GENOMIC RESOURCES FOR *CRYPTOSPORIDIUM* SPECIES, HUMAN PATHOGENS OF PUBLIC HEALTH SIGNIFICANCE IN DEVELOPING COUNTRIES

by

Olukemi O. Ifeonu

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Bioinformatics and Computational Biology

Committee:

_____	Dr. Iosif Vaisman, Committee Chair
_____	Dr. Joana C. Silva, Dissertation Director
_____	Dr. Donald Seto, Committee Member
_____	Dr. Iosif Vaisman, Acting Director, School of Systems Biology
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science
Date: _____	Spring Semester 2017 George Mason University Fairfax, VA

Genomic Resources For *Cryptosporidium* Species, Human Pathogens Of Public Health Significance In Developing Countries

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Olukemi O. Ifeonu
Master of Science
Johns Hopkins University, 2010
Bachelor of Arts
Hood College, 2007

Director: Joana C. Silva, Associate Professor
Department of School of Systems Biology

Spring Semester 2017
George Mason University
Fairfax, VA

Dedication

This is dedicated to my loving husband Mike. Thanks for your constant support and encouragement throughout this journey.

Acknowledgements

I would like to express my sincere appreciation to my advisor Dr. Joana Silva for her amazing guidance and mentorship, and for setting an example of excellence in research. I thank and acknowledge the members of my committee for their input and invaluable help throughout my research. I thank every researcher and collaborator who contributed to this research. I thank and appreciate my parents and my amazing family and friends for their enthusiasm and encouragement. I thank the staff of the College of Science and School of Systems Biology for their guidance throughout my graduate school experience. Finally, thanks to the Fenwick Library for guiding me through the process of putting together and submitting this dissertation.

Table of Contents

List of Tables

List of Figures

List of Abbreviations and/or Symbols

Base pair.....BP
Cryptosporidium hominis Gene Catalog..... ChGC
Global Enteric Multicenter Study ..... GEMS
Moderate-to-serve diarrhea.....MSD

Abstract

GENOMIC RESOURCES FOR *CRYPTOSPORIDIUM* SPECIES, HUMAN PATHOGENS OF PUBLIC HEALTH SIGNIFICANCE IN DEVELOPING COUNTRIES

Olukemi O. Ifeonu, M.S.

George Mason University, 2017

Dissertation Director: Dr. Joana C. Silva

Cryptosporidium species are intracellular protozoan parasites, members of the phylum Apicomplexa that can infect the intestinal or gastric epithelial tissue of vertebrates, causing diarrhea. In immunocompromised individuals and young children, cryptosporidiosis can be life-threatening. *Cryptosporidium* was recently identified as a major cause of diarrhea-induced death of young children in developing countries. Despite the immense public health impact of *Cryptosporidium* infections in developing countries, no significant progress has been made towards developing a vaccine. Biological and technical challenges have impeded traditional vaccinology approaches to identify novel targets for the development of vaccines against this pathogen. The availability of genomic resources for multiple species in the genus has the potential to make a reverse vaccinology approach feasible. This dissertation describes the development and availability of new genomic tools and resources that should prove a valuable resource for

the *Cryptosporidium* research community. This includes the annotated draft genome sequences of three species of *Cryptosporidium*, the *Cryptosporidium hominis* Gene Catalog, and the genome of a strain of anthroponotic *Cryptosporidium parvum*, and its analysis.

Chapter One: 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

Introduction

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). Despite the immense public health impact of *Cryptosporidium* infections in developing countries, several major knowledge gaps still exist regarding diagnosis, as well as the development of drugs and vaccines (Checkley *et al*. 2015). Nitazoxanide, the only FDA approved drug against cryptosporidiosis, is not effective in immunocompromised individuals, and there is no effective vaccine against *Cryptosporidium* (Checkley *et al.* 2015).

Some *Cryptosporidium* species are capable of zoonotic transmission (Ryan, Fayer and Xiao 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 *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 (Parkhomchuk *et al.* 2009). *De novo* assembly of the genomic reads was performed using MaSuRCA version1.9 (Zimin *et al.* 2013) (Table 1.1).

Table 1.1 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

Materials and Methods

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.*, 2016). 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 1.1. This Whole Genome Shotgun project has been deposited in DDBJ/EMBL/GenBank under the accession numbers listed in Table 1.1 and the sequences are accessible at CryptoDB (http://CryptoDB.org). These are the first versions of genome sequence assemblies and annotations for each isolate.

Discussion

The genome of *C. hominis* isolate TU502 had 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 is 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 lengths 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 (Figure 1.1). 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 1.1). 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 (Figure 1.1). 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.

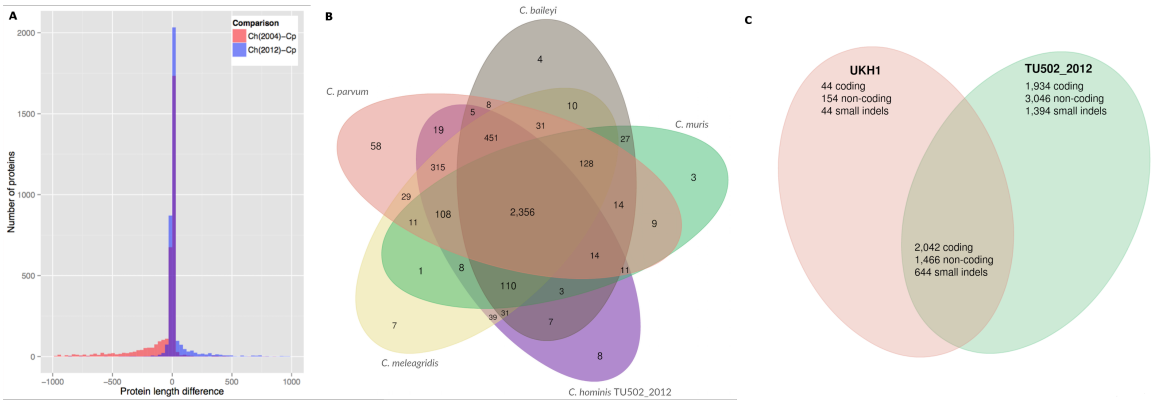

Figure 1.1 Inter- and intraspecies genome-wide comparisons of genome composition.
(A) Comparison of protein length between *C parvum* and the 2004 and 2012 versions of the *C. hominis* TU502. **(B)** Distribution of orthologous gene clusters in five *Cryptosporidium* species. **(C)** Distribution of 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.

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 (Figure 1.1). 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 1.1). Despite this difference, the relative expression values for each gene are remarkably similar for the two isolates ($r^2 \sim 0.96$; Figure 1.2), 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 1.1). Gene expression is also positively correlated between species ($r^2 \sim 0.51$; Figure 1.2), 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*.

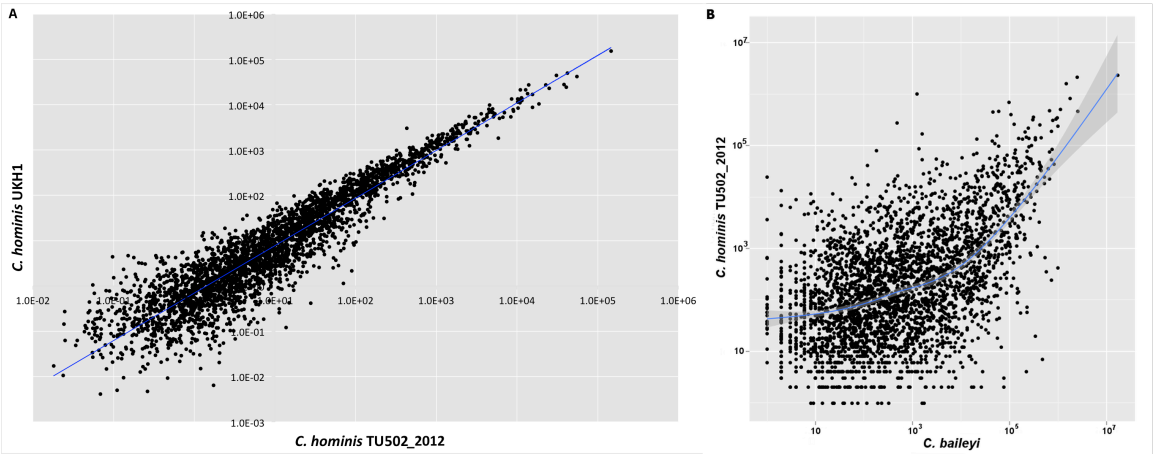

Figure 1.2 Gene expression in *Cryptosporidium* oocysts is correlated within and between species.
(A) Correlation in oocyst gene expression is highly correlated between two isolates of *C. hominis* ($r^2 \sim 96\%$). **(B)** Correlation in oocyst gene expression is correlated between *C. hominis* and *C. baileyi* ($r^2 \sim 51\%$), particularly among the most highly expressed genes. Each point represents the expression value (in RPKM) for each gene with respect to the isolate on the X and Y axis.

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.

Chapter Two: *Cryptosporidium hominis* gene catalog: a resource for the selection of novel *Cryptosporidium* vaccine candidates.

Human cryptosporidiosis, caused primarily by *Cryptosporidium hominis* and a subset of *Cryptosporidium parvum*, is a major cause of moderate-to-severe diarrhea in children under 5 years of age in developing countries and can lead to nutritional stunting and death. Cryptosporidiosis is particularly severe and potentially lethal in immunocompromised hosts. Biological and technical challenges have impeded traditional vaccinology approaches to identify novel targets for the development of vaccines against *C. hominis*, the predominant species associated with human disease. We deemed that the existence of genomic resources for multiple species in the genus, including a much-improved genome assembly and annotation for *C. hominis*, makes a reverse vaccinology approach feasible. To this end, we sought to generate a searchable online resource, termed *C. hominis* gene catalog, which registers all *C. hominis* genes and their properties relevant for the identification and prioritization of candidate vaccine antigens, including physical attributes, properties related to antigenic potential and expression data. Using bioinformatic approaches, we identified ~400 *C. hominis* genes containing properties typical of surface-exposed antigens, such as predicted glycosylphosphatidylinositol (GPI)-anchor motifs, multiple transmembrane motifs and/or signal peptides targeting the encoded protein to the secretory pathway. This set can be narrowed further, e.g. by focusing on potential GPI-anchored proteins lacking homologs in the human genome, but

with homologs in the other *Cryptosporidium* species for which genomic data are available, and with low amino acid polymorphism. Additional selection criteria related to recombinant expression and purification include minimizing predicted post-translation modifications and potential disulfide bonds. Forty proteins satisfying these criteria were selected from 3745 proteins in the updated *C. hominis* annotation. The immunogenic potential of a few of these is currently being tested.

Database URL: http://cryptogc.igs.umaryland.edu.

Introduction

Although young child mortality has dropped impressively since the millennium, almost six million deaths still occur annually in developing countries, with diarrheal diseases remaining the second most common cause of death after pneumonia (Liu *et al.* 2015). The Global Enteric Multicenter Study (GEMS), an enormous case-control study that investigated the burden, etiology and consequences of moderate-to-serve diarrhea (MSD) in children < 5 years of age in four sites in sub-Saharan Africa and three in South Asia (global regions where collectively 80% of young child diarrhea deaths occur) incriminated *Cryptosporidium* as one of the four predominant pathogens overall associated with MSD and as the second most common pathogen during the first two years of life, after rotavirus (Kotloff *et al*. 2013). GEMS also found that *Cryptosporidium* MSD was associated with linear growth stunting the ~60 days following the acute MSD episode and increased by 8.5-fold the risk of death over the ~60-day follow-up compared to matched control children. Although *Cryptosporidium*, a chlorine-resistant pathogen, also occurs in association with sporadic and outbreak water-related transmission in

industrialized countries, it is to address the burden of disease in developing countries that there have been calls to undertake vaccine development efforts.

Two main species of the apicomplexan genus *Cryptosporidium* are associated with human disease. GEMS revealed that 80% of *Cryptosporidium* associated with cases were human-restricted *C. hominis,* while the *C. parvum* strains were also mainly anthroponotic genotypes. The majority of human infections in non-GEMS developing countries is attributed to *C. hominis* and, to a lesser degree, *C. parvum* (Cama *et al*. 2003; Sulaiman *et al*. 2005; Tumwine *et al*. 2005; Xiao *et al*. 2001). Other *Cryptosporidium* species are found in all vertebrate groups, with a few occasionally isolated from humans with diarrhea (Xiao *et al*. 2001).

Vaccination remains one of the most successful and cost-effective methods of preventing the occurrence and spread of serious infectious diseases. The fact that only one parasitic vaccine has been licensed for human use (Mosquirix™ against *Plasmodium falciparum* malaria, approved only in 2015, for use in targeted groups) reflects the challenges associated with the design and development of effective anti-protozoal vaccines. Among the factors limiting the understanding of *Cryptosporidium hominis* biology and the development of anti-cryptosporidial vaccines has been the lack of a robust axenic *in vitro* culture system (Arrowood 2002), although successful *in vitro* cultivation of *C. parvum* has recently been demonstrated (Morada *et al*. 2016).

Reverse vaccinology takes advantage of annotated pathogen genomes to identify genes encoding proteins with properties predicted to induce a host immune response against the pathogen. This approach permits the rational selection of vaccine components

which can be subsequently validated experimentally to determine if they elicit immune responses and confer protection (Donati and Rappuoli 2013; Rappuoli and Covacci 2003; Sette and Rappuoli 2010). The reverse vaccinology approach was first used to successfully identify the four components of the *Neisseria meningitidis* B vaccine (Bexsero®)(Heinson *et al*. 2015; Pizza *et al*. 2000; Vernikos and Medini 2014), wherein the genome sequence of a virulent isolate (MC58) was used to predict candidate surface-exposed or exported proteins. Following a similar approach, Maione and colleagues identified four potential vaccine antigens against Group B streptococcus and demonstrated that a multivalent vaccine formulation using these antigens can confer broad serotype-independent protection (Maione *et al*. 2005). Reverse vaccinology is also being applied to other pathogens for which not licensed vaccines or other mature candidates exist, including *Porphyromonas gingivalis and Chlamydia pneumoniae* (Serruto and Rappuoli 2006). The reverse vaccinology approach is particularly promising for organisms that, like *Cryptosporidium*, are difficult to maintain under routine laboratory conditions (Kelly and Rappuoli 2005; Maione *et al*. 2005; Pizza *et al*. 2000; Pulendran 2009).

Advances in sequencing technologies and genome assembly and annotation methodologies have facilitated the generation of genomics resources for multiple species of *Cryptosporidium* (Heiges *et al*. 2006). *C. parvum* (isolate IOWA II) was the first species with a published genome (Abrahamsen *et al*. 2004). The genome was found to be 9.1 Mbp in length, and its eight chromosomes assembled into thirteen supercontigs, containing 3,807 predicted protein-coding genes with an average length of 1,795 base

pairs (bp). At about the same time the genome of *C. hominis* (isolate TU502) was published (Xu *et al*. 2004). It was sequenced to a much lower depth of coverage because of limitations of biological material and technology available at the time. For example, the lack of conventional animal models to propagate this species limited the amount of DNA that could be generated for sequencing. Consequently, this assembly is comparatively more fragmented, with the likely 8 chromosomes split among 1,413 contigs, grouped into ~240 scaffolds. Recently we generated a much-improved annotated genome assembly for *C. hominis*, isolate TU502_2012 (Ifeonu *et al*. 2016). Herein we report a comprehensive functional annotation, and targeted manual structural validation, of this new *C. hominis* TU502_2012 gene set, with a view to generate a complete list of genes predicted to potentially be sporozoite, and most likely merozoite, surface-expressed. In addition, we developed a searchable online catalog of all *C. hominis* genes and their characteristics of interest in the context of vaccine development, including physical attributes, properties related to antigenic potential and expression data (Figure 2.1). As an example of this approach, we identified a multitude of proteins that could be evaluated as protective immunogens.

Figure 2.1 *Cryptosporidium hominis* gene catalog (ChGC).
The landing page (http://cryptogc.igs.umaryland.edu) includes an overview of ChGC and links to related information and resources. Several data subsets are readily available for download (right hand bar), and the full dataset can be further queried with user-selected criteria (bottom button). Direct links to the definition of each criterion, as well as related publications, are also available (top right).

Materials and Methods

Genomic and transcriptomic data

This study relied on the use of the following genomics data:

- *Cryptosporidium hominis* TU502: WGS (AAEL00000000); assembly and annotation (GCA_000006425.1) (Xu *et al*. 2004)
- *Cryptosporidium hominis* TU502_2012: WGS (JIBM00000000); assembly and annotation (GCA_001593465.1); RNASeq data (SRX481527)
- *Cryptosporidium hominis* UKH1: WGS (JIBN00000000); assembly (GCA_001593475.1);

- *Cryptosporidium parvum* Iowa II: WGS (AAEE01000000); assembly and annotation (GCA_000165345.1) (Abrahamsen *et al*. 2004). Note: this genome was recently re-annotated (Isaza *et al*. 2015), but at the time of this study the updated annotation was not publicly available. Thus, all references to *C. parvum* Iowa II are based on the original annotation.
- *Cryptosporidium baileyi* TAMU-09Q1: WGS (JIBL00000000); assembly (GCA_001593455.1);
- *Cryptosporidium meleagridis* UKMEL1: WGS (JIBK00000000); assembly (GCA_001593445.1);
- *Cryptosporidium muris* RN66: WGS (AAZY02000000); assembly and annotation (GCA_000006515.1);
- *Homo sapiens*: WGS year 2014 (GRCh38.p1); assembly and annotation (GCA_000001405.16) (International Human Genome Sequencing 2004).

The first version of the annotation of the genomes of *C. hominis* UKH1, *Cryptosporidium baileyi* TAMU-09Q1 and *Cryptosporidium meleagridis* UKMEL1 will be released soon (Ifeonu *et al.* 2016).

Functional annotation

The structural and functional attributes of the 3,745 protein-coding genes in the updated *C. hominis* assembly were identified using a variety of approaches. These include BlastP (Altschul *et al*. 1997) searches against the proteome of other Apicomplexa, using the weight matrix BLOSUM62 and an E-value cutoff of $1e^{-5}$,

HMMer version 3.0 (Mistry *et al*. 2013) searches against the PFAM and TIGRfam databases of functional protein domains (Bateman *et al*. 2000), and searches against the InterPro (Apweiler *et al*. 2000) and CDD (Marchler-Bauer *et al*. 2005) databases. Results from these analyses were then parsed using a custom script to assign product names, gene symbols, Enzyme Commission (EC) numbers, and Gene Ontology (GO) terms, where available.

Characterization of surface-expressed/secreted proteins and epitope identification

The targets of protective antibodies on microbial pathogens are typically associated with the surface of the pathogen or the infected host cell. Accordingly, TargetP (Emanuelsson *et al*. 2000; H. Nielsen *et al*. 1997) was used to identify proteins predicted to be targeted to the secretory pathway with high reliability (reliability classes 1 or 2). Proteins were predicted to be GPI-anchored using GPI-SOM (Fankhauser and Maser 2005), PredGPI (Pierleoni *et al*. 2008) and FragAnchor (Poisson *et al*. 2007). The presence of five or more transmembrane helices is a strong indicator of a transmembrane protein; the presence of these transmembrane motifs was determined with TMHMM (Krogh *et al*. 2001; Sonnhammer *et al*. 1998). Prediction of antigens that may constitute robust immunogens was done by analysis of potential MHC class I and MHC class II epitopes with NetMHCpan and NetMHCIIpan, respectively (Hoof *et al*. 2009; Karosiene *et al*. 2013; M. Nielsen *et al*. 2007).

Manual curation of gene structure

Gene structure was manually validated for all genes predicted to be secreted or membrane-associated (determined by the presence of predicted GPI anchors or of at least five transmembrane motifs). The manually curated gene structural components included the location of the methionine start codon, and the location of all intron-exon boundaries. The following data was used as evidence: *C. hominis* strand-specific RNAseq data generated from the oocyst stage (GenBank: SRX481527), "TopHat junctions" (the set of reads predicted by TopHat (Trapnell *et al*. 2009) to span introns), homologous proteins from other *Cryptosporidium* species aligned against the *C. hominis* assembly using GMAP (Wu and Watanabe 2005), and CEGMA proteins, a set of highly conserved eukaryotic genes (Parra *et al.* 2007). Manual validation consisted of visual inspection of each gene model, comparison against all available evidence, and editing when necessary to conform to that evidence. Web Apollo (Lee *et al*. 2013) was used to visualize all evidence tracks and to modify gene models as necessary.

Protein physical attributes

The proteins were characterized according to several physical properties, including predicted isoelectric point (Walker 2005), molecular weight (Walker 2005), numbers of cysteine residues (assumed to reflect potential disulfide bonds), or of potential glycosylation sites. We predicted two types of glycosylation sites, O-glycosylation and N-glycosylation sites, by use of the software NetNGlyc, NetOGlyc and GlycoEP (Chauhan *et al*. 2013; Gupta *et al*. 2004; Steentoft *et al*. 2013).

Homology searches

C. parvum and human homologs were identified by running a BlastP search of *C. hominis* TU502_2012 proteins against the proteomes of *C. parvum* Iowa II (Abrahamsen *et al*. 2004) and human (Lander *et al*. 2001) respectively, with parameter values as described above. The presence of homologs of genes of interest was also determined in four other *Crytosporidium* genomes, namely, *C. parvum* Iowa II*, C. baileyi* TAMU 09Q1*, C. meleagridis* UKMEL1 *and C. muris* RN66. We computed homology clusters of *Cryptosporidium* proteins using the pipeline described by Crabtree and collaborators (Crabtree *et al*. 2007), and used the Sybil comparative platform (Crabtree *et al*. 2007) to visualize and analyze the results.

Identification of SNPs and small insertions/deletions (indels)

Sequence variants, in particular single nucleotide polymorphisms (SNPs) and small indels in *C. hominis* were identified based on the comparison of two strains: *C. hominis* TU502_2012 and *C. hominis* UKH1. In this case, the sequence reads of *C. hominis* UKH1 (SUB482088) were aligned to the new assembly of *C. hominis*, ChTU502_2012, using BWA (Li and Durbin 2009). Sequence data was formatted using SAM tools (Li *et al*. 2009) and Picard tools v.1.79 (http://broadinstitute.github.io/picard), and SNP variant calling and filtering using the Genome Analysis Toolkit GATK v2.2.5 (McKenna *et al*. 2010). Identified variants were filtered according to the following parameter values: (DP < 12) || (QUAL < 50) || (SB > -0.10) || (MQ0 >= 2 && (MQ0/(1.0 * DP)) > 0.1). SNPs that passed the filter were attributed to non-coding or

coding regions using VCFannotator (http://sourceforge.net/projects/vcfannotator) using as reference the annotation of ChTU502_2012.

Expression dataset

Given the lack of *C. hominis* sporozoite RNAseq data, we used transcriptomic data from *C. parvum*. From CryptoDB (Heiges *et al*. 2006), we extracted expression data representing transcriptomes of freshly excysted *C. parvum* sporozoites, as well as data for parasites collected 48- and 96- hours post-infection in HCT-8 cells. These data were generated using SOLiD, paired end, strand-specific RNA sequencing [Lippuner *et al*., unpublished]. In addition, we utilized amino acid data representing excysted sporozoite proteomes. These data originated from solubilized protein preparations analyzed by 2 dimensional electrophoresis LC-MS/MS (Sanderson *et al*. 2008).

Results

Generation of a comprehensive set of putative antigens

We recently completed the sequencing, assembly and annotation of the genome of *C. hominis* genome isolate TU502 from a DNA sample generated in 2012 at Tufts University, named *C. hominis* TU502_2012. The isolate is believed to be the same that was sequenced in 2004 (Xu *et al*. 2004), except that it has been maintained by serial propagation in pigs for an additional eight years. This sequencing effort resulted in a much improved draft genome assembly for *C. hominis*. The *C. hominis* TU502_2012 genome assembly, with 119 contigs, is much less fragmented than the 1,413-contig 2004 assembly (Xu *et al*. 2004), with the largest contig now the length of a chromosome. In

this more comprehensive genome assembly, the average length of protein-coding genes is 500 bp longer than in the original annotation (Ifeonu *et al*. 2016). The additional gene length resulted in a 25% increase in the fraction of the genome that encodes for proteins (Table 2.1). Based on this new gene set, we identified potential vaccine proteins using two bioinformatic approaches (Figure 2.2). In one approach, candidate antigens in *C. hominis* or *C. parvum* were identified from the literature (Bouzid *et al*. 2013; Cevallos *et al*. 2000b; Cevallos *et al*. 2000a; Forney *et al*. 1996; Khramtsov *et al*. 1995; Manque *et al*. 2011; O'Connor *et al*. 2009; O'Hara *et al*. 2004; Okhuysen *et al*. 1996; Perkins *et al*. 1999; Petersen *et al*. 1992; Riggs *et al*. 1997; Strong *et al*. 2000), and their homologs were identified in the new *C. hominis* annotation. In a complementary approach, we used the complete *C. hominis* gene set to identify novel candidate antigens. The structure of all genes identified through either approach was manually validated (see Methods).

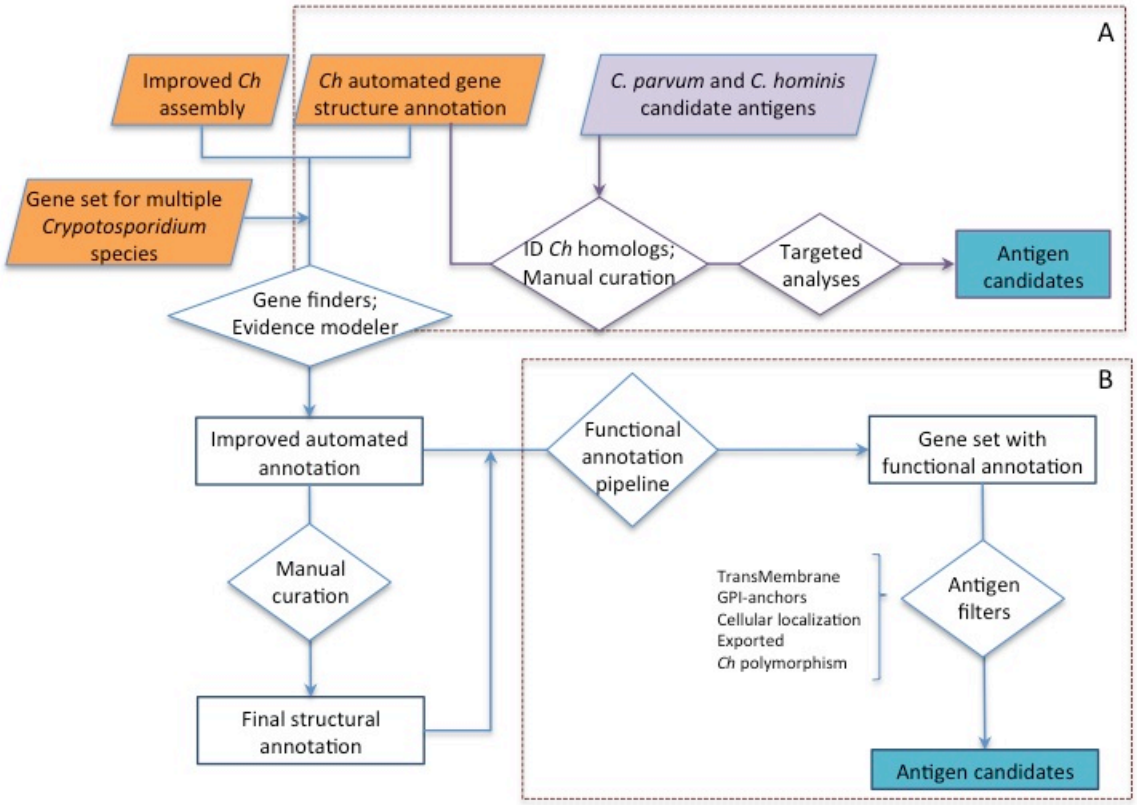

Figure 2.2 Approaches used for antigen identification.
(A) Genes homologous to previously proposed *C. hominis* (Ch) or *C. parvum* antigens (purple) were identified among the gene set from the new *C. hominis* TU502_2012 genome assembly. The structural annotation of the *C. hominis* TU502_2012 was improved using information from related species and several of gene finders. The resulting gene set was assigned functional annotation. This gene set was then screened from desired properties. The gene structure of antigen candidates was manually curated.

Table 2.1 Summary of assembly and annotation statistics for *Cryptosporidium* species.

Species	Isolate	Assembly length (bp)	No. contigs	Largest contig (bp)	No. protein-coding genes	Average gene length (bp)	Percent coding
C. hominis	TU502 (2004)	8,743,570	1413	90,444	3,886	1,360	60.4%
C. hominis	**TU502_2012**	**9,107,739**	**119**	**1,270,815**	**3,745**	**1,847**	**75.9%**
C. parvum	Iowa	9,103,320	13	1,278,458	3,807	1,795	75.3%

Identification of putative antigens by homology to 'known' antigens

The first approach we took was to manually curate the gene structure of all *C. hominis* TU502_2012 genes with homology to known or proposed surface antigens (Figure 2.2). Potential antigens were identified from the literature. Using reverse vaccinology strategies to analyze the *C. hominis* TU502 (2004) genome (Xu *et al*. 2004), Manque and collaborators (Manque *et al*. 2011) identified potential antigens by focusing on proteins associated with the parasite surface, including those possessing multiple transmembrane motifs, signal peptides, GPI signal anchors, and similarities with known pathogenic factors. Other studies have identified *Cryptosporidium* virulence factors using immunological and molecular methods. These virulence factors are predicted to be involved in processes such as adhesion, excystation, locomotion, invasion, membrane integrity, fatty acid metabolism and stress protection (Bouzid *et al*. 2013). Finally, some *Cryptosporidium* antigens were identified through a text search for "antigen" in the CryptoDB database (www.cryptodb.org) (Heiges *et al*. 2006). A total of 302 potential antigens were identified from these references. Of these, 132 proteins (44%) were reported as secreted, 185 (61%) as containing five or more transmembrane domains and 74 (24%) as containing GPI anchor motifs, with a few proteins possessing more than one of these attributes. We re-evaluated these assignments with new or improved methods and found that only 52 of the 74 genes are now predicted to have GPI-anchored domains. We manually curated the structure of all 302 genes in the new *C. hominis* genome assembly (Materials and Methods). In total, 94 of these genes needed to be corrected, resulting in more accurate gene structures than those published in 2004 (Xu *et al*. 2004).

Identification of novel vaccine candidates

Vaccines that elicit antibody-mediated immunity are based on secreted proteins, including toxins, and/or on highly-expressed, surface-exposed or membrane-associated proteins (Doro *et al*. 2009; Maione *et al*. 2005; Pizza *et al*. 2000). We sought to complement the gene set above by utilizing a variety of bioinformatics tools to identify additional genes encoding proteins with these properties, and which might have been missed in previous studies due to incorrect or missing gene models in the 2004 annotation properties (A 2B). Among the complete set of 3,745 protein-coding genes from the improved semi-automated annotation of *C. hominis* (Table 2.1), we identified 105 new antigen candidates, 41 of which have five or more transmembrane domains, 37 with GPI anchor motifs and 29 that are targeted to the secretory pathway. We confirmed that, relative to the original assembly, these 105 genes are either newly identified, genes with a considerably altered structure, or genes newly predicted using new software. The structure of these 105 new candidates was manually curated as described above.

A total of 407 potential antigens were identified using at least one approach: 209 of the 302 previously identified putative antigens were also detected using our bioinformatic screen (Figure 2.3A); of the remaining 93 genes, approximately half have altered gene structures that may change the region containing signal peptides, which likely explains why they are no longer selected according to the criteria used in our screen.

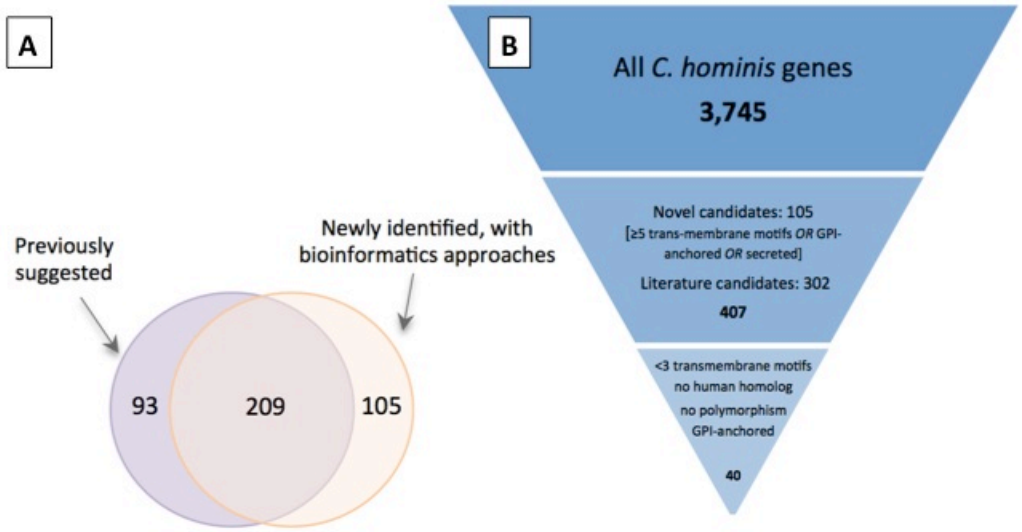

Figure 2.3 Selection of potential of *Cryptosporidium* vaccine candidates.
(A) Overlap between set of potential antigens, one collected from the literature (purple) and the other generated using a bioinformatic screen for genes with predicted GPI-anchor motifs, secretion signals or at least five transmembrane motifs (orange). Of the total 407 potential antigens, roughly one-half were identified with both approaches. (B) Down-selection of genes to be used in immunogenicity experiments. The complete gene complement was first reduced by 90% to 407 candidates from (A), and a further 90% reduction resulted from the use of stricter criteria.

Rational selection of candidate vaccine proteins

The two combined approaches resulted in a set of 407 manually curated, potential antigens. In order to prioritize these genes, we characterized them according to relevant polymorphic and physicochemical properties. These properties include the possibility that the encoded protein will undergo post-translational modifications, suggestive of an intricate process of protein folding. In addition, we considered homology information, both across the *Cryptosporidium* genus and relative to the human proteome as cross-reactive antigens may produce undesired adverse effects upon vaccination.

Antigens often evolve rapidly, as a result of the selective pressure imposed by the host's immune system (Holmes 2004; Yang and Bielawski 2000). Therefore, a relatively

high rate of non-synonymous polymorphism and evidence of balancing selection have been used as criteria to identify new vaccine antigens (Conway 2015; Mu *et al*. 2007). However, evidence is now mounting that high rate of polymorphism in vaccine antigens contributes to vaccine evasion (Neafsey *et al*. 2015; Ouattara *et al*. 2013; Takala *et al*. 2009). In order to identify, and possibly eliminate, polymorphic loci from the pool of potential vaccine candidates, we estimated the number of single nucleotide polymorphisms between publicly available *C. hominis* isolates TU502_2014 and UKH1. A total of 230 protein-encoding genes have amino-acid polymorphisms between these two isolates. In addition, we made use of publicly available gene expression data for *C. parvum*, to determine which genes are expressed during the sporozoite stage, since neutralizing antibodies are likely to target proteins expressed during this stage of development. Of the 3,745 predicted protein-coding genes, 3597 are predicted to be expressed in the sporozoite stage, even though transcript abundance varies widely among genes.

Several additional selection filters were created based on homology information. All proteins with detectable homology to the human proteome were identified. In addition, we determined the taxonomic distribution of each *C. hominis* gene across the genus. These filters allow the elimination of potential antigens that may induce cross-reactions with human genes, and the rapid assessment of the potential taxonomic breadth of specific antigens.

Since proteins are often expressed in bacterial systems, the number and type of post-translational modifications are important considerations when choosing adequate

vaccine candidates. Glycosylation is a type of posttranslational modification resulting from the addition of N- and O-linked oligosaccharides to proteins. It assists in protein structural folding, transport, and other functions (Schwarz and Aebi 2011; Van den Steen *et al*. 1998). Studies indicate that N-glycosylation of proteins is a rare event in apicomplexan parasites, even though it is an important post-translational modification in other eukaryotic phyla (Dieckmann-Schuppert *et al*. 1992; Dieckmann-Schuppert *et al*. 1994; Kimura *et al*. 1996; Luk *et al*. 2008; Odenthal-Schnittler *et al*. 1993). For the full set of proteins, the median number of predicted N- and O-glycosylation sites per protein was 5 and 8, respectively, but both distributions were highly skewed, with maximum values ≥100. For the subset of 407 potential antigens, the median number of predicted N- and O-glycosylation sites per protein was 5 and 3, respectively. The median number of cysteine residues per protein, which can also be modified post-translation, was 7, with a maximum number of 227. For the subset of 407 selected genes, the median number of cysteine residues was 9 per protein with a maximum number of 151. In most cases, the properties significant for the selection of candidate antigens have a higher rate of occurrence in the subset of 407 genes predicted to encode potential antigens compared to the full dataset (Table 2.2). Of these 407 genes, 33 were found to have amino acid polymorphism between the two *C. hominis* genomes and 216 had human homologs. Eliminating these, and further selecting genes with at most 2 predicted transmembrane motifs and genes predicted to be GPI-anchored, resulted in a list of 40 potential antigens, 39 of which have *C. parvum* homologs, that can be considered for further investigation as vaccine candidates (Figure 2.3). These can be further down-selected based on properties

relevant for protein expression and with consideration of the chosen expression system, such as optimal isoelectric point for biochemical purification or optimal molecular weight for expression.

Table 2.2 Distribution of properties significant for the selection of candidate antigens in the full dataset and subset of candidate antigens.

Desired properties	Full dataset (3,745)	Candidate antigens (407)
Cellular localization: secreted	1%	9%
Predicted GPI-anchored	2%	16%
>= 5 Transmembrane motifs	6%	56%
<= 6 Cysteine residues	44%	34%
No. N- glycosylation sites*	11%	9%
No. O-glycosylation sites*	19%	32%
No. SNPs (strains TU502_2012 vs. UKH1)	94%	92%
No. human homolog	52%	54%
Conserved in *C. hominis*, *C. meleagridis, C. parvum*	60%	65%

*Using NetNGlyc, NetOGlyc respectively

Cryptosporidium gene catalog

We created a *C. hominis* gene catalog based on all the properties described above. The catalog is freely available online (http://cryptogc.igs.umaryland.edu). It contains all *C. hominis* genes and their characteristics, including physical attributes, properties related to antigenic potential and expression data (Figure 2.4). Users can sort or filter the genes based on each characteristic. For example, a query for proteins targeted to the secretory pathway, with no human homologs and at most ten cysteine residues results in 14 hits (Figure 2.5). A quick query also shows that the estimated molecular weight for *C. hominis* proteins varies between 6.12 and 991.2 kDa, equivalent to 55 to 8,756 amino acid residues.

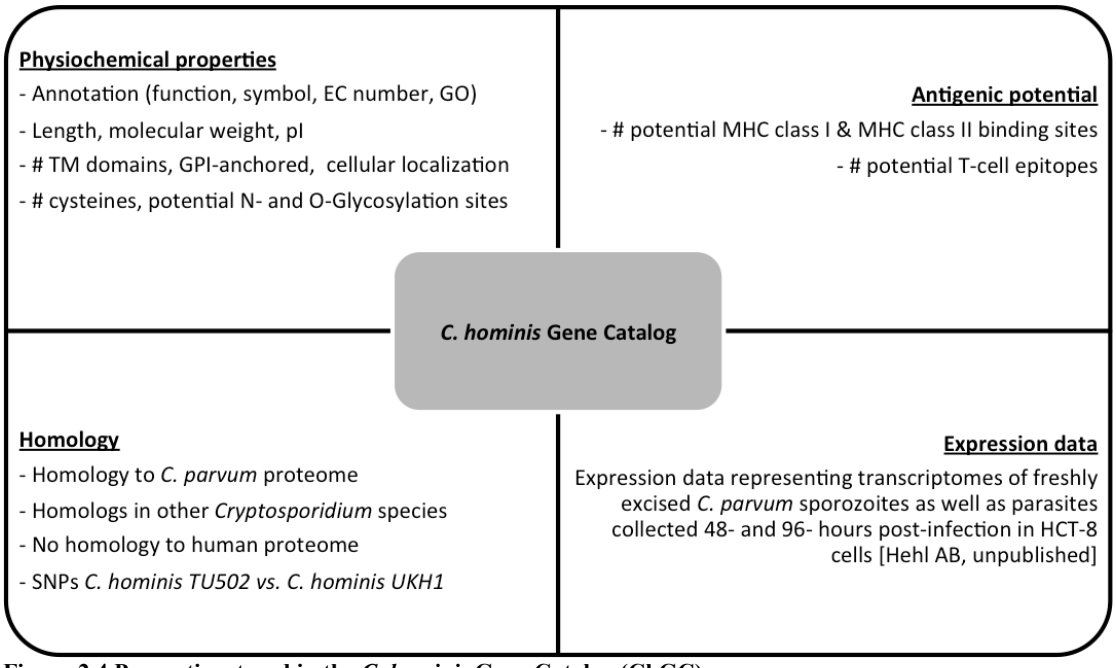

Figure 2.4 Properties stored in the *C. hominis* Gene Catalog (ChGC).
The database contains a variety of searchable properties for each gene, including physicochemical properties, gene expression data, presence of potential T-cell epitopes and distribution of detectable homologs across the *Cryptosporidium* genus and in the human genome.

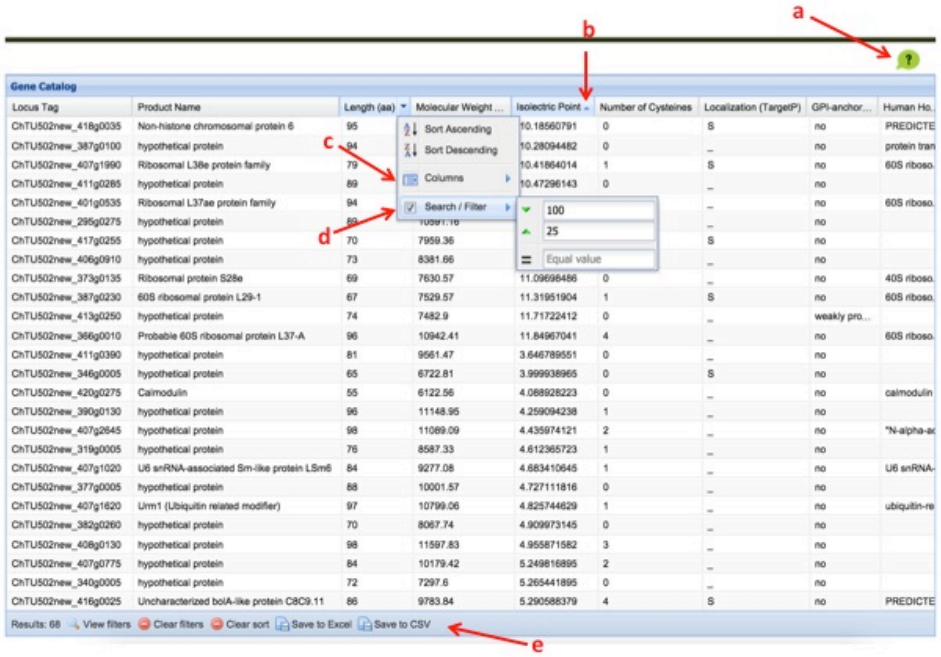

Figure 2.5 The ChGC interface.
Key elements: (a) 'Help' button; (b) click on a column header to sort by that column; (c) 'columns' menu available in the drop-down menu on any column header is used to add hidden, or remove visible, columns; (d) 'Sort/Filter': multiple columns can be filtered to generate customized datasets of interest; (e) filtered datasets can be downloaded as an Excel or a CSV file, using these buttons.

Three sets of genes readily available for download, both in nucleotide and amino acid sequence fasta formats include: all genes, genes that encode predicted GPI-anchored proteins and those whose products are predicted to be secreted. In addition, users can download the nucleotide and amino acid sequences of genes that meet specific user-defined criteria (Figure 2.5). The table of properties for all or a subset of filtered genes can also be downloaded in excel or comma separated values (CSV) format.

Discussion

The GEMS (Kotloff *et al*. 2013) was designed to measure the burden, identify the major etiologic agents and assess the consequences of moderate-to-severe diarrhea (MSD) in children < age 5 years in the developing world. One conclusion of the study was the recognition that targeting the top 4-5 ranked diarrheal pathogens with effective interventions could reduce considerably the global morbidity and mortality burden of MSD.

Surprising to many was the finding that *Cryptosporidium* ranked second as the most important attributable pathogen associated with MSD in children below the age of two years. Whereas vaccines against the other 3 major pathogens either exist (rotavirus) or are undergoing clinical evaluation (enterotoxigenic *E. coli* and shigellosis), efforts to develop a vaccine to protect humans against cryptosporidiosis have made little progress and no candidate has entered clinical trials. The advent of antiretroviral therapy and its widespread use in sub-Saharan Africa has markedly diminished the number of HIV-infected individuals that manifest overt immunodeficiency and as a result the frequency

of cryptosporidiosis has in turn diminished along with interest and funding to combat this infection. GEMS' revelation of the importance of *Cryptosporidium* has renewed interest in developing preventive as well as improved therapeutic measures to control in infants and toddlers in developing countries, including advocacy for developing vaccines. Given the practical obstacles associated with laboratory study of this parasite (Arrowood 2002), reverse vaccinology is an attractive option to identify and prioritize antigens that may prove useful for the development of a well tolerated and effective vaccine to prevent cryptosporidiosis.

With this in mind, our team has recently re-sequenced the TU502 isolate of *C. hominis*, assembled and annotated the genome, now designated TU502_2012 (Ifeonu *et al*. 2016). The improved gene set, consisting of 3,745 protein-coding genes, should provide the opportunity for new *in silico* analyses to identify potential immunogens. We are making this genomic database publicly available, with a view to stimulate additional investigators with expertise in reverse vaccinology to undertake research to develop *Cryptosporidium* vaccine candidates. Once *C. hominis* antigens of interest are identified, various vaccinology approaches can be adapted to assess their immunogenicity. Examples include assessment of the immune responses elicited in animal models or humans following immunization with protozoal antigens expressed in bacterial (C. Gonzalez *et al*. 1994; C. R. Gonzalez *et al*. 1998; Ruiz-Perez *et al*. 2002) or viral vectors (Biswas *et al*. 2014; de Barra *et al*. 2014; Stewart *et al*. 2007), as virus-like particles (Jones *et al*. 2013; Ord *et al*. 2014), as nanoparticles (Burkhard and Lanar 2015) or fused to carrier proteins, as has been done with *P. falciparum* and *Leishmania* proteins (Biswas

et al. 2014; Burkhard and Lanar 2015; de Barra *et al*. 2014; C. Gonzalez *et al*. 1994; C. R. Gonzalez *et al*. 1998; Jones *et al*. 2013; Ord *et al*. 2014; Ruiz-Perez *et al*. 2002; Stewart *et al*. 2007). Since *Cryptosporidium* is an intestinal protozoan, oral as well as parenteral routes of administration of the candidate vaccines should be studied, with and without adjuvants. Recent progress with a well-tolerated adjuvant for orally administered vaccines increases interest in a mucosal vaccine strategy (El-Kamary *et al*. 2013).

Recently, genome sequences of additional isolates of *C. parvum* and *C. hominis* have become publicly available in CryptoDB (Heiges *et al*. 2006). As annotation information for these genomes becomes available, a comparative analysis among *Cryptosporidium* species and isolates may help identify new antigens that will prove to have diagnostic value, since species identification currently entirely depends on cumbersome molecular genetic tools. The database may also help in the development of improved diagnostics of *Cryptosporidium* infection that may allow immunoassays that can identify the prevalent *Cryptosporidium* species in populations and geographic areas. Improved assays for species and sub-species differentiation can help elucidate the reservoirs of *Cryptosporidium*, likely modes of transmission and geographic spread, all of which can help formulate specific control measures.

Chapter Three: The Genome of Anthroponotic *Cryptosoridium parvum*

Introduction

Cryptosporidium spp. are known to infect a wide range of vertebrate hosts including fish, birds and mammals, and some species, such as *Cryptosporidium parvum,* are capable of both zoonotic and anthroponotic transmission. The existence of zoonotic *C. parvum* strains suggests that the infection originated from the primary reservoir of the species, most likely cattle, while the presence of anthroponotic strains suggests that the species can acquire the ability for human-to-human transmission.

The large case-control study, named Global Enteric Multicenter Study (GEMS), revealed that cryptosporidiosis infections in the developing world are caused primarily by the human pathogen *Cryptosporidium hominis,* followed by *C. parvum* (Sow *et al*. 2016). A subsequent, carefully controlled quantitative molecular study showed *Cryptosporidium* to be one of four most common causes of moderate to severe diarrhea (MSD) in young children in Africa and Asia (Liu *et al*. 2016). Of the *Cryptosporidium*-positive MSD cases, the human pathogen *C. hominis* was detected in about 78% of the cases, while *C. parvum*, primarily a pathogen of cattle, was detected in another 10% of the cases. Among the *C. parvum*-positive MSD cases tested, about 92% were caused by anthroponotic strains (Sow *et al*. 2016). Anthroponotic strains refer to strains that are restricted to the human host. Other studies also support this observation, having shown that, in developing countries, *C. parvum* infections in humans are mostly caused by anthroponotic strains,

while zoonotic infections seem to be dominant in developed countries (Xiao 2010). Anthroponotic *(Xiao 2010).* Anthroponotic *Cryptosporidium* species and strains are characterized by the presence of a distinct group of alleles of the gene that encodes the 60 kDa sporozoite surface glycoprotein, GP60 (Xiao 2010).

Despite the tremendous public health impact of anthroponotic *C. parvum* infections in developing countries, until now *C. parvum* genomic resources have been based on zoonotic isolates, particularly the strain IOWA II, the source of the reference genome for the species, and the first *Cryptosporidium* genome to be published (Abrahamsen *et al*. 2004). Seven other zoonotic *C. parvum* isolates, named UKP 2 – 8, were also recently published (Hadfield *et al*. 2015).

Here, we describe the first assembled genome of an anthroponotic *C. parvum* strain, isolate TU114. Comparison between the genomes of anthroponotic and zoonotic isolates may enable a more comprehensive understanding of transmission mechanisms of cryptosporidiosis in humans and other animals, and aid vaccine development research against anthroponotic subtypes of *C. parvum*.

Materials and Methods

Oocyst isolation and DNA extraction

C. parvum TU114 oocysts were originally isolated in 2003 from a Ugandan child and maintained by serial propagation in immunosuppressed mice (Widmer *et al*. 2012). To generate DNA for sequencing, immunosuppressed mice were orally inoculated with oocysts, and new oocysts resulting from this infection collected in the feces. The oocysts

were briefly suspended in a 10% solution of commercial bleach to destroy foreign DNA, including bacterial contaminants and free nucleic acid molecules. Subsequently, the oocysts were subjected to three cycles of freezing and thawing to lyse the thick oocyst shell and release the DNA. Genomic DNA was extracted using the HighPure DNA isolation kit (Roche Diagnostics, Indianapolis, IN).

Sequencing and assembly

DNA sequencing was performed using both Illumina and Pacific Biosciences (PacBio) sequencing technologies. An Illumina HiSeq 2000 platform was used to generate 27,565,702 raw 101 bp-long paired-end reads. A PacBio RS II platform was used to generate 920,622 raw sequence reads using PacBio's single molecule, real-time (SMRT) technology. SPAdes version 3.9.0 (Bankevich *et al*. 2012) was used to generate a hybrid assembly using both Illumina and PacBio reads, and with a read coverage cutoff value of 100. Contigs with highest sequence similarity to non-apicomplexan taxa were removed (see next section). Finally, we eliminated contigs less than 1,000 bp in length as these were partially or completely contained in larger contigs.

Removal of bacterial contamination

A preliminary analysis of the raw sequence reads revealed the presence of substantial bacterial contamination. About 60% of the reads had very high sequence similarity to bacterial genomes, including those of *Escherichia coli*, *Citrobacter* spp. and *Shigella* spp. To remove this contamination, we mapped the reads against a custom

database with the genome sequence of sixty prevalent bacterial taxa, using Bowtie2 (Langmead and Salzberg 2012) and BWA-MEM (Li and Durbin 2010) for the Illumina and PacBio reads, respectively. We removed reads that aligned to bacterial genome sequences, and assembled the remaining reads.

We ran a second round of contaminant filtering by searching the assembled sequences against the NCBI's Nucleotide collection database (NT) using MegaBLAST (Chen *et al*. 2015), and removing all contigs for which the top hit was to a non-apicomplexan sequence.

Annotation

Structural gene annotation was performed using a pipeline consisting of the following steps: first, the repeat regions of genome were masked with RepeatMasker version 4.0.2 (Tarailo-Graovac and Chen 2009); then, gene calling was conducted using GeneMark-ES (Borodovsky and Lomsadze 2011), a self-training program that, in our experience, has been by far the most accurate gene caller for the annotation of apicomplexan genes (Ifeonu *et al*. 2016b; Silva *et al*. 2016; Tretina *et al*. In preparation). In a parallel approach, we used GMAP (Wu and Watanabe 2005) to map the genes from the reference *C. parvum* Iowa II strain to this new *C. parvum* TU114 assembly. We plan to sequence RNA and run additional gene callers including BRAKER-1 (Hoff *et al*. 2016) and AUGUSTUS (Stanke *et al*. 2004). We will also perform manual curation based on these results, using Web Apollo (Lee *et al*. 2013).

Sequence variant identification

In order to identify sequence variants (single nucleotide polymorphisms, SNPs; insertion/deletions, indels) between *C. parvum* TU114 and *C. parvum* IOWA II, *C. parvum* TU114 Illumina reads were mapped to the reference *C. parvum* IOWA II assembly using Bowtie2 (Langmead and Salzberg 2012), and further processed using SAMtools (Li *et al*. 2009) and Picard tools v.1.79 (http://broadinstitute.github.io/picard), for format changes and removal of duplicate reads. SNPs and small indels were called using the GATK version 3.7 (McKenna *et al*. 2010), using Haplotypecaller with the following settings: --emitRefConfidence GVCF --sample_ploidy 1 --output_mode EMIT_VARIANTS_ONLY --pcr_indel_model NONE -nct 48 -dt NONE -A AlleleBalance --variant_index_type LINEAR --variant_index_parameter 128000 –useNewAFCalculator, and subsequence filtration of identified variants using VariantFiltration with settings: --filterExpression "DP < 12 || QUAL < 50 || FS > 14.5 || MQ0 >= 2 && (MQ0/(1.0 * DP)) > 0.1".

In order to establish that the *C. parvum* TU114 strain sequenced here was identical to that reported in 2012 by Widmer and colleagues (Widmer *et al*. 2012), and using the same methods for read mapping and variant identification described above, we compared the Illumina read data from both our *C. parvum* TU114 sample, as well as the Illumina reads generated for this same isolate in 2012 (Widmer *et al*. 2012), with sequence variants called against the new genome assembly we generated for *C. parvum* TU114 in 2017.

Results

Genome description

We generated a draft genome assembly of *C. parvum* isolate TU114. Once reads and contigs of likely bacterial provenance were removed, the resulting assembly consists of 9,115,361 base pairs, and its eight nuclear chromosomes are captured in 25 contigs. The properties of the genome assembly of *C. parvum* TU114 are shown in Table 3.1, together with similar information for the reference genomes of *C. parvum* and *C. hominis*. The newly generated genome assembly of the anthroponotic *C. parvum* strain is very similar in genome size and GC nucleotide content to that of the reference *C. parvum* IOWA II isolate. The slightly higher fragmentation of the TU114 isolate genome relative to that of Iowa II is reflected in the lower N_{50}. N_{50} is defined as the length of the smallest contig in the subset of the largest contigs that together contain 50% of the genome, and is a common statistic used to assess genome fragmentation. However, the length of the largest contig, corresponding to one of the eight chromosomes, is nearly identical between all three assemblies.

Table 3.1 Summary of assembly and annotation statistics for three *Cryptosporidium* genomes.

	C. hominis **TU502 2012**	***C. parvum*** **Iowa II**	***C. parvum*** **TU114**
Assembly length (bp)	9,107,739	9,102,324	9,115,361
No. contigs	119	18	25
Contig N_{50}	238,509	1,014,526	888,861
Largest contig (bp)	1,270,815	1,278,458	1,280,921
G+C content (%)	30.1	30.4	30.2

No. protein-coding genes	3,745	3,805	GeneMark-ES 3,350	GMAP 3,816
Average gene length (bp)	1,892	1,795	2,194	1,756
Percent coding	77.8%	75.3%	80.2%	73.5%

The preliminary gene structural annotation of the genome assembly of *C. parvum* TU114 was conducted using two approaches (see Methods). The outcome of these two approaches shows significant differences. GeneMark-ES predicted fewer genes than those generated by mapping the reference genes using GMAP. But the GeneMark-ES gene predictions are much longer, resulting in a higher percentage of the genome it predicts to represent protein-coding genes. We plan to sequence mRNA, use additional gene prediction software such as BRAKER1 (Hoff *et al*. 2016) and AUGUSTUS (Stanke *et al*. 2004), and perform manual curation, in order to generate a reliable consensus gene structural annotation.

Mutation accumulation in *C. parvum* TU114 over five years of propagation

Previously, questions arose about the rate of genetic change as a *Cryptosporidium* isolate is serially propagated over a number of years (Ifeonu *et al*. 2016b; Isaza *et al*. 2015). Since Illumina data had been generated for this same *C. parvum* TU114 isolate from a DNA sample collected in 2011, and published in 2012 (Widmer *et al*. 2012), we compared those data with the genomic data we generated for this same isolate, for a DNA sample obtained in 2016. Using the new anthroponotic *C. parvum* TU114 assembly as

reference, we compared SNP calls using the Illumina data generated at the two time points. In each case we found less than 100 coding SNPs and small indels, most of which are common to both datasets. This indicates that we sequenced the same isolate as that sequenced in 2012 and suggests that there have been no significant genetic changes due to serial propagation of the isolate. In addition, there has been no shift in the composition of the isolate, a possibility that has been raised for a *C. hominis* isolate before (Ifeonu *et al*. 2016b, Widmer *et al.* 2015).

Comparison of genome structure and gene content between *C. parvum* IOWA II and *C. parvum* TU114

The structure for both genomes was broadly compared using MUMmer 3.0 (Kurtz *et al*. 2004). These analyses show that the draft genome assembly of *C. parvum* TU114 is fairly complete, with homology to most of the reference *C. parvum* IOWA II genome (Figure 3.1). Each of the eight *C. parvum* nuclear chromosomes is represented by one scaffold in the genome assembly of the isolate IOWA II; in the TU114 isolate assembly some chromosomes are also in one contig while others are represented by 2 to 13 contigs (Figure 3.1). The two genomes are mostly syntenic; however the TU114 genome assembly shows rearrangements in three chromosomes, compared to the genome of IOWA II. It remains to be determined whether these are real or instead represent assembly artifacts.

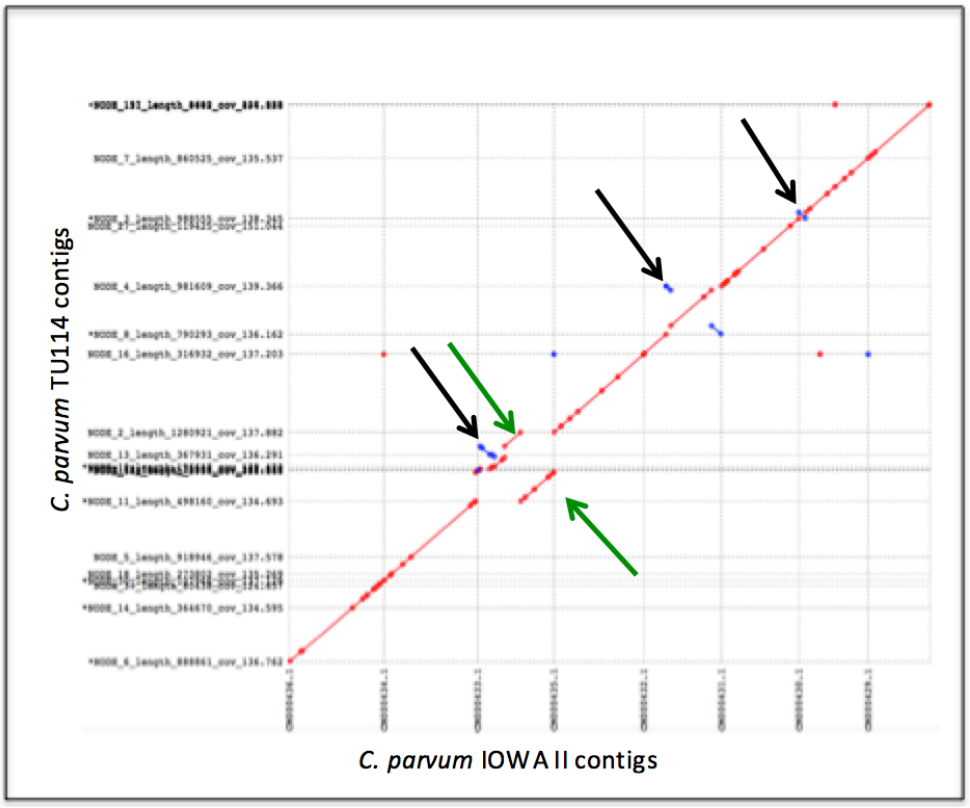

Figure 3.1 Comparison of genome assemblies between *C. parvum* strains.
Each chromosome of *C. parvum* is represented by a scaffold, shown concatenated, in the IOWA II strain (X axis). The genome assembly of the TU114 isolate is composed of 25 contigs, varying in size between 1,642 bp and 1,280,921 bp (Y axis). This plot was generated using the MUMmer3.5 software (Delcher *et al.* 2003).The aligned regions are represented by dots or lines. Red indicates matches in the same orientation while blue indicated reverse complement matches. Potential inversions (black arrows) and potential chromosomal translocations (green arrows) are shown.

In order to identify differences in gene content between *C. parvum* IOWA II and *C. parvum* TU114, we mapped *C. parvum* Iowa II genes onto the *C. parvum* TU114 assembly using GMAP, version 2014-06-10 (Wu and Watanabe 2005). Of the 3,805 reference genes, all but one gene mapped to the new *C. parvum* TU114 assembly. The missing gene (EAZ51361.1, cgd1_3860) encodes a hypothetical protein, which is partial (missing the five prime end) in *C. parvum* Iowa II. A full-length homolog exists in *C. hominis* TU502_2012 (ChTU502y2012_376g0005) and in other *Cryptosporidium* species

(Ifeonu *et al*. 2016b; Ifeonu *et al*. 2016a), as well as in other apicomplexans and more distantly related organisms. We eliminated the possibility that this gene is missing due to an incomplete assembly or contaminant filtering, as the missing gene was not present in the pre-filtered *C. parvum* TU114 assembly. Therefore, it is possible that this gene is truly absent from the TU114 genome.

Of the genes present, 3,792 mapped to a single place in the genome, while 12 mapped to two regions of the assembly. We confirmed that the genes that mapped twice were not due to gene family expansions. These genes were either one gene broken into two partials (three cases) or a full-length gene and a partial copy (three cases). We are investigating the possibility that these differences are due to misassembly of the TU114 genome or to erroneous assembly or annotation in the IOWA II genome.

Identification of sequence variants between *C. parvum* IOWA II and *C. parvum* TU114

We identified sequence variants, including both single nucleotide polymorphisms (SNPs) and insertions/deletions (indels), between the anthroponotic *C. parvum* TU114 and the reference *C. parvum* Iowa II from livestock. We found 4,000 small indels and about 21,000 SNPs of which roughly 2,000 are synonymous and 3,000 non-synonymous (Table 3.2), and resulting in an average SNP density of 2.3 SNPs per Kb. These differences, even though not trivial, are much smaller than those observed between *Cryptosporidium hominis* TU502_2012 and *C. parvum* Iowa II, which are very closely related species (Table 3.2).

Table 3.2 Sequence variants between reference *C. parvum* Iowa II and either *C. parvum* TU114 or *C. hominis* TU502_2012. The count of small indels and total number of SNPs, are shown, as well and a break down of coding SNPs into synonymous and non-synonymous.

	C. parvum TU114 vs. Iowa II	*C. parvum* Iowa II vs. *C. hominis*
Small Indels	4,000	12,879
Total SNPs	21,192	186,803
Synonymous SNPs	3,419	46,681
Non-Synonymous SNPs	3,175	80,506

We characterized the distribution of non-synonymous SNPs across the genome between the two *C. parvum* strains. These 3,175 amino acid-changing SNPs fall within 988 genes. We show that 70% of the affected genes have just one or two non-synonymous SNPs each. However, there are a few genes with a large number of amino acid changes between strains. A list of ten genes with the most non-synonymous SNPs reveals that many contain signal peptides, are predicted to be secreted, and/or have regions of low complexity, all properties often found over-represented in proteins involved in host-pathogen interactions (Table 3.3). Therefore, it is tempting to speculate that these genes may be antigenic, and that the large number of non-synonymous SNPs is due to positive selection imposed by the host's immune system.

Table 3.3 Ten genes with highest number of non-synonymous (NS) mutations.

Gene ID	Product name	No. NS SNPs
cgd7_1270	large hypothetical protein	64
cgd7_4500	signal peptide, large secreted protein	47
cgd3_1160	conserved hypothetical protein with a signal peptide	44
cgd7_420	protein with DEXDc plus ring plus HELIC-possible SNF2 domain	40
cgd3_720	very large probable mucin, 11700 aa long protein with signal peptide...	36
cgd7_1010	very large low complexity protein	35
cgd3_3370	large hypothetical protein with signal peptide	34
cgd6_830	protein with 2 pleckstrin homology (PH) domains	34
cgd6_3930	hypothetical protein	26
cgd6_3050	hypothetical protein with a signal peptide	25

Discussion

A broad comparison of genome structure between *C. parvum* IOWA II and TU114 shows the two genomes to be mostly syntenic, and their gene content to be nearly identical. The draft genome of *C. parvum* TU114 is fairly complete with the presence of only a few potential structural re-arrangements (Figure 3.1). Structural re-arrangements are also rare within other genera in the phylum Apicomplexa, including *Plasmodium* (Carlton *et al*. 2005) and *Theileria* (Pain *et al*. 2005), suggesting that is a characteristic of the phylum.

SNP density between these strains, as measured in SNPs/Kb, is over twice as high as that observed between strains of *Plasmodium vivax* and even higher than that observed between *Plasmodium falciparum* strains (Neafsey *et al*. 2012), suggesting that the common ancestor of *C. parvum* strains may be older than those of *Plasmodium* species.

The gene content of *C. parvum* IOWA II and TU114 are essentially identical. Further analysis of sequence variants may reveal functional differences. A next step is to

investigate the origin of anthroponotic *C. parvum* strains. *C. hominis* is very closely related to *C. parvum*, and the two differ only by 3 to 5% at the nucleotide level genome-wide (Mazurie *et al*. 2013; Xu *et al*. 2004). Despite the high genome-wide sequence similarity, *C. hominis* is a human parasite, unlike *C. parvum*, which typically infects bovines. The ability of the anthroponotic *C. parvum* TU114 strain to be transmissible among humans could be due to mutations acquired independently by this isolate. Alternatively, the possibility exists that this ability to infect human hosts has been acquired from *C. hominis*, through introgression of the genomic segment(s) containing the traits that facilitate transmission in this new mammalian host. Similar events have been observed in other apicomplexan parasites, which have facilitated a host switch (Sundararaman *et al*. 2016). This possibility could be investigated by identifying the presence of regions of potential genome introgression from *C. hominis* to the anthroponotic *C. parvum* TU114 strain, and performing a comprehensive analysis of genes located in these regions, to determine if they encode proteins that could have facilitated a host-switching event from cattle to humans.

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.

Altschul, S. F., et al. (1997), 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Res,* 25 (17), 3389-402.

Apweiler, R., et al. (2000), 'InterPro--an integrated documentation resource for protein families, domains and functional sites', *Bioinformatics,* 16 (12), 1145-50.

Arrowood, M. J. (2002), 'In vitro cultivation of *Cryptosporidium* species', *Clin Microbiol Rev,* 15 (3), 390-400.

Bankevich, A., et al. (2012), 'SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing', *J Comput Biol,* 19 (5), 455-77.

Bateman, A., et al. (2000), 'The Pfam protein families database', *Nucleic Acids Res,* 28 (1), 263-6.

Biswas, S., et al. (2014), 'Assessment of humoral immune responses to blood-stage malaria antigens following ChAd63-MVA immunization, controlled human malaria infection and natural exposure', *PLoS One,* 9 (9), e107903.

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.

Bouzid, M., et al. (2013), '*Cryptosporidium* pathogenicity and virulence', *Clin Microbiol Rev,* 26 (1), 115-34.

Burkhard, P. and Lanar, D. E. (2015), 'Malaria vaccine based on self-assembling protein nanoparticles', *Expert Rev Vaccines,* 14 (12), 1525-7.

Cama, V. A., et al. (2003), '*Cryptosporidium* species and genotypes in HIV-positive patients in Lima, Peru', *J Eukaryot Microbiol,* 50 Suppl, 531-3.

Carlton, J., Silva, J., and Hall, N. (2005), 'The genome of model malaria parasites, and comparative genomics', *Curr Issues Mol Biol,* 7 (1), 23-37.

Cevallos, A. M., et al. (2000a), 'Molecular cloning and expression of a gene encoding *Cryptosporidium parvum* glycoproteins gp40 and gp15', *Infect Immun,* 68 (7), 4108-16.

Cevallos, A. M., et al. (2000b), 'Mediation of *Cryptosporidium parvum* infection in vitro by mucin-like glycoproteins defined by a neutralizing monoclonal antibody', *Infect Immun,* 68 (9), 5167-75.

Chauhan, J. S., Rao, A., and Raghava, G. P. (2013), '*In silico* platform for prediction of N-, O- and C-glycosites in eukaryotic protein sequences', *PLoS One,* 8 (6), e67008.

Checkley, W., et al. (2015), 'A review of the global burden, novel diagnostics, therapeutics, and vaccine targets for *Cryptosporidium*', *Lancet Infect Dis,* 15 (1), 85-94.

Chen, Y., et al. (2015), 'High speed BLASTN: an accelerated MegaBLAST search tool', *Nucleic Acids Res,* 43 (16), 7762-8.

Conway, D. J. (2015), 'Paths to a malaria vaccine illuminated by parasite genomics', *Trends Genet,* 31 (2), 97-107.

Crabtree, J., et al. (2007), 'Sybil: methods and software for multiple genome comparison and visualization', *Methods Mol Biol,* 408, 93-108.

de Barra, E., et al. (2014), 'A phase Ia study to assess the safety and immunogenicity of new malaria vaccine candidates ChAd63 CS administered alone and with MVA CS', *PLoS One,* 9 (12), e115161.

Delcher, A. L., Salzberg, S. L., and Phillippy, A. M. (2003), 'Using MUMmer to identify similar regions in large sequence sets', *Curr Protoc Bioinformatics,* Chapter 10, Unit 10 3.

Dieckmann-Schuppert, A., Bause, E., and Schwarz, R. T. (1994), 'Glycosylation reactions in *Plasmodium falciparum, Toxoplasma gondii,* and *Trypanosoma brucei brucei* probed by the use of synthetic peptides', *Biochim Biophys Acta,* 1199 (1), 37-44.

Dieckmann-Schuppert, A., et al. (1992), 'Apparent lack of N-glycosylation in the asexual intraerythrocytic stage of *Plasmodium falciparum*', *Eur J Biochem,* 205 (2), 815-25.

Donati, C. and Rappuoli, R. (2013), 'Reverse vaccinology in the 21st century: improvements over the original design', *Ann N Y Acad Sci,* 1285, 115-32.

Doro, F., et al. (2009), 'Surfome analysis as a fast track to vaccine discovery: identification of a novel protective antigen for Group B *Streptococcus* hypervirulent strain COH1', *Mol Cell Proteomics,* 8 (7), 1728-37.

El-Kamary, S. S., et al. (2013), 'Safety and immunogenicity of a single oral dose of recombinant double mutant heat-labile toxin derived from enterotoxigenic *Escherichia coli*', *Clin Vaccine Immunol,* 20 (11), 1764-70.

Emanuelsson, O., et al. (2000), 'Predicting subcellular localization of proteins based on their N-terminal amino acid sequence', *J Mol Biol,* 300 (4), 1005-16.

Fankhauser, N. and Maser, P. (2005), 'Identification of GPI anchor attachment signals by a Kohonen self-organizing map', *Bioinformatics,* 21 (9), 1846-52.

Forney, J. R., et al. (1996), 'Efficacy of serine protease inhibitors against *Cryptosporidium parvum* infection in a bovine fallopian tube epithelial cell culture system', *J Parasitol,* 82 (4), 638-40.

Gonzalez, C., et al. (1994), '*Salmonella typhi* vaccine strain CVD 908 expressing the circumsporozoite protein of *Plasmodium falciparum*: strain construction and safety and immunogenicity in humans', *J Infect Dis,* 169 (4), 927-31.

Gonzalez, C. R., et al. (1998), 'Immunogenicity of a Salmonella typhi CVD 908 candidate vaccine strain expressing the major surface protein gp63 of *Leishmania mexicana mexicana*', *Vaccine,* 16 (9-10), 1043-52.
Gupta, R., Jung, E., and Brunak, S. (2004), 'Prediction of N-glycosylation sites in human proteins.'.
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.
Hadfield, S. J., et al. (2015), 'Generation of whole genome sequences of new *Cryptosporidium hominis* and *Cryptosporidium parvum* isolates directly from stool samples', *BMC Genomics,* 16, 650.
Heiges, M., et al. (2006), 'CryptoDB: a *Cryptosporidium* bioinformatics resource update', *Nucleic Acids Res,* 34 (Database issue), D419-22.
Heinson, A. I., Woelk, C. H., and Newell, M. L. (2015), 'The promise of reverse vaccinology', *Int Health,* 7 (2), 85-9.
Hoff, K. J., et al. (2016), 'BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS', *Bioinformatics,* 32 (5), 767-9.
Holmes, E. C. (2004), 'Adaptation and immunity', *PLoS Biol,* 2 (9), E307.
Hoof, I., et al. (2009), 'NetMHCpan, a method for MHC class I binding prediction beyond humans', *Immunogenetics,* 61 (1), 1-13.
Ifeonu, O. O., et al. (2016a), '*Cryptosporidium hominis* gene catalog: a resource for the selection of novel *Cryptosporidium* vaccine candidates', *Database (Oxford),* 2016.
Ifeonu, O. O., et al. (2016b), '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', *Pathog Dis*.
International Human Genome Sequencing, Consortium (2004), 'Finishing the euchromatic sequence of the human genome', *Nature,* 431 (7011), 931-45.
Isaza, J. P., et al. (2015), 'Revisiting the reference genomes of human pathogenic *Cryptosporidium* species: reannotation of *C. parvum* Iowa and a new *C. hominis* reference', *Sci Rep,* 5, 16324.
Jones, R. M., et al. (2013), 'A plant-produced Pfs25 VLP malaria vaccine candidate induces persistent transmission blocking antibodies against *Plasmodium falciparum* in immunized mice', *PLoS One,* 8 (11), e79538.
Karosiene, E., et al. (2013), 'NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ', *Immunogenetics,* 65 (10), 711-24.
Kelly, D. F. and Rappuoli, R. (2005), 'Reverse vaccinology and vaccines for serogroup B *Neisseria meningitidis*', *Adv Exp Med Biol,* 568, 217-23.
Khramtsov, N. V., et al. (1995), 'Cloning and analysis of a *Cryptosporidium parvum* gene encoding a protein with homology to cytoplasmic form Hsp70', *J Eukaryot Microbiol,* 42 (4), 416-22.
Kimura, E. A., et al. (1996), 'N-linked glycoproteins are related to schizogony of the intraerythrocytic stage in *Plasmodium falciparum*', *J Biol Chem,* 271 (24), 14452-61.

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.
Krogh, A., et al. (2001), 'Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes', *J Mol Biol,* 305 (3), 567-80.
Kurtz, S., et al. (2004), 'Versatile and open software for comparing large genomes', *Genome Biol,* 5 (2), R12.
Lander, E. S., et al. (2001), 'Initial sequencing and analysis of the human genome', *Nature,* 409 (6822), 860-921.
Langmead, B. and Salzberg, S. L. (2012), 'Fast gapped-read alignment with Bowtie 2', *Nat Methods,* 9 (4), 357-9.
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.
Li, H. and Durbin, R. (2010), 'Fast and accurate long-read alignment with Burrows-Wheeler transform', *Bioinformatics,* 26 (5), 589-95.
Li, H., et al. (2009), 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics,* 25 (16), 2078-9.
Liu, J., et al. (2016), 'Use of quantitative molecular diagnostic methods to identify causes of diarrhoea in children: a reanalysis of the GEMS case-control study', *Lancet,* 388 (10051), 1291-301.
Liu, L., et al. (2015), 'Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: an updated systematic analysis', *Lancet,* 385 (9966), 430-40.
Luk, F. C., Johnson, T. M., and Beckers, C. J. (2008), 'N-linked glycosylation of proteins in the protozoan parasite *Toxoplasma gondii*', *Mol Biochem Parasitol,* 157 (2), 169-78.
Maione, D., et al. (2005), 'Identification of a universal Group B *Streptococcus* vaccine by multiple genome screen', *Science,* 309 (5731), 148-50.
Manque, P. A., et al. (2011), 'Identification and immunological characterization of three potential vaccinogens against *Cryptosporidium* species', *Clin Vaccine Immunol,* 18 (11), 1796-802.
Marchler-Bauer, A., et al. (2005), 'CDD: a Conserved Domain Database for protein classification', *Nucleic Acids Res,* 33 (Database issue), D192-6.
Mazurie, A. J., et al. (2013), 'Comparative genomics of *Cryptosporidium*', *Int J Genomics,* 2013, 832756.
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.
Mistry, J., et al. (2013), 'Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions', *Nucleic Acids Res,* 41 (12), e121.
Morada, M., et al. (2016), 'Continuous culture of *Cryptosporidium parvum* using hollow fiber technology', *Int J Parasitol,* 46 (1), 21-9.

Mu, J., et al. (2007), 'Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome', *Nat Genet,* 39 (1), 126-30.
Neafsey, D. E., et al. (2012), 'The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*', *Nat Genet,* 44 (9), 1046-50.
Neafsey, D. E., et al. (2015), 'Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine', *N Engl J Med,* 373 (21), 2025-37.
Nielsen, H., et al. (1997), 'Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites', *Protein Eng,* 10 (1), 1-6.
Nielsen, M., et al. (2007), 'NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence', *PLoS One,* 2 (8), e796.
O'Connor, R. M., et al. (2009), 'Polymorphic mucin antigens CpMuc4 and CpMuc5 are integral to *Cryptosporidium parvum* infection in vitro', *Eukaryot Cell,* 8 (4), 461-9.
O'Hara, S. P., Yu, J. R., and Lin, J. J. (2004), 'A novel *Cryptosporidium parvum* antigen, CP2, preferentially associates with membranous structures', *Parasitol Res,* 92 (4), 317-27.
Odenthal-Schnittler, M., et al. (1993), 'Evidence for N-linked glycosylation in *Toxoplasma gondii*', *Biochem J,* 291 (Pt 3), 713-21.
Okhuysen, P. C., et al. (1996), '*Cryptosporidium parvum* metalloaminopeptidase inhibitors prevent in vitro excystation', *Antimicrob Agents Chemother,* 40 (12), 2781-4.
Ord, R. L., et al. (2014), 'A malaria vaccine candidate based on an epitope of the *Plasmodium falciparum* RH5 protein', *Malar J,* 13, 326.
Ouattara, A., et al. (2013), 'Molecular basis of allele-specific efficacy of a blood-stage malaria vaccine: vaccine development implications', *J Infect Dis,* 207 (3), 511-9.
Pain, A., et al. (2005), 'Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*', *Science,* 309 (5731), 131-3.
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.
Perkins, M. E., et al. (1999), 'CpABC, a *Cryptosporidium parvum* ATP-binding cassette protein at the host-parasite boundary in intracellular stages', *Proc Natl Acad Sci U S A,* 96 (10), 5734-9.
Petersen, C., et al. (1992), 'Characterization of a > 900,000-M(r) *Cryptosporidium parvum* sporozoite glycoprotein recognized by protective hyperimmune bovine colostral immunoglobulin', *Infect Immun,* 60 (12), 5132-8.
Pierleoni, A., Martelli, P. L., and Casadio, R. (2008), 'PredGPI: a GPI-anchor predictor', *BMC Bioinformatics,* 9, 392.
Pizza, M., et al. (2000), 'Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing', *Science,* 287 (5459), 1816-20.

Poisson, G., et al. (2007), 'FragAnchor: a large-scale predictor of glycosylphosphatidylinositol anchors in eukaryote protein sequences by qualitative scoring', *Genomics Proteomics Bioinformatics,* 5 (2), 121-30.
Pulendran, B. (2009), 'Learning immunology from the yellow fever vaccine: innate immunity to systems vaccinology', *Nat Rev Immunol,* 9 (10), 741-7.
Rappuoli, R. and Covacci, A. (2003), 'Reverse vaccinology and genomics', *Science,* 302 (5645), 602.
Riggs, M. W., et al. (1997), 'Protective monoclonal antibody defines a circumsporozoite-like glycoprotein exoantigen of *Cryptosporidium parvum* sporozoites and merozoites', *J Immunol,* 158 (4), 1787-95.
Ruiz-Perez, F., et al. (2002), 'Expression of the *Plasmodium falciparum* immunodominant epitope (NANP)(4) on the surface of *Salmonella enterica* using the autotransporter MisL', *Infect Immun,* 70 (7), 3611-20.
Ryan, U., Fayer, R., and Xiao, L. (2014), '*Cryptosporidium* species in humans and animals: current understanding and research needs', *Parasitology,* 141 (13), 1667-85.
Sanderson, S. J., et al. (2008), 'Determining the protein repertoire of *Cryptosporidium parvum* sporozoites', *Proteomics,* 8 (7), 1398-414.
Schwarz, F. and Aebi, M. (2011), 'Mechanisms and principles of N-linked protein glycosylation', *Curr Opin Struct Biol,* 21 (5), 576-82.
Serruto, D. and Rappuoli, R. (2006), 'Post-genomic vaccine development', *FEBS Lett,* 580 (12), 2985-92.
Sette, A. and Rappuoli, R. (2010), 'Reverse vaccinology: developing vaccines in the era of genomics', *Immunity,* 33 (4), 530-41.
Silva, J. C., et al. (2016), 'Genome-wide diversity and gene expression profiling of *Babesia microti* isolates identify polymorphic genes that mediate host-pathogen interactions', *Sci Rep,* 6, 35284.
Sonnhammer, E. L., von Heijne, G., and Krogh, A. (1998), 'A hidden Markov model for predicting transmembrane helices in protein sequences', *Proc Int Conf Intell Syst Mol Biol,* 6, 175-82.
Sow, S. O., et al. (2016), 'The Burden of *Cryptosporidium* Diarrheal Disease among Children < 24 Months of Age in Moderate/High Mortality Regions of Sub-Saharan Africa and South Asia, Utilizing Data from the Global Enteric Multicenter Study (GEMS)', *PLoS Negl Trop Dis,* 10 (5), e0004729.
Stanke, M., et al. (2004), 'AUGUSTUS: a web server for gene finding in eukaryotes', *Nucleic Acids Res,* 32 (Web Server issue), W309-12.
Steentoft, C., et al. (2013), 'Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology', *EMBO J,* 32 (10), 1478-88.
Stewart, V. A., et al. (2007), 'Priming with an adenovirus 35-circumsporozoite protein (CS) vaccine followed by RTS,S/AS01B boosting significantly improves immunogenicity to *Plasmodium falciparum* CS compared to that with either malaria vaccine alone', *Infect Immun,* 75 (5), 2283-90.
Strong, W. B., Gut, J., and Nelson, R. G. (2000), 'Cloning and sequence analysis of a highly polymorphic *Cryptosporidium parvum* gene encoding a 60-kilodalton

glycoprotein and characterization of its 15- and 45-kilodalton zoite surface antigen products', *Infect Immun,* 68 (7), 4117-34.
Sulaiman, I. M., et al. (2005), 'Unique endemicity of cryptosporidiosis in children in Kuwait', *J Clin Microbiol,* 43 (6), 2805-9.
Sundararaman, S. A., et al. (2016), 'Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria', *Nat Commun,* 7, 11078.
Takala, S. L., et al. (2009), 'Extreme polymorphism in a vaccine antigen and risk of clinical malaria: implications for vaccine development', *Sci Transl Med,* 1 (2), 2ra5.
Tarailo-Graovac, M. and Chen, N. (2009), 'Using RepeatMasker to identify repetitive elements in genomic sequences', *Curr Protoc Bioinformatics,* Chapter 4, Unit 4 10.
Trapnell, C., Pachter, L., and Salzberg, S. L. (2009), 'TopHat: discovering splice junctions with RNA-Seq', *Bioinformatics,* 25 (9), 1105-11.
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.
Tretina, K., et al. (In preparation).
Tumwine, J. K., et al. (2005), 'Cryptosporidiosis and microsporidiosis in ugandan children with persistent diarrhea with and without concurrent infection with the human immunodeficiency virus', *Am J Trop Med Hyg,* 73 (5), 921-5.
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.
Van den Steen, P., et al. (1998), 'Concepts and principles of O-linked glycosylation', *Crit Rev Biochem Mol Biol,* 33 (3), 151-208.
Vernikos, G. and Medini, D. (2014), 'Bexsero(R) chronicle', *Pathog Glob Health,* 108 (7), 305-16.
Walker, John M. (2005), *The proteomics protocols handbook* (Totowa, N.J.: Humana Press) xviii, 988 p.
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. (2012), 'Comparative genome analysis of two *Cryptosporidium parvum* isolates with different host range', *Infect Genet Evol,* 12 (6), 1213-21.
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.
Wu, T. D. and Watanabe, C. K. (2005), 'GMAP: a genomic mapping and alignment program for mRNA and EST sequences', *Bioinformatics,* 21 (9), 1859-75.

Xiao, L. (2010), 'Molecular epidemiology of cryptosporidiosis: an update', *Exp Parasitol,* 124 (1), 80-9.
Xiao, L., et al. (2001), 'Identification of 5 types of *Cryptosporidium* parasites in children in Lima, Peru', *J Infect Dis,* 183 (3), 492-7.
Xu, P., et al. (2004), 'The genome of *Cryptosporidium hominis*', *Nature,* 431 (7012), 1107-12.
Yang, Z. and Bielawski, J. P. (2000), 'Statistical methods for detecting molecular adaptation', *Trends Ecol Evol,* 15 (12), 496-503.
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.

Biography

Olukemi O. Ifeonu graduated from Igbinedion Secondary School, Benin City, Nigeria, in 2001. She received her Bachelor of Arts in Biochemistry from Hood College in 2007. She received her Master of Science in Bioinformatics from Johns Hopkins University in 2010.