Long-read assembly and comparative evidence-based reanalysis of *Cryptosporidium* genome sequences reveals expanded transporter repertoire and duplication of entire chromosome ends including subtelomeric regions

Rodrigo P. Baptista, Yiran Li, Adam Sateriale, et al.

Genome Res. published online November 11, 2021
Access the most recent version at doi:10.1101/gr.275325.121

Title: Long-read assembly and comparative evidence-based reanalysis of *Cryptosporidium* genome sequences reveals expanded transporter repertoire and duplication of entire chromosome ends including subtelomeric regions

Authors: Rodrigo P. Baptista[1,2], Yiran Li[2], Adam Sateriale[3], Mandy J. Sanders[4], Karen L. Brooks[4], Alan Tracey[4], Brendan R. E. Ansell[5], Aaron R. Jex[5], Garrett W. Cooper[6], Ethan D. Smith[6], Rui Xiao[2], Jennifer E. Dumaine[3], Peter Georgeson[6,7,9], Bernard J. Pope[6,7,8,10], Matthew Berriman[4], Boris Striepen[3], James A. Cotton[4] and Jessica C. Kissinger[1,2,11]

Affiliation: [1]Center for Tropical and Emerging Global Diseases; [2]Institute of Bioinformatics, University of Georgia, Athens, GA, USA; [3]Department of Pathology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA, USA; [4]The Wellcome Sanger Institute, Hinxton, UK; [5]Faculty of Veterinary and Agricultural Sciences, The University of Melbourne and Population Health and Immunity Division, the Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia; [6]Department of Clinical Pathology; [7]Melbourne Bioinformatics; [8]Department of Surgery (Royal Melbourne Hospital), Melbourne Medical School, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Australia; [9]University of Melbourne Centre for Cancer Research, Victorian Comprehensive Cancer Centre, Melbourne, Australia; [10]Department of Medicine, Central Clinical School, Faculty of Medicine Nursing and Health Sciences, Monash University, Australia; [11]Department of Genetics, University of Georgia, Athens, GA, USA

Key words: Reassembly, Long reads, genome rearrangement, subtelomeric duplication, subtelomere, reannotation

Corresponding Author: Jessica C. Kissinger jkissing@uga.edu

Manuscript Type: Research

Running Title: New Comparative *Cryptosporidium* Genomic Insights

ABSTRACT

Cryptosporidiosis is a leading cause of waterborne diarrheal disease globally and an important contributor to mortality in infants and the immunosuppressed. Despite its importance, the *Cryptosporidium* community has only had access to a good, but incomplete, *Cryptosporidium parvum* IOWA reference genome sequence. Incomplete reference sequences hamper annotation, experimental design and interpretation. We have generated a new *C. parvum* IOWA genome assembly supported by PacBio and Oxford Nanopore long-read technologies and a new comparative and consistent genome annotation for three closely related species *C. parvum*, *Cryptosporidium hominis* and *Cryptosporidium tyzzeri*. We made 1,926 *C. parvu*m annotation updates based on experimental evidence. They include new transporters, ncRNAs, introns and altered gene structures. The new assembly and annotation revealed a complete *Dnmt2* methylase ortholog. Comparative annotation between *C. parvum, C. hominis* and *C. tyzzeri* revealed that most “missing” orthologs are found suggesting that the biological differences between the species must result from gene copy number variation, differences in gene regulation and single nucleotide variants (SNVs). Using the new assembly and annotation as reference, 190 genes are identified as evolving under positive selection, including many not detected previously. The new *C. parvum* IOWA reference genome assembly is larger, gap free and lacks ambiguous bases. This chromosomal assembly recovers all 16 chromosome ends, 13 of which are contiguously assembled. The three remaining chromosome ends are provisionally placed. These ends represent duplication of entire chromosome ends including subtelomeric

regions revealing a new level of genome plasticity that will both inform and impact future research.

INTRODUCTION

Cryptosporidium spp. are parasitic apicomplexans that cause moderate-to-severe diarrhea in humans and animals. Studies have revealed that *Cryptosporidium* is one of the most common causes of waterborne disease in humans and the second leading cause of diarrheal etiology in children < 2 years (Kotloff et al. 2013; GBDDD-Collaborators 2017). In 2016, acute infections caused > 48,000 global deaths and more than 4.2 million disability-adjusted life years lost (Khalil et al. 2018).

Currently, 38 species of *Cryptosporidium* are recognized (Slapeta 2013; Feng et al. 2018). Most species have preferred hosts, and hosts range from fish to mammals. 15 species have an assembled genome sequence, however, only 8 are annotated (Supplemental Table S1). Most genomic sequence data are from the zoonotic *C. parvum* and anthroponotic *C. hominis,* the species primarily detected in humans (Chalmers et al. 2011; Zahedi et al. 2016; Khan et al. 2017). These two species are only 3-5% divergent at the DNA level (Mazurie et al. 2013). *Cryptosporidium* genome sequences are shorter than most other apicomplexans at around 9 Mbp distributed over 8 chromosomes and containing < 4,000 protein-encoding genes in the species examined. Most genome reduction consists of gene and intron loss, intron shortening and very short intergenic regions (Abrahamsen et al. 2004; Xu et al. 2004; Kissinger and DeBarry 2011).

As the *Cryptosporidium* field is exploding with new-found interest and much needed breakthroughs in genetics and culturing (Vinayak et al. 2015; Morada et al. 2016; DeCicco RePass et al. 2017; Heo et al. 2018; Wilke et al. 2019), the limitations of existing reference genome sequences need to be addressed. The *C. parvum* IOWA II reference genome sequence, (*Cp*IRef), was assembled with a limited physical map (Abrahamsen et al. 2004) and

a few hundred ESTs, for training gene finders. Genomic, transcriptomic and proteomic work has been lacking due to the obligate quasi-intracellular nature of portions of the parasite's life cycle, the historical lack of a continuous *in vitro* tissue culture system, the parasite's small size relative to host cells and difficult animal models. The physical map for the *Cp*IRef assembly was generated using a genome-wide HAPPilly anchored physical mapping technique (Piper et al. 1998; Bankier et al. 2003). Despite the cutting-edge approaches, some regions, especially chromosome ends, lacked support or were poorly resolved. Subsequent whole genome sequencing data remain unassembled or in a large number of contigs.

In 2015, the *Cp*IRef was re-annotated using new RNA-seq evidence and a new *C. hominis* sequence from a recent human isolate (UdeA01) was generated (Isaza et al. 2015). Many ambiguities in gene models were improved but the new *C. hominis* UdeA01 genome is fragmented. Incomplete, misassembled (i.e. gapped sequence, indels, frameshifts, compressed repetitive regions, artifactual inversions) and independently annotated reference genome sequences as are discussed in (Guo et al. 2015) can mislead analyses of the differences between isolates and species owing to these artifacts rather than the biology. Comparative analyses require additional assays to confirm indels and copy number variations (CNVs). Since incomplete and misassembled sequences are usually caused by repetitive and complex sequence regions, it is imperative to revisit reference genome sequences with new long-read technologies.

Long-read sequence technologies (PacBio and Oxford Nanopore) are becoming an essential tool to close full genome sequence assemblies across the tree of life (Vembar et al. 2016; Berna et al. 2018; Miga et al. 2020). They can be used to resolve complex regions such as repetitive content, structural variants (SVs) such as inversions, translocations and duplications, or for use as scaffolding evidence for existing fragmented genome assemblies (Mahmoud et al. 2019). They are proving crucial for completing pathogen genome sequences that are often riddled with large virulence-related gene families that may have been improperly

assembled in shorter-read assemblies (Xia et al. 2021). Here we provide a new *de novo* hybrid long-read assembly for *C. parvum* strain IOWA (*Cp*IA), and new consistent, comparative genome annotations for *Cp*IA, *C. hominis* 30976 (*Ch*30976) and *C. tyzzeri* UGA55 (*Ct*UGA55). The new data were used to assess genome-level species differences and assess rapidly evolving genes.

RESULTS

An improved long-read genome assembly for *Cryptosporidium parvum* (IOWA-ATCC)

The *Cp*IRef genome assembly, generated in 2004, has only 10 physical gaps of unknown size, but it has 18,558 ambiguous bases and is missing 6 telomeres. Alignment of 54,882,187 Illumina 100 bp paired-end reads (Supplemental Fig. S1) to this reference sequence, revealed many regions that had become collapsed during assembly (Supplemental Table S2). To resolve these issues, we generated a new PacBio+Illumina+Nanopore hybrid genome assembly (Table 1; Supplemental Fig. S1) for the *C. parvum* strain IOWA (ATCC®PRA-67DQ™), *Cp*IA. To minimize strain variation differences, we performed our analysis on the IOWA strain. However, because there is a 14-yr window of propagation between these two isolates, and cryopreservation has only recently been developed (Jaskiewicz et al. 2018), we modified the strain name to IOWA-ATCC (*Cp*IA).

The new, *Cp*IA genome assembly is compared to the current *Cp*IRef sequence and two closely related species with different host preferences and pathogenicity, *C. hominis* (*Ch*30976) and *C. tyzzeri* (*Ct*UGA55) (Slapeta 2013; Nader et al. 2019; Sateriale et al. 2019) (Table 1). These particular assemblies were selected because they are the best available. The new *Cp*IA long-read assembly increases the genome size by 19,939 bases (~152 kb when including new proposed subtelomeric regions) and putatively identifies all 16 telomeres. There are no gaps

and no ambiguous bases. As expected, the *Cp*IA genome sequence has diverged slightly but shares 99.93% average pairwise identity with the 2004 assembly in regions that are comparable (Supplemental Table S3). The main *Cryptosporidium* subtyping marker, the 60 kDa surface protein (*gp60* locus subtype IIa) shows 4 amino acid differences (2 in the serine repeat region) between *Cp*IA and *Cp*IRef (Supplemental Fig. S2; Supplemental Methods).

Structural differences between the *C. parvum* IOWA assemblies are confirmed

The 2004 *Cp*IRef genome assembly used Sanger reads combined with available HAPPY-map data to scaffold the contigs. We compared the *Cp*IRef and *Cp*IA assemblies to identify potential rearrangements. Inversions and relocations were detected in Chromosomes 2, 4 and 5 (Fig. 1A). These inversions may be previous assembly artifacts or represent genuine differences between the isolates. We investigated the synteny between *Cp*IRef, *Cp*IA, *Ch*30976 and *Ct*UGA55 and observed that *C. hominis* and *C. tyzzeri* also share the Chr4 and Chr5 inversions. Examination of the inverted region boundaries in *Cp*IRef revealed regions of ambiguous nucleotide bases or physical gaps (Fig. 1A). To further investigate, PCR primers were designed to test each possible inversion arrangement in genomic DNA from *C. parvum* KSU-1strain 2006 and Bunch Grass farms IOWA (*Cp*BGF) (Fig. 1B, Supplemental Fig S3, Supplemental Table S4). The results support the revised assembly orientation. Long-read Oxford Nanopore, ONT, data also support the *Cp*IA assembly (Supplemental Fig. S4). Better assemblies for the other species will be needed to determine the true level of synteny across these species.

Consistent structural gene annotation resolves inconsistencies and improves functional annotation

We consistently annotated and compared *Cp*IRef, *Cp*IA, *Ch*30976*, and Ct*UGA55 which have > 95% genome identity to assess differences in gene content. The new annotation for each

154 species was generated with three *de novo* approaches and evidence-based manual annotation.
155 Curation of the annotation was performed pairwise between each assembly to take full
156 advantage of syntenic regions. Data from one species could be used to assess computational
157 predictions in others. Using this approach, fragments of genes that were previously missing in *C.*
158 *hominis* were identified. This approach resulted in 1,926 gene alterations in *Cp*IA when
159 compared to *Cp*IRef, resulting in improved functional annotation. These changes increase the
160 overall number of predicted genes, introns (100% supported by RNA-seq data) and exons
161 (Table 2; Supplemental Results). The average mRNA length increased. These structural fixes
162 led to the repair of the N-terminus of the methylase ortholog, *Dnmt2* (Supplemental Fig. S5) as
163 well as 523 other genes and 113 fragmented genes previously annotated as pseudogenes.
164 *Cryptosporidium* has a very compact genome with 76.88% covered by protein coding
165 sequences, CDS. As a result, RNA-seq data, which is the best evidence for annotation, contains
166 reads that overlap adjacent genes creating false fusions of exons belonging to different genes.
167 Available strand-specific RNA-seq was used to characterize some of these regions but
168 expression data were not available for all predicted genes (87% of the annotated genes were
169 covered), thus, genes of unknown function in close proximity on the same strand remain
170 problematic. The expression data also revealed 3 putative alternative spliced genes
171 (CPATCC_0027530; CPATCC0027960; CPATCC_0035590) and 474 potential non-coding
172 RNAs (ncRNAs) predominantly anti-sense lncRNAs with differential expression as reported in
173 (Li et al. 2020).
174
175 **Comparative analysis reveals few gene content differences between closely related**
176 ***Cryptosporidium* species**
177 There is a cluster of species. *C. parvum, C. hominis, C. tyzzeri, C. meleagridis and C. ubiquitum*
178 that are highly syntenic relative to species outside of this cluster. The syntenic species are
179 biologically distinct and largely host-adapted with the main zoonotic exceptions being *C. parvum*

and *C. ubiquitum*. A synteny analysis of the clustered species and *Cryptosporidium muris* as an outgroup reveals high synteny (99.4% to 87%) within the cluster and only 4% synteny to *C. muris* (Supplemental Fig. S6; Supplemental Tables S5 and S14).

The consistent annotation of the species closest to *Cp*IA (GCA_015245375.1), *Ch*30976 (GCA_001483515.1) and *Ct*UGA55 (GCA_007210665.1), permitted the analysis of differences in CDS content and CNVs. Orthology analysis revealed that 94% of the genes were conserved among all species. Of the 4,008 ortholog groups identified, most gene families were maintained with a similar number of paralogs (max = 6) detected in the same ortholog group, but variation was detected among singletons (Fig. 2A; Supplemental Table S6). Some of these gene differences appear to be unique to a particular species (Supplemental Table S7). Of the 224 singletons detected, we observed only 0, 1 and 1 potential truly species-specific genes in *Cp*IA, *Ch*30976 and *Ct*UGA55, respectively following manual inspection (Fig. 2B; Fig. 2D). Both species-specific genes are uncharacterized proteins. The remaining 253 singletons are detected but incomplete in the fragmented assemblies of *Ch*30976 and *Ct*UGA55, appearing as split genes, frame-shifts, missed calls near a gap and missing subtelomeric regions or contig break and putative false gene predictions in small contigs (Fig. 2C). The major protein-encoding gene content differences between these species are gene copy number variations and not gene presence or absence.

To identify and assess putatively overly collapsed repetitive regions within the genome assemblies analyzed in this study, i.e. repetitive regions represented by only a single repeat in the assembly, we mapped Illumina reads from *Cp*IA, to the new *Cp*IA, *Cp*IRef, *Ch*30976 and *Ct*UGA55 genome assemblies (Supplemental Table S2; Supplemental Fig. S1). Our pipeline detected 12 compressions of at least 2× read depth and > 100 bp in length in the *Cp*IRef genome assembly compared to 6 in the new *Cp*IA assembly. The 6 compressed regions drops to 4 if the three putative new subtelomeric regions proposed in this study are included (See below). The *Ch*30976 and *Ct*UGA55 genome assemblies contain > 20 compressions mostly

due to the short reads used to generate these assemblies. The *Cp*IA collapsed regions have 2 hits in regions with gene annotations in Chr1 and Chr2. Both genic regions are composed of rRNA genes, some uncharacterized proteins, GMP synthase, aspartate-ammonia ligase, tryptophan synthase beta and MEDLE genes, all associated with complex subtelomeric regions discussed below. The four intergenic compressions all match small simple repeat regions (Supplemental Table S8).

Functional annotation identifies new protein features

Several approaches to assess function were applied including InterProScan and I-TASSER among others (see Methods). As a result, 138 new *C. p*arvum protein annotations were generated or modified. The percentage of *Cp*IA genes annotated as uncharacterized proteins was reduced from 40% to 33% in all reannotated sequences (Supplemental Table S9). Many new features including domain and repeat content were added to 738 previously uncharacterized proteins. 729 predicted *Cp*IA CDSs have signal peptides and 1,990 have GO assignments (Supplemental Results). 1,414 CDSs were further assessed for confidence using I-TASSER protein structure searches and 1,008 predicted structures were assigned as high-confidence by random forest categorization. 143 previously uncharacterized proteins in *Cp*IRef were assigned high confidence GO terms.

New transporter genes were identified

We further characterized transporter genes using three different prediction methods. A total of 152 proteins in *Cp*IA and *Ch*30976 were identified as transporters including 128 confident candidates and 24 putative candidates (Supplemental Table S10). This represents an increase of 53 transporters relative to the *Cp*IRef GO annotation (CryptoDB v36) (Heiges et al. 2006). Most identifiable transporters are related to purine metabolism, peptidoglycan biosynthesis,

oxidative phosphorylation and N-Glycan biosynthesis pathways (Fig. 3). Six translocases were also identified.

Entire subtelomeric regions are duplicated

As shown in the read depth coverage analysis and in Table 1, Supplemental Table S2 and Supplemental Figure S1, the new *Cp*IA assembly was able to recover ~2.3kb cumulative length in collapsed regions relative to *Cp*IRef. One subtelomeric region on Chr1 in *Cp*IA, previously reported on Chr5 in *Cp*IRef (but not linked to Chr5 in the HAPPY Map), still shows signs of sequence compression suggesting that most of the genes present in this region have more than one copy (Fig. 4A, Supplemental Fig. S7). This region reveals at least 13 genes which vary in copy number between different *Cryptosporidium* species (Supplemental Fig. S8). The genes contained in this region are 18S rRNA, 5S rRNA and 28S rRNA, uncharacterized proteins, a GMP synthase, an aspartate-ammonia ligase, tryptophan synthase beta and a cluster of several MEDLE genes. Some of these genes, such as the tryptophan synthase beta and the MEDLE's are the focus of considerable research since they may be related to parasite survival and are potentially involved in parasite invasion, respectively (Sateriale and Striepen 2016; Li et al. 2017; Fei et al. 2018). The predicted number of copies of rRNAs and MEDLE's are underrepresented as they also have paralogs on Chr2 and Chr5, respectively. The Illumina pileup of ~1,350 reads on Chr2, positions 681,607 to 686,953 (Supplemental Table S2, Supplemental Fig. S1, is the region where the 18S/28S rRNA gene(s) are located on this chromosome. The five 18S genes are identical and 28S rRNAs have three gaps (Supplemental Fig. S9). Thus, alignment competition explains why the read coverage varies relative to the equivalent 18S/28S rRNA Chr1 pileups. Regions with pileups on inner portions of Chr5, 7 and 8 are low complexity regions composed by tandem repeats (Supplemental Table S8). In Chr5, we have one uncharacterized protein (CPATCC_0023030), full of tandem repeats and good RNA-

seq support for its expression. On Chr7 and 8, these regions are smaller than 100 bp and do not contain any annotated genes.

Since there is an apparent compression in a subtelomeric region assembly with no gaps and good PacBio long-read coverage, we hypothesized that these extra copies might derive from additional copies of this region. The *Cp*IA assembly was only missing three telomeric regions, both ends of Chr7 and one telomere of Chr8. Using existing PacBio long reads we were able to identify a few reads that extended into rRNA regions on the chromosomes missing telomeres. We attempted re-assembly with only PacBio reads and we could not resolve the missing regions. Thus, we generated very deep (2260×) ONT single molecule reads from *Cp*BGF, (ATCC DNA was not available, only 143 SNVs are detected between the strains of which 108 are Indels)(Supplemental Table S11). The ONT reads revealed related, yet unique subtelomeric regions linked to the chromosomes missing their telomeres, in addition to Chr1 (Fig. 4B). We found good ONT long-read support for these regions (Supplemental Fig. S7). Each distinct subtelomeric region begins with chromosome-specific sequences followed by a conserved ribosomal RNA cluster which is followed by the duplicated subtelomeric region and telomere. There are many ONT and PacBio reads that link the unique chromosomal regions and the beginning of the subtelomeric gene families but only a few span the entire chromosome end. We also note that there is slight variation observed among the reads for each subtelomeric region distal to the rRNA cluster.

New positively selected genes are identified in *C. parvum*

The new gapless *Cp*IA genome assembly and annotation presented an opportunity to revisit the prediction of genes evolving under positive selection in this species. We performed a Single Nucleotide Variant, SNV, analysis using 136 different *C. parvum* WGS data sets obtained from GenBank (Supplemental Table S12) using the new *Cp*IA assembly and annotation. A total of 24,407 positions were found to contain at least one high-confidence bi-allelic variant. Multiallelic

calls were removed to guard against mixed infections. The biallelic variants reflect 3,892 genes, 342 of which show a πN/πS ratio of non-synonymous/synonymous rates of > 1.5 (Supplemental Table S13). Of the 342, 17 genes were previously identified and 145 are classified as uncharacterized proteins, 105 of which are annotated as having a signal peptide or being secreted. All previously identified genes evolving under positive selection were detected, including: Insulinase-like protein (CPATCC_0017080), an uncharacterized secreted protein (CPATCC_0010380), *gp60* (CPATCC_0012540) and others (Strong et al. 2000; Sanderson et al. 2008; Nader et al. 2019; Zhang et al. 2019). Of the top 10 genes by πN/πS ratio, nine appear to be new to this study. Gene family members such as MEDLEs, FLGN and SKSR were also detected but, new members of each of these families are identified as also evolving under positive selection. Since the putative new subtelomeric repeats (Fig. 4) were not included in these analysis (they were identified in a different strain) evolution of the MEDLE genes may be an over estimation. A family of WYLE proteins (Sanderson et al. 2008) is also identified as being positively selected.

DISCUSSION

The first genome sequence assembly of *Cryptosporidium parvum* IOWA II referred to as *Cp*IRef here, was excellent given the technology at the time. As a result, the community has relied on this genome assembly and annotation to this day to design their experiments. However, gaps and ambiguous bases remain, and there was little available expression and orthology evidence at the time to facilitate the annotation. We used PacBio and Illumina sequencing technologies to generate a new complete genome assembly of *C. parvum* strain IOWA-ATCC. We then applied de novo computational and evidence-based annotation approaches with manual curation of two additional species to generate consistent annotation that can be used to detect differences between species and strains. *Cp*IA DNA was not available for Nanopore sequencing or PCR validation of the assembly, so *Cp*BGF DNA (which differs by < 200 SNVs) was used instead.

However, all results are consistent when strains can be compared, e.g. compressions of *Cp*IA Chr1 detected with *Cp*IA Illumina reads are the same when *Cp*BGF is used lending strength to the broader applicability of the findings.

The first, expected, finding was that the *C. parvum* IOWA strain is continuing to evolve (Cama et al. 2006) as it is maintained by passage through cattle in a few different locations for research use. Some natural *Cryptosporidium* isolates have been propagated in unnatural hosts before sequencing. Thus, potential selection during the move to a non-natural host and subsequent drift in propagated and naturally circulating parasites has led to accumulated differences. This phenomenon has been observed in other protozoan parasites (Akiyoshi et al. 2002; Cama et al. 2006; Chan et al. 2015; Isaza et al. 2015). Genomic DNA for the 2004 *Cp*IRef and *Cp*IA were obtained from the same source, but many years apart. We note small differences in the *gp60* sequence, and an overall genome average difference of 0.07% in identity (Supplemental Table S3).

We detect chromosomal inversions in *Cp*IA relative to *Cp*IRef which have also been detected by others (Guo et al. 2015; Isaza et al. 2015). Chromosomal inversions are known to affect rates of adaptation, speciation, and the evolution of chromosomes (Rieseberg 2001; Guo et al. 2015), but they can also represent assembly artifacts. PCR spanning the genomic regions flanking each major inversion in each orientation using genomic DNA from an unsequenced isolate from 2006, *C. parvum* KSU-1 and *Cp*BGF from 2019 validated the long-read *Cp*IA assembly. Since the other species still lack physical evidence for their chromosomal structures, further long-read sequencing or chromosome conformation capture sequencing, such as Hi-C, will be needed to detect and validate species-specific genomic structural variations for the other *Cryptosporidium* species.

C. hominis and *C. tyzzeri*, which are 95–97% identical at the nucleotide level to *C. parvum* show incongruences in annotated genes with respect to the new *Cp*IA genome assembly. The differences result in part from numerous sequence gaps and a lack of

experimental evidence (e.g. RNA-seq data) to facilitate annotation. Assembly gaps can lead to frame-shift artifacts, fragments of genes split on different contigs and missing genes. These differences affect similarity-based analyses such as ortholog detection, giving the impression that some of these partially annotated genes are unique to a species when, they are not. These mis-interpretations can sabotage some experimental designs and analysis (Baptista and Kissinger 2019). The re-annotation of the original assembly had 114 pseudogenes, now reduced to just one. This improvement facilitated ortholog and functional identification of the genes involved. Assembly gap regions are usually complex, with repetitive sequence patterns, or hypervariable regions in the population analyzed and some have high polymorphism rates. False assumptions regarding species-specific genes can affect many downstream analyses including the detection of highly polymorphic loci.

In this study we were able to improve the structural and functional annotation for three *Cryptosporidium* species using two different approaches: (i) inclusion of seven full-length stranded cDNA libraries derived from three time points (0h, 24h and 48h post infection (Tandel et al. 2019), covering ~ 90% of the protein-encoding genes in *C. parvum*; and (ii) by using synteny information to construct a consistent genome annotation between three different closely related species. This approach facilitated a new comparative analysis of genome content between species. Our analyses reveal that *C. parvum, C. hominis* and *C. tyzzeri* show few differences in gene content for the regions that can be compared. Most differences are related to slight structural variation, such as small translocations and inversions, and copy number variation as revealed by read depth coverage analysis.

Apicomplexans have streamlined genomes that range from ~8.5 to ~125 Mbp (Woo et al. 2015) relative to the only sequenced free-living ancestor, *Chromera velia* at, 194 Mbp (Kissinger and DeBarry 2011). *Cryptosporidium* species have among the most compact apicomplexan genomes with ~3,900 protein-encoding genes and ~77% of the genome sequence being protein encoding. They also lack a mitochondrial genome and apicoplast

organelle (Keeling 2004) a finding that holds with our deep sequencing. Thus, the higher number of transporters found in our re-analysis makes biological sense and adds to growing work in this area, e.g. *Cryptosporidium* may have adapted a novel type of nucleotide transporter for ATP uptake from the host (Striepen et al. 2004; Pawlowic et al. 2019). The new *Cp*IA assembly and annotation reveals a complete ortholog of the *Dnmt2* methylase family. The *C. parvum Dnmt2* sequence was previously annotated as truncated and lacking a DNMT-specific motif containing a prolyl-cysteinyl dipeptide (Abrahamsen et al. 2004; Ponts et al. 2013; Isaza et al. 2015). DNMT2 proteins share high sequence and structural similarity with DNA methyltransferases, however they appear to function primarily as RNA methyltransferases in plants and animals (Goll et al. 2006). Substrates for DNMT2 in protozoa remain unclear.

The lack of three telomeres in the new *Cp*IA long-read assembly was an intriguing result that could be explained by the detection of three putative similar but not identical copies of subtelomeric regions containing genes including tryptophan synthase beta, the MEDLE genes and 18S, 5.8S and 28S rRNAs among others. This finding raises the possibility that *Cryptosporidium* has recombination between telomeres by break-induced replication, like some yeasts (McEachern and Iyer 2001; McEachern and Haber 2006), or telomere maintenance by recombination as is observed in human cancers (Natarajan et al. 2006). Some genes in this region may be essential for parasite survival (Sateriale and Striepen 2016). It is possible, but remains to be proven, that extra or altered copies of these genes may confer an advantage to individual parasites or the population as a whole. In fact, we have not shown that all 4 subtelomeric regions coexist in the same cell, but 4× coverage of these sequences are present in the population sequenced. We have support from single ONT reads indicating that this region is detected on 4 different chromosome ends. The ONT reads also prove that these structures are varying within the sequenced *Cp*BGF population (Fig. 4), raising the possibility of recombination or gene conversion during sexual reproduction. This subtelomeric plasticity in which transfer or duplication of important gene sequences between homologous and

nonhomologous chromosome ends, may affect genetic manipulations of the parasite and their resulting phenotype. Currently, cloning does not exist for *Cryptosporidium* and oocysts, which can be sequenced (Troell et al. 2016), must still be considered a population of four haploid meiotic progeny (sporozoites). Single-cell sporozoite sequencing will facilitate recombination and sub-telomeric plasticity studies but currently is still impossible in the absence of genome amplification.

Cryptosporidium species are usually typed and characterized by a small number of genetic markers: 18S, COWP, HSP70, and *gp60* (Ghaffari et al. 2014). As shown in this study *gp60* which is a gene evolving under positive selection used for *Cryptosporidium* subtyping characterization, had small differences between *Cp*IRef and *Cp*IA. Using a single marker to characterize an obligately sexual organism with 8 chromosomes is problematic. In this study, we confirm an existing group of genes evolving under positive selection and identify 325 additional potential candidates distributed across all 8 chromosomes. Some of these genes belong to gene families so to avoid artifacts only uniquely mapped reads were used for the SNV analysis. The genes identified here can be used to help the community develop additional markers for typing parasite isolates. However, the global diversity of *Cryptosporidium* is yet to be characterized. Only 136 isolates from a small geographic region have been sampled here. Newer techniques such as hybrid capture bait set techniques (Mamanova et al. 2010) are a powerful future alternative to characterize and select *Cryptosporidium* population variants and better characterize genetic diversity.

The new *C. parvum* long-read assembly combined with a consistent comparative annotation has proven powerful. The species analyzed here have different host preferences and pathogenicity. Comparisons of previous sequences and annotation suggested numerous gene content differences. However, this systematic study reveals that the primary differences between the zoonotic *C. parvum*, the anthroponotic *C. hominis* and the rodent-infecting *C. tyzzeri* are SNVs and CNVs rather than differences in unique gene content. Finally, new findings

related to within parasite and/or within population subtelomeric amplification and variation events in *C. parvum* reveal a new level of genome plasticity that will complicate some genetic manipulations and may affect the organisms' phenotype.

METHODS

Sample DNA sources

C. parvum IOWA-ATCC (*Cp*IA) DNA from oocysts/sporozoites was purchased from the ATCC. The source was the University of Arizona, Sterling Parasitology Laboratory. It is a GP60 subtype (IIa) like the current *C. parvum* IOWA II reference (*Cp*IRef) genome sequence. *C. parvum* IOWA DNA was also prepared from oocysts obtained in 2018 from Bunch Grass Farms, Deary, ID, referred to as *Cp*BGF in this study. *C. parvum* KSU-1 genomic DNA was also prepared in 2006 from oocysts obtained from Steve Upton. Public sequence data were accessed from NCBI BioProjects PRJNA252787, PRJEB3213, PRJNA388495 and PRJEB10000. Accession numbers for the 136 *C. parvum* sequences used for evolutionary analysis are detailed in Supplemental Table S12.

Cryptosporidium parvum IOWA-ATCC sequencing and genome assembly

PacBio RSII and Illumina HiSeq 2000 sequencing were performed at the Wellcome Sanger Institute, UK. The *Cp*IA reads were first assembled using the PacBio open source SMRTlink v6.0 from 9 PacBio SMRT cells, with ~75× mean genome coverage. The resulting assembly was then submitted to the accuracy improver tool Sprai 0.9.9.23 (https://sprai-doc.readthedocs.io/en/latest/index.html) and then gaps were filled using PBJelly 15.24.8 (English et al. 2014) with PacBio reads and IMAGE 2.4.1 (Swain et al. 2012) with Illumina reads. A manual inspection and improvement using GAP5 (Bonfield and Whitwham 2010) was applied,

and the final scaffolded genome assembly was polished with Illumina reads using iCORN2 0.95 (Otto et al. 2010) and Pilon 1.22 (Walker et al. 2014).

ONT single-molecule long-read sequencing was performed on DNA from *Cp*BGF (ATCC®PRA-67DQ™ was out of stock) following the recommended R9.4.1 flow cell protocol. MinION ONT sequencing was performed at the Georgia Genomics Bioinformatics Core at the University of Georgia, USA, using an R.9.4 flow cell and the Ligation Sequencing kit (SQK-LSK109). The ONT long reads generated > 2000× coverage of the *C. parvum* genome. This high coverage complemented the PacBio data to confirm and resolve several complex regions (Supplemental Methods). The final assembly was submitted along with *Cp*IRef, *Ch*30976 and *Ct*UGA55 to QUAST v.5.02 (Gurevich et al. 2013) to compare and evaluate the quality of *Cp*IA. All sequencing statistics are in Supplemental Table S12.

Cryptosporidium genome reannotation

Genome annotation was generated with: ab initio prediction using GeneMark-ES 4.57 (Lomsadze et al. 2005); evidence-trained predictions using SNAP/MAKER (Cantarel et al. 2008; Johnson et al. 2008) and AUGUSTUS (Stanke and Morgenstern 2005). For training, we used publicly available data from each respective species: RNA-seq, ESTs, previously predicted proteins and MassSpec proteomics data when available. In parallel we also generated transcriptome assemblies using HISAT2 v.2.1.0 (Kim et al. 2015) and StringTie v.1.3.4 (Pertea et al. 2015), and non-coding RNA predictions were generated for *C. parvum* as described (Li et al. 2020). Manual curation of all genes in the context of existing molecular evidence was performed using WebApollo2 (Lee et al. 2013).

We performed comparative genome annotation using the Artemis Comparison tool 17.0.1 (Carver et al. 2005). OrthoFinder v.2.3.7 (Emms and Kelly 2015) was used to detect paralogs, orthologs and singletons. All singletons were then manually compared using MCScanX 0.8 (Wang et al. 2012) and JBrowse (Buels et al. 2016) to verify their uniqueness and

assess the contribution of sequence gaps or misassembly to the findings. We considered the following error types: Split genes caused by frameshifts or early stop-codons, lack of stranded RNA-seq to confirm the gene model, and the presence of a gapped region in the genome assembly. All genes that did not fall into one of these categories were identified as unique.

Functional annotation

Following structural annotation, the predicted protein sequences were used to search the Swiss-Prot, Trembl and the NCBI non-redundant Protein database with BLASTP using an e-value threshold at the superfamily level of 1×10^{-6}. Protein structure similarity was explored using I-TASSER (Roy et al. 2010) as in (Ansell et al. 2019) and Supplemental Methods. Blast2GO (Conesa et al. 2005) version 4.1.9 was used to assign Enzyme Code (E.C) and Gene Ontology (GO) terms. We compared the existing protein product names to the new functional results. Some structural information, such as protein domain and repeat pattern content were added to some uncharacterized proteins and nomenclature errors were corrected according to the NCBI guidelines.

Transporter prediction

Predicted proteins were submitted to four different transporter prediction methods: (i) BLASTP against TCDB (Saier et al. 2009) with a threshold e-value of 1×10^{-5} cutoff; (ii) TMHMM (Server v. 2.0) (Krogh et al. 2001) and SignalP (Server 4.1) (Bendtsen et al. 2004) to reduce false positives from the TCDB BLASTP results. Transporter candidates with no transmembrane domains or candidates with only one transmembrane prediction while having signal peptides predicted were removed; (iii) TransAAP (Ren et al. 2007), a Transporter Classification tool at TransportDB v2.0 (http://www.membranetransport.org/transportDB2/index.html), was used to provide information about potential transporter identity and substrate; and (iv) a structural proof for candidate transporters using Phyre2.0 (Kelley et al. 2015). Final candidate transporters were

checked according to above results as well as annotations obtained from InterProScan 5.44 (Jones et al. 2014).

Comparative analyses

Structural variation sites were calculated using NucDiff v2.0.3 (Khelik et al. 2017) and the major inversions observed between *Cp*IRef and *Cp*IA were verified by PCR. Primers to test both ends of each orientation were designed using Primer3 v0.4.0 (Untergasser et al. 2012) (Supplemental Table S4). Primers designed to be specific and conserved among the species were tested using an in silico PCR amplification tool (San Millan et al. 2013). These regions contain repeats so the amplicons range from 2kb to 9kb to avoid them. PrimeSTAR GXL DNA polymerase (TAKARA) was used with Long-PCR conditions: initial 3 min hot start at 98 °C, 35 cycles of: 10 sec denaturation at 98 °C; 15 sec primer annealing at 55.4 °C; and 10 min elongation at 68 °C; followed by 10 min elongation at 72 °C. PCR products were separated in a 0.8% agarose gels and stained with ethidium bromide.

The consistency of annotation and potential gene family CNVs, were determined with OrthoFinder v.2.2.7. CNVs were also determined by aligning Illumina sequence reads from each species studied to the new *Cp*IA genome sequence to check for variations in read depth coverage. Alignment was performed using BWA-MEM 0.7.17 (Li and Durbin 2009) with default options and the alignment depth per base was calculated using BEDTools genomecov 2.29.2 (Quinlan and Hall 2010) and SAMtools depth 1.6 (Li et al. 2009). Read depth coverage plots were generated using the reshape R package (Wickham 2007; R Development Core Team 2011). To avoid the interference of multiply-mapped regions, only mapped reads were kept for this analysis. For plotting purposes the *Ch*30976 genome was scaffolded using the *Cp*IA chromosomal genomic structure using RagTag v.2.0.1 (Alonge et al. 2019).

Resolving the structure of repetitive subtelomeric regions

Following the CNV analysis, the sequence content of the subtelomeric compressed regions and their *Cp*IA assembly non-compressed chromosomal sequence boundaries containing at least ten genes of Chr1, 7 and 8 were used to build a BLAST database. We then used this database and BLASTN 2.10.0 (Camacho et al. 2009) to detect *Cp*BGF ONT reads capable of aligning to both subtelomeric and chromosomal boundary regions. The few reads meeting these criteria were evaluated and visualized by aligning all subtelomeric ONT reads to the unique pre-subtelomeric regions of Chromosomes 1, 7 & 8 using the Geneious mapper 2019.1.3 (https://www.geneious.com) with medium-sensitivity and minimap2 v.2.22 (Li 2018). Finally, the longest ONT reads were polished with Illumina reads and annotated as previously described for gene content analysis.

Variant analysis, selection prediction and populational analysis

Illumina sequence reads from 136 different isolates of *C. parvum* from different geographical locations (Supplemental Table S8) as well as *Cp*BGF were aligned against the *Cp*IA reference genome sequence using BWA-MEM. The BAM files were parsed to select uniquely mapped reads using Picard (Broad_Institute) and then submitted to the GATK 3.8 Haplotypecaller (McKenna et al. 2010). The results were filtered by mapping quality > 40 and depth coverage > 10. Because mixed infections exist, we restricted analysis to biallelic sites. The individual VCF files were combined into one GVCF file using the GATK tool GenotypeGVCF. After selecting only SNVs from this data, the combined gvcf file was annotated using SnpEff v.4.3 (Cingolani et al. 2012). The SNV variants from the combined annotated gvcf file had their $\pi N/\pi S$ ratio (Nei and Gojobori 1986) estimated using SNPGenie 1.0 (Nelson et al. 2015). To avoid noise in the data and identify top candidates, genes with ratios > 1.5 were detected and denoted as evolving under positive selection in the *C. parvum* population analyzed. The higher threshold of 1.5 was

chosen based on known genes evolving under positive selection, such as *gp60* (Strong et al. 2000) and Inslulinase-like (Zhang et al. 2019).

DATA ACCESS

The sequence data, and annotation generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA573722 and PRJEB3213. Subtelomeric sequences from Chr7 and 8 have GenBank accessions MZ892386, MZ892387 and MZ892388.

ACKNOWLEDGMENTS

We would like to thank Dr, Lihua Xiao for sharing *C. hominis* 30976 and permitting us to update the annotation. This work was supported by Bill and Melinda Gates Foundation grant OPP1151701 to JCK, The Wellcome Trust via its core funding of the Wellcome Sanger Institute (grant WT206194) and NHMRC Investigator Grant (APP1194330) to ARJ.

AUTHORS CONTRIBUTIONS

RPB and JCK designed research; RPB and JCK performed research; AS, JD and BS contributed with new reagents and samples; BA and AJ contributed with analytical tools; MS, KB, AT, MB and JAC contributed Illumina and PacBio sequencing; RPB, YL, KB, AT, RX, EDS, GWC and JCK analyzed data; RPB and JCK wrote the paper and ARJ, BREA, BS, AS and JAC provided feedback.

COMPETING INTEREST STATEMENT

JCK has a financial interest in PacBio.

References

566 Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C,
567 Widmer G, Tzipori S et al. 2004. Complete genome sequence of the apicomplexan,
568 *Cryptosporidium parvum*. *Science* **304**: 441-445.
569 Akiyoshi DE, Feng X, Buckholt MA, Widmer G, Tzipori S. 2002. Genetic analysis of a
570 *Cryptosporidium parvum* human genotype 1 isolate passaged through different host
571 species. *Infect Immun* **70**: 5670-5675.
572 Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz
573 MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes.
574 *Genome Biol* **20**: 224.
575 Ansell BRE, Pope BJ, Georgeson P, Emery-Corbin SJ, Jex AR. 2019. Annotation of the *Giardia*
576 proteome through structure-based homology and machine learning. *Gigascience* **8**.
577 Bankier AT, Spriggs HF, Fartmann B, Konfortov BA, Madera M, Vogel C, Teichmann SA, Ivens
578 A, Dear PH. 2003. Integrated mapping, chromosomal sequencing and sequence analysis
579 of *Cryptosporidium parvum*. *Genome Res* **13**: 1787-1799.
580 Baptista RP, Kissinger JC. 2019. Is reliance on an inaccurate genome sequence sabotaging your
581 experiments? *PLoS Pathog* **15**: e1007901.
582 Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides:
583 SignalP 3.0. *J Mol Biol* **340**: 783-795.
584 Berna L, Rodriguez M, Chiribao ML, Parodi-Talice A, Pita S, Rijo G, Alvarez-Valin F, Robello
585 C. 2018. Expanding an expanded genome: long-read sequencing of *Trypanosoma cruzi*.
586 *Microb Genom* **4**.
587 Bonfield JK, Whitwham A. 2010. Gap5--editing the billion fragment sequence assembly.
588 *Bioinformatics* **26**: 1699-1703.
589 Broad_Institute. Picard Tools.
590 Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG,
591 Lewis SE, Stein L et al. 2016. JBrowse: a dynamic web platform for genome
592 visualization and analysis. *Genome Biol* **17**: 66.
593 Cama VA, Arrowood MJ, Ortega YR, Xiao L. 2006. Molecular Characterization of the
594 *Cryptosporidium parvum* IOWA Isolate Kept in Different Laboratories. *J Eukaryot*
595 *Microbiol* **53 Suppl 1**: S40-42.
596 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
597 BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.
598 Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell
599 M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model
600 organism genomes. *Genome Res* **18**: 188-196.
601 Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. 2005. ACT:
602 the Artemis Comparison Tool. *Bioinformatics* **21**: 3422-3423.
603 Chalmers RM, Smith R, Elwin K, Clifton-Hadley FA, Giles M. 2011. Epidemiology of
604 anthroponotic and zoonotic human cryptosporidiosis in England and Wales, 2004-2006.
605 *Epidemiol Infect* **139**: 700-712.
606 Chan ER, Barnwell JW, Zimmerman PA, Serre D. 2015. Comparative analysis of field-isolate
607 and monkey-adapted *Plasmodium vivax* genomes. *PLoS Negl Trop Dis* **9**: e0003566.
608 Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012.
609 A program for annotating and predicting the effects of single nucleotide polymorphisms,
610 SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*
611 *(Austin)* **6**: 80-92.

612 Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal
613 tool for annotation, visualization and analysis in functional genomics research.
614 *Bioinformatics* **21**: 3674-3676.
615 DeCicco RePass MA, Chen Y, Lin Y, Zhou W, Kaplan DL, Ward HD. 2017. Novel
616 Bioengineered Three-Dimensional Human Intestinal Model for Long-Term Infection of
617 *Cryptosporidium parvum*. *Infect Immun* **85**.
618 Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome
619 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**: 157.
620 English AC, Salerno WJ, Reid JG. 2014. PBHoney: identifying genomic variants via long-read
621 discordance and interrupted mapping. *BMC Bioinformatics* **15**: 180.
622 Fei J, Wu H, Su J, Jin C, Li N, Guo Y, Feng Y, Xiao L. 2018. Characterization of MEDLE-1, a
623 protein in early development of *Cryptosporidium parvum*. *Parasit Vectors* **11**: 312.
624 Feng Y, Ryan UM, Xiao L. 2018. Genetic Diversity and Population Structure of *Cryptosporidium*.
625 *Trends in Parasitology* **34**: 997-1011.
626 GBDDD-Collaborators. 2017. Estimates of global, regional, and national morbidity, mortality,
627 and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of
628 Disease Study 2015. *Lancet Infect Dis* **17**: 909-948.
629 Ghaffari S, Kalantari N, C AH. 2014. A Multi-Locus Study for Detection of *Cryptosporidium*
630 Species Isolated from Calves Population, Liverpool; UK. *Int J Mol Cell Med* **3**: 35-42.
631 Goll MG, Kirpekar F, Maggert KA, Yoder JA, Hsieh CL, Zhang X, Golic KG, Jacobsen SE,
632 Bestor TH. 2006. Methylation of tRNAAsp by the DNA methyltransferase homolog
633 Dnmt2. *Science* **311**: 395-398.
634 Guo Y, Tang K, Rowe LA, Li N, Roellig DM, Knipe K, Frace M, Yang C, Feng Y, Xiao L. 2015.
635 Comparative genomic analysis reveals occurrence of genetic recombination in virulent
636 *Cryptosporidium hominis* subtypes and telomeric gene duplications in *Cryptosporidium*
637 *parvum*. *BMC Genomics* **16**: 320.
638 Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome
639 assemblies. *Bioinformatics* **29**: 1072-1075.
640 Heiges M, Wang H, Robinson E, Aurrecoechea C, Gao X, Kaluskar N, Rhodes P, Wang S, He CZ,
641 Su Y et al. 2006. CryptoDB: a *Cryptosporidium* bioinformatics resource update. *Nucleic*
642 *Acids Res* **34**: D419-422.
643 Heo I, Dutta D, Schaefer DA, Iakobachvili N, Artegiani B, Sachs N, Boonekamp KE, Bowden G,
644 Hendrickx APA, Willems RJL et al. 2018. Modelling *Cryptosporidium* infection in
645 human small intestinal and lung organoids. *Nat Microbiol* **3**: 814-823.
646 Isaza JP, Galvan AL, Polanco V, Huang B, Matveyev AV, Serrano MG, Manque P, Buck GA,
647 Alzate JF. 2015. Revisiting the reference genomes of human pathogenic Cryptosporidium
648 species: reannotation of *C. parvum* Iowa and a new *C. hominis* reference. *Sci Rep* **5**:
649 16324.
650 Jaskiewicz JJ, Sandlin RD, Swei AA, Widmer G, Toner M, Tzipori S. 2018. Cryopreservation of
651 infectious *Cryptosporidium parvum* oocysts. *Nat Commun* **9**: 2883.
652 Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. 2008. SNAP: a
653 web-based tool for identification and annotation of proxy SNPs using HapMap.
654 *Bioinformatics* **24**: 2938-2939.
655 Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A,
656 Nuka G et al. 2014. InterProScan 5: genome-scale protein function classification.
657 *Bioinformatics* **30**: 1236-1240.

Keeling PJ. 2004. Reduction and compaction in the genome of the apicomplexan parasite *Cryptosporidium parvum*. *Dev Cell* **6**: 614-616.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**: 845-858.

Khalil IA, Troeger C, Rao PC, Blacker BF, Brown A, Brewer TG, Colombara DV, De Hostos EL, Engmann C, Guerrant RL et al. 2018. Morbidity, mortality, and long-term consequences associated with diarrhoea from *Cryptosporidium* infection in children younger than 5 years: a meta-analyses study. *Lancet Glob Health* **6**: e758-e768.

Khan A, Shaik JS, Grigg ME. 2017. Genomics and molecular epidemiology of *Cryptosporidium* species. *Acta Trop* doi:10.1016/j.actatropica.2017.10.023.

Khelik K, Lagesen K, Sandve GK, Rognes T, Nederbragt AJ. 2017. NucDiff: in-depth characterization and annotation of differences between two sets of DNA sequences. *BMC Bioinformatics* **18**: 338.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**: 357-360.

Kissinger JC, DeBarry J. 2011. Genome cartography: charting the apicomplexan genome. *Trends Parasitol* **27**: 345-354.

Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, Wu Y, Sow SO, Sur D, Breiman RF et al. 2013. 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.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**: 567-580.

Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE. 2013. Web Apollo: a web-based genomic annotation editing platform. *Genome Biol* **14**: R93.

Li B, Wu H, Li N, Su J, Jia R, Jiang J, Feng Y, Xiao L. 2017. Preliminary Characterization of MEDLE-2, a Protein Potentially Involved in the Invasion of *Cryptosporidium parvum*. *Front Microbiol* **8**: 1647.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)* **34**: 3094-3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li Y, Baptista RP, Sateriale A, Striepen B, Kissinger JC. 2020. Analysis of Long Non-Coding RNA in *Cryptosporidium parvum* Reveals Significant Stage-Specific Antisense Transcription. *Front Cell Infect Microbiol* **10**: 608298.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* **33**: 6494-6506.

Mahmoud M, Gobet N, Cruz-Davalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant calling: the long and the short of it. *Genome Biol* **20**: 246.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. 2010. Target-enrichment strategies for next-generation sequencing. *Nat Methods* **7**: 111-118.

Mazurie AJ, Alves JM, Ozaki LS, Zhou S, Schwartz DC, Buck GA. 2013. Comparative genomics of *Cryptosporidium*. *Int J Genomics* **2013**: 832756.

McEachern MJ, Haber JE. 2006. Break-induced replication and recombinational telomere elongation in yeast. *Annu Rev Biochem* **75**: 111-135.

McEachern MJ, Iyer S. 2001. Short telomeres in yeast are highly recombinogenic. *Mol Cell* **7**: 695-704.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**: 79-84.

Morada M, Lee S, Gunther-Cummins L, Weiss LM, Widmer G, Tzipori S, Yarlett N. 2016. Continuous culture of *Cryptosporidium parvum* using hollow fiber technology. *Int J Parasitol* **46**: 21-29.

Nader JL, Mathers TC, Ward BJ, Pachebat JA, Swain MT, Robinson G, Chalmers RM, Hunter PR, Oosterhout C, Tyler KM. 2019. Evolutionary genomics of anthroponosis in *Cryptosporidium*. *Nature Microbiology*.

Natarajan S, Nickles K, McEachern MJ. 2006. Screening for telomeric recombination in wild-type *Kluyveromyces lactis*. *FEMS Yeast Res* **6**: 442-448.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**: 418-426.

Nelson CW, Moncla LH, Hughes AL. 2015. SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. *Bioinformatics (Oxford, England)* **31**: 3709-3711.

Otto TD, Sanders M, Berriman M, Newbold C. 2010. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26**: 1704-1707.

Pawlowic MC, Somepalli M, Sateriale A, Herbert GT, Gibson AR, Cuny GD, Hedstrom L, Striepen B. 2019. Genetic ablation of purine salvage in *Cryptosporidium parvum* reveals nucleotide uptake from the host cell. *Proc Natl Acad Sci U S A* **116**: 21160-21165.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290-295.

Piper MB, Bankier AT, Dear PH. 1998. A HAPPY map of *Cryptosporidium parvum*. *Genome Res* **8**: 1299-1307.

Ponts N, Fu L, Harris EY, Zhang J, Chung DW, Cervantes MC, Prudhomme J, Atanasova-Penichon V, Zehraoui E, Bunnik EM et al. 2013. Genome-wide mapping of DNA methylation in the human malaria parasite *Plasmodium falciparum*. *Cell Host Microbe* **14**: 696-706.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

748 R Development Core Team. 2011. R: A language and environment for statistical computing. R
749 Foundation for Statistical Computing, Vienna, Austria.
750 Ren Q, Chen K, Paulsen IT. 2007. TransportDB: a comprehensive database resource for
751 cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids*
752 *Res* **35**: D274-279.
753 Rieseberg LH. 2001. Chromosomal rearrangements and speciation. *Trends Ecol Evol* **16**: 351-
754 358.
755 Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein
756 structure and function prediction. *Nat Protoc* **5**: 725-738.
757 Saier MH, Jr., Yen MR, Noto K, Tamang DG, Elkan C. 2009. The Transporter Classification
758 Database: recent advances. *Nucleic Acids Res* **37**: D274-278.
759 San Millan RM, Martinez-Ballesteros I, Rementeria A, Garaizar J, Bikandi J. 2013. Online
760 exercise for the design and simulation of PCR and PCR-RFLP experiments. *BMC Res*
761 *Notes* **6**: 513.
762 Sanderson SJ, Xia D, Prieto H, Yates J, Heiges M, Kissinger JC, Bromley E, Lal K, Sinden RE,
763 Tomley F et al. 2008. Determining the protein repertoire of *Cryptosporidium parvum*
764 sporozoites. *Proteomics* **8**: 1398-1414.
765 Sateriale A, Slapeta J, Baptista R, Engiles JB, Gullicksrud JA, Herbert GT, Brooks CF, Kugler
766 EM, Kissinger JC, Hunter CA et al. 2019. A Genetically Tractable, Natural Mouse Model
767 of Cryptosporidiosis Offers Insights into Host Protective Immunity. *Cell Host Microbe* **26**:
768 135-146 e135.
769 Sateriale A, Striepen B. 2016. Beg, Borrow and Steal: Three Aspects of Horizontal Gene
770 Transfer in the Protozoan Parasite, *Cryptosporidium parvum*. *PLoS Pathog* **12**: e1005429.
771 Slapeta J. 2013. Cryptosporidiosis and *Cryptosporidium* species in animals and humans: a thirty
772 colour rainbow? *Int J Parasitol* **43**: 957-970.
773 Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes
774 that allows user-defined constraints. *Nucleic Acids Res* **33**: W465-467.
775 Striepen B, Pruijssers AJ, Huang J, Li C, Gubbels MJ, Umejiego NN, Hedstrom L, Kissinger JC.
776 2004. Gene transfer in the evolution of parasite nucleotide biosynthesis. *Proc Natl Acad*
777 *Sci U S A* **101**: 3154-3159.
778 Strong WB, Gut J, Nelson RG. 2000. Cloning and sequence analysis of a highly polymorphic
779 *Cryptosporidium parvum* gene encoding a 60-kilodalton glycoprotein and
780 characterization of its 15- and 45-kilodalton zoite surface antigen products. *Infection and*
781 *Immunity* **68**: 4117-4134.
782 Swain MT, Tsai IJ, Assefa SA, Newbold C, Berriman M, Otto TD. 2012. A post-assembly
783 genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nat*
784 *Protoc* **7**: 1260-1284.
785 Tandel J, English ED, Sateriale A, Gullicksrud JA, Beiting DP, Sullivan MC, Pinkston B,
786 Striepen B. 2019. Life cycle progression and sexual development of the apicomplexan
787 parasite C*ryptosporidium parvum*. *Nat Microbiol* **4**: 2226-2236.
788 Troell K, Hallstrom B, Divne AM, Alsmark C, Arrighi R, Huss M, Beser J, Bertilsson S. 2016.
789 *Cryptosporidium* as a testbed for single cell genome characterization of unicellular
790 eukaryotes. *BMC Genomics* **17**: 471.
791 Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012.
792 Primer3--new capabilities and interfaces. *Nucleic Acids Res* **40**: e115.

Vembar SS, Seetin M, Lambert C, Nattestad M, Schatz MC, Baybayan P, Scherf A, Smith ML. 2016. Complete telomere-to-telomere de novo assembly of the *Plasmodium falciparum* genome through long-read (>11 kb), single molecule, real-time sequencing. *DNA Res* **23**: 339-351.

Vinayak S, Pawlowic MC, Sateriale A, Brooks CF, Studstill CJ, Bar-Peled Y, Cipriano MJ, Striepen B. 2015. Genetic modification of the diarrhoeal pathogen *Cryptosporidium parvum*. *Nature* **523**: 477-480.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**: e112963.

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**: e49.

Wickham H. 2007. Reshaping Data with the reshape Package. *Journal of Statistical Software* **21**.

Wilke G, Funkhouser-Jones LJ, Wang Y, Ravindran S, Wang Q, Beatty WL, Baldridge MT, VanDussen KL, Shen B, Kuhlenschmidt MS et al. 2019. A Stem-Cell-Derived Platform Enables Complete *Cryptosporidium* Development In Vitro and Genetic Tractability. *Cell Host Microbe* **26**: 123-134 e128.

Woo YH, Ansari H, Otto TD, Klinger CM, Kolisko M, Michalek J, Saxena A, Shanmugam D, Tayyrov A, Veluchamy A et al. 2015. Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *Elife* **4**: e06974.

Xia J, Venkat A, Bainbridge RE, Reese ML, Le Roch KG, Ay F, Boyle JP. 2021. Third-generation sequencing revises the molecular karyotype for *Toxoplasma gondii* and identifies emerging copy number variants in sexual recombinants. *Genome Res* **31**: 834-851.

Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ et al. 2004. The genome of *Cryptosporidium hominis*. *Nature* **431**: 1107-1112.

Zahedi A, Monis P, Aucote S, King B, Paparini A, Jian F, Yang R, Oskam C, Ball A, Robertson I et al. 2016. Zoonotic *Cryptosporidium* Species in Animals Inhabiting Sydney Water Catchments. *PLoS One* **11**: e0168169.

Zhang S, Wang Y, Wu H, Li N, Jiang J, Guo Y, Feng Y, Xiao L. 2019. Characterization of a Species-Specific Insulinase-Like Protease in *Cryptosporidium parvum*. *Front Microbiol* **10**: 354.

TABLES

Table 1 Comparative *Cryptosporidium* Genome Assembly Statistics

	***Cp*IReF**	***CpIA***	***Ch*30976**	***Ct*UGA55**
Scaffolds	8	8	53	11
Gaps in assembly	10	0	25	97
Total length bp	9,102,324	9,122,263	9,059,225	9,015,713
Compressed regions*	12	6	21	26
Ambiguous nt	18,558	0	1,699	78,408
Telomere #	10	16**	7	8
N50	1,104,417	1,108,396	470,636	1,108,290
GC%	30.23	30.18	30.13	30.25

*Number of compressed regions > 100 nt and > 2× average depth.

**BioProject PRJNA573722 and sequence records MZ892386, MZ892387 and MZ892388

Table 2 - Reannotation Summary Statistics.

	C. parvum IOWA II			*C. hominis*		*C. tyzzeri*
Strains	IOWA II Before*	IOWA II After**	IOWA-ATCC***	UdeA01***	30976***	UGA55***
Total sequence length (bp)	9,102,324	9,102,324	9,122,263	9,043,938	9,059,225	9,015,884
Number of genes	3,886	4,020	4,424	3,863	3,996	4,037
Number of CDSs	3,805	3,944	3,897	3,818	3,959	3,986
Average CDS length	1,794	1,765	1,799	1,785	1,755	1,735
Number of exons	4,104	5,043	5,800	4,546	5,045	5,136
Number of introns	238	1,020	1,370	683	1,040	1,089
Shortest intron (bp)	9	36	36	36	36	22
Pseudogenes	74	114	1	45	88	62
% of genome covered by CDS	75.4	82.1	76.88	76.1	83.6	79.2

*Before refers to the 2007 annotation version available from CryptoDB downloads v.35

**After refers to the 2018 annotation version submitted by our group available from CryptoDB v.36

***Version of the annotation available in CryptoDB v.50

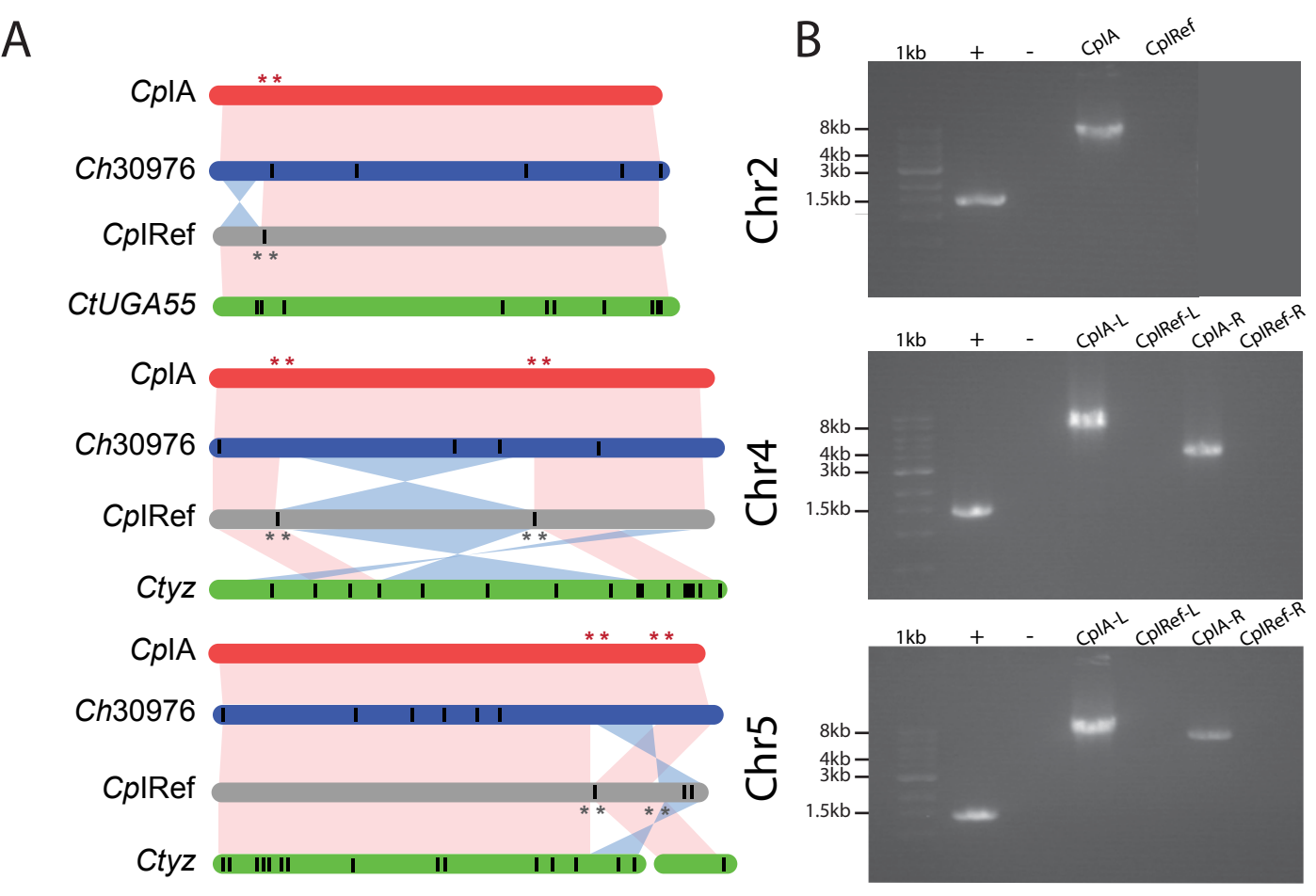
FIGURE LEGENDS

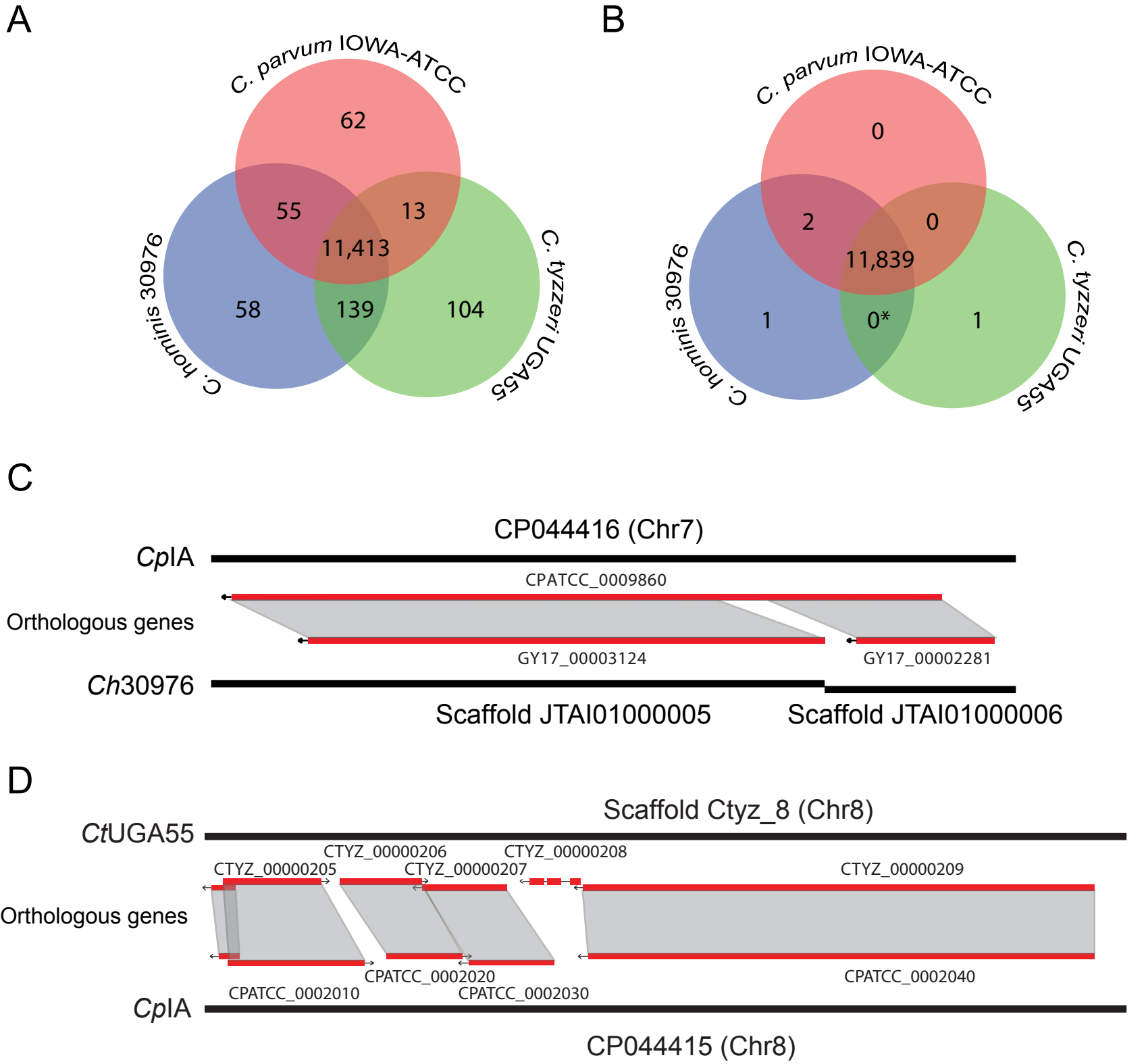
Figure 1. Syntenic relationships between select *Cryptosporidium* chromosome assemblies. (A) Synteny between Chromosomes 2, 4 and 5. Vertical black lines within a chromosome represent known physical gaps. Synteny between chromosomes is shown in pink and inversions in blue. (B) PCR validation using *C. parvum* KSU-1 DNA, (Supplemental Table S4). Lanes 1, 2 & 3 in all gels are 1kb ladder, positive control *Dnmt*2 gene and no template control respectively. The remaining lanes test each orientation of the Left, L, and Right, R, inversion boundaries. Red stars indicate the location of primers designed based on the *Cp*IA assembly and grey stars indicate the same on the *Cp*IRef assembly.

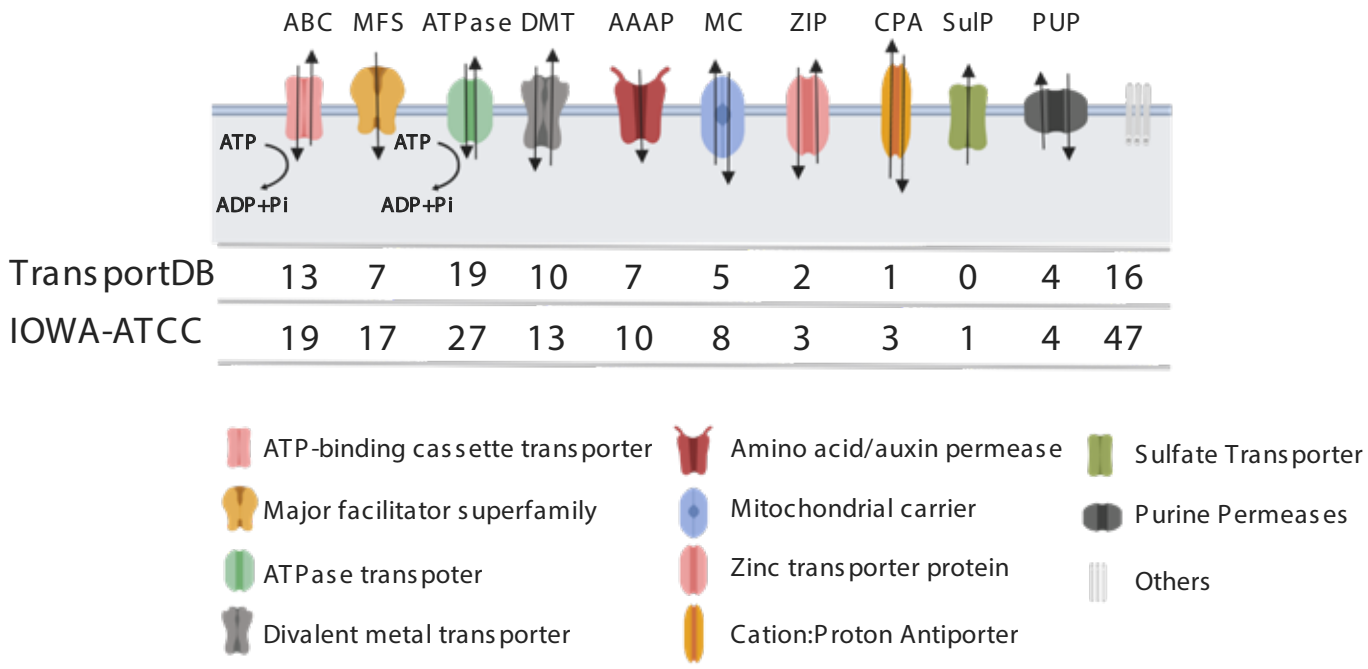
Figure 2. Ortholog distribution of protein-encoding genes reveals few differences between the species. (A) Venn diagram of automated protein orthology assignment between *Cp*IA, *Ch*30976 and *Ct*UGA55; (B) Venn diagram of the same orthologous genes following manual investigation and removal of putative false positives. *The 139 genes shared between *C. hominis* and *C. tyzzeri* in panel A are in complex regions with repeats and gaps and do not have enough evidence to prove their uniqueness at this stage (Supplemental Table S7); (C) Example of a false positive paralog count caused by gene fragments on different scaffolds. (D) Putative unique uncharacterized gene found on Chr8 in *Ct*UGA55.

Figure 3. *Cp*IA assembly and annotation reveal new transporters. Numbers of transporters corresponds to the counts of genes encoding each type of transporter protein. ABC: ATP-binding cassette transporter; MFS: Major facilitator superfamily; DMT: Divalent metal transporter; AAAP: amino acid/auxin permease; MC: mitochondrial carrier; ZIP: Zinc transporter protein; CPA: Cation/Proton Antiporter; SulP: Sulfate Transporter; and PUP: Purine Permeases.

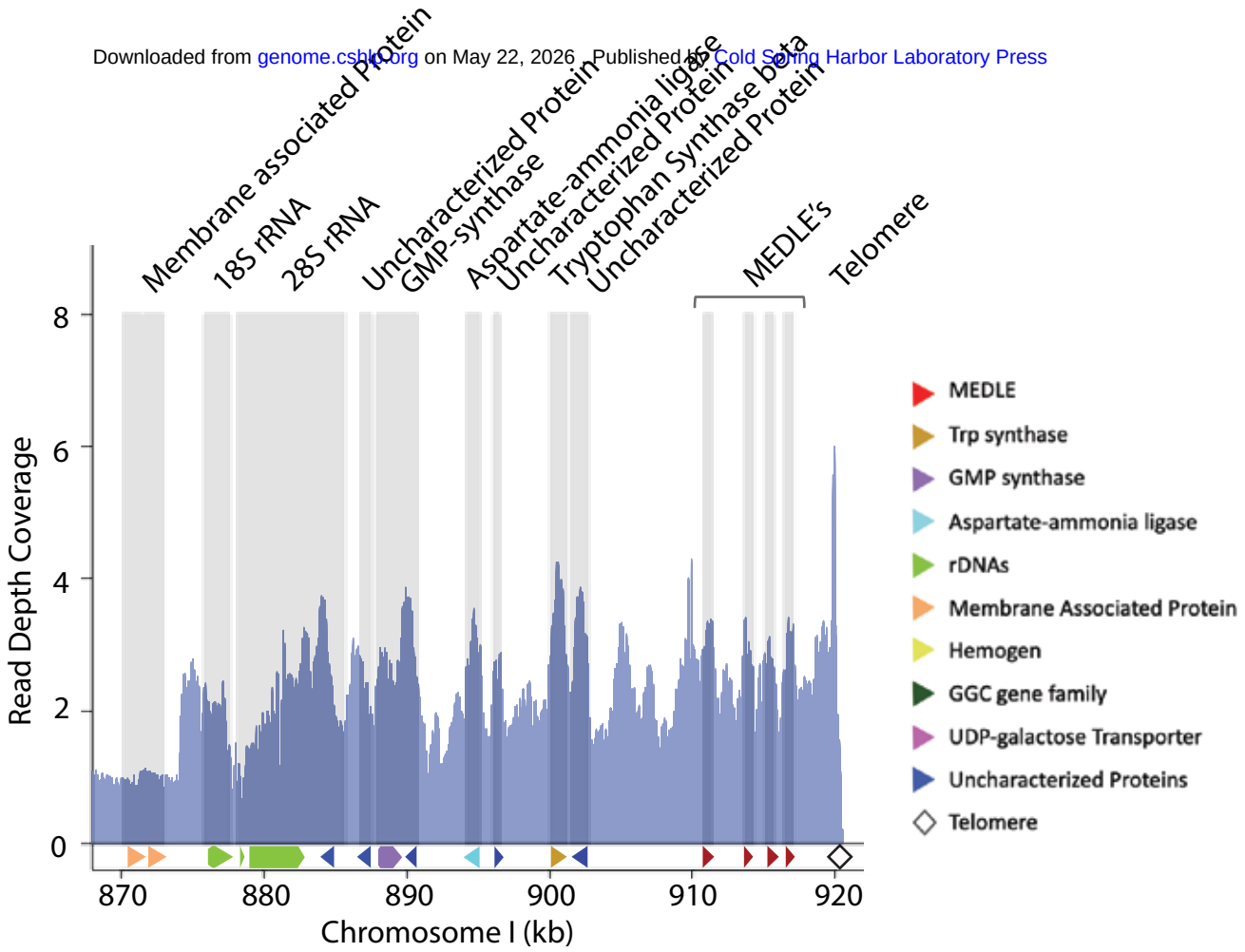
Figure 4. Resolution of repetitive subtelomeric regions found on Chr 1 identifies missing telomeres on Chromosomes 7 and 8. (A) Illumina reads from *Cp*IA are mapped to the *Cp*IA Chr1 long-read assembly subtelomeric region to identify read pileups and estimates of sequence copy number by normalizing against the average genomics Illumina read depth. Vertical grey areas indicate regions with annotated genes. Annotated genes are represented below the shaded regions, the 5.8S rRNA is present but not indicated; (B) Subtelomeric variaton observed on different *Cp*IA chromosomes is supported by *Cp*BGF ONT long reads. Individual ONT long reads provide evidence of at least four different, yet related, subtelomeric regions that extend into the chromosomes that were missing telomeres in *Cp*IA (Chr7 and Chr8) in addition to Chr1. The white and black reference bar above each collection of annotated ONT reads identify the resolved subtelomeric regions (white) and linkage to existing assembly (black). The penultimate read on the Chr7 3' end panel indicates a unique region of insertion (nucleotide positions 1191705-1217462). This region contains mostly uncharacterized proteins and two transferases. Each ONT read is annotated as indicated in the Fig. 5A key.


A
CplA
Ch30976
CplRef
CtUGA55
CplA
Ch30976
CplRef
Ctyz
CplA
Ch30976
CplRef
Ctyz
B
1kb
+
-
CplA
CplRef
8kb
4kb
3kb
1.5kb
Chr2
1kb
+
-
CplA-L
CplRef-L
CplA-R
CplRef-R
8kb
4kb
3kb
1.5kb
Chr4
1kb
+
-
CplA-L
CplRef-L
CplA-R
CplRef-R
8kb
4kb
3kb
1.5kb
Chr5


A
C. parvum IOWA-ATCC
C. hominis 30976
C. tyzzeri UGA55
62
55
13
11,413
58
139
104
B
C. parvum IOWA-ATCC
C. hominis 30976
C. tyzzeri UGA55
0
2
0
11,839
1
0*
1
C
CP044416 (Chr7)
CpIA
CPATCC_0009860
Orthologous genes
GY17_00003124
GY17_00002281
Ch30976
Scaffold JTAI01000005
Scaffold JTAI01000006
D
Scaffold Ctyz_8 (Chr8)
CtUGA55
CTYZ_00000206
CTYZ_00000208
CTYZ_00000205
CTYZ_00000207
CTYZ_00000209
Orthologous genes
CPATCC_0002020
CPATCC_0002040
CPATCC_0002010
CPATCC_0002030
CpIA
CP044415 (Chr8)


ABC
MFS
ATPase
DMT
AAAP
MC
ZIP
CPA
SulP
PUP
ATP
ADP+Pi
ATP
ADP+Pi
TransportDB 13 7 19 10 7 5 2 1 0 4 16
IOWA-ATCC 19 17 27 13 10 8 3 3 1 4 47
ATP-binding cassette transporter
Major facilitator superfamily
ATPase transpoter
Divalent metal transporter
Amino acid/auxin permease
Mitochondrial carrier
Zinc transporter protein
Cation:Proton Antiporter
Sulfate Transporter
Purine Permeases
Others

A



B