# Draft Hybrid Genome Assembly of a Canadian *Cyclospora cayetanensis* Isolate

Christine A. Yanta,[a,b] Stephen M. J. Pollo,[c] John R. Barta,[b] Sarah J. Reiling,[d] James D. Wasmuth,[c] Brent R. Dixon,[d] ⓘ Rebecca A. Guy[a]

aDivision of Enteric Diseases, National Microbiology Laboratory at Guelph, Public Health Agency of Canada, Guelph, Ontario, Canada
bDepartment of Pathobiology, University of Guelph, Guelph, Ontario, Canada
cDepartment of Ecosystem and Public Health, Faculty of Veterinary Medicine, University of Calgary, Calgary, Alberta, Canada
dBureau of Microbial Hazards, Food Directorate, Health Canada, Ottawa, Ontario, Canada

**ABSTRACT** The apicomplexan parasite *Cyclospora cayetanensis* causes foodborne gastrointestinal disease in humans. Here, we report the first hybrid assembly for *C. cayetanensis*, which uses both Illumina MiSeq and Oxford Nanopore Technologies MinION platforms to generate genomic sequence data. The final genome assembly consists of 44,586,677 bases represented in 313 contigs.

*C*yclospora cayetanensis is an emerging human pathogen worldwide that causes the gastrointestinal disease cyclosporiasis. This coccidian parasite is becoming more prevalent in foodborne outbreaks across North America due to contaminated farm-produced crops imported from regions of endemicity (1). There are several inherent challenges in gathering genomic data for *C. cayetanensis*, including limited parasite material present in standard diagnostic samples and an inability to propagate the organism in a laboratory setting (2). The objective of this work was to generate the first whole-genome assembly from a *C. cayetanensis* specimen using both long- and short-read sequencing platforms. This hybrid approach will improve both the genome assembly available, similar to findings we described previously for *Giardia intestinalis* (3), and the identification of markers to refine current subtyping schemes to aid in outbreak investigations (4). We sequenced *C. cayetanensis* isolated from a positive human fecal specimen identified by the Newfoundland and Labrador Public Health Laboratory (NPHL) in 2020. Oocysts were purified from 5 g of stool as described previously (5) and incubated in 3% sodium hypochlorite for 10 min on ice prior to DNA extraction with the QIAamp Fast DNA Stool Mini Kit (Qiagen, Hilden, Germany). A bead beating step using lysing matrix Y (MP Biomedicals, OH, USA) at 6 m/s for 1 min on the Bead Ruptor 24 (Omni International, GA, USA) was included after the InhibitEx incubation to break the oocyst walls. The extracted DNA was purified using AMPure XP beads (Beckman Coulter, CA, USA). Two libraries were prepared with the Nextera DNA Flex kit and rapid PCR barcoding kit (SQK-RPB004) using 1 ng of purified DNA as the starting material. The resulting libraries were sequenced on the MiSeq platform (Illumina, Inc., CA, USA) using the v3 reagent kit (600 cycles) with a 2 × 300-bp paired-end protocol and on the MinION platform (Oxford Nanopore Technologies, Oxford, UK) using the flow cell type R9.4.1 (FLO-MIN106D). Sequencing on the Illumina MiSeq system yielded 11 million paired-end reads (5.2 Gbp), and sequencing on the Oxford Nanopore Technologies MinION system generated 1.0 million reads (2.9 Gbp), with an average fragment length of 3.3 kbp. The quality of the raw short reads was examined using FastQC (6), and the reads were filtered and trimmed using BBDuk v37.62 from the BBTools suite (7). Long reads were base called using Oxford Nanopore Technologies Guppy v4.0.11, and reads with average quality scores of <7 were removed. The resulting MiSeq short reads and MinION long reads were *de novo* assembled using hybridSPAdes v3.11.1 (8, 9) using

default parameters. A similarity search using BLAST v2.9.0+ (10) was performed against the RefSeq genome database for coccidian parasites with the initial assembly, and all contaminating contigs were removed. Polishing of the assembly was performed by mapping the short and long reads back to the assembly with Minimap2 v2.17-r941 (11) and correcting them with Pilon v1.23 (12). The final assembly is 313 contigs, with a total genome size of 44,586,677 bases (the largest contig is 1,973,156 bases), a G+C content of 51.9%, an $N_{50}$ value of 526,712 bases, and an $L_{50}$ value of 24 contigs. Based on these statistics, this is the most continuous genome assembly described to date of the 37 draft genomes available for *C. cayetanensis*.

**Data availability.** This whole-genome shotgun project has been deposited in DDBJ/EMBL/GenBank under the accession no. JAJEWR000000000. The version described in this paper is the first version, JAJEWR010000000. Raw sequence reads are available under the BioProject no. PRJNA772675.

## ACKNOWLEDGMENTS

## REFERENCES

1. Hadjilouka A, Tsaltas D. 2020. *Cyclospora cayetanensis*: major outbreaks from ready to eat fresh fruits and vegetables. Foods 9:1703. https://doi.org/10.3390/foods9111703.
2. Nascimento FS, Wei-Pridgeon Y, Arrowood MJ, Moss D, da Silva AJ, Talundzic E, Qvarnstrom Y. 2016. Evaluation of library preparation methods for Illumina next generation sequencing of small amounts of DNA from foodborne parasites. J Microbiol Methods 130:23–26. https://doi.org/10.1016/j.mimet.2016.08.020.
3. Pollo SMJ, Reiling SJ, Wit J, Workentine ML, Guy RA, Batoff GW, Yee J, Dixon BR, Wasmuth JD. 2020. Benchmarking hybrid assemblies of *Giardia* and prediction of widespread intra-isolate structural variation. Parasit Vectors 13:108. https://doi.org/10.1186/s13071-020-3968-8.
4. Nascimento FS, Barratt J, Houghton K, Plucinski M, Kelley J, Casillas S, Bennett C, Snider C, Tuladhar R, Zhang J, Clemons B, Madison-Antenucci S, Russell A, Cebelinski E, Haan J, Robinson T, Arrowood MJ, Talundzic E, Bradbury R, Qvarnstrom Y. 2020. Evaluation of an ensemble-based distance statistic for clustering MLST datasets using epidemiologically defined clusters of cyclosporiasis. Epidemiol Infect 148:E172. https://doi.org/10.1017/S0950268820001697.
5. Qvarnstrom Y, Wei-Pridgeon Y, van Roey E, Park S, Srinivasamoorthy G, Nascimento FS, Moss DM, Talundzic E, Arrowood MJ. 2018. Purification of *Cyclospora cayetanensis* oocysts obtained from human stool specimens

for whole genome sequencing. Gut Pathog 10:45. https://doi.org/10.1186/s13099-018-0272-7.
6. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc.
7. Bushnell B. 2014. BBMap short read aligner, and other bioinformatics tools. https://sourceforge.net/projects/bbmap.
8. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes de novo assembler. Curr Protoc Bioinformatics 70:e102. https://doi.org/10.1002/cpbi.102.
9. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. 2016. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics 32:1009–1015. https://doi.org/10.1093/bioinformatics/btv688.
10. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
11. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
12. 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.