AMERICAN SOCIETY FOR MICROBIOLOGY
genomeAnnouncements

High-Quality Genome Assembly and Annotation for *Plasmodium coatneyi*, Generated Using Single-Molecule Real-Time PacBio Technology

Jung-Ting Chien,[a,f] Suman B. Pakala,[b,f] Juliana A. Geraldo,[c] Stacey A. Lapp,[a,f] Jay C. Humphrey,[b,f] John W. Barnwell,[d,f] Jessica C. Kissinger,[b,f] Mary R. Galinski[a,e,f]

International Center for Malaria Research, Education and Development, Emory Vaccine Center, Yerkes National Primate Research Center, Emory University, Atlanta, Georgia, USA[a]; Department of Genetics, Institute of Bioinformatics, Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, Georgia, USA[b]; Biosystems Informatics & Genomics, René Rachou Research Center (CPqRR-FIOCRUZ), Belo Horizonte, Minas Gerais, Brazil[c]; Malaria Branch, Division of Parasitic Diseases and Malaria, Centers for Disease Control and Prevention, Atlanta, Georgia, USA[d]; Division of Infectious Diseases, Department of Medicine, Emory University, Atlanta, Georgia, USA[e]; Malaria Host–Pathogen Interaction Center, Emory University, Atlanta, Georgia, USA[f]

***Plasmodium coatneyi* is a protozoan parasite species that causes simian malaria and is an excellent model for studying disease caused by the human malaria parasite, *P. falciparum*. Here we report the complete (nontelomeric) genome sequence of *P. coatneyi* Hackeri generated by the application of only Pacific Biosciences RS II (PacBio RS II) single-molecule real-time (SMRT) high-resolution sequence technology and assembly using the Hierarchical Genome Assembly Process (HGAP). This is the first *Plasmodium* genome sequence reported to use only PacBio technology. This approach has proven to be superior to short-read only approaches for this species.**

Received 1 July 2016 **Accepted** 6 July 2016 **Published** 1 September 2016

Citation Chien J-T, Pakala SB, Geraldo JA, Lapp SA, Humphrey JC, Barnwell JW, Kissinger JC, Galinski MR. 2016. High-quality genome assembly and annotation for *Plasmodium coatneyi*, generated using single-molecule real-time PacBio technology. Genome Announc 4(5):e00883-16. doi:10.1128/genomeA.00883-16.

Address correspondence to Mary R. Galinski, mgalins@emory.edu.

A *Plasmodium* genome sequence was published initially in 2002, for *P. falciparum* (1). Genome sequences for several other primate *Plasmodium* species have followed (1–6), but none has yet been generated using only PacBio technology (6). *Plasmodium coatneyi*, which infects *Macaca mulatta* (rhesus macaques) and serves as a model of *P. falciparum* (7, 8), was chosen as a test platform for PacBio sequencing. A preliminary draft of the *P. coatneyi* genome based on short-read (<500 bp) sequence technology is available in the NCBI database (PRJNA233970). Although the overall "big picture" can be gained from this genome assembly, there are over 500 sequence gaps distributed throughout the parasite's estimated 14 nuclear chromosomes. Like *P. falciparum*, the *P. coatneyi* genome has numerous repetitive sequences and complex multi-gene families which present major difficulties that have prohibited nontelomeric genome assembly with closure using only short-read technologies. Gaps prevent reliable gene content analysis, genetics and reference-based gene expression analyses, all of which are critical for understanding *Plasmodium* and disease progression. We have implemented PacBio (RSSMRT) sequence technology to tackle these issues.

Genomic DNA was extracted from *ex vivo* matured schizont-stage parasites with a Qiagen DNA blood midi kit. The gDNA was further purified with a PowerClean DNA cleanup kit (Mo Bio Laboratories). Five micrograms of gDNA were subsequently used for library preparation. SMRTbell DNA libraries (Pacific Biosciences) were constructed according to the PacBio standard protocol with the BluePippin size-selection system (Sage Science). Sequence was generated on a PacBio RSII instrument using P6-C4 chemistry. Following cleaning, the mean assembled subread length is 5,824 bp; the N_{50} is 7,257; the total number of bases is 1,792,197,364 and the total number of reads is 257,557. HGAP3 (9) *de novo* assembly was performed using the Amazon EC2 cloud SMRT portal. The error correction module was defined as minimum subread length of 100 bp, a minimum read quality of 0.80, and a minimum read length of 6,000 bp. Following host (*M. mulatta*) contig removal, 15 nuclear contigs, one mitochondrial contig, and one apicoplast contig remained (51.42× average coverage). Contig identity and synteny were evaluated via BLASTn and progressive MAUVE algorithms (10) using the *P. knowlesi* genome from GeneDB (3) as the reference. Two suspected interchromosomal rearrangements occurring within gene family sequences located on Chr4/Chr13 and Chr12/Chr14 could not be validated by PCR, suggesting these sequences may in fact be correct as presented here.

De novo gene prediction was performed using SNAP (11) and Augustus (12) for gene calls in the MAKER2 (13) genome annotation tool. The *P. vivax* and *P. knowlesi* predicted proteomes were included as evidence. In total, 5,516 protein-encoding genes were predicted, including up to 112 *SICAvar* genes. The complete annotated mitochondrial and apicoplast genomes are also included in this report. The annotation was validated with *P. coatneyi* RNA-Seq data, Uniprot, KEGG and OrthoMCL Orthology, and InterProScan5 (14–18). 5,060 genes have strong evidence of synteny.

Accession number(s). The fourteen chromosome sequences were deposited at NCBI (BioProject PRJNA315987) under accession numbers CP016239 to CP016252 and provided to PlasmoDB (19).

ACKNOWLEDGMENTS

This study was supported in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia's Office of the Vice President for Research and Office of the Vice President for Information Technology.

FUNDING INFORMATION

This work, including the efforts of Mary R. Galinski, was funded by HHS | NIH | National Institute of Allergy and Infectious Diseases (NIAID) (HHSN272201200031C).

This project was funded in part by federal funds from the U.S. National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract no. HHSN272201200031C (PI, Mary R. Galinski), which supports the Malaria Host-Pathogen Interaction Center (MaHPIC), and supported in part by the Office of Research Infrastructure Programs/OD P51OD011132. J.A.G. was supported by CNPq---Conselho Nacional de Desenvolvimento Científico e Tecnológico, of the Ministry of Science, Technology and Innovation of Brazil (process, 203585/2015-1).

REFERENCES

1. **Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DMA, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, Mcfadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B.** 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum* Nature **419:**498–511. http://dx.doi.org/10.1038/nature01097.
2. **Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, Crabtree J, Angiuoli SV, Merino EF, Amedeo P, Cheng Q, Coulson RM, Crabb BS, Del Portillo HA, Essien K, Feldblyum TV, Fernandez-Becerra C, Gilson PR, Gueye AH, Guo X, Kang'a S, Kooij TW, Korsinczky M, Meyer EV, Nene V, Paulsen I, White O, Ralph SA, Ren Q, Sargeant TJ, Salzberg SL, Stoeckert CJ, Sullivan SA, Yamamoto MM, Hoffman SL, Wortman JR, Gardner MJ, Galinski MR, Barnwell JW, Fraser-Liggett CM.** 2008. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature **455:**757–763. http://dx.doi.org/10.1038/nature07327.
3. **Pain A, Böhme U, Berry AE, Mungall K, Finn RD, Jackson AP, Mourier T, Mistry J, Pasini EM, Aslett MA, Balasubrammaniam S, Borgwardt K, Brooks K, Carret C, Carver TJ, Cherevach I, Chillingworth T, Clark TG, Galinski MR, Hall N, Harper D, Harris D, Hauser H, Ivens A, Janssen CS, Keane T, Larke N, Lapp S, Marti M, Moule S, Meyer IM, Ormond D, Peters N, Sanders M, Sanders S, Sargeant TJ, Simmonds M, Smith F, Squares R, Thurston S, Tivey AR, Walker D, White B, Zuiderwijk E, Churcher C, Quail MA, Cowman AF, Turner CM, Rajandream MA, Kocken CH, Thomas AW, Newbold CI, Barrell BG, Berriman M.** 2008. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. Nature **455:**799–803. http://dx.doi.org/10.1038/nature07306.
4. **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.** 2012. *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. Nat Genet **44:**1051–1055. http://dx.doi.org/10.1038/ng.2375.
5. **Otto TD, Rayner JC, Böhme U, Pain A, Spottiswoode N, Sanders M, Quail M, Ollomo B, Renaud F, Thomas AW, Prugnolle F, Conway DJ, Newbold C, Berriman M.** 2014. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. Nat Commun **5:**4754. http://dx.doi.org/10.1038/ncomms5754.
6. **Pinheiro MM, Ahmed MA, Millar SB, Sanderson T, Otto TD, Lu WC, Krishna S, Rayner JC, Cox-Singh J.** 2015. *Plasmodium knowlesi* genome sequences from clinical isolates reveal extensive genomic dimorphism. PLoS One **10:**e0121303. http://dx.doi.org/10.1371/journal.pone.0121303.
7. **Aikawa M, Brown A, Smith CD, Tegoshi T, Howard RJ, Hasler TH, Ito Y, Perry G, Collins WE, Webster K.** 1992. A primate model for human cerebral malaria: *Plasmodium coatneyi*-infected rhesus monkeys. Am J Trop Med Hyg **46:**391–397.
8. **Moreno A, Cabrera-Mora M, Garcia A, Orkin J, Strobert E, Barnwell JW, Galinski MR.** 2013. *Plasmodium coatneyi* in rhesus macaques replicates the multisystemic dysfunction of severe malaria in humans. Infect Immun **81:**1889–1904. http://dx.doi.org/10.1128/IAI.00027-13.
9. **Chin C.-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J.** 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods **10:**563–569. http://dx.doi.org/10.1038/nmeth.2474.
10. **Darling AE, Mau B, Perna NT.** 2010. Progressive mauve: multiple genome alignment with Gene gain, loss and rearrangement. PLoS One **5:**e11147. http://dx.doi.org/10.1371/journal.pone.0011147.
11. **Korf I.** 2004. Gene finding in novel genomes. BMC Bioinformatics **5:**1–9. http://dx.doi.org/10.1186/1471-2105-5-59.
12. **Stanke M, Diekhans M, Baertsch R, Haussler D.** 2008. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Sequence Anal **24:**637–644.
13. **Holt C, Yandell M.** 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics **12:**491. http://dx.doi.org/10.1186/1471-2105-12-491.
14. **Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LL.** 2004. UniProt: the universal protein knowledgebase. Nucleic Acids Res **32:**D115–D119. http://dx.doi.org/10.1093/nar/gkh131.
15. **Nakaya A, Katayama T, Itoh M, Hiranuka K, Kawashima S, Moriya Y, Okuda S, Tanaka M, Tokimatsu T, Yamanishi Y, Yoshizawa AC, Kanehisa M, Goto S.** 2013. KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. Nucleic Acids Res **41:**D353–D357. http://dx.doi.org/10.1093/nar/gks1239.
16. **Jones P, Binns D, Chang H.-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S.-Y, Lopez R, Hunter S.** 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics **30:**1236–1240. http://dx.doi.org/10.1093/bioinformatics/btu031.
17. **Li L, Stoeckert CJ, Jr, Roos DS.** 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res **13:**2178–2189. http://dx.doi.org/10.1101/gr.1224503.
18. **Chen F, Mackey AJ, Stoeckert CJ, Jr, Roos DS.** 2006. OrthoMCL-DB: querying a comprehensive multi-speciescollection of ortholog groups. Nucleic Acids Res **34:**D363–D368. http://dx.doi.org/10.1093/nar/gkj123.
19. **Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Miller JA, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Stoeckert CJ, Treatman C, Wang H.** 2009. PlasmoDB: a functional genomic database for malaria parasite. Nucleic Acids Res **37:**D539–D543. http://dx.doi.org/10.1093/nar/gkn814.