De Novo Genome Assembly of a *Plasmodium falciparum* NF54 Clone Using Single-Molecule Real-Time Sequencing

Jessica M. Bryant,[a,b,c] Sebastian Baumgarten,[a,b,c] Audrey Lorthiois,[d] Christine Scheidig-Benatar,[a,b,c] Aurélie Claës,[a,b,c] Artur Scherf[a,b,c]

[a]Unité de Biologie des Interactions Hôte-Parasite, Institut Pasteur, Paris, France
[b]INSERM U1201, Paris, France
[c]CNRS ERL9195, Paris, France
[d]Center for Production and Infection of Anopheles (CEPIA), Institut Pasteur, Paris, France

ABSTRACT   *Plasmodium falciparum* is the species of human malaria parasite that causes the most severe form of the disease. Here, we used single-molecule real-time (SMRT) sequencing technology from Pacific Biosciences (PacBio) to sequence, assemble *de novo*, and annotate the genome of a *P. falciparum* NF54 clone.

Plasmodium falciparum is a protozoan parasite with a complex life cycle in human and mosquito hosts (1). Its 23-Mb haploid genome is extremely AT rich (~80%) and contains stretches of highly repetitive sequences, especially in telomeric and subtelomeric regions. As genome editing techniques advance, the need has arisen for a complete and accurate genome sequence of laboratory strains used to investigate pathogenesis. The genome of the *P. falciparum* 3D7 strain, which is widely used to study the intraerythrocytic life cycle *in vitro*, has been sequenced and annotated (2). 3D7 is a clone of the NF54 isolate, which is believed to have originated in Africa (3, 4). However, 3D7 produces fewer gametocytes under *in vitro* culture conditions than the original NF54 strain (5). Here, we present the nuclear genome sequence and annotation of a new NF54 clone. We have shown that this clone can be easily grown in culture and transmitted to the mosquito, producing an amount of salivary gland sporozoites similar to that of its parent NF54 strain. This strain and its genome sequence will be useful for performing functional studies of gametocytogenesis, mosquito transmission, and mosquito stage development.

Late-stage parasites were grown in 6 ml of concentrated blood at 3.5% parasitemia and enriched by Plasmagel flotation. DNA was extracted with the Genomic-tip 500/G kit (catalog number 10262; Qiagen) and further purified with phenol-chloroform. Library preparation and sequencing were performed by GATC Biotech (Constance, Germany) on a PacBio RS II sequencer, with an average read length of 16,291 bp and genome coverage of 100× postfiltering. The same genomic DNA was sonicated, and libraries were prepared with the NEBNext Ultra II DNA library preparation kit (catalog number E7645S; New England Biolabs) and sequenced with a NextSeq 500 platform (Illumina) with an average genome coverage of 600×.

The genome was assembled using the PacBio RS_Assembly_HGAP.3 protocol (included in SMRT Portal version 2.3.0), with default settings and a target genome size of 23 Mb (6). Scaffolding of the initial assembly was performed using pyScaf (https://github.com/lpryszcz/pyScaf) and the 3D7 genome as a reference. The resulting gaps were closed with GapFiller (7) using the Illumina short reads. The assembly was further polished with the PacBio reads using BLASR (8) and Quiver (6). Illumina short reads were subsequently mapped to the assembly using Bowtie2 (9), and read alignments were used to fix remaining single-nucleotide polymorphisms (SNPs), short indels, and breaks using Pilon (10).

The final assembly is composed of 19 contigs, with a total size of 23,435,585 bases and an average GC content of 19.34%. All 14 chromosomes found in the 3D7 strain are present in our NF54 assembly, with 9 contigs representing putatively full-length chromosomes. The remaining 5 chromosomes are assembled in no more than two contigs per chromosome.

A LiftOver annotation was performed using FLO (https://github.com/wurmlab/flo) with the curated 3D7 annotation provided at PlasmoDB version 32 (http://plasmodb .org) (11). In total, 5,587 loci were identified, of which 5,015 loci are protein coding, 399 loci are putative pseudogenes, and 173 loci are non-protein-coding RNA. All multigene families associated with antigenic variation were annotated, including *var* (59 genes), *rifin* (158 genes), and *stevor* (31 genes).

Accession number(s). This whole-genome shotgun project has been deposited at DDBJ/ENA/GenBank under the accession number NYMT00000000. The version described in this paper is version NYMT01000000.

REFERENCES

1. Miller LH, Ackerman HC, Su XZ, Wellems TE. 2013. Malaria biology and disease pathogenesis: insights for new treatments. Nat Med 19:156–167. https://doi.org/10.1038/nm.3073.

2. 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. https://doi.org/10.1038/nature01097.

3. Ponnudurai T, Leeuwenberg AD, Meuwissen JH. 1981. Chloroquine sensitivity of isolates of *Plasmodium falciparum* adapted to *in vitro* culture. Trop Geogr Med 33:50–54.

4. Walliker D, Quakyi IA, Wellems TE, McCutchan TF, Szarfman A, London WT, Corcoran LM, Burkot TR, Carter R. 1987. Genetic analysis of the human malaria parasite *Plasmodium falciparum*. Science 236:1661–1666. https://doi.org/10.1126/science.3299700.

5. Delves MJ, Straschil U, Ruecker A, Miguel-Blanco C, Marques S, Dufour AC, Baum J, Sinden RE. 2016. Routine *in vitro* culture of *P. falciparum* gametocytes to evaluate novel transmission-blocking interventions. Nat Protoc 11:1668–1680. https://doi.org/10.1038/nprot.2016.096.

6. Chin CS, 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. https://doi.org/10.1038/nmeth.2474.

7. Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. Genome Biol 13:R56. https://doi.org/10.1186/gb-2012-13-6 -r56.

8. Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13:238. https://doi.org/10.1186/1471-2105-13-238.

9. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.

10. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963. https://doi.org/10.1371/journal .pone.0112963.

11. Bahl A, Brunk B, Crabtree J, Fraunholz MJ, Gajria B, Grant GR, Ginsburg H, Gupta D, Kissinger JC, Labo P, Li L, Mailman MD, Milgram AJ, Pearson DS, Roos DS, Schug J, Stoeckert CJ, Whetzel P. 2003. PlasmoDB: the *Plasmodium genome* resource. A database integrating experimental and computational data. Nucleic Acids Res 31:212–215. https://doi.org/10 .1093/nar/gkg081.