite’s basic biology, in part owing to the inability to maintain this species in continuous \emph{ex vivo} culture. Genetic studies provide an alternative approach to gain novel insights into the parasite from which epidemiological tools and therapeutic approaches can be developed for clinical application\(^3\)-\(^6\). The rapidly declining costs of massively parallel sequencing technologies have made it feasible to undertake whole genome sequencing of hundreds of \emph{Plasmodium} isolates, with recent population genomic studies of \emph{P. vivax} revealing novel antimalarial drug resistance and vaccine candidates amongst other biological features of the parasite\(^6\)-\(^9\). However, in order to achieve a comprehensive understanding of the structure and composition of the \emph{P. vivax} genome, and to improve read mapping efforts to characterise genetic polymorphisms, a high quality reference genome(s) representative of naturally occurring patient isolates is essential.

The sequences of 5 monkey-adapted strains including the Salvador-I reference\(^10\) and drafts of Brazil-I, India-VII, North Korea and Mauritania\(^1\) have provided important resources for the \emph{vivax} research community to investigate the core genome of \emph{P. vivax}. However, over 60% of the genes in the published Salvador-I reference\(^1\) (prior to curation by the authors) had unknown function, limiting insight into underlying biological mechanisms. Furthermore, assembly of the subtelomeric regions is highly fragmented in these strains, with Salvador-I comprising >2500 scaffolds. A subsequent draft assembly of a Cambodian patient isolate (C127) revealed 792 genes not present in Salvador-I, including 366 new \textit{pir} (\emph{Plasmodium} interspersed repeat) genes\(^11\). The \textit{pir} genes are a highly variable multigene family present in all \emph{Plasmodium} genomes investigated to date\(^12\). The function of \textit{pir}-encoded proteins (PIRs) remains poorly understood, although recent studies suggest roles in mechanisms associated with virulence. In \emph{vitro} studies of \emph{P. vivax} have demonstrated \textit{pir} encoded protein mediated cytoadherence to endothelial cells\(^9,10\) and a \emph{P. chabaudi} mouse malaria model demonstrated red blood cell-binding properties consistent with roles in invasion and/or rosette formation\(^11\). A further \emph{P. chabaudi} study demonstrated that changes in the expression of the \textit{pir} gene repertoire following mosquito passage may attenuate virulence\(^2\). The sequence diversity amongst the \textit{pir} genes in \emph{P. vivax} suggests that different subfamilies may have different functions\(^14\). The published Salvador-I reference sequence revealed 346 \textit{pir} genes, including 80 fragments and/or pseudogenes, 10 subfamilies and 84 unassigned genes\(^14\). In the most recent computational classification, Lopez et al. re-classified the Salvador-I \textit{pir} genes, excluding members of 3 major subfamilies (A, D and H) but including previously unassigned genes, and re-defining 39 genes as encoding PIRs rather than hypothetical proteins\(^15\). However, given the limited number of PIRs in Salvador-I, further characterisation is required using a reference(s) with a more complete set of genes.

To address the need of the \emph{vivax} research community for a \emph{P. vivax} reference with more comprehensive assembly and annotation, we used Illumina genomic data to establish a reference from a Papua Indonesian patient isolate (PvP01). Since \emph{P. vivax} exhibits marked regional variation in phenotypes such as duration of the dormant liver-stage, drug resistance and disease severity, we compared PvP01 to C127 and the 5 monkey-adapted strains, and generated draft assemblies of patient isolates from Thailand (PvT01) and central China (PvC01). Our sampling focuses on the Asia-Pacific region, where a large burden of \emph{P. vivax} infection lies\(^16\). The Indonesian reference provides representation of the island of Papua - the epicentre of multidrug resistance emergence in \emph{P. vivax}\(^3\). The draft references from Thailand and Central China provide respective representation of the Mekong region, and the temperate north where long latency phenotypes prevail\(^15\).

Methods

Samples

Three \emph{P. vivax} field isolates that were judged to be clonal infections following preliminary genomic analysis within the framework of a separate study\(^17\) were selected for assembly. The isolates were sourced from a patient presenting at hospital in northern Australia in December 2012 with a recent travel history to Mimika Regency, Papua Indonesia (strain PvP01), and patients presenting with symptomatic infection to local clinics in Nan Province, Thailand in May 2011 (strain PvT01) and Anhui Province, China, in September 2010 (strain PvC01). Patient blood samples were leukodepleted\(^18\), and DNA extracted using the QIAamp blood midi kit (Qiagen). All samples were collected with written informed consent from the patients within the framework of previous studies.

Ethical approval

Ethical approval was provided by the Human Research Ethics Committee of NT Department of Health and Families and Menzies School of Health Research, Darwin, Australia (HREC-09/83), the Mahidol University Faculty of Medical Technology Ethics Committee, Bangkok, Thailand (MUTM 2011-043-03), and the Institutional Review Board of Jiangsu Institute of Parasitic Diseases, Wuxi, China (IRB00004221).

Sequencing, assembly and annotation

Library preparation and sequencing was performed at the Wellcome Trust Sanger Institute. Genomic DNA was sheared into 300–500 base pair (bp) fragments using ultrasonication (Covaris). Amplification-free Illumina libraries were prepared\(^17\) and 75 bp, 100 bp and 250 bp paired end reads were generated on the Illumina GAIL, Hi-Seq 2000 v3 and MiSeq platforms respectively, following the manufacturer’s standard cluster generation and sequencing protocols\(^19\). Mate-pair libraries with 2–3 kilobase (kb) inserts were additionally prepared for PvP01 and PvT01, using the Illumina mate-pair library preparation kit (v2), and sequenced on the Illumina HiSeq 2500 platform. Prior to assembly, contaminating host-derived sequences were excluded by mapping against the human reference genome (GRCh37: ftp://ftp1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/) using BWA\(^20\) (version 0.7.4). Assemblies were prepared using velvet (version 1.2.07, parameters: -exp_cov auto -ins_length 450 -ins_length_sd 30-cov_cutoff 8, and using for a kmer of 71) and MaSuRCA\(^21\) (version 2.0.3.1, default
parameters). Post-assembly genome improvements were undertaken using a range of automated configuration tools including ABACAS (version 2), IMAGE (version 2, iterating k-mers from 71 down 31, 7 iterations), Gapfiller (version 1–11, 14 iteration, parameter n=31) and iCORN (version 2, 7 iterations). PAGIT (version 1) and REAPR (version 1.0.17) were employed to detect assembly errors. This was followed by visual inspection using ACT to identify any further assembly anomalies. Annotation was undertaken initially using the automated algorithms, RATT (version 1) and Augustus (version 2.7, trained on 500 manually curated gene models) and further improved by detailed manual inspection performed by an experienced genome curator. PvT01 and PvC01 were annotated using Companion, a new automated annotation tool. RNA-Seq data from asexual blood stage preparations of 4 P. vivax patient isolates from Cambodia (unpublished report, Jessica Hostetler, Lia Chappell, Chanaki Amaratunga, Seila Suon, Thomas D. Otto, Rick Fairhurst and Julian C. Rayner; Accession number ERP017542) was used as supporting evidence to aid the improvement of gene models in PvP01 by manual curation.

For comparative analyses, genome assemblies and gene annotations were sourced for 6 additional P. vivax strains; Salvador-I, C127, Brazil-I, India-VII, Mauritania-I and North Korea. The published version of Salvador-I presented in PlasmoDB release 9 was selected for comparison of gene annotations as the additional improvements in release 10 reflected curations performed by the authors. Companion was also used to update the annotation of the four previously published genomes (Brazil-I, India-VII, Mauritania-I and North Korea).

OrthoMCL and pir analysis
Comparisons of predicted protein-coding genes between the 9 P. vivax assemblies and P. falciparum 3D7 (Pf3D7) (geneDB.org) were undertaken using OrthoMCL version 1.4 using the default parameter settings. We determined core genes as 1-1 orthologous between P. vivax P01 and Pf3D7, in total 4465.

Cluster analysis based on structural and sequence homology was undertaken to compare the subfamily organization of the pirs in the partial (Salvador-I) versus more complete (PvP01) reference. All PIR encoded protein sequences in Salvador-I and PvP01 with length greater than 150 amino acids and not flagged as pseudogenes were included in the analysis. Low complexity regions were excluded using the SEG program. The relatedness between sequences was assessed using BLASTp (parameters -F F -e 1e-6), and the results were visualized as a network constructed in Gephi. After provisional assessment of cluster resolution at different thresholds, a cutoff of 25% of the global similarity was selected for distinguishing different clusters (subfamilies). To aid comparison against the new PIRs identified in PvP01, the Salvador-I PIRs were colour-coded according to the subfamily classification proposed by Lopez et al.

Further investigation of the diversity and relatedness amongst the PIRs was undertaken using the PIR sets from PvP01, PvT01, PvC01, Salvador-I and Brazil-I. Exclusion of proteins with less than 150 amino acids, filtering of low complexity sequences and relatedness analysis using BLASTp were performed as described above. A network was constructed from the BLAST output using tribeMCL with an inflation of 1.5. To aid visualization, clusters with less than 15 PIRs were excluded.

Dataset validation
The PvP01 assembly was generated as a new reference sequence and is thus a higher quality, more accurately annotated assembly than PvC01 and PvT01, which were both created as draft assemblies for comparative purposes. The PvP01 assembly quality is greatly improved over the previous Salvador-I reference genome, with fragmentation reduced to <250 scaffolds amongst other features (Table 1). At 29 megabases (Mb), the assembly is notably larger than Salvador-I (27 Mb), mainly due to newly assembled subtelomeric sequences. A complete mitochondrial sequence (5 kb) and partial apicoplast sequence (29.6 kb) are also available. As in P. falciparum, the apicoplast reference will facilitate efforts to identify geographic surveillance markers for P. vivax.

<table>
<thead>
<tr>
<th>Genome features</th>
<th>PvP01</th>
<th>PvC01</th>
<th>PvT01</th>
<th>Salvador-I</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nuclear genome</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assembly size (Mb)</td>
<td>29.0</td>
<td>30.2</td>
<td>28.9</td>
<td>26.8</td>
</tr>
<tr>
<td>Coverage (fold)</td>
<td>212</td>
<td>56</td>
<td>89</td>
<td>10</td>
</tr>
<tr>
<td>G + C content (%)</td>
<td>39.8</td>
<td>39.2</td>
<td>39.7</td>
<td>42.3</td>
</tr>
<tr>
<td>No. scaffolds assigned to chrom.</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>No. unassigned scaffolds</td>
<td>226</td>
<td>529</td>
<td>359</td>
<td>2745</td>
</tr>
<tr>
<td>No. genes a</td>
<td>6,642</td>
<td>6,690</td>
<td>6,464</td>
<td>5,433</td>
</tr>
<tr>
<td>No. pir genes</td>
<td>1,212</td>
<td>1,061</td>
<td>867</td>
<td>346</td>
</tr>
<tr>
<td><strong>Mitochondrial genome</strong> d</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assembly size (bp)</td>
<td>5,989</td>
<td>-</td>
<td>-</td>
<td>5,990</td>
</tr>
<tr>
<td>G + C content (%)</td>
<td>30.5</td>
<td>-</td>
<td>-</td>
<td>30.5</td>
</tr>
<tr>
<td><strong>Apicoplast genome</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assembly size (kb)</td>
<td>29.6</td>
<td>27.6</td>
<td>6.6</td>
<td>5.1</td>
</tr>
<tr>
<td>G + C content (%)</td>
<td>13.3</td>
<td>12.7</td>
<td>19.7</td>
<td>17.1</td>
</tr>
<tr>
<td>No. genes</td>
<td>30</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

a Genome version 1.09.2016
b Published reference sequence
c Including pseudogenes and partial genes, excluding non-coding RNA genes.
d T. vivax chloroplast genome is not present in PvP01 and PvC01.

Further investigation of the diversity and relatedness amongst the PIRs was undertaken using the PIR sets from PvP01, PvT01, PvC01, Salvador-I and Brazil-I. Exclusion of proteins with less than 150 amino acids, filtering of low complexity sequences and relatedness analysis using BLASTp were performed as described above. A network was constructed from the BLAST output using tribeMCL with an inflation of 1.5. To aid visualization, clusters with less than 15 PIRs were excluded.
Whilst the assembly quality in the core region is high in Salvador-I\textsuperscript{14}, PvP01 displays improved gene models and has more complete subtelomeres. Figure 1 provides a schematic of the right-hand end of chromosome 12 from PvP01 and Salvador-I, illustrating the generally greater extension into the subtelomeric regions of chromosomes in PvP01. Furthermore, owing to detailed manual curation and continuous maintenance within the GeneDB framework, the level of gene annotation in the core genome of PvP01 greatly exceeds that of the other available \textit{P. vivax} assemblies. The asexual stage \textit{P. vivax} RNA-Seq data enabled correction of the structure of 377 genes. Of the 4577 core \textit{P. vivax} genes with 1:1 orthologues in \textit{P. falciparum}, 3318 genes were transcribed with RPKM (reads per kilobase of transcript per million mapped reads) values greater than 15, and contained a total of 4887 splice sites. Of these splice sites, a total of 4845 (99.1\%) were confirmed by \( \geq 10 \) reads, highlighting the high quality of the structural annotation. Whereas the published Salvador-I reference includes functions attributed to a total of 1783 (38.0\%) core genes\textsuperscript{14}, we have been able to expand this to 2848 (58.6\%) in PvP01, as of the latest GeneDB release (1st September 2016). Ongoing curation on PvP01 will yield further improvements to the annotation statistics, and progress is highlighted in Table 2, which summarizes annotation changes over a 12 month period between GeneDB releases in 2015 and 2016. To date, a total of 1209 genes have been identified in PvP01 that were either completely absent from Salvador-I or have arisen by splitting gene structures that were falsely joined previously (Table 1). Although the majority of newly identified genes belong to subtelomeric gene families, we confirmed the recently identified EBP2 (erythrocyte binding protein 2, PVP01\textsubscript{0102300}) and RBP2e (reticulocyte binding protein 2e, PVP01\textsubscript{0700500}) genes\textsuperscript{41}. These genes are members of families encoding proteins implicated in host cell recognition during red blood cell (RBC) invasion, and present potential vaccine targets\textsuperscript{48–51}.

As summarised in Table 3, the comparatively high assembly quality in the subtelomeres of PvP01 greatly expanded the repertoire of genes belonging to multigene families in these chromosome regions. Notably, more than 1200 pir genes were identified in PvP01 versus 346 in Salvador-I. To generate a snapshot of the diversity and structural organization of this expanded gene family in \textit{P. vivax}, we conducted cluster analysis of the PIRs in PvP01 with comparison to previous homology classifications performed by Lopez \textit{et al} on the partial set of PIRs from Salvador-I\textsuperscript{23}. As illustrated in the network diagram in Figure 2a, the main subfamily clusters defined in earlier classifications are expanded but, on addition of the new PvP01 PIRs, the clusters remained moderately stable with no pooling between or sub-structure within subfamilies. However, the new PvP01 PIRs reveal several large subfamilies containing just 1–4 Salvador-I genes that were

\begin{table}
\centering
\caption{Annotation changes in \textit{P. vivax} P01 from 1\textsuperscript{st} of September 2015 until 27\textsuperscript{th} of September 2016.}
\begin{tabular}{|l|l|}
\hline
Annotation event type & PvP01\textsuperscript{*} \\
\hline
Assigned or updated product & 408 \\
Product updated from “conserved Plasmodium protein, unknown function” & 107 \\
Updated GO term & 597 \\
Linked to publication & 291 \\
All unique genes with new functional annotations, e.g. EC number, gene name & 608 \\
All unique genes with new structural annotations & 50 \\
\hline
\end{tabular}
\textsuperscript{*}Genome version 1.09.2016
\end{table}

\textbf{Figure 1.} Organization of the subtelomeric regions of chromosome 12 of the PvP01 and Salvador-I \textit{P. vivax} references illustrating the higher assembly quality of PvP01. The order and orientation of the genes in the 3’ subtelomeric region of chromosomes 12 of PvP01 (top) and Salvador-I (bottom) are shown. Exons are shown in coloured boxes, with introns illustrated by linking lines. Gaps in PvP01 are indicated with a forward slash (“/”). The blue box indicates the start of the telomeric heptamer repeats. The shaded (grey) areas mark the start of the conserved core of the chromosome that shares synteny with other \textit{Plasmodium} species (e.g. \textit{P. falciparum}). The black box shows the syntenic area of PvP01 and Salvador-I. The last gene in this syntenic area is fragmented in Salvador-I.
Table 3. Number of most abundant genes in the subtelomeres in the genomes of Salvador-I, PvP01, PvT01 and PvC01.

<table>
<thead>
<tr>
<th>Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multigene family</td>
<td></td>
</tr>
<tr>
<td>PIR protein</td>
<td>346</td>
</tr>
<tr>
<td>tryptophan-rich protein</td>
<td>34</td>
</tr>
<tr>
<td>lysophospholipase</td>
<td>11</td>
</tr>
<tr>
<td>STP1 protein</td>
<td>9</td>
</tr>
<tr>
<td>early transcribed membrane protein (ETRAMP)</td>
<td>10</td>
</tr>
<tr>
<td>Plasmodium exported protein (PHIST), unknown function</td>
<td>64</td>
</tr>
<tr>
<td>reticulocyte binding protein (RBP)</td>
<td>9</td>
</tr>
<tr>
<td>Other genes</td>
<td></td>
</tr>
<tr>
<td>Plasmodium exported proteins of unknown function</td>
<td>23</td>
</tr>
<tr>
<td>Total</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Numbers include pseudogenes and partial genes

a Published reference sequence

b Genome version 1.09.2016

c Other names include VIR protein and Pv-fam-c protein

d Other names include Pv-fam-a, trag and tryptophan-rich antigen

e Other names include PST-A protein

f Other names include PvSTP1

g Other names include Phist protein (Pf-fam-b) and RAD protein (Pv-fam-e)
h Includes RBP2e (PVP01_0700500) that was not present in the Salvador-I assembly. RBP1b (PVP01_0701100) is complete in PvP01. In Salvador-I RBP1b consists of two partial genes (PVX_098582, PVX_125738)
i Other names include Pv-fam-d protein and Pv-fam-c protein

previously unclassified (Figure 2a). Additional investigation with the PvC01, PvT01 and Brazil-I assemblies using tribeMCL (also used in Lopez et al) confirmed the stability of the new subfamilies identified in PvP01 across a geographically divergent collection of isolates (Figure 2b). The analysis conducted here provides a broad overview of the diversity and relatedness amongst the expanded P. vivax pir gene sets, however further investigation beyond the scope of this study will be required to provide detailed characterisation of this family and its contribution to virulence and pathophysiology.

The PvP01 reference is an important new resource for the vivax research community. It will support studies of the complex subtelomeric regions and provide insights into the mechanisms by which the gene families in this region contribute to virulence-associated functions. It will also allow investigation of an array of other biological functions that will expand with continual improvements in annotation in the core genome. PvP01, PvC01 and PvT01 add new geographic locations to the collection of P. vivax assemblies, facilitating biological studies of the diversity of this phenotypically divergent species.

Data availability

The raw sequence data for PvP01, PvT01 and PvC01 can be retrieved from the European Nucleotide Archive; sample accession numbers PvP01 ERS017708, ERS312161 3kb ERS328510, PvT01 ERS055881, ERS312160 3kb ERS328509 and PvC01 ERS407449. The assemblies can be found under the study PRJEB14589. The individual accession numbers are PvP01 (chromosomes: currently in submission to EBI, files on ftp, contigs: FLZR01000001-FLZR01000226), PvT01 (chromosomes LT615239-LT615252, contigs: FLYH01000001-FLYH01000360) and PvC01 (chromosomes LT615256-LT615269, contigs: FLYI01000001-FLYI01000530). PvP01 is maintained in GeneDB: http://www.genedb.org/Homepage/PvivaxP01 and updates are synchronized to PlasmoDB.
Figure 2. Cluster analysis illustrating the relatedness between the PIR proteins in PvP01 versus Salvador-I (a), and the stability of the major clusters in several other P. vivax assemblies (b). Panel a) presents a network illustrating the relatedness between the 1063 PIR proteins of PvP01 and 341 PIRs of Salvador-I (Sal-I) with length greater than 150 amino acids. The PvP01 PIRs are illustrated by black dots (nodes). The Sal-I PIRs are illustrated by coloured dots with colour-coding according to the subfamily classification of Lopez et al\textsuperscript{23} as follows; purple = A, pink = B, pale green = C, red = D, pale blue = E, orange = G, green = H, blue = I, white = J, yellow = K, and grey = unassigned genes. Two nodes (PIRs) are connected if they have a global similarity of at least 25%. With the exception of a few proteins, the majority of Sal-I PIRs demonstrate clustering consistent with the classification of Lopez et al. Five new, interconnected clusters comprising previously unassigned Sal-I PIRs are denoted with a white “X”. In Panel b, a heat map summarises the number of PIRs assigned to the 27 major clusters (minimum 15 PIRs in total) in five geographically divergent P. vivax strains; PvP01 (Papua Indonesia), PvT01 (Thailand), PvC01 (Central China), Sal-I (El Salvador) and Brazil-I (Brazil). With the exception of Sal-I, which displayed fewer genes than the other isolates in several of the major clusters, the isolates demonstrated similar numbers of genes in most clusters.
This section will be updated with accession numbers for PvP01 chromosomes once available.

Author contributions
SA, CIN, MB, RNP and TDO conceived the study. QG and FN provided essential resources for the data generation. MS managed the sequencing of the samples. SA, HT and TDO performed analyses. SS performed automated annotation and UB is maintaining the manual annotation and generated statistics on the annotation. JH generated the RNA-Seq data. SA and TDO prepared the first draft of the manuscript. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

Competing interests
No competing interests were disclosed.

Grant information
This work was supported by Wellcome Trust [098051], [099198], [091625].

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

Acknowledgments
We would like to thank the patients who contributed samples and the health workers who assisted with the sample collections as well as Julian Rayner, Rick Fairhurst, Chanaki Amaratunga, Lia Chappell and Seila Suon for use of unpublished P. vivax RNA-Seq data. We would also like to thank staff from the Illumina Bespoke Sequencing Team at the Wellcome Trust Sanger Institute for their contribution.

References


This work describes de novo assemblies of three new P. vivax genomes and comparison with the reference Sal I genome and other Pv genomes. Compared with previous annotation of the reference genome, the new assembly of the PvP01 genome for an isolate from Papua Indonesia has reduced the total scaffolds from over 2500 in SalI to 226 (+14). Major improvements are in the subtelomeric regions, where a significantly increased number of pir genes have been discovered. This more in-depth study of the Pv genome and manual curation of genes provide a better resource for biological studies of the vivax parasite.

Comments:
1. Abstract: The quality of the PvP01 assembly is improved greatly over Salvador-I, with fragmentation reduced to 226 scaffolds. Perhaps “with fragmentation reduced to 226 unassigned scaffolds in addition to the 14 chromosomal scaffolds” will be more accurate?

2. Does the “results” section begin at “Dataset validation’’?

3. Table 1 presented comparison of the genomes the three new sequences with that of Sal I. The PvC01 and PvT01 sequences contained more assigned scaffolds – are these located mostly in the telomeric regions?

4. A more detailed comparison of the temperate strain PvC01 with the tropical strains would be more useful. A big-picture type perspective on the C01 and T01 would be nice.

5. Figure 1 illustrates the extension of the assembled sequences in the subtelomeric region of chrom12 as compared to that in SalI. Are the gap junctions verified by PCR? Also, the PvP01 also has quite some gaps – how are these assembled and verified?

6. The network presentation of the Pir genes is interesting – A link to the alignment of the sequences or a phylogenetic tree-type of presentation (as supplements) would be very useful.
Competing Interests: No competing interests were disclosed.

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

Reviewer Report 12 December 2016

https://doi.org/10.21956/wellcomeopenres.10647.r18268

© 2016 Bártfai R. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Richárd Bártfai
Department of Molecular Biology, Radboud University Nijmegen, Nijmegen, The Netherlands

In this manuscript, Auburn and colleagues describes the generation and initial analysis of a new *P. vivax* reference sequence. The authors extensively sequenced a single-clone field isolate from Papua Indonesia using Illumina technology. Assembly of these sequence reads lead to a much less fragmented and better covered core genome sequence. Furthermore, the improved assembly and annotation resulted in a much more complete overview on the subtelomeric multi gene families in this and two other isolates from China and Thailand.

Reference genomes are the foundation for all genomic, transcriptomic and proteomic studies. Therefore, this new reference genome sequence is very welcome and will undoubtedly fuel the exploration of the biology and pathogenesis of *P. vivax*. While it is always difficult to access the quality of such assemblies based on description only it is conceivable that the 20x increase in coverage and the use of various post-assembly improvement tools have resulted in considerably better genome sequence. Furthermore, the manual curated gene models and functional classifications bring substantial added value to this work.

Overall this study is well executed and the manuscript is well-written. I have only some minor suggestions for improvement:

- It would be important to clarify in the manuscript why PvP01 has been chosen to be the new reference “strain”.

- Sequencing and annotation of multigene families is challenging. To fully exclude the possibility that the 5 new clusters of PIR proteins identified in this study are the result of incorrect sequence assembly it would be relevant to PCR amplify and sequence a representative member from each of these families.

- In the abstract the authors state that the new reference genome contains 226 scaffolds, while according to table 1 it appears to be 226+14. Please double-check.

- I do not find Table 2 particularly useful/informative. It is basically a tribute to a huge amount of work.
It might not be formally required to include a subheading “Results” in Welcome Open Research data notes, but nonetheless it would be nice to know where the description of the results begins.

- It would bring added value to this article if Table 3 would be extended by description of all and not only the subtelomeric gene families (Table 2 in Tachibana et al., 20121 could provide a nice example). Instead of the extensive footnotes an extra column could be included for alternative names.

- In Table 3 it is unclear if other genes includes only “Plasmodium exported protein of unknown function” or also other proteins. If there are indeed couple of hundred of these proteins encoded in the PvP01 genome and they localize to the subtelomeric regions as Figure 1 suggests, it would be perhaps relevant to discuss them as a gene family. It could even be worthwhile to perform a cluster analysis on this “gene family” similar to the one performed on PIR proteins.

- On Figure 2B it would be useful to indicate the correspondence between the cluster numbers of this study and the former classification (A-K). Similarly it would be informative to indicate the cluster numbers on Figure 2A.

- From Figure 2B it seems that cluster 5 PIR gene subfamily has expanded (substantially more numerous) in the Brazilian isolate. Something perhaps worthwhile mentioning/discussing.

References

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

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