Pathogens and Disease Advance Access published August 12, 2016

Annotated draft genome sequences of three species of *Cryptosporidium*: *C. meleagridis* isolate UKMEL1*, C. baileyi* isolate TAMU-09Q1, and *C. hominis* isolates TU502_2012 and UKH1

Olukemi O. Ifeonu[a], Marcus Chibucos[a], Joshua Orvis[a], Qi Su[a], Kristin Elwin[b], Fengguang Guo[c], Haili Zhang[c], Lihua Xiao[d], Mingfei Sun[e], Rachel M. Chalmers[b], Claire M. Fraser[a], Guan Zhu[c], Jessica C. Kissinger[f], Giovanni Widmer[g], Joana C. Silva[a,f]

Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, USA[a]; *Cryptosporidium* Reference Unit, Public Health Wales Microbiology, Singleton Hospital, Swansea, UK[b]; Department of Veterinary Pathobiology, Texas A&M University, College Station, Texas, USA[c]; Division of Foodborne, Waterborne and Environmental Diseases, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA[d]; Institute of Animal Health, Guangdong Academy of Agricultural Sciences, Guangzhou, Guangdong, China[e]; Center for Tropical and Emerging Global Diseases, Institute of Bioinformatics & Department of Genetics, University of Georgia, Athens, Georgia, USA[f]; Department of Infectious Disease and Global Health, Tufts University Cummings School of Veterinary Medicine, North Grafton, Massachusetts, USA[g]; Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, Maryland, USA[f]

#Address correspondence to JCS, jcsilva@som.umaryland.edu

Keywords: *Cryptosporidium*, *C. hominis* TU502_2012, *Cryptosporidium meleagridis*, *Cryptosporidium baileyi*, genome assembly, annotation

[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.

ABSTRACT

Human cryptosporidiosis is caused primarily by *Cryptosporidium hominis*, *Cryptosporidium parvum* and *Cryptosporidium 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 *Cryptosporidium 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.

[MAIN TEXT]

Cryptosporidium parasites (Phylum: Apicomplexa) infect a wide range of vertebrates, from fish to humans, and are the causative agents of cryptosporidiosis in humans (Tzipori 1988; Upton and Current 1985; 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 et al. 2014). Comparative analysis of genomes from diverse

Cryptosporidium species and related protists is essential to fully understand the biology, pathology, host specificity, and evolution of this genus.

The reference *Cryptosporidium parvum* IOWA II genome (Abrahamsen et al. 2004) is essentially complete, with its 8 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 1,413 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. Additionally, 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. *C. baileyi* can complete its entire life cycle in embryonated chicken eggs, making it a useful laboratory model to address some aspects of *Cryptosporidium* biology. *C. meleagridis* appears to lack host specificity, as it is known to infect both avian and mammalian species (Akiyoshi et al. 2003).

C. hominis UKH1 and *C. meleagridis* UKMEL1 oocysts were isolated from fecal samples of naturally infected humans. *C. meleagridis* oocysts were propagated in immunosuppressed adult CD-1 mice, and *C. hominis* UKH1 in neonatal gnotobiotic pigs. *C. hominis* TU502_2012 originates from *C. hominis* TU502 isolate maintained by serial propagation in gnotobiotic pigs (Tzipori et al. 1994; Xu et al. 2004). *C. baileyi* oocysts were extracted from experimentally infected embryonated chicken eggs. Prior to isolating DNA, extracted oocysts were purified on density gradients (Widmer et al. 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 less than 4 months old, and sequenced to high coverage using strand-specific RNASeq (Parkhomchuk et al. 2009). *De novo* assembly of the genomic 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 (CEGMA) genes (Parra et al. 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 prep). The *C. hominis* TU502_2012 genes 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 re-sequenced 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 ten-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 500bp longer than their counterparts in the original 2004 annotation, resulting in an increase of 25% 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 re-annotation indeed has lower variance, with an additional 500 genes similar in length between the two species (Figure 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 (Figure 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 3,741 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 (Figure 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 approach used, rather than to true instances of gene gain/loss. 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 4,394 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 (Figure 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 et al. 2016). The quantity of RNAseq data generated for *C. hominis* UKH1 was six times 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 ~ 0.96; Figure 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 ~ 0.51; Figure 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

This work was funded by the National Institute for 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. We thank Saul Tzipori for facilitating the use of gnotobiotic piglets and to the staff of the animal facility for managing the animals.

REFERENCES

Abrahamsen, M. S., et al. (2004), 'Complete genome sequence of the apicomplexan, Cryptosporidium parvum', *Science,* 304 (5669), 441-5.

Akiyoshi, D. E., et al. (2003), 'Characterization of Cryptosporidium meleagridis of human origin passaged through different host species', *Infect Immun,* 71 (4), 1828-32.

Borodovsky, M. and Lomsadze, A. (2011), 'Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES', *Curr Protoc Bioinformatics,* Chapter 4, Unit 4 6 1-10.

Grinberg, A. and Widmer, G. (2016), 'Cryptosporidium within-host genetic diversity: systematic bibliographical search and narrative overview', *Int J Parasitol.*

Haas, B. J., et al. (2008), 'Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments', *Genome Biol,* 9 (1), R7.

Kotloff, K. L., 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 (9888), 209-22.

Lee, E., et al. (2013), 'Web Apollo: a web-based genomic annotation editing platform', *Genome Biol,* 14 (8), R93.

Li, H. and Durbin, R. (2009), 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics,* 25 (14), 1754-60.

McKenna, A., et al. (2010), 'The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Res,* 20 (9), 1297-303.

Parkhomchuk, D., et al. (2009), 'Transcriptome analysis by strand-specific sequencing of complementary DNA', *Nucleic Acids Res,* 37 (18), e123.

Parra, G., Bradnam, K., and Korf, I. (2007), 'CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes', *Bioinformatics,* 23 (9), 1061-7.

Ryan, U., Fayer, R., and Xiao, L. (2014), 'Cryptosporidium species in humans and animals: current understanding and research needs', *Parasitology,* 141 (13), 1667-85.

Stanke, M., et al. (2004), 'AUGUSTUS: a web server for gene finding in eukaryotes', *Nucleic Acids Res,* 32 (Web Server issue), W309-12.

Tretina, K., Pelle, R., and Silva, J. C. (2016), 'Cis regulatory motifs and antisense transcriptional control in the apicomplexan Theileria parva', *BMC Genomics,* 17, 128.

Tzipori, S. (1988), 'Cryptosporidiosis in perspective', *Adv Parasitol,* 27, 63-129.

Tzipori, S., et al. (1994), 'Evaluation of an animal model system for cryptosporidiosis: therapeutic efficacy of paromomycin and hyperimmune bovine colostrum-immunoglobulin', *Clin Diagn Lab Immunol,* 1 (4), 450-63.

Upton, S. J. and Current, W. L. (1985), 'The species of Cryptosporidium (Apicomplexa: Cryptosporidiidae) infecting mammals', *J Parasitol,* 71 (5), 625-9.

Widmer, G. and Sullivan, S. (2012), 'Genomics and population biology of Cryptosporidium species', *Parasite Immunol,* 34 (2-3), 61-71.

Widmer, G., Feng, X., and Tanriverdi, S. (2004), 'Genotyping of Cryptosporidium parvum with microsatellite markers', *Methods Mol Biol,* 268, 177-87.

Widmer, G., et al. (2015), 'Population structure of natural and propagated isolates of Cryptosporidium parvum, C. hominis and C. meleagridis', *Environ Microbiol,* 17 (4), 984-93.

Xu, P., et al. (2004), 'The genome of Cryptosporidium hominis', *Nature,* 431 (7012), 1107-12.

Zhang, H., et al. (2012), 'Transcriptome analysis reveals unique metabolic features in the Cryptosporidium parvum Oocysts associated with environmental survival and stresses', *BMC Genomics,* 13, 647.

Zimin, A. V., et al. (2013), 'The MaSuRCA genome assembler', *Bioinformatics,* 29 (21), 2669-77.

Table 1. Summary statistics of whole-genome sequence and transcriptome data, assemblies and annotation.

	C. hominis			*C. meleagridis*	*C. baileyi*
Isolate: DNA	**TU502**[a]	**TU502_2012**	**UKH1**	**UKMEL1**	**TAMU-09Q1**
gDNA Illumina library fragment size (bp)	N/A	460	461	517	654
No. MiSeq reads	N/A	6,871,858	7,596,410	22,862,044	6,240,960
No. base pairs	N/A	1,724,836,358	1,906,698,910	6,881,475,244	1,566,480,960
Assembly size (bp)	8,743,570	9,107,739	9,156,091	8,973,200	8,493,640
No. of contigs	1413	119	156	57	145
Contig N_{50}	14,504	238,509	179,408	322,908	203,018
Largest contig (bp)	90,444	1,270,815	542,781	732,862	702,637
G + C content (%)	31.7	30.14	30.13	30.97	24.27
No. protein-coding genes	3,994	3,745	3,765	3,758	3,692
Average gene length (bp)	1,360	1,847	1,830	1,844	1,778
Percent coding	60.4%	75.9%	75.2%	77.2%	77.3%
Accession no.	AAEL00000000	JIBM00000000	JIBN00000000	JIBK00000000	JIBL00000000
SNPs relative to TU502[a] synonymous : non-syn		1303 : 2,567	718 : 1336	N/A	N/A
SNPs relative to TU502_2012 synonymous : non-syn		N/A	143 : 339	N/A	N/A
Isolate: RNA		**TU502_2012**	**UKH1**	**UKMEL1**	**TAMU-10GZ1**
No. HiSeq read pairs		16,568,115	92,878,236	N/A	55,829,305
No. expressed genes[b]		1,868	2,454	N/A	2,235
Accession no.		SRX481527	SRX481475	N/A	SRX481530

[a] 2004 assembly (Xu et al. 2004)
[b] Minimum 5X CDS coverage

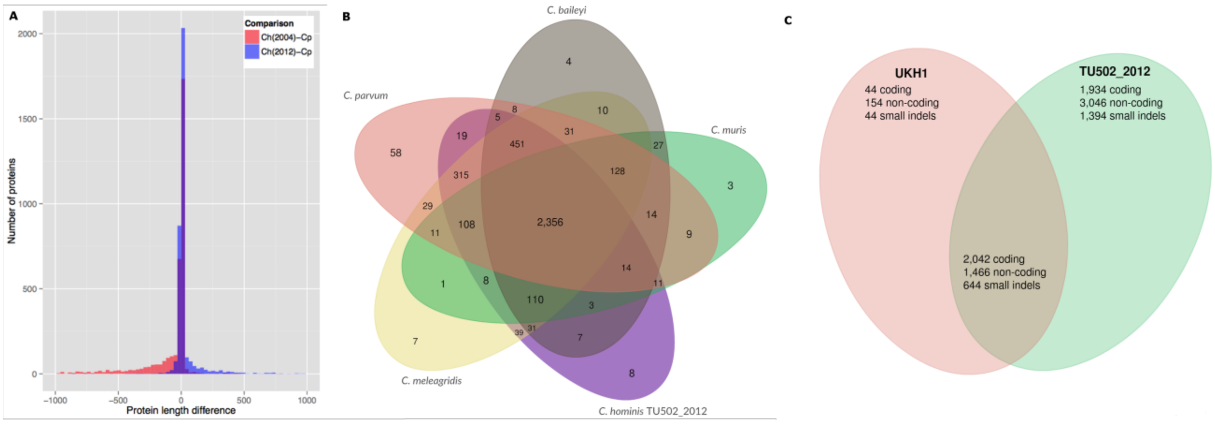

Figure 1. Inter- and intra-species genome-wide comparisons of genome composition. A) Comparison of protein length between *C parvum* and the 2004 and 2012 versions of the *C*. *hominis* TU502. B) Distribution of orthologous gene clusters in five *Cryptosporidium* species. C) Distribution of SNPs and 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.

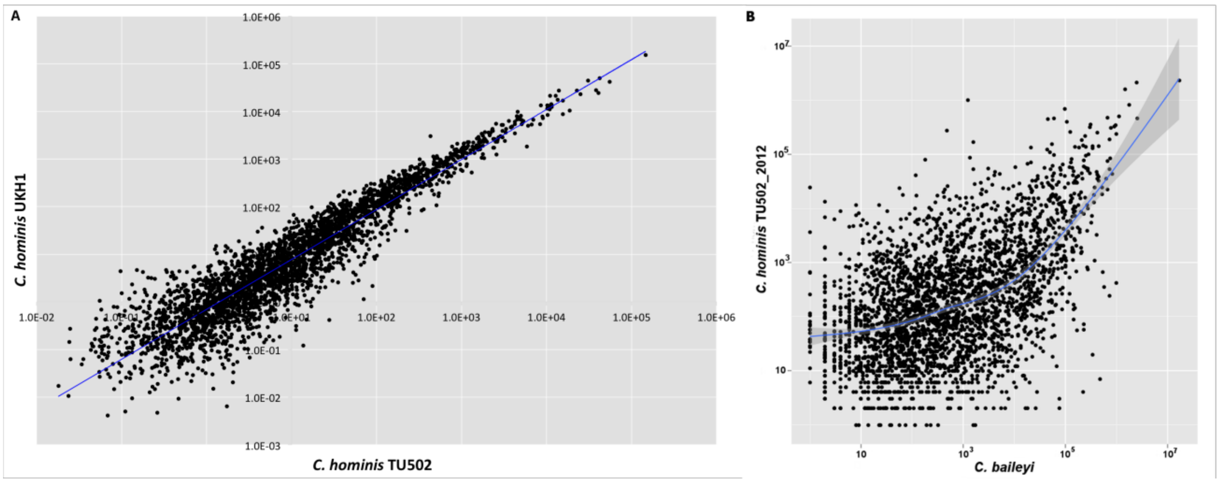

Figure 2. Gene expression in *Cryptosporidium* oocysts is correlated within and between species. A) Correlation in oocyst gene expression is highly correlated between two isolates of *C*. *hominis* (r^2 ~ 96%). B) Correlation in oocyst gene expression is correlated between *C*. *hominis* and *C*. *baylei* (r^2 ~ 51%), particularly among the most highly expressed genes.