Fully resolved assembly of Cryptosporidium parvum

Vipin K. Menon[1,*], Pablo C. Okhuysen[2], Cynthia Chappell[3], Medhat Mahmoud[1], Qingchang Meng[1], Harsha Doddapaneni[1], Vanesa Vee[1], Yi Han[1], Sejal Salvi[1], Sravya Bhamidipati[1], Kavya Kottapalli[1], George Weissenberger[1], Hua Shen[1], Matthew C. Ross[4], Kristi L. Hoffman[4], Sara Javornik Cregeen[4], Donna M. Muzny[1], Ginger A. Metcalf, Richard A. Gibbs[1], Joseph F. Petrosino[4], Fritz J. Sedlazeck[1,*]

Corresponding authors*: menon@bcm.edu, fritz.sedlazeck@bcm.edu

1: Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, United States of America
2: Department of Infectious Diseases, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America
3: The University of Texas School of Public Health, Houston, Texas, United States of America
4: Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, Texas, United States of America

Abstract

Background

Cryptosporidium *parvum* are apicomplexan parasites commonly found across many species with a global infection prevalence of 7.6%. As such it is important to understand the diversity and genomic makeup of this prevalent parasite to prohibit further spread and to fight an infection. The general basis of every genomic study is a high-quality reference genome that has continuity and completeness and is of high quality and thus enables comprehensive comparative studies.

Findings

Here we provide a highly accurate and complete reference genome of Cryptosporidium spp.. The assembly is based on Oxford Nanopore reads and was improved using Illumina reads for error correction. The assembly encompasses 8 chromosomes and includes 13 telomeres that were resolved. Overall, the assembly shows a high completion rate with 98.4% single copy Busco genes. This is also shown by the identification of 13 telomeric regions across the 8 chromosomes. The consensus accuracy of the established reference genome was further validated by sequence alignment of established genetic markers for C.*parvum*.

Conclusions

This high-quality reference genome provides the basis for subsequent studies and comparative genomic studies across the Cryptosporidium clade.

Keywords: Assembly, Cryptosporidium, nanopore, canu

Introduction

Cryptosporidium spp. are apicomplexan parasites of public health and veterinary significance, with a recent analysis reporting a global infection prevalence of 7.6% [1]. Historically, limited government and private funding was available to study the epidemiology and molecular dynamics of the organism, but this has recently shifted [2].

Cryptosporidium spp. have been found in 155 species of mammals, including primates [3,4]. Among humans, twenty species of Cryptosporidium spp. have been identified [5]. Although the parasite can be transmitted in a variety of ways, the most common method is via water (drinking water and recreational water). In the United States, Cryptosporidium is the most common cause of waterborne disease in humans [6]. Studies have shown that *Cryptosporidium* is responsible for a large proportion of all cases of moderate-to-severe diarrhea in children under the age of two [7,8]. There is currently no vaccine available, and the only approved drug for the treatment of Cryptosporidium-related diarrhea is nitazoxanide (NTZ), which has limited activity in immunocompromised patients.

The inability to grow Cryptosporidium *in vitro* hampered progress in understanding pathogenesis and exploring new treatment modalities. The use of human organoids recapitulates *in vivo* physiology of their original tissues [9][10,11] and the molecular mechanisms and pathways used by Cryptosporidium during infection. However, it became apparent that a high-quality reference genome is needed to facilitate any genomic or association studies.

C.*parvum* was included in early genome-sequencing projects due to its public health importance and high global prevalence. The first reported complete genome assembly for C.*parvum* Iowa II became available in 2004 [12], generated by random shotgun sequencing approach, resulting in roughly 13x genome coverage totaling 9.1 Mb of DNA sequence across all the eight chromosomes. The reference had a reduced coverage across the genome, with gaps and was not adequate to represent the full breadth of genes present which could result in misleading interpretations of the isolates being studied. Online repositories such as GenBank, CryptoDB and the Wellcome Trust Sanger Institute FTP servers also have a range of unassembled, unprocessed raw read sequences.

Long read sequencing technology has advanced to enable read lengths of 15 - 20 Kb (PacBio) and 2 - 3 Mb (Oxford Nanopore(ONT)) with low error rates and is frequently utilized to improve reference genome assembly [5,13–18]. Thus, enabling long continuous assemblies without gaps even across highly repetitive regions[19]. In the current study, we have generated a reference genome for Cryptosporidium *parvum* by using long read sequencing on the ONT PromethION supplemented with short-read data generated on NovaSeq 6000 for error correction (see **Figure 1**). This complete reference generated includes all chromosomes and represents a gap-less representation of this important pathogen. Furthermore, it includes almost all telomeric sequences. The assembly is available at PRJNA744539.

Results

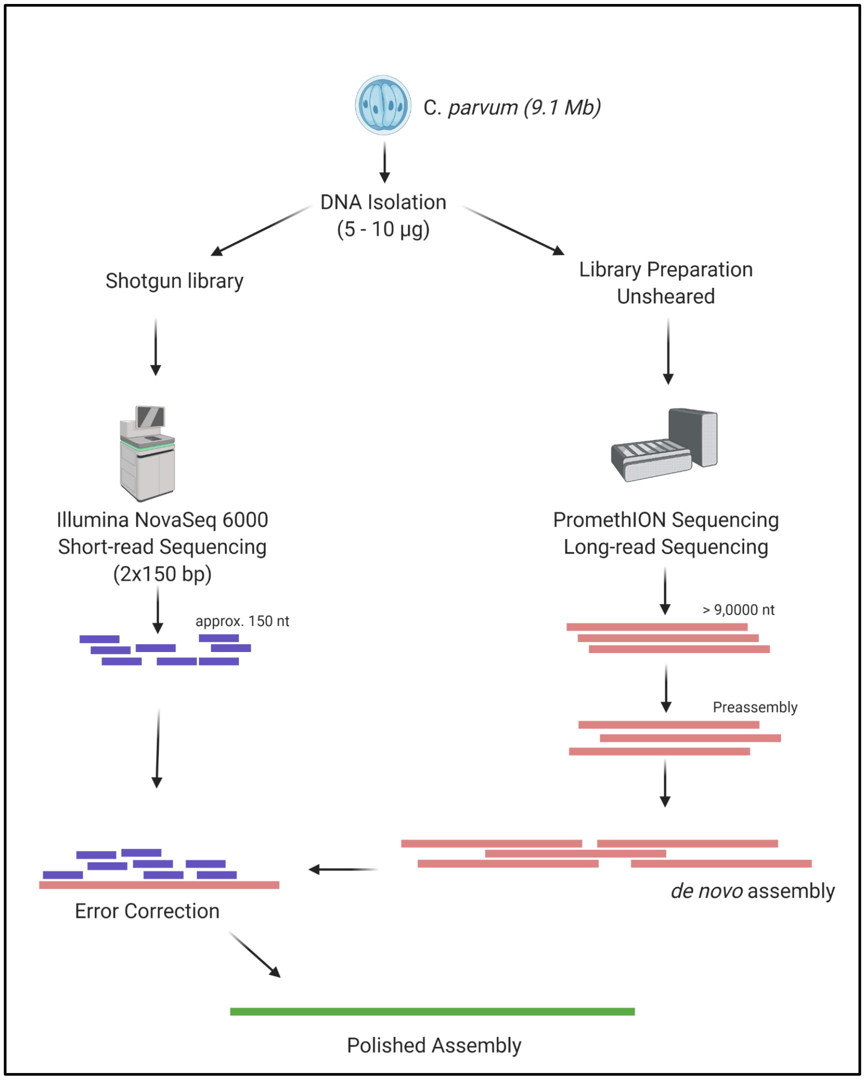


Figure 1: *Workflow for the generation of Cryptosporidium parvum assembly* (figure was created by BioRender).

We sequenced the Cryptosporidium genome with Oxford Nanopore long reads (see methods) and obtained a total of ~480Mbp of sequence (**Figure 1**). This is equivalent to 53x coverage for this genome (~9Mbp genome size). **Figure 2** shows overall statistics on read length and coverage. The N50 read length is 15.3 kbp with 10x coverage of reads with ≥30kbp length. Our longest read detected was 808kbp. In addition, we sequenced the genome using the Illumina NovaSeq 6000 to produce 352x coverage of 150bp paired end reads.

Using these short reads, we ran a genome estimation using GenomeScope [20] to obtain a genome size estimate using a polyploidy of 1. Doing so resulted in an estimate of 9.9Mbp with an 89.24% model fit (see **Supplementary Figure 1**). Inspection of the resulting data shown in the figure highlights that this is a potential overestimation of the genome size itself and thus fits in the realm of the previous reported reference assembly of ~9.1Mbp.

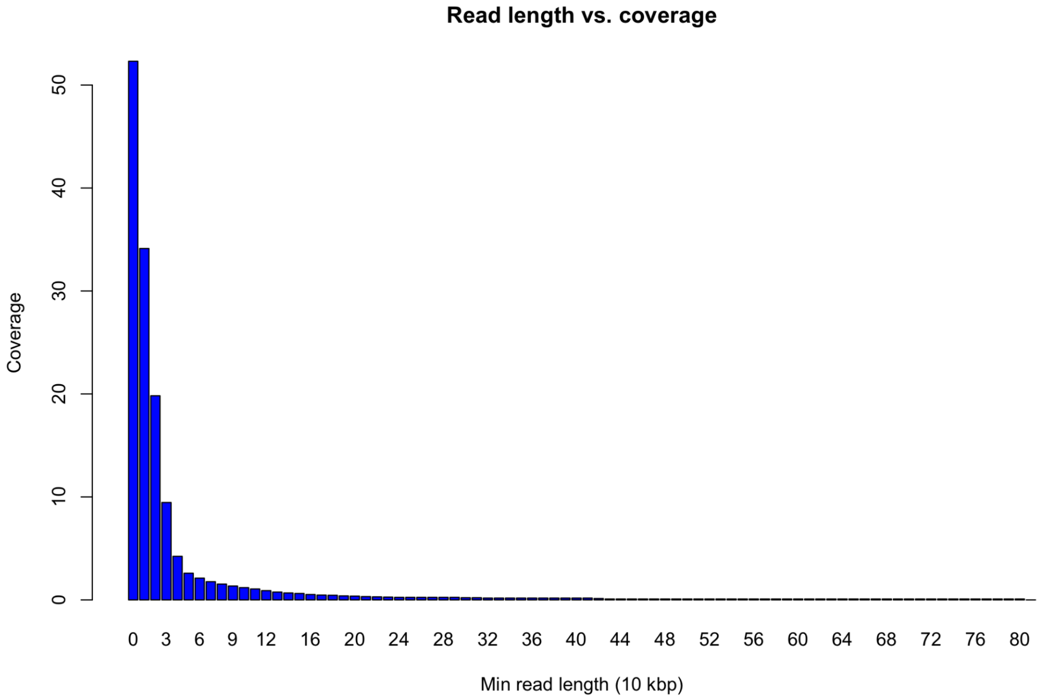


Figure 2: Read length distribution and cumulative coverage over the Oxford Nanopore. Sequencing. We obtained a total of 53x coverage with long reads and even 10x coverage with reads larger than 30kbp (x axis). The longest read measured was 808kbp.

Assembly

The initial assembly was carried out with only the ONT reads only using Canu[21] (see methods) and resulted in 25 contigs, with 8 contigs representing the entire chromosomes. **Table1** shows the general statistics obtained. We obtained a total genome length of 9.19Mbp of genome length across 8 assembled contigs with an average N50 size of 1.11Mbp. The largest contig was 1.4 Mbp. Our assembly as well as the previous assembly from 2004 shows a similar NG50 (see **Supplementary Figure 2**).

We also ran an assembly with Flye[22] (see methods), which leads to a total of 8 contigs as well. However, one contig was only 62160bp in length. Over a genomic alignment between the two assemblies, we identified a merge of two potential chromosomes in the Flye assembly which resulted in this short contig. We selected the Canu assembly as a correct representation of the genome, based on this result and interpretation from the short read analysis.

Next, we improved the quality of the Canu assembly using the short reads (see methods) over 2 rounds of assembly polishing. After the first round the number of corrections were reduced to ~20 along the entire genome. After these improvements, we extracted the 8 largest contigs and compared them over a genomic alignment (see methods) to the previously published Cryptosporidium reference[12]. This confirmed that the eight contigs represent the previously published chromosomes, whereas the other contigs seem to be repeats at the start or the end of the contigs. Our assembled eight chromosomes fill 14,669 bp of unresolved sequences (ie N) with our assembly. Our assembly also showed a similar GC content (30.11%) than the previous version (30.18%) again attesting to the overall quality.

To assess the completeness of our assembly further we used Busco[23] with the coccidia_odb10 linkage set (see methods). This analysis confirmed further the high quality of our assembly showing 98.4% complete re-identified genes from a total of 502. All 494 genes found have single copies, indicating that the new assembly is error-free. Apart from these single-copy genes, three genes have been reported as fragmented, and five genes have been reported as missing from the busco run.

A further comparison to the reference genome (BioProject PRJNA144)[12] revealed a high consistency with only a total of 4 Structural Variations (1 insertion, 1 deletion, 1 tandem expansion and 1 tandem contraction) between the two assemblies. This was done based on the genomic alignment and using Assemblytics[24].

	2007 Assembly	Current Assembly
Total sequence length	9,102,324	9,197,619
Total ungapped length	9,087,655	9,197,619
Unresolved sequences	14,669	0
N50	1,104,417	1,108,772
N90	985,969	993,129
L50	4	4
Total number of chromosomes	8	8

Table1: Overall assembly statistics and comparison of our assembly and the previous established assembly.

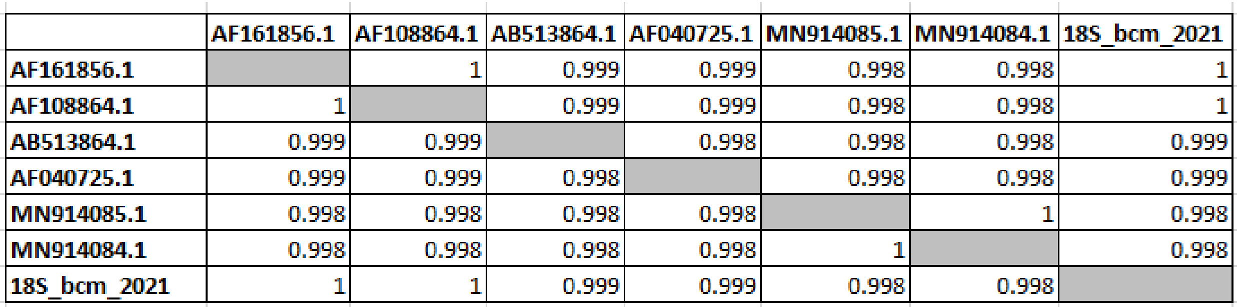
Telomere identification

Since our canu assembly directly reported chromosomes, we next sought to identify telomeric ends on either side of each chromosome (see methods). To search for the telomeres, we identified matching sequences in our assemblies of “TTTAGG” repeats [25] (see methods). We required at least 100 matches within a region towards the start and end of the contigs. Given these conservative thresholds to avoid other repeats we identified a total of 13 telomeric regions. For the majority of chromosomes (3,4,5,6 and 7) both sides showed telomeric regions. Thus, fully representing the chromosomes from telomere to telomere including the centromere. On chromosomes 2 and 8, the telomere was only found at the beginning of the chromosome. We only found the telomeric sequence at the end of chromosome 1. We further cross checked the other contigs that were filtered out previously. These highlighted telomeric sequences but couldn’t be placed automatically to the other chromosomes (chr 1,2 or 8). Overall, the identification of the telomeric sequences on the fast majority of the contigs highlights the overall high quality and continuity of our newly established C. parvum genome.

Comparative genomics

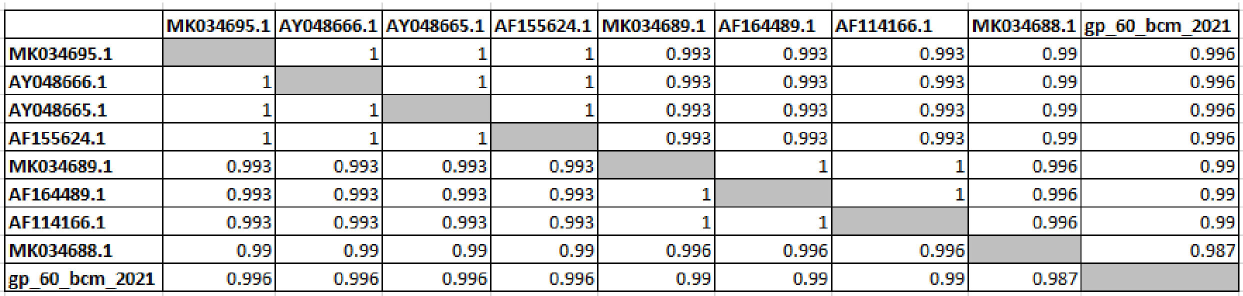
Cryptosporidium spp. are usually typed and characterized widely by using a small set of genetic markers including *gp60*, COWP, HSP70 and 18S [26]. Most of the genetic marker data available in GenBank are generated from short-read amplification and sequencing by Sanger, thus providing a better resolution, but still contain errors arising from manual curation.

A nucleotide identity matrix was generated post ClustalW alignment for 18S (Table 2) and *gp60* (Table 3) genetic markers, which were downloaded from GenBank. No gaps or large mismatches between the assembled genome and the genetic markers were observed (see **Supplementary Figure 3 & 4**). There were 0.2% identity mismatches observed in some of the references, which represent variations observed within the species.



	AF161856.1	AF108864.1	AB513864.1	AF040725.1	MN914085.1	MN914084.1	18S_bcm_2021
AF161856.1		1	0.999	0.999	0.998	0.998	1
AF108864.1	1		0.999	0.999	0.998	0.998	1
AB513864.1	0.999	0.999		0.998	0.998	0.998	0.999
AF040725.1	0.999	0.999	0.998		0.998	0.998	0.999
MN914085.1	0.998	0.998	0.998	0.998		1	0.998
MN914084.1	0.998	0.998	0.998	0.998	1		0.998
18S_bcm_2021	1	1	0.999	0.999	0.998	0.998	

Table 2: Sequence identity matrix of 18S genes from *Cryptosporidium parvum species* showing the sequence identity on 0-1 scale between data from GenBank to the current assembly.



	MK034695.1	AY048666.1	AY048665.1	AF155624.1	MK034689.1	AF164489.1	AF114166.1	MK034688.1	gp_60_bcm_2021
MK034695.1		1	1	1	0.993	0.993	0.993	0.99	0.996
AY048666.1	1		1	1	0.993	0.993	0.993	0.99	0.996
AY048665.1	1	1		1	0.993	0.993	0.993	0.99	0.996
AF155624.1	1	1	1		0.993	0.993	0.993	0.99	0.996
MK034689.1	0.993	0.993	0.993	0.993		1	1	0.996	0.99
AF164489.1	0.993	0.993	0.993	0.993	1		1	0.996	0.99
AF114166.1	0.993	0.993	0.993	0.993	1	1		0.996	0.99
MK034688.1	0.99	0.99	0.99	0.99	0.996	0.996	0.996		0.987
gp_60_bcm_2021	0.996	0.996	0.996	0.996	0.99	0.99	0.99	0.987	

Table 3: Sequence identity matrix of *gp60* genes from *Cryptosporidium parvum species* showing the sequence identity on 0-1 scale between data from GenBank to the current assembly.

Conclusion

The current work shows how next-generation sequencing, including third-generation long-read sequencing, can be used to generate a high-quality entire genome assembly, complete with centromeric regions and numerous telomeres. The genome assembly generated provides a gapless reference as compared to the previously published reference[12]. The study was able to boost the fidelity and robustness of the results by combining short and long reads.

Studies of C.spp. are based on genetic markers previously identified for some of the 8 chromosomes, and are not able to provide a better understanding of the genetic variation and recombination occurring within the species. Establishing stronger marker genes and perhaps enabling improved recovery of Cryptosporidium-specific sequencing reads by mapping to a high-resolution reference genome will enable better understanding of Cryptosporidium transmission.

Published studies have shown the presence of contingency genes in C.spp., which are responsible for surmounting challenges from the host and are subject to spontaneous mutation rates [27–29]. The majority of these genes are located in the telomere regions of the chromosomes, which are prime sites that evolve and mediate host-parasite interactions. In the current assembly, we are able to resolve 13 of the estimated 16 telomeres. The ability to resolve the telomeres and any sub telomeres across different chromosomes in C.spp. provides the ability to understand the adaptability of the organism to a wide range of environmental and host conditions.

The assembly will be a helpful resource to study this important pathogen further and investigate its complexity during growth in development in vitro and will allow for the study of genetic diversity among different isolates.

Methods

DNA extraction: Cryptosporidium *parvum* oocysts were obtained from Bunchgrass Farm (Lot #22-20, shed date, 10/2/20). Purified oocysts (10^8) were washed in PBS and treated with diluted bleach for 10 minutes on ice to allow for sporozoite excystation. Parasites were pelleted, washed in PBS, and DNA was extracted using Utrapure™ phenol:chloroform:isoamyl alcohol (Thermo Scientific) followed by ethanol precipitation. Glycoblue™ co-precipitant (Thermo Scientific) was used to facilitate visualization of DNA during extraction and purification steps.

ONT Library preparation & sequencing

NEBNext FFPE DNA Repair Mix was used to repair 620ng of genomic DNA, which was then followed by end-repair and dA-tailing with NEBNext Ultra II reagents. The dA-tailed insert molecules were further ligated with an Oxford Nanopore adaptor via ligation kit SQK-LSK110. Purification of the library was carried out with AMPure XP beads (Beckman, Cat# A63880), the final library of 281ng was loaded to one PromethION 24 flow cell (FLO-PRO002) and the sequencing data was collected for 24 hours.

Illumina Library preparation & sequencing

DNA (100 ng) was sheared into fragments of approximately 300-400 bp in a Covaris E210 system (96 well format, Covaris, Inc. Woburn, MA) followed by purification of the fragmented DNA using AMPure XP beads. DNA end repair, 3'-adenylation, ligation to

Illumina multiplexing dual-index adaptors, and ligation-mediated PCR (LM-PCR) were all completed using automated processes. The KAPA HiFi polymerase (KAPA Biosystems Inc.) was used for PCR amplification (10 cycles), which is known to amplify high GC and low AT rich regions at greater efficiency. A fragment analyzer (Advanced Analytical Technologies, Inc) electrophoresis system was used for library quantification and size estimation. The libraries were 630 bp (including adapter and barcode), on average. The library was pooled with other internal samples with adjustment carried out to yield 3 Gbp of data on a NovaSeq 6000 S4 flow cell.

Genome size estimation

We used Jellyfish (version 2.3.0) to generate a k-mer based histogram of our raw reads in order to estimate the genome size based on our short read data. To obtain this we ran Jellyfish[30,31] with " jellyfish count -C -m 21 -s 1000000000 -t 10" and subsequently the "histo" module with default parameters. The obtained histogram was loaded into GenomeScope[30] given the appropriate parameter (k-mer size of 21) and haploid genome. GenomeScope provided the overall statistics across the short reads.

Assembly and polishing

We utilized Canu[21] (v2.0) for the assembly, which was based only on Nanopore pass data and a genome size estimate of 9Mbp. On the Nanopore pass reads, we also ran the assembly using Flye[22] (version 2.8.1-b1676) with the default parameters. Subsequently, we aligned the short reads using bwa mem (version 0.7.17-r1188) with -M -t 10 parameters. Samtools[32] (v1.9) was used to compress and sort the alignments. The so generated alignment was used by Pilon[33] (v 1.24) with the parameters "--fix bases " by correcting one chromosome after another of the raw assembly. This process was repeated two times achieving a high concordance of the reads and the long read assembly at the 2nd polishing step.

Busco assessment

We ran Busco[23] (v5.1.3) to assess the completeness of our assembly using default parameter and --auto-lineage, but lineage was detected to coccidia_odb10 (Creation date: 2020-08-05, number of genomes: 20, number of BUSCOs: 502). We reported the stats from the output summary file from Busco in this manuscript.

Telomere Identification

We used the sequence "TTTAGGTTTAGGTTTAGG" to identify telomeric sequences at the start and end of every contig from our assembly. To do so we used Bowtie[34] (version 1.2.3) to align the telomeric sequence back to the assembly with -a parameter. Subsequently we counted the matches across regions using a custom script. In short, we used 10kbp windows to count the number of reported hits, align the genome and compare the locations with the expected start/end locations. The identified regions were filtered for at least a 100 hits to guarantee a robust match. This way we counted the number of times each chromosome was listed.

Regional comparison

Two genetic markers 18S and gp60 were used to determine any significant gaps or mismatches against available GenBank genomes for Cryptosporidium *parvum*. The 18S and *gp60* coding regions downloaded from GenBank were aligned using ClustalW against the current assembly. For each gene, a similarity identity matrix table was created.

Additional Files

Supplemental Figure 1. Genomescope estimation of genome size
Supplemental Figure 2. NG50 comparison of the previous reference
Supplemental Figure 3. ClustalW alignment of the 18S coding sequence with the assembly.
Supplemental Figure 4. ClustalW alignment of the *gp60* coding sequence with the assembly.

Competing Interests

The corresponding author of the paper has presented at both ONT and PacBio sponsored conferences.

Funding

This work was supported by the National Institute of Allergy and Infectious Diseases (Grant#1U19AI144297).

Authors' Contributions

F.J.S and V.K.M : Conceptualization, Analysis and Writing-Original Draft Preparation
C.C and G.A.M : Conceptualization and Writing-Review & Editing
P.C.O : Conceptualization, Resources and Writing-Review & Editing
H.D.; Q.M. and D.M.M. : Conceptualization, Writing-Review & Editing
S.S.; S.B.; K.K.; G. W.; H.S.; V.V.; Y.H. : Methodology, Investigation
M.C.R.; K.L.H.; S.J.C. : Conceptualization
M.M.: Analysis
R.A.G.; J.F.P. : Conceptualization, Funding Acquisition

References:

1. Dong S, Yang Y, Wang Y, Yang D, Yang Y, Shi Y, et al.. Prevalence of Cryptosporidium Infection in the Global Population: A Systematic Review and Meta-analysis. *Acta Parasitol*. 65:882–92020;

2. Head MG, Brown RJ, Newell M-L, Scott JAG, Batchelor J, Atun R. The allocation of USdollar;105 billion in global funding from G20 countries for infectious disease research between 2000 and 2017: a content analysis of investments. *Lancet Glob Health*. 8:e1295–3042020;

3. Fayer R, Morgan U, Upton SJ. Epidemiology of Cryptosporidium: transmission, detection and identification. *Int J Parasitol*. 30:1305–222000;

4. Fayer R. Cryptosporidium: a water-borne zoonotic parasite. *Vet Parasitol*. 126:37–562004;

5. Xiao L, Feng Y. Molecular epidemiologic tools for waterborne pathogens Cryptosporidium spp. and Giardia duodenalis. *Food Waterborne Parasitol*. 8-9:14–322017;

6. : Parasites - Cryptosporidium (also known as “Crypto”). https://www.cdc.gov/parasites/crypto/index.html (2019). Accessed 2021 May 20.

7. Platts-Mills JA, Babji S, Bodhidatta L, Gratz J, Haque R, Havt A, et al.. Pathogen-specific burdens of community diarrhoea in developing countries: a multisite birth cohort study (MAL-ED). *Lancet Glob Health*. 3:e564–752015;

8. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, et al.. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *Lancet*. 382:209–222013;

9. Heo I, Dutta D, Schaefer DA, Iakobachvili N, Artegiani B, Sachs N, et al.. Modelling Cryptosporidium infection in human small intestinal and lung organoids. *Nat Microbiol*. 3:814–232018;

10. Cardenas D, Bhalchandra S, Lamisere H, Chen Y, Zeng X-L, Ramani S, et al.. Two- and Three-Dimensional Bioengineered Human Intestinal Tissue Models for Cryptosporidium. *Methods Mol Biol*. 2052:373–4022020;

11. Vinayak S, Pawlowic MC, Sateriale A, Brooks CF, Studstill CJ, Bar-Peled Y, et al.. Genetic modification of the diarrhoeal pathogen Cryptosporidium parvum. *Nature*. 523:477–802015;

12. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, et al.. Complete genome sequence of the apicomplexan, Cryptosporidium parvum. *Science*. 304:441–52004;

13. Dong L, Wang X, Guo H, Zhang X, Zhang M, Tang W. Chromosome-level genome assembly of the endangered humphead wrasse Cheilinus undulatus: Insight into the expansion of opsin genes in fishes. *Mol Ecol Resour*. 2021; doi: 10.1111/1755-0998.13429.

14. Brancaccio RN, Robitaille A, Dutta S, Rollison DE, Tommasino M, Gheit T. MinION nanopore sequencing and assembly of a complete human papillomavirus genome. *J Virol Methods*. 294:1141802021;

15. Espiritu HM, Mamuad LL, Jin S-J, Kim S-H, Lee S-S, Cho Y-I. High quality genome sequence of Treponema phagedenis KS1 isolated from bovine digital dermatitis. *Hanguk Tongmul Chawon Kwahakhoe Chi*. 62:948–512020;

16. Cuscó A, Pérez D, Viñes J, Fàbregas N, Francino O. Long-read metagenomics retrieves complete single-contig bacterial genomes from canine feces. *BMC Genomics*. 22:3302021;

17. Sun F, Sun S, Ye W, Duan C, Li B, Shan W, et al.. Genome Sequence Data of three formae speciales of Phytophthora vignae Causing Phytophthora Stem Rot on different Vigna species. *Plant Dis*. 2021; doi: 10.1094/PDIS-11-20-2546-A.

18. De Coster W, Weissensteiner MH, Sedlazeck FJ. Towards population-scale long-read sequencing. *Nat Rev Genet*. 2021; doi: 10.1038/s41576-021-00367-3.

19. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet*. 19:329–462018;

20. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al.. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. 33:2202–42017;

21. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 27:722–362017;

22. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*. 37:540–62019;

23. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 31:3210–22015;

24. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics*. 32:3021–32016;

25. Liu C, Schroeder AA, Kapur V, Abrahamsen MS. Telomeric sequences of Cryptosporidium parvum. *Mol Biochem Parasitol*. 94:291–61998;

26. Widmer G, Sullivan S. Genomics and population biology of Cryptosporidium

species. *Parasite Immunol*. 34:61–712012;

27. Bouzid M, Tyler KM, Christen R, Chalmers RM, Elwin K, Hunter PR. Multi-locus analysis of human infective Cryptosporidium species and subtypes using ten novel genetic loci. *BMC Microbiol*. 10:2132010;

28. Widmer G, Lee Y, Hunt P, Martinelli A, Tolkoff M, Bodi K. Comparative genome analysis of two Cryptosporidium parvum isolates with different host range. *Infect Genet Evol*. 12:1213–212012;

29. Moxon ER, Lenski RE, Rainey PB. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Perspect Biol Med*. 42:154–51998;

30. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 11:14322020;

31. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 27:764–702011;

32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al.. The Sequence Alignment/Map format and SAMtools. Bioinformatics.

33. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al.. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 9:e1129632014;

34. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 9:357–92012;