scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A new chromosome-level genome assembly and annotation of *Cryptosporidium meleagridis*

Lasya R. Penumarthi[1,2], Rodrigo P. Baptista[1,2,3,6], Megan S. Beaudry[4,7], Travis C. Glenn[1,4,5] & Jessica C. Kissinger[1,2,6] ✉

Cryptosporidium **spp. are medically and scientifically relevant protozoan parasites that cause severe diarrheal illness in infants, immunosuppressed populations and many animals. Although most human *Cryptosporidium* infections are caused by *C. parvum* and *C. hominis*, there are several other human-infecting species including *C. meleagridis*, which are commonly observed in developing countries. Here, we annotated a hybrid long-read Oxford Nanopore Technologies and short-read Illumina genome assembly for *C. meleagridis* (*Cm*TU1867) with DNA generated using multiple displacement amplification. The assembly was then compared to the previous *C. meleagridis* (*Cm*UKMEL1) assembly and annotation and a recent telomere-to-telomere *C. parvum* genome assembly. The chromosome-level assembly is 9.2 Mb with a contig N50 of 1.1 Mb. Annotation revealed 3,919 protein-encoding genes. A BUSCO analysis indicates a completeness of 96.6%. The new annotation contains 166 additional protein-encoding genes and reveals high synteny to *C. parvum* IOWA II (*Cp*BGF). The new *C. meleagridis* genome assembly is nearly gap-free and provides a valuable new resource for the *Cryptosporidium* community and future studies on evolution and host-specificity.**

Background & Summary

Cryptosporidium is an apicomplexan protozoan parasite of global medical, scientific, and veterinary significance that can cause moderate-to-severe diarrhea in humans and animals[1]. It is the leading cause of waterborne disease outbreaks in the United States[2,3]. Though cryptosporidiosis occurs in both immunocompromised and immunocompetent individuals, it is especially severe in immunocompromised and elderly populations as well as in children, resulting in persistent infection, malnutrition, and, in some cases, death[3–5]. In 2019, the Global Burden of Disease study found 133,422 global deaths and an annual loss of 8.2 million disability-adjusted life years (DALYs) due to *Cryptosporidium*[6]. *C. meleagridis* is an avian and mammalian-infecting *Cryptosporidium* species that was first described in turkeys[7,8]. Human infections with *Cryptosporidium* are caused predominantly by *C. parvum* and *C. hominis*, but species such as *C. meleagridis* can also infect humans. In fact, *C. meleagridis* is the third most common human-infecting *Cryptosporidium* species following *C. parvum* and *C. hominis*[9]. Though generally less common, *C. meleagridis* infection has been reported to be as common as *C. parvum* in some parts of the world and can lead to death in rare cases[10,11].

Currently, 17 of the >30 reported *Cryptosporidium* species have assembled genome sequences. Twelve species have annotated genome sequences including *C. andersoni, C. bovis, C. canis, C. felis, C. hominis, C. meleagridis, C. muris, C. parvum, C. ryanae, C. tyzzeri, C. ubiquitum* and *C. sp. Chipmunk genotype*[12,13]. *Cryptosporidium* spp. have eight chromosomes and genome sizes of ~9 Mb. The only *C. meleagridis* genome sequence, strain UKMEL1 (*Cm*UKMEL1), contains gaps and is assembled into 57 contigs. Historically, it has been challenging to sequence the genome of *Cryptosporidium* parasites. Sustainable *in vitro* culture and cloning are not possible. Thus, sequencing a bulk population of parasites, when enough can be isolated, has been the preferred approach. Recently, a new method was implemented to generate genome sequences for *Cryptosporidium* using multiple

[1]Institute of Bioinformatics, University of Georgia, Athens, GA, 30602, USA. [2]Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, GA, 30602, USA. [3]Department of Medicine, Weill Cornell Medical College, New York City, NY, 10065, USA. [4]Department of Environmental Health Science, University of Georgia, Athens, GA, 30602, USA. [5]Department of Genetics, University of Georgia, Athens, GA, 30602, USA. [6]Present address: Division of Infectious Diseases, Houston Methodist Research Institute, Houston, TX, 77030, USA. [7]Present address: Genomics, Daicel Arbor Biosciences, Ann Arbor, MI, 48103, USA. ✉e-mail: jkissing@uga.edu

Statistics	CmTU1867	CmUKMEL1[34]	CpBGF[15]
# of contigs	13	57	8
Largest contig (bp)	1,365,597	732,862	1,379,419
# T2T chromosomes	0	0	8
Total length (bp)	9,178,485	8,973,200	9,259,183
N50 (bp)	1,105,563	322,908	1,107,426
GC (%)	30.9	31.0	30.04
# N's per 100 kbp	0	0	0
# of telomeres identified	1	10	16

Table 1. Statistics of the *C. meleagridis and Cp*BGF genome assemblies.

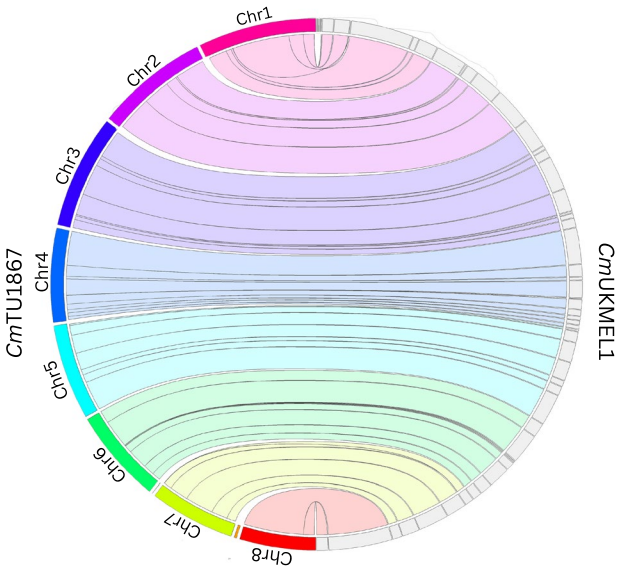

Fig. 1 DNA synteny plot mapping the contigs of *Cm*UKMEL1 to the eight chromosome-level contigs of *Cm*TU1867. Jupiterplot between the previous *Cm*UKMEL1 genome sequence and the new *Cm*TU1867 genome sequence. Ribbons are colored with respect to the reference *Cm*TU1867 chromosome.

displacement amplification, a whole genome amplification (WGA) approach. It was tested on 10 ng of genomic DNA from *C. meleagridis* strain TU1867 (*Cm*TU1867) which provided sufficient DNA for library construction and generation of a high-quality genome sequence using Oxford Nanopore Technologies (ONT) long-read sequencing[14]. Here we share a chromosome-level assembly and reannotation of the *C. meleagridis* genome.

Whole genome sequencing and assembly. The newly generated *Cm*TU1867 genome assembly contains an additional 201,275 base pairs (bp) of sequence relative to *Cm*UKMEL1. The largest contig in the new assembly is 632,735 bp longer than the largest contig in *Cm*UKMEL1. We also note a larger N50 value in the new assembly (Table 1). For comparison, a recent telomere-to-telomere, T2T, genome assembly for the closely related and highly syntenic species, *C. parvum* IOWA II (*Cp*BGF) is provided[15]. This high-quality *C. meleagridis* genome assembly results from a new experimental approach designed to help generate long-read genome sequences from limiting quantities of genomic DNA and is an important resource that will facilitate our understanding of *Cryptosporidium* evolution and host specificity.

The initial genome assembly contained 8 chromosomes and 5 contigs ranging in size from 681–30,300 bp, 2 of which were later identified as contamination and removed. Two additional contigs were manually created ("contig_10" and "contig_11") from the beginnings of chromosome 2 and chromosome 6 due to detection of assembly artifacts in these chromosomes. The final assembly contains 8 chromosomes and contigs 9–13 (Table 1). Contig_9 and contig_13 have regions of sequence identical to parts of chromosomes 1 and 3, respectively, but assembled separately from the chromosomes. The chromosomes of *C. meleagridis* are numbered and oriented according to their homology with the highly syntenic *C. parvum*. The new assembly is highly syntenic to the previous *Cm*UKMEL1 assembly at the nucleotide level (Fig. 1).

In comparison to *Cm*UKMEL1, the new *Cm*TU1867 assembly lacks telomeres, except for chromosome 5, which has one assembled telomere. A search of the ONT long-reads, revealed several reads with telomere sequences that did not assemble. Though these reads did not assemble, regions of the read that did not contain the telomere pattern matched unique sequences in the assembled chromosomes. By mapping these reads back to the genome assembly, we identified three additional telomeres that could be placed manually at the 5′ and 3′

	Protein-encoding sequences	Average CDS length	tRNA	18SrRNA	5.8SrRNA	28SrRNA	5SrRNA
*Cm*TU1867	3,919	1,824	45	5	5	5	6
*Cm*UKMEL1	3,753	1,885	45	0	0	1	7
*Cp*BGF*	3,923	2,145	45	5	0	5	5

Table 2. Annotated genes and RNAs in *Cm*BEI, *Cp*BGF, and *Cm*UKMEL1. *Protein totals do not include putative protein isoforms related to alternative splicing.

ends of chromosome 3, and the 5′ end of chromosome 4. At least 4 telomere-containing long-reads mapped to these regions with at least 1 long (>1 kb) read that extended into unique regions of the chromosome. However, due to the low number of reads in support of these telomeres, we did not extend the ends of chromosomes in the assembly with these telomere-containing reads.

Genome annotation. The new *Cm*TU1867 genome assembly was annotated using the previous *Cm*UKMEL1 annotation, a recent *Cp*BGF annotation, orthology analysis, and *de novo* gene prediction. Gene expression data for *Cm*TU1867 are not available to assist with the annotation, so UTRs are not predicted. Annotation of *Cm*TU1867 reveals 166 additional protein-encoding genes and numerous additional ribosomal RNAs (Table 2).

A comparison of the synteny of the protein-encoding genes and rRNAs between *Cm*TU1867 and *Cp*BGF revealed highly syntenic chromosomes (Fig. 2). The new *Cm*TU1867 genome sequence has 16 additional ribosomal RNA genes compared to *Cm*UKMEL1 (Table 2). The 5 small, 5.8S rRNA units are found on chromosomes 1, 2, 7, 8 (Fig. 2). The six 5S rRNAs in *Cm*TU1867 are in 2 clusters of 3, on chromosome 3 and 7 (Fig. 2). In *Cp*BGF the cluster of 5S rRNAs on chromosome 3 contains 2 rRNAs whereas in *Cp*IOWA-ATCC and *Cm*TU1867, the cluster of 5S rRNAs on chromosome 3 contains 3 rRNAs. These patterns may arise because of variation in the copy number of the 5S rRNA within a population of parasites or among different species of *Cryptosporidium* or compressions during genome assembly. When *Cm*TU1867 reads were mapped to the assembly at regions where there are 5S rRNA clusters in chromosomes 3 and 7, we saw relatively even coverage throughout the region. However, *Cp*BGF shows 2-3X read compression at this locus on chromosome 3 and 7 suggestive of population variation. One of the unassembled contigs, contig_9, has an additional 18S/28S cluster. However, due to the fact that we are not able to find a chromosomal location for it despite long-read sequencing and since our sample is not clonal, we do not have sufficient evidence to conclude its status.

While annotating, we noticed several genes that encoded a single long protein in *Cm*UKMEL1 but were annotated as two distinct genes in *Cp*BGF. Upon investigation, we discovered that these gene annotations vary in size in several *Cryptosporidium* species. In *Cm*TU1867, we have kept the long protein annotation when it is observed. There are 20 cases where the single long protein in *C. meleagridis* does not appear to exist as a single open reading frame in *Cp*BGF, (Table 3). A lack of RNAseq evidence for *C. meleagridis* makes it challenging to validate the existence of these long open reading frames whereas *C. parvum* has a large quantities of expression data available. We made a note in the submitted *Cm*TU1867 annotation if the gene is annotated as two or three distinct genes in other species. Two of the 20 *Cm*TU1867 proteins are annotated as 3 distinct proteins in *Cp*BGF (Table 3).

Interestingly, each of the 5 annotated 18S and 28S rRNAs has a putative protein-encoding gene within it (Table 4). Our submitted annotation does not contain any of these putative ORFs as their presence would be so unusual it cannot be accepted by the NCBI GenBank. However, we note, they may exist. The 18S rRNA genes encode a putative intron-encoded homing endonuclease. While we detect the presence of this putative protein, we do not detect an intron in the 18S rRNA. The six putative homing endonuclease protein sequences in the 18S rRNAs are not identical due to a guanine deletion at position 1061 in two of the five 18S rRNAs (Chr1 and Chr7). This results in a premature stop codon in three of the putative homing endonuclease sequences (Fig. 3). This indel is likely due to an ONT homopolymer sequencing artifact. BLASTp searches of other *Cryptosporidium* species revealed annotations of this gene in *C. ubiquitum* and *C. felis*. We note that annotation in other species does not make these genes real, only proteomics can confirm them, thus they are not included in our submitted annotation.

The six putative senescence associated proteins encoded in the 28S rRNAs are identical. This protein is found in BLASTp searches in *C. hominis* TU502, *C. canis*, *C. ubiquitum*, and *C. muris*. This protein has an ART2/RRT15 domain according to InterPro. As was the case with the putative intron homing endonuclease in the 18S genes above, given the location, we have not included these proteins in the submitted annotation due to a lack of evidence for their existence.

Comparison with previous assemblies. The annotation was assessed by comparing the *Cm*TU1867, *Cm*UKMEL1 and *Cp*BGF protein-coding sequence gene content using orthology-based algorithms. Several putative species-specific single copy genes were identified (Table 5). We identified 23 species-specific genes in *Cm*TU1867, 11 in *Cm*UKMEL1, and 39 in *Cp*BGF (Table 5). This finding makes sense because CpBGF, which is a T2T assembly, is the most complete of the three assemblies and *Cm*TU1867 is a more complete genome assembly than *Cm*UKMEL1. To assess whether species-specific genes were located in sub-telomeric regions, the first and last 25 genes of each chromosome were assessed for the presence of species-specific genes. We observe that the putative species-specific genes are not enriched in sub-telomeric regions (bolded gene names in Table 5), rather, they are scattered throughout the genome. If real, the evolutionary origin of these genes is intriguing. However,

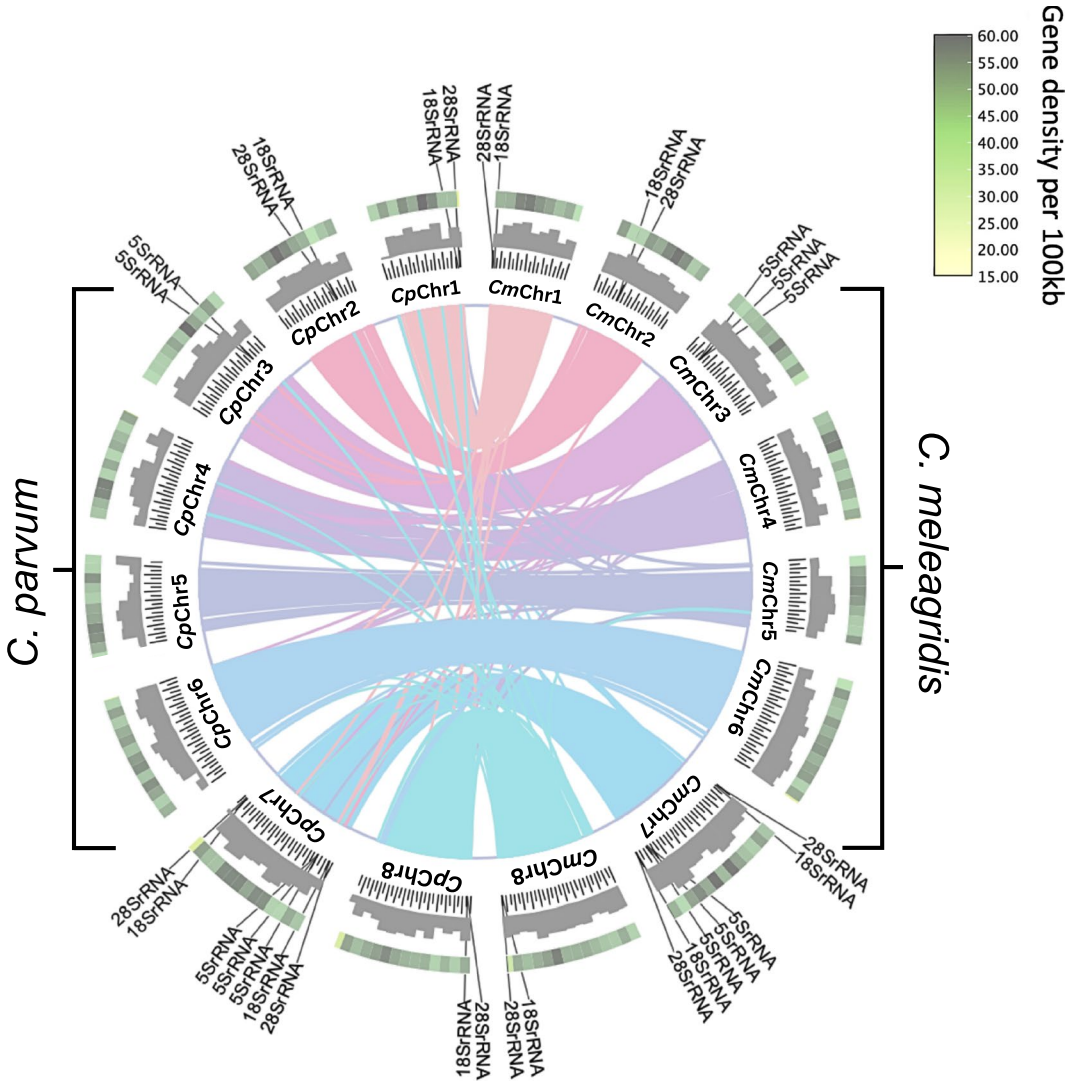

Fig. 2 Protein synteny analysis of the eight chromosome-level contigs of *Cm*TU1867 and *Cryptosporidium parvum, Cp*BGF. Circos plot rings, moving from the center to the exterior illustrate shared ortholog clusters between *Cm*TU1867 and *Cp*BGF, number of base pairs in 50,000 bp increments, GC content histogram, and gene density. Locations of rRNA genes are as indicated.

these results are derived from as of yet, incomplete genome assemblies for *C. meleagridis* and require further validation.

All orthogroups (multiple shared derived genes – orthologs or paralogs) as opposed to the single-copy genes in Table 5, that were not shared by all three genome assemblies were investigated. Some of the orthogroups fall at the ends of chromosomes in *C. parvum* that extended beyond the ends of the *Cm*TU1867 and *Cm*UKMEL1 chromosomes. Other times they were unannotated in one species or the other but present in the syntenic genome sequence region(s). When we found unannotated proteins that were not initially detected in *Cm*TU1867, we manually added these genes to the submitted annotation. Ultimately, we found very few orthogroups that were unique to a subset of species (Fig. 4). The manual validation of the orthogroups is presented in Supplementary Table 1.

Methods

Whole genome sequencing and assembly. *C. meleagridis* isolate TU1867 genomic DNA was obtained from BEI Resources (cat. number NR-2521 ATCC, Manassas, VA). A total of 10 ng of *C. meleagridis* DNA was amplified through whole genome amplification using multiple displacement amplification (MDA), followed by T7 endonuclease debranching yielding 400 ng debranched DNA[14]. Following sequence generation and assembly, polishing and annotation proceeded as in (Fig. 5). The sequence was polished with existing reads from the NCBI GenBank Sequence read archive accession SRR793561[16].

ONT library preparation used the SQK-RBK004 Rapid Barcoding Sequencing Kit (Oxford Nanopore Technologies, Oxford, UK) as per the manufacturer's instructions. Sequencing was performed on an ONT MinION device with R9.4.1 flow cells and bases were called by guppy v.6.4.2 using the high-accuracy base

Chr	Gene ID in *Cm*TU1867	Protein in *Cm*TU1867	Gene IDs in *Cp*BGF
1	cmbei_100150	Uncharacterized secreted protein (SKSR gene family)	cpbgf_100150, cpbgf_100160
	cmbei_100730	Glutamine cyclotransferase domain containing protein	cpbgf_100730, cpbgf_100733
	cmbei_1002800	Methyltransferase TRM13, MED7, Zinc finger domain-containing protein	cpbgf_1002800, cpbgf_1002810
2	cmbei_20010	SFI domain containing protein	cpbgf_20010, cpbgf_200470
3	cmbei_3002310	RNA recognition motif and AAA-type ATPase core domain containing protein	cpbgf_3002310, cpbgf_3002300, cpbgf_3002290
	cmbei_3002700	Transport protein particle (TRAPP) domain containing protein	cpbgf_3002700, cpbgf_3002706
4	cmbei_4002100	PIG-A GPI anchor and glucosyltransferase domain containing protein	cpbgf_4002100, cpbgf_4002093
	cmbei_4002180	Peptidase A1 and Dpy-19/Dpy19-like domain-containing protein	cpbgf_4002180, cpbgf_4002190
5	cmbei_500340	Signal peptide containing protein	cpbgf_500340, cpbgf_500350
	cmbei_500470	Peptidase S9, prolyl oligopeptidase, catalytic domain containing protein	cpbgf_500470, cpbgf_500466
	cmbei_5002280	Signal peptide and transmembrane domain containing protein	cpbgf_5002280, cpbgf_5002290
	cmbei_5002830	Vacuolar protein sorting-associated protein 13 domain containing protein	cpbgf_5002830, cpbgf_5002840
	cmbei_5004500	Vacuolar protein sorting-associated protein 13 domain containing protein	cpbgf_5004500, cpbgf_5004490, cpbgf_5004480
	cmbei_5003110	AAA + ATPase and VWFA domain containing protein	cpbgf_5003110, cpbgf_5005540
6	cmbei_600540	Serine/threonine protein kinase domain containing protein	cpbgf_600540, cpbgf_600530
	cmbei_6001250	Uncharacterized protein	cpbgf_6001250, cpbgf_6001260
	cmbei_6002100	Uncharacterized protein	cpbgf_6002100, cpbgf_6002110
	cmbei_6002140	Potassium channel domain containing protein	cpbgf_6002140, cpbgf_6002143
7	None		
8	cmbei_800690	Signal peptide containing putative Formin J protein	cpbgf_800690, cpbgf_800680
	cmbei_8002510	Putative cyclin dependent kinase	cpbgf_8002510, cpbgf_8002500

Table 3. Single large, annotated genes in *Cm*TU1867 that are annotated as two or three distinct sequential genes in *Cp*BGF and/or other *Cryptosporidium* spp.

rRNA	Chr	Encoded rRNA ORF	ORF start	ORF stop	Strand
18S	1	Putative intron encoded homing endonuclease	15,894	16,160	–
	2	"	303,089	303,412	+
	7	"	22,619	22,885	–
	7	"	1,294,026	1,294,349	+
	8	"	1,350,046	1,350,369	+
28S	1	Putative senescence associated protein	11,232	11,552	+
	2	"	307,697	308,017	–
	7	"	17,956	18,276	+
	7	"	1,298,636	1,298,956	–
	8	"	1,354,656	1,364,976	–

Table 4. Putative ORFs (not submitted in the *Cm*TU1867 GenBank record) encoded with the 18S and 28S rRNAs in *Cm*BEI and their coordinates.

call model. The long-read fastq reads were assembled using Flye v.2.8.2[17] with the–nano-raw option and -g 9 m. The draft long-read genome assembly was polished with PolyPolish v.0.5.0[18] using default parameters to increase the accuracy of the base calls by using *C. meleagridis* strain TU1867 Illumina sequences (NCBI accession SRX253214) generated elsewhere. Intermediate files needed for PolyPolish were generated using BWA v.0.7.17[19]. The resulting contigs were ordered and oriented to match the *Cp*IOWA-BGF T2T genome assembly[15] (GCA_035232765.1), called *Cp*BGF in this manuscript, using AGAT[20] v. 1.1.0 PERL script agat_sq_reverse_complement.pl, GenomeTools[21], and the progressive Mauve alignment v 1.1.3[22] in Geneious Prime v 2023.2.1[23]. Contamination was detected by searching the NCBI nr database using BLAST[24] (BLASTx default parameters) and FCS-GX[25] (Fig. 5). Contaminant contigs were removed from further analysis. Telomeres were identified, as in *Cp*BGF[15] using the telomere-locating python script FindTelomeres to find the *Cryptosporidium*

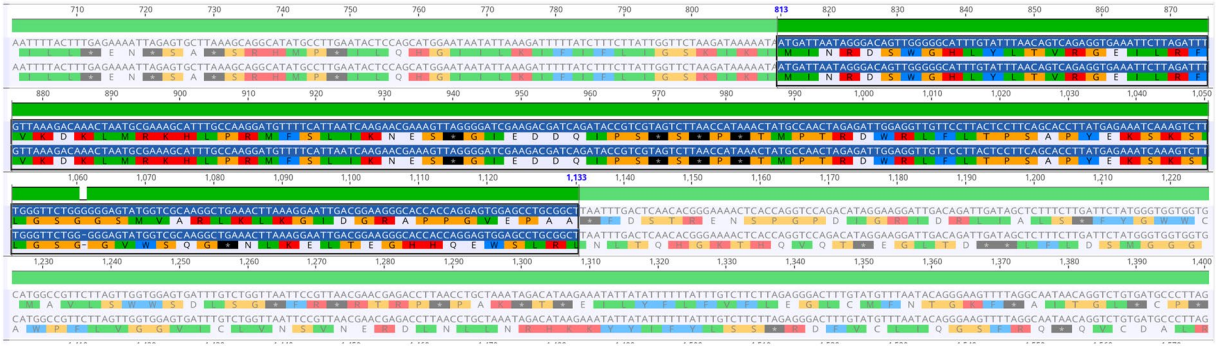

Fig. 3 Portion of the 18SrRNA gene sequence and the putative ORF contained within it. Multiple sequence alignment of 18S rRNAs representing the guanine SNV and the putative ORFs contained in these sequences.

Species and Strain	Singleton IDs*	Coordinates	Singleton IDs	Coordinates
Cryptosporidium meleagridis TU1867	**cmbei_100110**	41716–42914	**cmbei_7005530**	24550–24696
	cmbei_1002880	642026–642454	cmbei_7002513	606698–607360
	cmbei_2002120	429587–429678	cmbei_7004665	1076111–1076549
	cmbei_2003280	683302–683667	cmbei_7005006	1168993–1169022
	cmbei_300370	104263–104549	cmbei_7005103	1195849–1197118
	cmbei_3001410	398831–399157	cmbei_8001763	493894–494274
	cmbei_3002380	622418–622869	cmbei_8004050	1022925–1023056
	cmbei_3002720	682534–682922	**cmbei_8005515**	1362994–1364327
	cmbei_40023	1100198–1100721	**cmbei_8005530**	1364036–1364251
	cmbei_4004070	104218–104513	cmbei_9000040	12454–13368
	cmbei_6004980	1164441–1164829	cmbei_120010	596–1065
	cmbei_6004993	1166119–1166454		
Cryptosporidium meleagridis UKMEL1	CmeUKMEL1_00600 (JIBK01000002)	283378–283705	CmeUKMEL1_10305 (JIBK01000034)	58–677
	CmeUKMEL1_01660 (JIBK01000003)	23133–23597	CmeUKMEL1_14005 (JIBK01000048)	8–503
	CmeUKMEL1_06150 (JIBK01000010)	229933–231820	CmeUKMEL1_15125 (JIBK01000048)	526051–526433
	CmeUKMEL1_06155 (JIBK01000011)	204–1517	CmeUKMEL1_16475 (JIBK01000050)	54883–55439
	CmeUKMEL1_08805 (JIBK01000020)	259–555	CmeUKMEL1_17330 (JIBK01000051)	75631–76029
	CmeUKMEL1_09820 (JIBK01000027)	33768–33989		
Cryptosporidium parvum IOWA BGF	**cpbgf_100150**	83322–86482	cpbgf_600530	120556–122982
	cpbgf_100733	222029–222922	cpbgf_6001250	299808–301409
	cpbgf_1002800	655776–656402	cpbgf_5004500	721443–722609
	cpbgf_1002887	672381–674544	cpbgf_5004480	730965–737726
	cpbgf_1002880	672573–674537	cpbgf_5005540	925181–931260
	cpbgf_20010	101359–102994	cpbgf_4002090	1100521–1101041
	cpbgf_2003690	831932–832447	cpbgf_6002110	501537–505218
	cpbgf_300370	106882–107196	cpbgf_6002143	511286–512806
	cpbgf_3001410	404488–404820	cpbgf_6004100	952703–953008
	cpbgf_3002290	608641–609101	cpbgf_6004980	1179722–1180105
	cpbgf_3002310	611727–612195	**cpbgf_7005800**	25041–26437
	cpbgf_3002700	684838–685643	cpbgf_7002513	631350–632023
	cpbgf_4004070	106195–106509	cpbgf_7004665	1102082–1102518
	cpbgf_4002093	263323–263861	cpbgf_7005103	1222657–1224777
	cpbgf_4002190	282619–284271	cpbgf_8001763	473062–473454
	cpbgf_40023	1100198–1100721	cpbgf_8002510	656383–657238
	cpbgf_500470	140861–141394	cpbgf_8003056	790111–790462
	cpbgf_5002280	513213–513842	cpbgf_8003550	895329–896309
	cpbgf_5002830	644601–651482	cpbgf_8004050	1104236–1004847

Table 5. Putative species-specific genes identified in *Cm*TU1867, *Cm*UKMEL1 and *Cp*BGF. *Gene IDs in bold are presumed subtelomeric genes because they exist within the first or last 25 annotated genes on a chromosome (non-chromosomal contigs, i.e *Cm*UKMEL1, were excluded from this analysis since the ends of chromosomes are unknown).

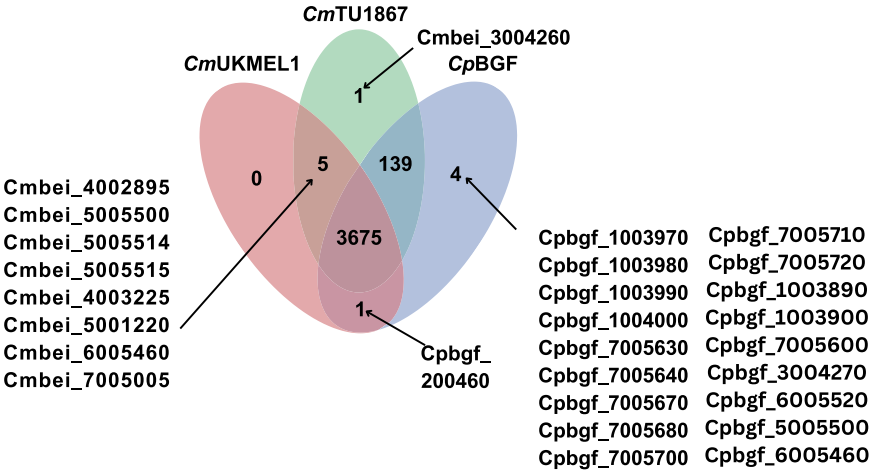

Fig. 4 Venn diagram of orthogroup search results following manual validation. Orthogroup comparison among the new *Cm*TU1867, the previous *Cm*UKMEL1, and the newly released reference genome, *Cp*BGF. See Fig. 6 for the pre-validation results. Arrows link the list of gene IDs found in the smaller orthogroups that are unique to a species or shared by two species.

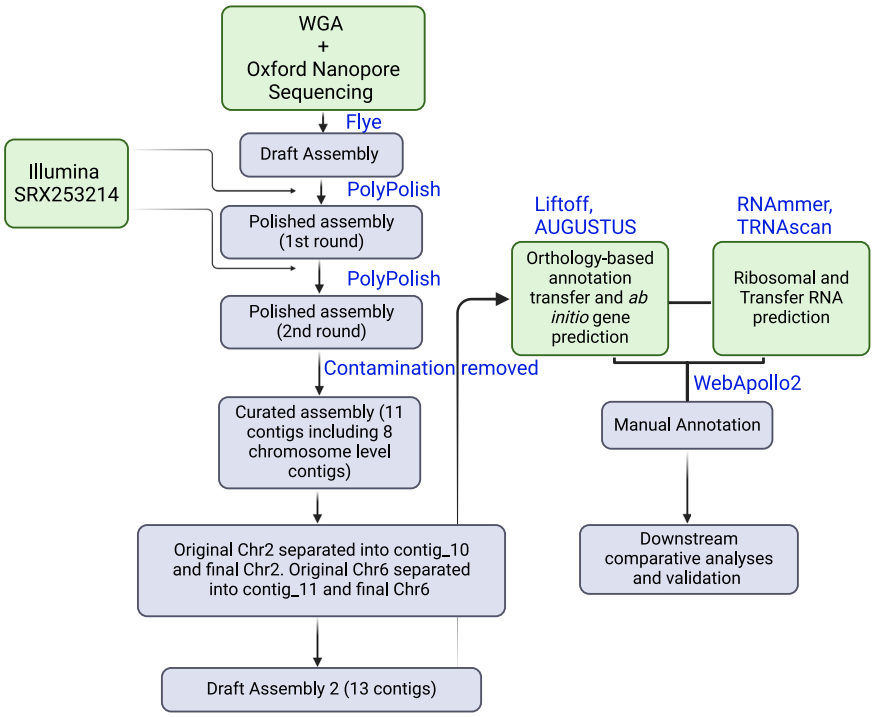

Fig. 5 Experimental workflow for genome sequencing, assembly, annotation, and validation. Bioinformatics workflow for assembly and annotation of the DNA derived from *Cm*TU1867 WGA. Green boxes represent initial steps as well as new data used for parts of the pipeline and blue boxes represent subsequent downstream analyses of the data generated. The illustration was generated in BioRender.

telomere repeat 5'-CCTAAA-3' and its complement at the ends of assembled contigs (https://github.com/JanaSperschneider/FindTelomeres). The unassembled ONT long-reads were also searched for this telomere repeat with FindTelomeres and reads with telomeres were mapped back to the genome assembly. Read-mapping to the whole genome assembly was performed using minimap2 v.2.26 with the option –secondary = no to prevent multi-mapping. Genome statistics were generated using the GenomeTools v.1.6.2[21] programs gt stat and gt seqstat. AGAT v.1.1.0[20] PERL scripts agat_sq_stat_basic.pl and agat_sp_statistics.pl were used to generate statistical information with default parameters.

Genome annotation. Tracks for manual annotation were generated using a local Apollo2 server[26] using two approaches: (1) an orthology based annotation transfer using the tool Liftoff[27] and (2) an *ab initio* gene

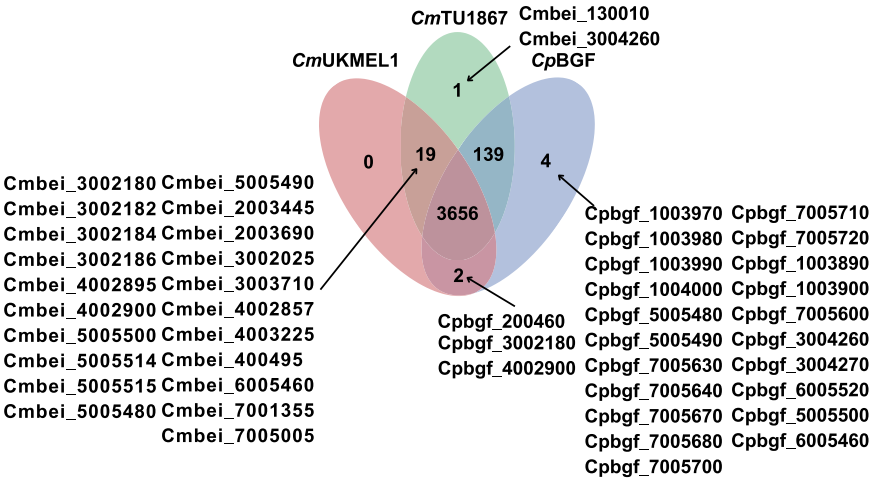

Fig. 6 Ortholog search results shown in a Venn diagram. Orthogroup comparison among the new *Cm*TU1867, the previous *Cm*UKMEL1, and *Cp*BGF prior to validation and correction. Arrows link the list of gene IDs found in the smaller orthogroups that are unique to a species or shared by two species.

prediction using Augustus[28] trained with *C. parvum* IOWA-ATCC[12] (GCA_015245375.1) and *Cm*UKMEL1 (GCA_001593445.1) protein sequences from CryptoDBv.50[29]. Annotation Liftoff tracks were created from the current *Cm*UKMEL1, *Cp*BGF, and *Cp*IOWA-ATCC annotated genes with the -copies flag to look for extra gene copies. In situations where AUGUSTUS and Liftoff gene structures disagreed, the conflicting gene models were searched using BLASTp in CryptoDB to check for the gene structure that was most abundant in existing annotations. As there are no available RNA-seq data for *C. meleagridis* there is no evidence to confirm gene predictions and annotate UTRs. Tracks for prediction and manual annotation of rRNAs were created using barrnap[30] with the parameters–kingdom euk–outseq Cmel_barrnap.fasta–evalue 1e-06–lencutoff 0.8 Cmel_genome.fasta. TRNAscan 2.0[31] was used to predict tRNAs using default parameters. Functional annotation was generated with Blast2GO[32] (using BLASTp, the nr database, word size 5, and e-value 1e-5) and compared with results from the reference T2T *Cp*BGF genome functional annotation. Edits to the *Cm*TU1867 gff file gene names were performed with basic bash and awk commands. InterPro[33] was used for classification of protein families investigated in Table 3 and for the genes encoded within rRNAs.

Comparative genomics. A comparison of orthologous genes between the new *C. meleagridis* assembly and the previous *C. meleagridis* assembly[34] was completed using Orthofinder v2.5.5 which was run in a conda environment using the latest Anaconda release[35] (2024.02-1) with default parameters (latest Diamond algorithm[36]) and visualized using OrthoVenn3[37]. Figure 4 represents the orthology results following extensive manual validation (Fig. 6 and Supplementary Table 1) of each orthogroup difference. Manual analyses utilized BLASTp searches of both NCBI and CryptoDB[29]. Orthology, genome, and rRNA comparisons were created using Circos[38]. The configuration file for Circos was created following the Circos documentation and run using the command: circos -conf config_file.conf. Additionally, TBTools[39] was used to visualize the circos plot. In TBTools, the "Advanced Circos" feature was selected. The ChrLen File and the Links File were generated manually following the Circos format. The rRNA features were added to the plot in the "Set Input Genome Feature List" option on TBTools. Gene density and GC content were created in TBTools following the TBTools documentation with default parameters. The genome comparisons between *Cm*UKMEL1 and *Cm*TU1867 were created using JupiterPlot[40]. The JupiterPlot was made using the command: jupiter name = $prefix ref = $reference fa = $scaffolds where the ref variable is set to the reference genome in FASTA format and the fa variable is set to the set of scaffolds in FASTA format. The general parameters were set to the default (t = 4) and karyotype options were slightly modified (m = 10000, ng=0, i=0, g = 1, gScaff=100000, labels = ref). Link options followed the default options (maxGap = 100000, minBundleSize = 50000, MAPQ = 50, linkAlpha=5). *Cm*TU1867 long reads were mapped back to contig regions containing 5S rRNA clusters using minimap2 with --secondary = no to account for multi-mapping. The raw orthogroups were analyzed and validated extensively to create the final Venn diagram (Fig. 4). Any protein that was syntenic to proteins in the other two genome sequences was moved into an orthogroup with those proteins. Proteins that were found in the sequence of the other two genome sequences but were not annotated in one or more genome(s) were also moved into an orthogroup and added to the Venn diagram.

Data Records

The genomic sequence, reads SRR27282542[41], and metadata for the *Cryptosporidium meleagridis* TU1867 strain have been deposited in the National Center for Biotechnology Information, NCBI GenBank under BioProject accession number PRJNA1022047[42]. This whole genome shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JBCHVM000000000[43]. The version described in this paper contains WGS scaffolds JBCHVM010000001-JBCHVM0100000013.

Technical Validation

*Cm*TU1867 assembly completeness was evaluated using the Benchmarking Universal Single-Copy Orthologs (BUSCO) software v.5.5.0[44] to search against apicomplexan databases (apicomplexa_odb10) which contain 446 orthologous single-copy genes in total. The results showed an overall completeness score of 96.6% (n = 446). Of these, 430 (96.4%) single-copy genes were retrieved of which 1 (0.224%) was duplicated. These results indicate high completeness of the genome assembly.

Further analysis of the assembly and annotated protein encoding regions utilized an orthology comparison of *Cm*TU1867, *Cm*UKMEL1, and *Cp*BGF with the OrthoFinder algorithm and the results were visualized in OrthoVenn3 as described in the methods (Fig. 5). Orthogroups belonging to *Cm*TU1867 only, *Cp*BGF only, *Cm*UKMEL and *Cp*BGF only, and *Cm*UKMEL1 and *Cm*TU1867 only were extensively analyzed (Supplementary Table 1). Several genes found only in *Cp*BGF were shown to be subtelomeric in both *Cm*UKMEL1 and *Cm*TU1867 and thus likely missing from the incomplete chromosome ends of CmTU1867. Several genes encoding short <100 amino acid proteins found in both *Cm*TU1867 and *Cm*UKMEL1 exist in *Cp*BGF but are unannotated. Following these analyses, a new Venn diagram (Fig. 4) was created that represents the revised, validated findings.

We additionally found one putative open-reading frame (ORF) predicted by AUGUSTUS in *Cm*TU1867 Chr4 region 828518–828700 bp that was not in *Cm*UKMEL1, *Cp*BGF, or any other species according to BLASTp and BLASTn searches. We removed this putative gene from our annotation since we could not validate it with RNAseq data or orthology, but it may be a gene unique to *C. meleagridis* detected by the improved assembly.

Code availability

Pipelines and code involved in processing the data were executed by following the respective manuals of the bioinformatics software programs used. A custom script was generated to convert OrthoFinder output into the ClusterVenn input format on OrthoVenn3.

Genome pre-assembly parameters:

1. **Guppy**
   guppy_basecaller –i./fast5_pass –s./guppy_out –c dna_r9.4.1_450bps_hac.cfg --num_callers 2 --cpu_threads_per_caller 1

Assembly and gene calling parameters:

1. **Flye**
   flye --nano-raw ../Cmel.fastq -o Cmel_flye -g 9 m

2. **PolyPolish**
   bwa index polished_1.fasta
   bwa mem -t 16 -a polished_1.fasta SRR793561_1.fastq.gz > alignments_1.sam
   bwa mem -t 16 -a polished_1.fasta SRR793561_2.fastq.gz > alignments_2.sam
   polypolish_insert_filter.py --in1 alignments_1.sam --in2 alignments_2.sam --out1 filtered_1.sam --out2 filtered_2.sam
   polypolish polished_1.fasta filtered_1.sam filtered_2.sam > polished_2.fasta

3. **AGAT: agat_sq_reverse_complement.pl (for reorienting annotations)**
   agat_sq_reverse_complement.pl --gff Cmel_annotations_Chr4_Chr6.gff3 --fasta Cmel_genome.fasta -o Cmel_annotations_reoriented.gff3

4. **GenomeTools (for validating after reorienting chromosomes)**
   gt gff3validator Cmel_annotations_reoriented.gff3

5. **BLASTx**
   blastx -query Cmel_genome.fasta -db nr -out results_blastx.txt -outfmt 6 -evalue 1e-5 -num_threads 4

6. **FCS-GX**
   python $EBROOTNCBIMINFCS/fcs.py screen genome --fasta Cmel_genome.fasta --out-dir./fcs_gx_out/ --gx-db "$GXDB_LOC/gxdb" --tax-id 93969

7. **FindTelomeres.py (on the reads)**
   https://github.com/JanaSperschneider/FindTelomeres
   python FindTelomeres_Crypto_Repeat.py Cmel_pool.fasta

8. **Minimap2**
   minimap2 -a --secondary = no Cmel_genome.fasta reads_with_telomeres.fastq > telomere_map.sam

9. **GenomeTools (for statistics)**
   gt stat Cmel.gff3

10. **AGAT: agat_sq_stat_basic.pl and agat_sp_statistics.pl (for statistics)**
    agat_sq_stat_basic.pl -i Cmel.gff3
    agat_sp_statistics.pl -gff Cmel.gff3

11. **AUGUSTUS**
    perl ~/Augustus/scripts/autoAugTrain.pl --cpus=10 --trainingset=CryptoDB-64_CparvumIOWA-ATCC_AnnotatedProteins.fasta --species=trained_species -g=genome.fa --workingdir=./autoAug --optrounds = 1
    augustus --gff3=on --stopCodonExcludedFromCDS=false --species=trained_species–softmasking=0–AUGUSTUS_CONFIG_PATH=./augustus/config --strand=both–genemodel=partial Cmel_genome.fasta>augustus.gff

12. **Liftoff**
    liftoff -g CpBGF.gff3 -o bgf -infer_genes -infer_transcripts -polish Cmel_genome.fasta bgf.fa

13. **Blastp in CryptoDB (default)**
    https://cryptodb.org/cryptodb/app/workspace/blast/new
    Expectation value: 10; Max Target Sequences: 100; Max matches in query range: 0; Word Size: 6; Scoring Matrix: BLOSUM62; Gap Costs (Open/Extension): 11,1; Compositional adjustments: Conditional compositional score matrix adjustment Low complexity regions: no filter; Mask for lookup table: false; Mask lower case letters: false

14. **Barrnap**
    barrnap --kingdom euk –outseq Cmel_barrnap.fasta --evalue 1e-06 --lencutoff 0.8 Cmel_genome.fasta

15. **TRNAscan**
    tRNAscan-SE Cmel_genome.fasta

16. **Blast2GO**
    Installed Blast2GO locally
    Parameters: BLASTp, the nr database, word size 5, and e-value 1e-5

17. **InterPro (online website used – default)**
    https://www.ebi.ac.uk/interpro/
    All Member databases selected

Comparative genomics parameters:

1. **OrthoFinder in Anaconda (raw figure in technical validation):**
    source activate /orthology_analysis/env-ortho
    conda install -c bioconda diamond
    conda install orthofinder
    orthofinder -f FASTAS

2. **Blastp NCBI (default):**
    https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins
    Database: nr; Max target sequences = 100; Short queries (checked); Expect threshold: 0.05; Word size = 5; Max matches in a query range: 0; Matrix BLOSUM62; Gap Costs: Existence: 11 Extension: 1; Compositional adjustments: Conditional compositional score matrix adjustment

3. **Blastp CryptoDB (default):**
    https://cryptodb.org/cryptodb/app/workspace/blast/new
    Expectation value: 10; Max Target Sequences: 100; Max matches in query range: 0; Word Size: 6; Scoring Matrix: BLOSUM62; Gap Costs (Open/Extension): 11,1; Compositional adjustments: Conditional compositional score matrix adjustment Low complexity regions: no filter; Mask for lookup table: false; Mask lower case letters: false

4. **Circos:**
    circos -conf config_file.conf

5. **ChrLen file for Circos** (*Cm*TU1867 contigs indicated by "1–8" and *Cp*BGF contigs indicated by "9–16"):
    1880501
    2980476
    31087121
    41105563
    51085659
    61311173
    71306104
    81365597

9919856
10992060
111102418
121107426
131085856
141308482
151363666
161379419

6. **TBTools:**
"Advanced Circos" selected; ChrLen File created manually, Links File generated manually; rRNA features added in "Set Input Genome Feature List" option; Gene density: "Sequence Toolkits" - > "GFF3/GTF Manipulate" - > "Gene Density Profile", Input File: Cmel.gff3, Output File: CmelGeneDensity.profile (repeat for CpBGF); GC content: "Sequence Toolkits" - > "Fasta Tools" - > "Fasta Window Stat", Input Genome Sequence File: Cmel_genome.fasta, Output file Prefix: Cmel_genome.genome.Window.Stat

7. **Jupiterplot:**
jupiter name = $prefix ref = $reference fa = $scaffolds

Technical validation parameters:

1. **BUSCO:**
busco -i < genome.fasta > -l./apicomplexa_odb10 -o BUSCO_CM.txt -m genome

2. **OrthoFinder in Anaconda (see above)**

Formatting OrthoFinder Result for OrthoVenn3's ClusterVenn:

```
awk -F': ' '{for (i=2; i<=NF; i++) if ($i!~ /^OG/) printf "%s%s", sep, $i; sep="\n"}'
Orthogroups.txt | tr -d ':' |
awk '{
for (i=1; i<=NF; i++) {
if ($i ~ /^CmeUKMEL1/)
$i="CryptoDB-55_CmeleagridisUKMEL1_AnnotatedProteins|" $i;
else if ($i ~ /^cmbei/)
$i="CmBEI_proteins_file|" $i;
else if ($i ~ /^cpbgf/)
$i="CpBGF_protein_file|" $i;
}
print
}' |
awk '{
for (i=1; i<=NF; i++) {
if ($i ~ /^CmBEI/) {
cmbei=cmbei $i " "
} else if ($i ~ /^CpBGF/) {
cpbgf=cpbgf $i " "
} else if ($i ~ /^Crypto/) {
crypto=crypto $i " "
}
}
print cmbei cpbgf crypto
cmbei=""
cpbgf=""
crypto=""
}' > Orthogroups2.txt
```

Configuration File:

```
#Add this to run circos faster
#alias circos=…
#Append this line to the ~/.bashrc to load when starting a new session
# Chromosome name, size, and color definition
karyotype=ChromosomeContigLabels.txt
<ideogram> <spacing> default=0.005r</spacing> radius=0.50r
thickness=20p
fill=yes
stroke_color=dgrey
stroke_thickness=2p
```

```
show_label = yes

#see etc/fonts.conf for list of font names
label_font = default
label_radius = 1r + 75p
label_size = 60
label_parallel = no < /ideogram > show_ticks = yes
show_tick_labels = yes < ticks > radius = 1r + 10p
color = black
thickness = 3p
# the tick label is derived by multiplying the tick position
# by 'multiplier' and casting it in 'format':
#
# sprintf(format,position*multiplier)
#
multiplier = 1e-6
# %d - integer
# %f - float
# %.1 f - float with one decimal
# %.2 f - float with two decimals
#
# for other formats, see http://perldoc.perl.org/functions/sprintf.html
format = %d
< tick >
spacing=70000 u
size = 50p
< /tick >
# < tick >
#spacing = 25000 u
#size = 15p
#show_label = yes
#label_size = 20p
#label_offset = 10p
#format = %d
# < /tick >
< /ticks >
########### NEW
< links >
< link >
file = Links.txt
#color = black_a5
radius = 0.95r
bezier_radius = 0.1r
thickness = 15
ribbon = yes
< /link >
< /links >
##########

< image >
# Included from Circos distribution.
<< include etc/image.conf >>
#To modify the size of the output image, default is 1500
#radius* = 3000p
< /image>
<<include etc/colors_fonts_patterns.conf>>
<<include etc/housekeeping.conf>>
```

Received: 17 May 2024; Accepted: 4 December 2024;
Published online: 18 December 2024

References

1. Ryan, U., Fayer, R. & Xiao, L. *Cryptosporidium* species in humans and animals: current understanding and research needs. *Parasitology* **141**, 1667–1685, https://doi.org/10.1017/S0031182014001085 (2014).
2. Hlavsa, M. C. *et al.* Outbreaks Associated with Treated Recreational Water - United States, 2000–2014. *MMWR Morb Mortal Wkly Rep* **67**, 547–551, https://doi.org/10.15585/mmwr.mm6719a3 (2018).
3. Kotloff, K. L. *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**, 209–222, https://doi.org/10.1016/S0140-6736(13)60844-2 (2013).

4. Girma, M., Teshome, W., Petros, B. & Endeshaw, T. Cryptosporidiosis and Isosporiasis among HIV-positive individuals in south Ethiopia: a cross sectional study. *BMC Infect Dis* **14**, 100, https://doi.org/10.1186/1471-2334-14-100 (2014).
5. Investigators, M.-E. N. The MAL-ED study: a multinational and multidisciplinary approach to understand the relationship between enteric pathogens, malnutrition, gut physiology, physical growth, cognitive development, and immune responses in infants and children up to 2 years of age in resource-poor environments. *Clin Infect Dis* **59**(Suppl 4), S193–206, https://doi.org/10.1093/cid/ciu653 (2014).
6. Gilbert, I. H. *et al*. Safe and effective treatments are needed for cryptosporidiosis, a truly neglected tropical disease. *BMJ Glob Health* **8** https://doi.org/10.1136/bmjgh-2023-012540 (2023).
7. Akiyoshi, D. E. *et al*. Characterization of *Cryptosporidium meleagridis* of human origin passaged through different host species. *Infect Immun* **71**, 1828–1832, https://doi.org/10.1128/IAI.71.4.1828-1832.2003 (2003).
8. Slavin, D. *Cryptosporidium meleagridis* (sp. nov.). *J Comp Pathol* **65**, 262–266, https://doi.org/10.1016/s0368-1742(55)80025-2 (1955).
9. Fayer, R. Taxonomy and species delimitation in *Cryptosporidium*. *Exp Parasitol* **124**, 90–97, https://doi.org/10.1016/j.exppara.2009.03.005 (2010).
10. Stensvold, C. R., Beser, J., Axen, C. & Lebbad, M. High applicability of a novel method for gp60-based subtyping of *Cryptosporidium meleagridis*. *J Clin Microbiol* **52**, 2311–2319, https://doi.org/10.1128/JCM.00598-14 (2014).
11. Cama, V. A. *et al*. *Cryptosporidium* species and genotypes in HIV-positive patients in Lima, Peru. *J Eukaryot Microbiol* **50**(Suppl), 531–533, https://doi.org/10.1111/j.1550-7408.2003.tb00620.x (2003).
12. Baptista, R. P. *et al*. Long-read assembly and comparative evidence-based reanalysis of *Cryptosporidium* genome sequences reveal expanded transporter repertoire and duplication of entire chromosome ends including subtelomeric regions. *Genome Res* **32**, 203–213, https://doi.org/10.1101/gr.275325.121 (2022).
13. Agyabeng-Dadzie, F., Xiao, R. & Kissinger, J. C. *Cryptosporidium* Genomics - Current Understanding, Advances, and Applications. *Current Tropical Medicine Reports*. https://doi.org/10.1007/s40475-024-00318-y (2024)
14. Agyabeng-Dadzie, F. *et al*. Evaluating the benefits and limits of multiple displacement amplification with whole-genome Oxford Nanopore Sequencing. *bioRxiv* https://doi.org/10.1101/2024.02.09.579537 (2024).
15. Baptista, R. P., Xiao, R., Li, Y., Glenn, T. C. & Kissinger, J. C. New T2T assembly of *Cryptosporidium parvum* IOWA annotated with reference genome gene identifiers. *bioRxiv* https://doi.org/10.1101/2023.06.13.544219 (2023).
16. Keely, S. P. *Cryptosporidium meleagridis* clinical isotate TU1867 isolated from gnotobiotic piglets. *NCBI Sequence Read Archive* http://identifiers.org/insdc.sra:SRR793561 (2011).
17. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540–546, https://doi.org/10.1038/s41587-019-0072-8 (2019).
18. Wick, R. R. & Holt, K. E. Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLoS Comput Biol* **18**, e1009802, https://doi.org/10.1371/journal.pcbi.1009802 (2022).
19. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
20. Dainat, J. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format.
21. Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput Biol Bioinform* **10**, 645–656, https://doi.org/10.1109/TCBB.2013.68 (2013).
22. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5**, e11147, https://doi.org/10.1371/journal.pone.0011147 (2010).
23. Kearse, M. *et al*. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649, https://doi.org/10.1093/bioinformatics/bts199 (2012).
24. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2 (1990).
25. Astashyn, A. *et al*. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *bioRxiv* https://doi.org/10.1101/2023.06.02.543519 (2023).
26. Lee, E. *et al*. Web Apollo: a web-based genomic annotation editing platform. *Genome Biol* **14**, R93, https://doi.org/10.1186/gb-2013-14-8-r93 (2013).
27. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa1016 (2020).
28. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* **33**, W465–467, https://doi.org/10.1093/nar/gki458 (2005).
29. Warrenfeltz, S., Kissinger, J. C. & EuPath, D. B. T. Accessing *Cryptosporidium* Omic and Isolate Data via CryptoDB.org. *Methods Mol Biol* **2052**, 139–192, https://doi.org/10.1007/978-1-4939-9748-0_10 (2020).
30. Barrnap -Bacterial ribosomal RNA predictor v. 28 Apr 2018 (GitHub, 2013).
31. Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33**, W686–689, https://doi.org/10.1093/nar/gki366 (2005).
32. Conesa, A. *et al*. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676, https://doi.org/10.1093/bioinformatics/bti610 (2005).
33. Paysan-Lafosse, T. *et al*. InterPro in 2022. *Nucleic Acids Res* **51**, D418–D427, https://doi.org/10.1093/nar/gkac993 (2023).
34. Ifeonu, O. O. *et al*. Annotated draft genome sequences of three species of Cryptosporidium: *Cryptosporidium meleagridis* isolate UKMEL1, *C. baileyi* isolate TAMU-09Q1 and *C. hominis* isolates TU502_2012 and UKH1. *Pathog Dis* **74** https://doi.org/10.1093/femspd/ftw080 (2016).
35. Anaconda Software Distribution v. 2-2.4.0 (2016).
36. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60, https://doi.org/10.1038/nmeth.3176 (2015).
37. Sun, J. *et al*. OrthoVenn3: an integrated platform for exploring and visualizing orthologous data across genomes. *Nucleic Acids Res* **51**, W397–W403, https://doi.org/10.1093/nar/gkad313 (2023).
38. Krzywinski, M. *et al*. Circos: An information aesthetic for comparative genomics. *Genome Res* gr.092759.109 [pii] (2009).
39. Chen, C. *et al*. TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol Plant* **13**, 1194–1202, https://doi.org/10.1016/j.molp.2020.06.009 (2020).
40. Chu, J. JupiterPlot: A Circos-based tool to visualize genome assembly consistency (1.0). *Zenodo* (2018).
41. Penumarthi, L. R., Baptista, R. P., Beaudry, M. S., Glenn, T. C. & Kissinger, J. C. A new chromosome-level genome assembly and annotation of *Cryptosporidium meleagridis NCBI SRA*. http://identifiers.org/insdc.sra:SRR27282542 (2024).
42. Penumarthi, L. R., Baptista, R. P., Beaudry, M. S., Glenn, T. C. & Kissinger, J. C. A new chromosome-level genome assembly and annotation of *Cryptosporidium meleagridis NCBI BioProject* http://identifiers.org/bioproject:PRJNA1022047 (2024).
43. Penumarthi, L. R., Baptista, R. P., Beaudry, M. S., Glenn, T. C. & Kissinger, J. C. A new chromosome-level genome assembly and annotation of *Cryptosporidium meleagridis NCBI Nucleotide* http://identifiers.org/insdc:JBCHVM000000000 (2024).
44. Hulsen, T., Huynen, M. A., de Vlieg, J. & Groenen, P. M. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol* **7**, R31 (2006).

Acknowledgements
This work was funded by NIH R01AI14866 to JCK and TCG.

Author contributions
L.R.P. performed analyses and wrote the manuscript; J.C.K., R.P.B. and T.C.G. conceived the study; R.P.B., M.S.B., and L.R.P. generated the genome assembly and L.R.P. and R.P.B. performed annotation; L.R.P., J.C.K., R.P.B. and M.S.B. edited the manuscript; T.C.G. and J.C.K. provided oversight and funding; All authors reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-04235-7.

Correspondence and requests for materials should be addressed to J.C.K.