SHORT COMMUNICATION

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

Olukemi O. Ifeonu1, Marcus C. Chibucos1, Joshua Orvis1, Qi Su1, Kristin Elwin2, Fengguang Guo3, Haili Zhang3, Lihua Xiao4, Mingfei Sun5, Rachel M. Chalmers2, Claire M. Fraser1, Guan Zhu3, Jessica C. Kissinger6, Giovanni Widmer7 and Joana C. Silva1,8,*,†

1Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA, 2Cryptosporidium Reference Unit, Public Health Wales Microbiology, Singleton Hospital, Swansea SA2 8QA, UK, 3Department of Veterinary Pathobiology, Texas A&M University, College Station, TX 77843, USA, 4Division of Foodborne, Waterborne and Environmental Diseases, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30329, USA, 5Institute of Animal Health, Guangdong Academy of Agricultural Sciences, Guangzhou, Guangdong, China 510640, 6Center for Tropical and Emerging Global Diseases, Institute of Bioinformatics and Department of Genetics, University of Georgia, Athens, GA 30602, USA, 7Department of Infectious Disease and Global Health, Tufts University Cummings School of Veterinary Medicine, North Grafton, MA 01536, USA and 8Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD 21201, USA

*Corresponding author: Institute for Genome Sciences, University of Maryland School of Medicine, 801 W Baltimore St, Biopark II, Rm 645, Baltimore, MD 21201, USA. Tel: +4107060817; E-mail: jcsilva@som.umaryland.edu

One sentence summary: The release of the draft genome sequence, and corresponding annotation, of Cryptosporidium baileyi, C. hominis isolates TU502_2012 and UKH1, and C. meleagridis, will accelerate research on Cryptosporidium parasites.

Editor: Nicholas Thomson

ABSTRACT

Human cryptosporidiosis is caused primarily by Cryptosporidium hominis, C. parvum and C. meleagridis. To accelerate research on parasites in the genus Cryptosporidium, we generated annotated, draft genome sequences of human C. hominis isolates TU502_2012 and UKH1, C. meleagridis UKMEL1, also isolated from a human patient, and the avian parasite C. baileyi TAMU-09Q1. The annotation of the genome sequences relied in part on RNAseq data generated from the oocyst stage of both C. hominis and C. baileyi. The genome assembly of C. hominis is significantly more complete and less fragmented than that available previously, which enabled the generation of a much-improved gene set for this species, with an increase in...
average gene length of 500 bp relative to the protein-encoding genes in the 2004 C. hominis annotation. Our results reveal that the genomes of C. hominis and C. parvum are very similar in both gene density and average gene length. These data should prove a valuable resource for the Cryptosporidium research community.

**Keywords**: Cryptosporidium; C. hominis TU502,2012; Cryptosporidium meleagridis; Cryptosporidium baileyi; genome assembly; annotation

_Cryptosporidium_ parasites (Phylum: Apicomplexa) infect a wide range of vertebrates, from fish to humans, and are the causative agents of cryptosporidiosis in humans (Upton and Current 1985; Tzipori 1988; Widmer and Sullivan 2012). A recent, large, multicenter study of the etiology of moderate-to-severe diarrhea (MSD) in infants in the developing world found _Cryptosporidium hominis_ to be among the four predominant pathogens associated with MSD in children under 5 years of age (Kotloff et al. 2013). Some _Cryptosporidium_ species are capable of zoonotic transmission (Ryan, Fayer and Xiao 2014). Comparative analysis of genomes from diverse _Cryptosporidium_ species and related protoists is essential to fully understand the biology, pathology, host specificity and evolution of this genus.

The reference assembly of _C. parvum_ IOWA II genome (Abrahamsen et al. 2004) is essentially complete, with its eight chromosomes distributed among 18 contigs, including full-length chromosomes. In contrast, the reference assembly of _C. hominis_, based on isolate TU502, published in 2004 (Xu et al. 2004), is a highly fragmented draft genome consisting of 1422 contigs. To accelerate research on these pathogens of public health and veterinary significance, we sequenced, assembled and annotated four _Cryptosporidium_ genome sequences belonging to three species as part of a community White Paper undertaking. Two sequences were generated from a species infective to humans, _C. hominis_ isolates TU502,2012 and UKH1. In addition, sequences were generated from the generalist species _C. meleagridis_, isolate UKMEL1, and from the TAMU-09Q1 isolate of _C. baileyi_, an avian-infecting parasite. All three species are enteric parasites. _Cryptosporidium baileyi_ can complete its entire life cycle in embryonated chicken eggs, making it a useful laboratory model to address some aspects of _Cryptosporidium_ biology. _Cryptosporidium meleagridis_ appears to lack host specificity, as it is known to infect both avian and mammalian species (Akiyoshi et al. 2003).

_Cryptosporidium hominis_ UKH1 and _C. meleagridis_ UKMEL1 oocysts were isolated from fecal samples of naturally infected humans. _Cryptosporidium meleagridis_ oocysts were propagated in immunosuppressed adult CD-1 mice, and _C. hominis_ UKH1 in neonatal gnotobiotic pigs. _Cryptosporidium hominis_ TU502,2012 originates from _C. hominis_ TU502 isolate maintained by serial propagation in gnotobiotic pigs (Tzipori et al. 1994; Xu et al. 2004). _Cryptosporidium baileyi_ oocysts were extracted from experimentally infected embryonated chicken eggs. Prior to isolating DNA, extracted oocysts were purified on density gradients (Widmer, Feng and Tanriverdi 2004) and surface-sterilized with bleach to minimize contamination with host and bacterial DNA. RNA samples were obtained from _C. hominis_ TU502,2012 and _C. baileyi_ TAMU-10GZ1 oocysts ~4 months old, and sequenced to high coverage using strand-specific RNAseq (Parkhomichek et al. 2009). De novo assembly of the gonomic reads was performed using MaSuRCA version1.9 (Zimin et al. 2013) (Table 1).

All the genomes except _C. hominis_ UKH1 were annotated using a semi-automated approach. We trained Augustus (Stanke et al. 2004) using a set of previously manually curated genes. Consensus predictor EVidence Modeler, EVM (Haas et al. 2008), was used to generate annotations based on predictions from Augustus and GeneMark-ES (Borodovsky and Lomsadze 2011), transcripts assembled from RNAseq reads and matches to a set of highly conserved eukaryotic genes—the Core Eukaryotic Genes Mapping Approach genes (Parra, Bradnam and Korf 2007). In addition, 394 genes (~10% of all genes) in the _C. hominis_ TU502,2012 genome were manually annotated using Web Apollo (Lee et al. 2013). The manually curated genes are thought to encode antigens (Ifeonu et al., in preparation). The _C. hominis_ genes TU502,2012 were mapped to the _C. hominis_ UKH1 assembly using GMAP (v2015-12-31), and filtered to include only matches that extend at least over 95% of the sequences and have ≥95% alignment identity at the amino acid level. The final assembly attributes are listed in Table 1. This Whole Genome Shotgun project has been deposited in DDBJ/EMBL/GenBank under the accession numbers listed in Table 1 and the sequences are accessible at CryptoDB (http://CryptoDB.org). These are the first versions of genome sequence assemblies and annotations for each isolate.

The genome of _C. hominis_ isolate TU502 has been sequenced previously (Xu et al. 2004). We resequenced the genome of this isolate, after multiple passages, in an attempt to improve the reference genome assembly and gene set for this species. The resulting _C. hominis_ TU502,2012 genome assembly consists of only 119 contigs, a 10-fold reduction relative to the 2004 assembly. The genome assembly is now more complete, and roughly the same size as that of _C. parvum_, which is also 9.1 Mbp in length (Abrahamsen et al. 2004). The genes in the new annotation are on average 500 bp longer than their counterparts in the original 2004 annotation, resulting in an increase of 17% in the fraction of the genome that encodes for proteins. In order to determine if this gene structural annotation is more accurate than the one published in 2004, we compared the length of all _C. parvum_ IOWA II proteins with their orthologs in either _C. hominis_ TU502 or _C. hominis_ TU502,2012. The distribution of length differences based on the comparison to the 2012 reannotation indeed has lower variance, with an additional 500 genes similar in length between the two species (Fig. 1). Also, there are 538 _C. parvum_ genes without orthologs in the _C. hominis_ TU502 2004 annotation compared to only 288 such cases in the 2012 annotation. Interestingly, while the original _C. hominis_ annotation had a preponderance of genes shorter than their _C. parvum_ orthologs, the current gene set is skewed in the opposite direction (Fig. 1). Whether this difference is real, or a result of remaining gene structure errors in one or both species, remains to be determined. The _C. hominis_ TU502,2012 annotation contains 206 predicted protein-coding genes with no orthologs in _C. parvum_ IOWA II. Of the 3745 predicted protein-coding genes in _C. hominis_ TU502,2012, only 63% are also found in all other annotated _Cryptosporidium_ genomes available to date: _C. parvum_ IOWA II, _C. meleagridis_ UKMEL1, _C. baileyi_ TAMU-09Q1 and _C. muris_ RN66 (Fig. 1). Finally, 110 predicted protein-coding genes are present in the three newly sequenced genomes, but homologs are absent in the current _C. parvum_ predicted proteome. These significant differences in gene content among species are, in all likelihood, due mostly to the limitations of the semi-automated annotation
Table 1. Summary statistics of whole-genome sequence and transcriptome data, assemblies and annotation.

| Isolate: DNA | Cryptosporidium hominis | Cryptosporidium meleagridis | Cryptosporidium baileyi |
|-------------|-------------------------|-----------------------------|------------------------|
| gDNA Illumina library fragment size (bp) | N/A | 460 | 461 |
| No. MiSeq reads | N/A | 6,871,858 | 7,596,410 |
| No. base pairs | N/A | 1,724,836,358 | 1,906,698,910 |
| Assembly size (bp) | 8,743,570 | 9,107,739 | 9,156,091 |
| No. of contigs | 1422 | 119 | 156 |
| Contig N50 | 14,504 | 238,509 | 179,408 |
| Largest contig (bp) | 90,444 | 1,270,815 | 542,781 |
| G + C content (%) | 30.9 | 30.1 | 30.1 |
| No. protein-coding genes | 3,885 | 3,745 | 3,765 |
| Average gene length (bp) | 1,360 | 1,892 | 1,830 |
| Percent coding | 60.4% | 77.8% | 75.2% |
| Accession no. | AAEL00000000 | JIBM00000000 | JIBN00000000 |
| SNP(s) relative to TU502 | 1303 : 2,567 | 718 : 1336 | N/A |
| Isolate: RNA | TU502, 2012 | UKH1 | UKMEL1 |
| No. HiSeq read pairs | 16,568,115 | 92,878,236 | N/A |
| No. expressed genes | 1,868 | 2,454 | N/A |
| Accession no. | SRX481527 | SRX481475 | N/A |

*2004 assembly (Xu et al. 2004).
*Minimum 5X CDS coverage.

Figure 1. Inter- and intraspecies genome-wide comparisons of genome composition. (A) Comparison of protein length between C. parvum and the 2004 and 2012 versions of the C. hominis TU502. (B) Distribution of orthologous gene clusters in five Cryptosporidium species. (C) Distribution of SNPs and short indels among three C. hominis isolates, TU502, TU502_2012 and UKH1. DNA sequence reads from the C. hominis TU502_2012 and UKH1 were mapped against the reference genome assembly of C. hominis TU502, as well as against each other, using BWA (Li and Durbin 2009). SNPs and small indels were identified using GATK (McKenna et al. 2010). Identified variants were further filtered for reliability, according to the following parameter values: (DP < 12) || (QUAL < 50) || (SB > -0.10) || (MQ0 = 2 && (MQ0/(1.0 * DP)) > 0.1). SNPs were categorized as coding and non-coding, given the assembly and the annotation, using VCFtools.

An intense, manual curation effort of the genome annotation of each species is ongoing, and will be essential to validate these results.

Genetic differences among C. hominis isolates were identified by read mapping, followed by calling and filtering of single nucleotide polymorphisms (SNPs) and small insertions/deletions (indels). A total of 10 526 sequence variants were identified in C. hominis TU502_2012 relative to the reference C. hominis TU502 assembly; in contrast, only 4394 sequence variants were found between C. hominis UKH1 and the reference C. hominis. Interestingly, the vast majority of the differences relative to the reference TU502 genome are shared between the two new isolates (Fig. 1). A plausible explanation, which remains to be verified, is that these SNPs common to both new isolates are in fact sequencing errors in the original C. hominis TU502 assembly, which was based on low-coverage Sanger sequencing. This, however, does not explain the fact C. hominis TU502_2012 has more differences relative to TU502 than does UKH1. It is possible that during the approximate 20 passages in gnotobiotic pigs which C. hominis TU502_2012 isolate has experienced between 2004 and 2012, the make-up of the parasite population has shifted. In the absence of methods for cloning and expanding single Cryptosporidium sporozoites, the isolates sequenced to date are likely to be heterogeneous populations (Grinberg and Widmer 2016). In fact, high-throughput sequencing of a polymorphic locus demonstrated the presence of multiple alleles in laboratory and natural Cryptosporidium isolates (Widmer et al. 2015).
We generated RNAseq data for two of the species, *C. hominis* and *C. baileyi*. These data are strand specific, a tremendous advantage when attempting to generate accurate gene-specific expression values in highly gene-dense genomes, where neighboring transcriptional units often overlap (Tretina, Pelle and Silva 2016). The quantity of RNAseq data generated for *C. hominis* UKH1 was six times than that for the TU502_2012 isolate (Table 1). Despite this difference, the relative expression values for each gene are remarkably similar for the two isolates ($r^2 \sim 0.96$; Fig. 2), which supports the strength of the relative expression results. The RNAseq data generated from oocysts indicate that ~50% and ~60% of protein-coding genes are expressed in *C. hominis* TU502_2012 and *C. baileyi*, respectively, during this stage of the life cycle (Table 1). Gene expression is also positively correlated between species ($r^2 \sim 0.51$; Fig. 2), with lactate/malate dehydrogenase (LDH), a GDP-fucose transporter, agrin and the ubiquitous heat shock protein 90 (HSP90) being among the most highly expressed genes in both species. LDH and HSP90 have been shown to be among the top nine most highly expressed genes in *C. parvum* oocysts (Zhang et al. 2012). Genes preferentially expressed in one or the other species may provide a good starting point to investigate biological differences between taxa. Among the genes that differ most in expression level between the two species are pyridine nucleotide-disulphide oxidoreductase, which has a higher level of expression in *C. hominis*, and AhpC/TSA family protein, WD repeat-containing protein 82 and DNA mismatch repair protein msh-2, all of which have higher expression levels in *C. baileyi*.

The work on *Cryptosporidium* genomes and their respective annotations with particular emphasis on the manual curation of the structure and function of all protein-coding genes is continuing. Together with the identification of genes unique to each species and genes with species-specific expression profiles, this work will facilitate the identification of genes responsible for host specificity and other phenotypes relevant to the understanding of cryptosporidiosis.

ACKNOWLEDGEMENTS

The authors would like to thank Saul Tzipori for facilitating the use of gnotobiotic piglets and to the staff of the animal facility for managing the animals.

FUNDING

This work was funded by the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIAID, NIH), under contract number HHSN272200900009C. Genome sequencing was carried out by the Genome Resource Center at the Institute for Genome Sciences, University of Maryland School of Medicine.

Conflict of interest. None declared.

REFERENCES

Abrahamsen MS, Templeton TJ, Enomoto S et al. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. Science 2004;304:441–5.

Akiyoshi DE, Dilo J, Pearson C et al. Characterization of *Cryptosporidium meleagridis* of human origin passaged through different host species. *Infect Immun* 2003;71:1828–32.

Borodovsky M, Lomsadze A. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Curr Protoc Bioinform* 2011;6:1–10.

Grinberg A, Widmer G. Cryptosporidium within-host genetic diversity: systematic bibliographical search and narrative overview. *Int J Parasitol* 2016;46:465–71.

Haas BJ, Salzberg SL, Zhu W et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* 2008;9:R7.

Kotloff KL, Nataro JP, Blackwelder WC 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* 2013;382:209–22.

Lee E, Helt GA, Reese JT et al. Web Apollo: a web-based genomic annotation editing platform. *Genome Biol* 2013;14:R93.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–60.

McKenna A, Hanna M, Banks E et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303.

Parkhomchuk D, Borodina T, Amsistravskiy V et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* 2009;37:e123.
Ifeonu et al.

Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 2007;23:1061–7.

Ryan U, Fayer R, Xiao L. Cryptosporidium species in humans and animals: current understanding and research needs. Parasitology 2014;141:1667–85.

Stanke M, Steinkamp R, Waack S et al. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 2004;32:W309–12.

Tretina K, Pelle R, Silva JC. Cis regulatory motifs and antisense transcriptional control in the apicomplexan Theileria parva. BMC Genomics 2016;17:128.

Tzipori S. Cryptosporidiosis in perspective. Adv Parasitol 1988;27:63–129.

Tzipori S, Rand W, Griffiths J et al. Evaluation of an animal model system for cryptosporidiosis: therapeutic efficacy of paromomycin and hyperimmune bovine colostrum-immunoglobulin. Clin Diagn Lab Immunol 1994;1:450–63.

Upton SJ, Current WL. The species of Cryptosporidium (Apicomplexa: Cryptosporidiidae) infecting mammals. J Parasitol 1985;71:625–9.

Widmer G, Feng X, Tanriverdi S. Genotyping of Cryptosporidium parvum with microsatellite markers. Methods Mol Biol 2004;268:177–87.

Widmer G, Ras R, Chalmers RM et al. Population structure of natural and propagated isolates of Cryptosporidium parvum, C. hominis and C. meleagridis. Environ Microbiol 2015;17:984–93.

Widmer G, Sullivan S. Genomics and population biology of Cryptosporidium species. Parasite Immunol 2012;34:61–71.

Xu P, Widmer G, Wang Y et al. The genome of Cryptosporidium hominis. Nature 2004;431:1107–12.

Zhang H, Guo F, Zhou H et al. Transcriptome analysis reveals unique metabolic features in the Cryptosporidium parvum Oocysts associated with environmental survival and stresses. BMC Genomics 2012;13:647.

Zimin AV, Marcais G, Puig D et al. The MaSuRCA genome assembler. Bioinformatics 2013;29:2669–77.