DATA NOTE

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

Sarah Auburn[1], Ulrike Böhme[2], Sascha Steinbiss[2], Hidayat Trimarsanto[3], Jessica Hostetler[2,4], Mandy Sanders[2], Qi Gao[5], Francois Nosten[6,7], Chris I. Newbold[2,8], Matthew Berriman[2], Ric N. Price[1,7], Thomas D. Otto[2]

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

v1 **First published:** 15 Nov 2016, **1**:4 (doi: 10.12688/wellcomeopenres.9876.1)
Latest published: 15 Nov 2016, **1**:4 (doi: 10.12688/wellcomeopenres.9876.1)

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.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
version 1 published 15 Nov 2016	✓ report	✓ report

1 **Richárd Bártfai**, Radboud University Nijmegen Netherlands

2 **Liwang Cui**, Pennsylvania State University USA

Discuss this article

Comments (0)

Corresponding author: Thomas D. Otto (tdo@sanger.ac.uk)

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; referees: 2 approved]** Wellcome Open Research 2016, **1**:4 (doi: 10.12688/wellcomeopenres.9876.1)

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.

Competing interests: No competing interests were disclosed.

First published: 15 Nov 2016, **1**:4 (doi: 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[1–3]. 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[4–8]. 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 *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[9–17]. The rapidly declining costs of massively parallel sequencing technologies have made it feasible to undertake whole genome sequencing of hundreds of *Plasmodium* isolates, with recent population genomic studies of *P. vivax* revealing novel antimalarial drug resistance and vaccine candidates amongst other biological features of the parasite[16,17]. However, in order to achieve a comprehensive understanding of the structure and composition of the *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[14] and drafts of Brazil-I, India-VII, North Korea and Mauritania-I[13] have provided important resources for the vivax research community to investigate the core genome of *P. vivax*. However, over 60% of the genes in the published Salvador-I reference[14] (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 *pir* (*Plasmodium* interspersed repeat) genes[11]. The *pir* genes are a highly variable multigene family present in all *Plasmodium* genomes investigated to date[18]. The function of *pir*-encoded proteins (PIRs) remains poorly understood, although recent studies suggest roles in mechanisms associated with virulence. *In vitro* studies of *P. vivax* have demonstrated PIR encoded protein mediated cytoadherence to endothelial cells[19,20] and a *P. chabaudi* mouse malaria model demonstrated red blood cell-binding properties consistent with roles in invasion and/or rosette formation[21]. A further *P. chabaudi* study demonstrated that changes in the expression of the *pir* gene repertoire following mosquito passage may attenuate virulence[22]. The sequence diversity amongst the *pir* genes in *P. vivax* suggests that different subfamilies may have different functions[14]. The published Salvador-I reference sequence revealed 346 *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 *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[23]. 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 vivax research community for a *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 *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 *P. vivax* infection lies[24]. The Indonesian reference provides representation of the island of Papua - the epicentre of multidrug resistance emergence in *P. vivax*[8]. The draft references from Thailand and Central China provide respective representation of the Mekong region, and the temperate north where long latency phenotypes prevail[25].

Methods

Samples

Three *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[26], 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[27] and 75 bp, 100 bp and 250 bp paired end reads were generated on the Illumina GAII, Hi-Seq 2000 v3 and MiSeq platforms respectively, following the manufacturer's standard cluster generation and sequencing protocols[28]. 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://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/) using BWA[29] (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[30,31] (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[32–38]. This was followed by visual inspection using ACT[39] 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)[38,40,41] 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[42]. 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[9,13,14]. The published version of Salvador-I[14] 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 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[43] 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[44]. 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[45]. After provisional assessment of cluster resolution at different thresholds, a cut-off 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*[23].

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[46]. 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*[47], the apicoplast reference will facilitate efforts to identify geographic surveillance markers for *P. vivax*.

Table 1. Features of the new *P. vivax* assemblies against Salvador-I.

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

[a] Genome version 1.09.2016

[b] Published reference sequence[14]

[c] Including pseudogenes and partial genes, excluding non-coding RNA genes.

[d] Mitochdondrial genome is not present in PvT01 and PvC01

[e] scaffold PvC01_00_191

[f] scaffold PvT01_00_162

[g] Partial apicoplast sequence of Salavador-I reference assembly has been published (scaffolds AAKM01000417, AAKM01000371)

Whilst the assembly quality in the core region is high in Salvador-I[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 *P. vivax* assemblies. The asexual stage *P. vivax* RNA-Seq data enabled correction of the structure of 377 genes. Of the 4577 core *P. vivax* genes with 1:1 orthologues in *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 ≥ 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[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_0102300) and RBP2e (reticulocyte binding protein 2e, PVP01_0700500) genes[11]. These genes are members of families encoding proteins implicated in host cell recognition during red blood cell (RBC) invasion, and present potential vaccine targets[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 *P. vivax*, we conducted cluster analysis of the PIRs in PvP01 with comparison to previous homology classifications performed by Lopez *et al* on the partial set of PIRs from Salvador-I[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

Table 2. Annotation changes in *P. vivax* P01 from 1st of September 2015 until 27th of September 2016.

Annotation event type	PvP01[a]
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

[a] Genome version 1.09.2016

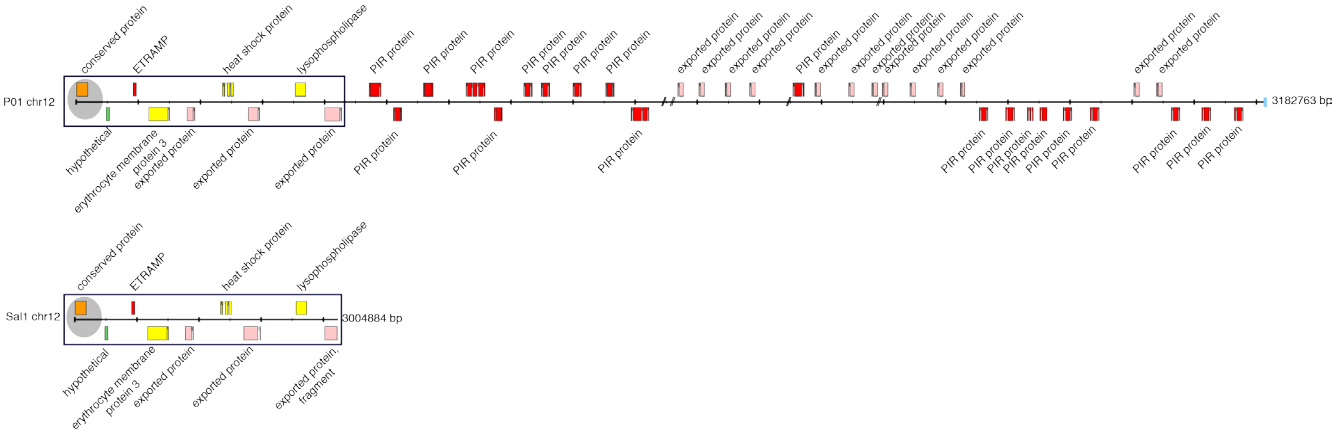


Figure 1. Organization of the subtelomeric regions of chromosome 12 of the PvP01 and Salvador-I *P. vivax* references illustrating the higher assembly quality of PvP01. The order and orientation of the genes in the 3' subtelomeric region of chromosomes 12 of PvP01 (top) and Salvador-I (bottom) are shown. 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 *Plasmodium* species (e.g. *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.

	Description	Sal-I [a]	PvP01 [b]	PvC01	PvT01
Multigene family	PIR protein [c]	346	1212	1061	867
	tryptophan-rich protein [d]	34	40	40	40
	lysophospholipase[e]	11	10	9	8
	STP1 protein [f]	9	10	11	3
	early transcribed membrane protein (ETRAMP)	10	9	9	9
	Plasmodium exported protein (PHIST), unknown function [g]	64	84	22	23
	reticulocyte binding protein (RBP)	9 [h]	9 [h]	9	8
Other genes	Plasmodium exported proteins of unknown function [i]	23	447	266	261
Total	n/a	497	1812	1427	1219

Numbers include pseudogenes and partial genes

[a] Published reference sequence[14]

[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

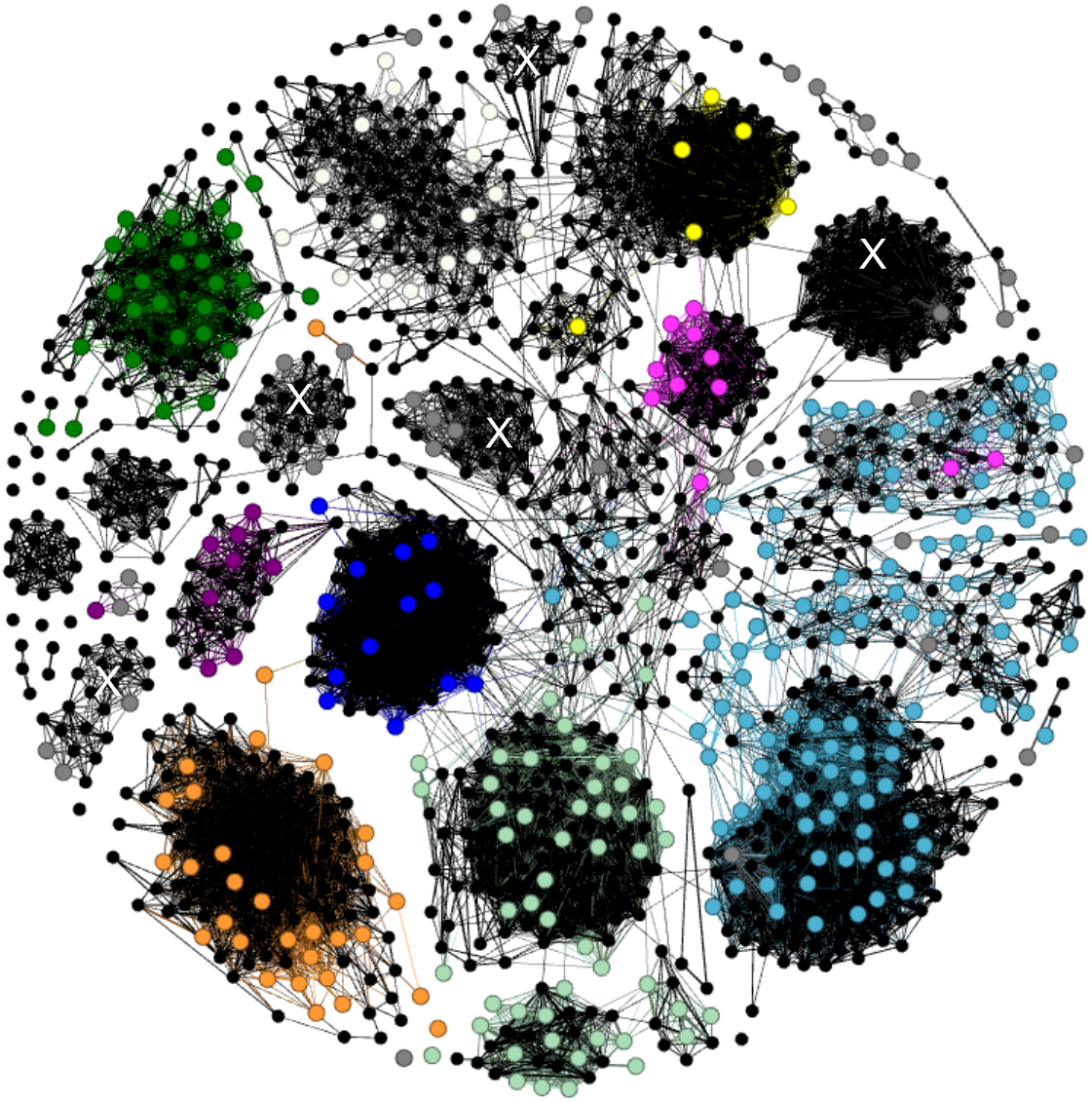
previously unclassified (Figure 2a). Additional investigation with the PvC01, PvT01 and Brazil-I assemblies using tribeMCL (also used in Lopez *et al)* confirmed the stability of the new subfamilies identified in PvP01 across a geographically divergent collection of isolates (Figure 2b). The analysis 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 (chromsomes LT615256-LT615269, contigs: FLYI01000001-FLYI01000530). PvP01 is maintained in GeneDB: http://www.genedb.org/Homepage/PvivaxP01 and updates are synchronized to PlasmoDB.

a



b

Cluster	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
PVP01	118	92	104	81	74	56	85	66	55	24	21	26	21	17	21	13	21	19	22	15	16	14	10	6	7	6	3
PVT01	92	88	72	76	77	60	55	47	31	29	17	25	11	15	13	18	8	10	12	12	6	10	9	6	5	3	4
PVC01	132	106	79	83	93	68	66	53	41	42	30	27	23	16	15	20	16	13	8	13	6	10	12	7	4	4	4
PVX	66	34	30	40	14	1	15	5	27	3	17	4	8	1	8	2	0	3	4	0	1	0	1	5	0	0	1
PVBB	94	83	62	65	62	104	49	37	36	25	22	16	10	18	9	12	15	11	9	6	11	5	6	6	4	4	3

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*[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 onces 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

1. Carlton JM, Sina BJ, Adams JH: **Why is *Plasmodium vivax* a neglected tropical disease?** *PLoS Negl Trop Dis.* 2011; **5**(6): e1160.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Mendis K, Sina BJ, Marchesini P, *et al.*: **The neglected burden of *Plasmodium vivax* malaria.** *Am J Trop Med Hyg.* 2001; **64**(1–2 Suppl): 97–106.
PubMed Abstract
3. Price RN, Tjitra E, Guerra CA, *et al.*: **Vivax malaria: neglected and not benign.** *Am J Trop Med Hyg.* 2007; **77**(6 Suppl): 79–87.
PubMed Abstract | Free Full Text
4. Karyana M, Burdarm L, Yeung S, *et al.*: **Malaria morbidity in Papua Indonesia, an area with multidrug resistant *Plasmodium vivax* and *Plasmodium falciparum*.** *Malar J.* 2008; **7**: 148.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Nosten F, McGready R, Simpson JA, *et al.*: **Effects of *Plasmodium vivax* malaria in pregnancy.** *Lancet.* 1999; **354**(9178): 546–9.
PubMed Abstract | Publisher Full Text
6. Poespoprodjo JR, Fobia W, Kenangalem E, *et al.*: **Vivax malaria: a major cause of morbidity in early infancy.** *Clin Infect Dis.* 2009; **48**(12): 1704–12.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Poespoprodjo JR, Fobia W, Kenangalem E, *et al.*: **Adverse pregnancy outcomes in an area where multidrug-resistant *Plasmodium vivax* and *Plasmodium falciparum* infections are endemic.** *Clin Infect Dis.* 2008; **46**(9): 1374–81.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Price RN, von Seidlein L, Valecha N, *et al.*: **Global extent of chloroquine-resistant *Plasmodium vivax*: a systematic review and meta-analysis.** *Lancet Infect Dis.* 2014; **14**(10): 982–91.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Chan ER, Menard D, David PH, *et al.*: **Whole genome sequencing of field isolates provides robust characterization of genetic diversity in *Plasmodium vivax*.** *PLoS Negl Trop Dis.* 2012; **6**(9): e1811.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Dharia NV, Bright AT, Westenberger SJ, *et al.*: **Whole-genome sequencing and microarray analysis of *ex vivo Plasmodium vivax* reveal selective pressure on putative drug resistance genes.** *Proc Natl Acad Sci U S A.* 2010; **107**(46): 20045–50.
PubMed Abstract | Publisher Full Text | Free Full Text
11. 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
12. Menard D, Chan ER, Benedet C, *et al.*: **Whole genome sequencing of field isolates reveals a common duplication of the Duffy binding protein gene in Malagasy *Plasmodium vivax* strains.** *PLoS Negl Trop Dis.* 2013; **7**(11): e2489.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Neafsey DE, Galinsky K, Jiang RH, *et al.*: **The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*.** *Nat Genet.* 2012; **44**(9): 1046–50.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Carlton JM, Adams JH, Silva JC, *et al.*: **Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*.** *Nature.* 2008; **455**(7214): 757–63.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Bright AT, Manary MJ, Tewhey R, *et al.*: **A high resolution case study of a patient with recurrent *Plasmodium vivax* infections shows that relapses were caused by meiotic siblings.** *PLoS Negl Trop Dis.* 2014; **8**(6): e2882.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Hupalo DN, Luo Z, Melnikov A, *et al.*: **Population genomics studies identify signatures of global dispersal and drug resistance in *Plasmodium vivax*.** *Nat Genet.* 2016; **48**(8): 953–8.
PubMed Abstract | Publisher Full Text
17. Pearson RD, Amato R, Auburn S, *et al.*: **Genomic analysis of local variation and recent evolution in *Plasmodium vivax*.** *Nat Genet.* 2016; **48**(8): 959–64.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Cunningham D, Lawton J, Jarra W, *et al.*: **The *pir* multigene family of *Plasmodium*: antigenic variation and beyond.** *Mol Biochem Parasitol.* 2010; **170**(2): 65–73.
PubMed Abstract | Publisher Full Text
19. Bernabeu M, Lopez FJ, Ferrer M, *et al.*: **Functional analysis of *Plasmodium vivax* VIR proteins reveals different subcellular localizations and cytoadherence to the ICAM-1 endothelial receptor.** *Cell Microbiol.* 2012; **14**(3): 386–400.
PubMed Abstract | Publisher Full Text
20. Carvalho BO, Lopes SC, Nogueira PA, *et al.*: **On the cytoadhesion of *Plasmodium vivax*-infected erythrocytes.** *J Infect Dis.* 2010; **202**(4): 638–47.
PubMed Abstract | Publisher Full Text
21. Yam XY, Brugat T, Siau A, *et al.*: **Characterization of the *Plasmodium* Interspersed Repeats (PIR) proteins of *Plasmodium chabaudi* indicates functional diversity.** *Sci Rep.* 2016; **6**: 23449.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Spence PJ, Jarra W, Lévy P, *et al.*: **Vector transmission regulates immune control of *Plasmodium* virulence.** *Nature.* 2013; **498**(7453): 228–31.
PubMed Abstract | Publisher Full Text | Free Full Text
23. 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
24. Guerra CA, Howes RE, Patil AP, *et al.*: **The international limits and population at risk of *Plasmodium vivax* transmission in 2009.** *PLoS Negl Trop Dis.* 2010; **4**(8): e774.
PubMed Abstract | Publisher Full Text | Free Full Text
25. White NJ: **Determinants of relapse periodicity in Plasmodium vivax malaria.** *Malar J.* 2011; **10**: 297.
PubMed Abstract | Publisher Full Text | Free Full Text
26. Auburn S, Marfurt J, Maslen G, *et al.*: **Effective preparation of *Plasmodium vivax***

field isolates for high-throughput whole genome sequencing. *PLoS One.* 2013; **8**(1): e53160.
PubMed Abstract | Publisher Full Text | Free Full Text

27. 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): 291–5.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Bentley DR, Balasubramanian S, Swerdlow HP, *et al.*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature.* 2008; **456**(7218): 53–9.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2009; **25**(14): 1754–60.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Zerbino DR, Birney E: **Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs.** *Genome Res.* 2008; **18**(5): 821–9.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Zimin AV, Marçais G, Puiu D, *et al.*: **The MaSuRCA genome assembler.** *Bioinformatics.* 2013; **29**(21): 2669–77.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Assefa S, Keane TM, Otto TD, *et al.*: **ABACAS: algorithm-based automatic contiguation of assembled sequences.** *Bioinformatics.* 2009; **25**(15): 1968–9.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Hunt M, Kikuchi T, Sanders M, *et al.*: **REAPR: a universal tool for genome assembly evaluation.** *Genome Biol.* 2013; **14**(5): R47.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Otto TD, Sanders M, Berriman M, *et al.*: **Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology.** *Bioinformatics.* 2010; **26**(14): 1704–7.
PubMed Abstract | Publisher Full Text | Free Full Text

35. 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): 1260–84.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Tsai IJ, Otto TD, Berriman M: **Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps.** *Genome Biol.* 2010; **11**(4): R41.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Boetzer M, Pirovano W: **Toward almost closed genomes with GapFiller.** *Genome Biol.* 2012; **13**(6): R56.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Otto TD: **From sequence mapping to genome assemblies.** *Methods Mol Biol.* 2015; **1201**: 19–50.
PubMed Abstract | Publisher Full Text

39. 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): 2672–6.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Otto TD, Dillon GP, Degrave WS, *et al.*: **RATT: Rapid Annotation Transfer Tool.** *Nucleic Acids Res.* 2011; **39**(9): e57.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Stanke M, Steinkamp R, Waack S, *et al.*: **AUGUSTUS: a web server for gene finding in eukaryotes.** *Nucleic Acids Res.* 2004; **32**(Web Server issue): W309–12.
PubMed Abstract | Publisher Full Text | Free Full Text

42. 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

43. Li L, Stoeckert CJ Jr, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res.* 2003; **13**(9): 2178–89.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Wootton JC, Federhen S: **Analysis of compositionally biased regions in sequence databases.** *Methods Enzymol.* 1996; **266**: 554–71.
PubMed Abstract | Publisher Full Text

45. Bastian M, Heymann S, Jacomy M: **Gephi: An Open Source Software for Exploring and Manipulating Networks**. In: *International AAAI Conference on Weblogs and Social Media.* 2009.
Reference Source

46. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res.* 2002; **30**(7): 1575–84.
PubMed Abstract | Publisher Full Text | Free Full Text

47. Preston MD, Campino S, Assefa SA, *et al.*: **A barcode of organellar genome polymorphisms identifies the geographic origin of *Plasmodium falciparum* strains.** *Nat Commun.* 2014; **5**: 4052.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Fang XD, Kaslow DC, Adams JH, *et al.*: **Cloning of the *Plasmodium vivax* Duffy receptor.** *Mol Biochem Parasitol.* 1991; **44**(1): 125–32.
PubMed Abstract | Publisher Full Text

49. Galinski MR, Medina CC, Ingravallo P, *et al.*: **A reticulocyte-binding protein complex of *Plasmodium vivax* merozoites.** *Cell.* 1992; **69**(7): 1213–26.
PubMed Abstract | Publisher Full Text

50. França CT, He WQ, Gruszczyk J, *et al.*: ***Plasmodium vivax* Reticulocyte Binding Proteins Are Key Targets of Naturally Acquired Immunity in Young Papua New Guinean Children.** *PLoS Negl Trop Dis.* 2016; **10**(9): e0005014.
PubMed Abstract | Publisher Full Text

51. Ntumngia FB, Thomson-Luque R, Torres Lde M, *et al.*: **A Novel Erythrocyte Binding Protein of *Plasmodium vivax* Suggests an Alternate Invasion Pathway into Duffy-Positive Reticulocytes.** *MBio.* 2016; **7**(4): pii: e01261-16.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 19 December 2016

doi:10.21956/wellcomeopenres.10647.r18200

✔ **Liwang Cui**
Department of Entomology, Pennsylvania State University, State College, PA, USA

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.

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.

Competing Interests: No competing interests were disclosed.

Referee Report 12 December 2016

doi:10.21956/wellcomeopenres.10647.r18268

Richárd Bártfai
Department of Molecular Biology, Radboud University Nijmegen, Nijmegen, 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*., 2012[1] 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

1. Tachibana S, Sullivan SA, Kawai S, Nakamura S, Kim HR, Goto N, Arisue N, Palacpac NM, Honma H, Yagi M, Tougan T, Katakai Y, Kaneko O, Mita T, Kita K, Yasutomi Y, Sutton PL, Shakhbatyan R, Horii T, Yasunaga T, Barnwell JW, Escalante AA, Carlton JM, Tanabe K: Plasmodium cynomolgi genome sequences provide insight into Plasmodium vivax and the monkey malaria clade.*Nat Genet*. 2012; **44** (9): 1051-5 PubMed Abstract | Publisher Full Text

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.

Competing Interests: No competing interests were disclosed.