HHS Public Access

Author manuscript

Int J Parasitol. Author manuscript; available in PMC 2018 January 19.

Published in final edited form as:

Int J Parasitol. 2017 April ; 47(5): 281–290. doi:10.1016/j.ijpara.2016.12.002.

Comparative genomic analysis of the IId subtype family of *Cryptosporidium parvum*☆

Yaoyu Feng[a,b,*], **Na Li**[a,c], **Dawn M. Roellig**[c], **Alyssa Kelley**[c], **Guangyuan Liu**[b], **Said Amer**[c,d], **Kevin Tang**[e], **Longxian Zhang**[f], and **Lihua Xiao**[c,*]

[a]College of Veterinary Medicine, South China Agricultural University, Guangzhou 510642, China

[b]State Key Laboratory of Veterinary Etiological Biology, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Lanzhou 730046, China

[c]Division of Foodborne, Waterborne and Environmental Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30329, USA

[d]Department of Zoology, Faculty of Science, Kafr El sheikh University, Kafr El Sheikh 33516, Egypt

[e]Division of Scientific Resources, Centers for Disease Control and Prevention, Atlanta, GA 30329, USA

[f]College of Animal Science and Veterinary Medicine, Henan Agricultural University, Zhengzhou 450002, China

Abstract

Host adaptation is known to occur in *Cryptosporidium parvum*, with IIa and IId subtype families preferentially infecting calves and lambs, respectively. To improve our understanding of the genetic basis of host adaptation in *Cryptosporidium parvum*, we sequenced the genomes of two IId specimens and one IIa specimen from China and Egypt using the Illumina technique and compared them with the published IIa IOWA genome. Sequence data were obtained for >99.3% of the expected genome. Comparative genomic analysis identified differences in numbers of three subtelomeric gene families between sequenced genomes and the reference genome, including those encoding SKSR secretory proteins, the MEDLE family of secretory proteins, and insulinase-like proteases. These gene gains and losses compared with the reference genome were confirmed by PCR analysis. Altogether, 5,191–5,766 single nucleotide variants were seen between genomes sequenced in this study and the reference genome, with most SNVs occurring in subtelomeric regions of chromosomes 1, 4, and 6. The most highly polymorphic genes between IIa and IId encode mainly invasion-associated and immunodominant mucin proteins, and other families of secretory proteins. Further studies are needed to verify the biological significance of these genomic differences.

☆*Note:* Nucleotide sequence data reported in this paper, including all Sequence Read Achieve (SRA) data and assembled contigs, are available in GenBank under the BioProject accession number PRJNA320419 and BioSample accession numbers SAMN04938568, SAMN04938569 and SAMN04938570.

*Corresponding authors at: College of Veterinary Medicine, South China Agricultural University, Guangzhou 510642, China. Fax: +86 21 6425 0664 (Y. Feng). Fax: +1 404 718 4197 (L. Xiao). yyfeng@scau.edu.cn (Y. Feng), lxiao@cdc.gov (L. Xiao).

Keywords

Cryptosporidium parvum; Genomics; Whole genome sequencing; Host adaptation; Transmission

1. Introduction

Cryptosporidium parvum is the *Cryptosporidium* sp. responsible for watery diarrhoea in pre-weaned ruminants (Santin, 2013). It is also the most important zoonotic *Cryptosporidium* sp. in humans (Ryan et al., 2014). Previous sequence characterizations of the 60 kDa glycoprotein (gp60) gene had shown the existence of host adaption in *C. parvum*, with the occurrence of the IIa subtype family mostly in cattle, IIc subtype family mostly in humans, and IId subtype family mostly in sheep and goats, although all three subtype families are human pathogens and IId subtypes have been found in calves in some areas (Xiao, 2010). More recent sequence characterizations at other genetic loci have confirmed the existence of host adapted *C. parvum* subtype families (Widmer and Lee, 2010).

The genetic basis for host adaptation in *C. parvum* is not clear. The genome of one *C. parvum* IIa isolate (IOWA) from a calf in the United States, propagated through calf passages, was among the first two *Cryptosporidium* isolates sequenced (Abrahamsen et al., 2004). More recently, the genome of one IIc isolate from a child in Uganda and propagated in immunosuppressed mice was sequenced (Widmer et al., 2012). Due to the existence of significant sequence differences between the two isolates across the entire genomes, more comparative genomic analysis of other host-adapted *C. parvum* subtypes, especially different subtypes from the same area, is needed in order to better understand the genetic determinants for host adaptation in *C. parvum*.

In this study, we sequenced the genomes of two IId specimens of *C. parvum* from China and Egypt. As a control, we also sequenced the genome of one IIa specimen. The comparative genomic analysis revealed some differences in the number of several subtelomeric gene families between specimens sequenced in this study and the reference *C. parvum* IOWA, and in the sequences of the invasion-associated and immunodominant mucin-type secretory glycoproteins between IIa and IId subtype families.

2. Materials and methods

2.1. Cryptosporidium specimens

The genomes of three *C. parvum* specimens were sequenced in the study: specimen 31727 of the IIdA19G1 subtype, 34902 of the IIdA20G1 subtype, and 35090 of the IIaA15G1R1 subtype. Specimen 31727 was collected from a 1 month old dairy calf with diarrhoea in Zhengzhou, Henan Province, China in November 2008 and maintained through animal passages in gerbils. Specimen 34902 was collected in January 2011 from a 3 week old buffalo calf with diarrhoea in Sakha, Kafr El Sheikh Province, Egypt. Specimen 35090 was collected in October 2011 from a 5 week old dairy calf with diarrhoea in Al Nubaria, El Beheira Province, Egypt. The two IId subtypes from China and Egypt were targeted for sequencing in this project because they are commonly found in calves in both countries. The

three subtypes chosen in this study represented the most common *C. parvum* subtypes in calves and humans in China and Egypt. For each specimen, faecal material was stored in 2.5% potassium dichromate at 4 °C for less than 6 months before use in *Cryptosporidium* oocyst isolation. *Cryptosporidium* species and subtypes were determined by PCR-RFLP analysis of the ssrRNA and sequence analysis of the gp60 genes, respectively (Xiao et al., 2009). The collection of faecal specimens used in the study was approved by the Institutional Committee of the Post-graduate Studies and Research at Kafr El Sheikh University, Egypt, and the Research Ethics Committee of Henan Agricultural University, Zhengzhou, China.

2.2. Oocyst isolation and whole genome sequencing

Cryptosporidium oocysts were isolated from stool specimens by sucrose and cesium chloride gradient centrifugation, and immunomagnetic separation as previously described (Guo et al., 2015a). They were subjected to treatment with 10% commercial bleach on ice for 10 min and five freezing-and-thawing cycles. DNA was extracted using the Qiagen DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA, USA), and amplified using the REPLI-g Midi Kit (Qiagen). The amplified DNA was sequenced on an Illumina Genome Analyzer IIx (Illumina, San Diego, CA, USA). For specimen 31727, a 100 bp paired-end technique was used whereas for specimens 34902 and 35090, a 100 bp single-end technique was used. The sequence reads were analysed for sequencing quality using CLC Genomic Workbench 8.5 (https://www.qiagenbioinformatics.com/products/clc-genomics-workbench). They were trimmed off by 10 nucleotides at the 5′ end and Phred score < 25 at both ends, using the error rate limit setting of 0.02, minimum read length of 65 nucleotides, and ambiguous trim setting of 2. Trimmed reads were assembled de novo using CLC Genomics Workbench with word size 50, bubble size of 400, mismatch cost of 2, insertion and deletion costs of 3, minimum contig length of 500 bp, and updating contigs after read mapping. The word size 50 was selected based on the outcome of de novo assemblies of several *Cryptosporidium* genomes using word sizes of 22 (default), 30, 40, 50 and 60, and assessment of assemblies using QUAST (http://bioinf.spbau.ru/quast). Raw sequence reads were also trimmed using BBDuk from the BBMap package (https://sourceforge.net/ projects/bbmap/). Reads were trimmed at both ends for Phred score < 30, phix adapters from BBMap package resources, and 10 bp from the 5′ end with paired-end reads trimmed to equal length and reads shorter than 65 bp removed. Sequence reads were analysed for sequence quality before and after the cleaning using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmed sequence reads were also de novo assembled using SPAdes (http://cab.spbu.ru/software/spades/), with word sizes 31, 41, 51 and 61.

2.3. Genome comparison and identification of gene insertions and deletions (indels)

For assessment of gene gains and losses among *C. parvum* isolates, contigs from each genome assembly were aligned with the eight assembled chromosome sequences of the *C. parvum* IOWA isolate of the IIaA15G2R1 subtype (version AAEE00000000.1) by using the progressive alignment algorithm of Mauve 2.3.1 (http://asap.genetics.wisc.edu/software/mauve/) with default options. This reference was also used in other analyses described below, and all analyses were conducted prior to the recent re-annotation of the IOWA genome, which increased the number of annotated genes from 3,805 to 3,865 (Isaza et al.,

2015). Major insertions and deletions (indels) in genomic fragments were identified by manual inspection of the genome alignment. Potential genes in major insertions were identified using FGENESH (http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind) or geneid (http://genome.crg.es/software/geneid/geneid.html) webservers. The identities of the genes were established by blastp analysis of the deduced amino acid sequences against the National Center for Biotechnology Information (NCBI), USA, non-redundant protein sequence database. Contigs from contaminants were removed from the assembly through BLAST analysis against the NCBI non-redundant nucleotide database, which produced the final genome. Contigs obtained from the study were further aligned with the reference IOWA genome using NUCmer within the MUMmer 3.0 package (http://mummer.sourceforge.net/), with a minimum cluster length of 100 bp. Indels in the alignments were detected by using the MUMmer show-snps utility and in-house scripts.

2.4. Variant analysis and identification of highly polymorphic genes

To identify highly polymorphic genes, sequence reads of each genome were mapped to the reference *C. parvum* IOWA genome using CLC Genomics Workbench 8.5, with a mismatch cost of 2, insertion and deletion costs of 3, length fraction of 0.5 and similarity fraction of 0.8. The outcome of the read mapping was analysed using the Basic Variant Detection tool in the software, with minimum coverage of 10, variant probability of 90, required variant count of 2. The outcome of the variant detection was exported into Excel and the number of single nucleotide variants (SNVs) present per 1,000 bp along the 9.1 Mb *C. parvum* genome was plotted by using the Pivot Table function of Excel. Due to the likely presence of mismapping of sequence reads to multicopy genes and sequence heterozygosity at some loci (especially cgd2_1370, cgd1_3290, cgd3_4230, cgd5_4510, cgd5_4520, cgd6-1000, cgd6_5510), only homozygous SNVs were considered in the identification of highly polymorphic genes. The number of SNVs present among genomes sequenced in this study was also compared using the same approach.

Alternatively, sequence reads were also mapped to the *C. parvum* IOWA genome using Burrows Wheeler Alignment (BWA) (https://www.msi.umn.edu/sw/bwa) with default parameters. All BAM files were passed to SAMtools mpileup (http://www.htslib.org/) with parameters -g for computing genotype likelihoods, -C50 to handle excessive read depth that could cause errors, -P for platform information, and -t to output read depth and allelic depth. The output was piped to BCFtools (http://www.htslib. org/download/) call algorithm to create a variant only VCF file. BCFtools filter was used to remove variants with a call quality of Phred score ≤20 and read depth ≤10. The SNVs identified were annotated and analysed using SnpEff (http://snpeff.sourceforge.net/) for variant type, function, and genes affected.

Nucleotide sequences from the highly polymorphic genes identified in the SNV analysis were concatenated together for each of the genomes under study. They were aligned using MAFFT 7.3 (http://mafft.cbrc.jp/alignment/software/) with default parameters and analysed with the neighbour-joining method implemented in MEGA6 (http://www.megasoftware.net/), based on genetic distances calculated with the Kimura 2-

parameter model and the inclusion of both transitions and transversions. The topology of the tree obtained was compared with the whole genome phylogeny based on SNV analysis. The latter was conducted using R package 'APE' (Analyses of Phylogenetics and Evolution) (https://cran.r-project.org/web/packages/ape/index.html). The two VCF files were merged and then transformed into a SNP matrix using VCFtools (http://vcftools.sourceforge.net/) and in house scripts. The matrix was imported into R where the distance and neighbour-joining algorithms of APE were used to create the phylogenetic tree, with *Cryptosporidium hominis* isolate TU502 as the outgroup.

2.5. Confirmation of major indels by PCR

PCR was used to confirm the presence of major insertions at the 3′ end of chromosomes 3 and 6. For confirmation of the 4,135–4,189 bp insertion in chromosome 3, a nested PCR was designed based on conserved sequences 245–327 bp upstream of the telomeric repeats in the reference IOWA genome and nucleotides 485–631 of the insertions in specimen 31727 sequenced in this study. The primers used were IId-Indel-C3-F1 (5′-TCG AGT ATG GAT AGC TGT AGT TC-3′) and IId-Indel-C3-R1 (5′-CTA ACT TCA GAC AGG ATA GAC TCT-3′) in primary PCR, and IId-Indel-C3-F2 (5′-ATG TAA CCT TCT CGG AAT CCG TT-3′) and IId-Indel-C3-R2 (5′-CTG CAA GCA TAA GAA AAG ATA CCC AT-3′) in secondary PCR, which amplify a fragment of 952 and 781 bp in *C. parvum* specimens containing the insertion, respectively. For confirmation of the 5,273 bp insertion in chromosome 6, nested PCR primers were designed based on conserved nucleotide sequences flanking the insertion (nucleotides 555–640 of cgd6_5490 and nucleotides 826–885 of cgd6_5500). The primers used were IId-Indel-C6-F1 (5′-ACA CGA TCA AAT AGA TTC AGG TCG AA-3′) and IId-Indel-C6-R1 (5′-GAA GAG GCA ATG ATA ACC GGT-3′) in primary PCR, and IId-Indel-C6-F2 (5′-AAG TCA AGG CCA AGA GGT TCT G-3′) and IId-Indel-C6-R2 (5′-GTC TTT CTA GAT TGA GAG GAT TAA G-3′) in secondary PCR, which amplify a fragment of 820 and 736 bp in *C. parvum* specimens without the insertion, respectively. The long insertion in other *C. parvum* specimens would have prevented amplification of the PCR target.

Three specimens of IIa subtypes (IIaA15G2R1, IIaA15G2R2, and IIaA18G3R1) from the United States and three specimens of IId subtypes from other countries (IIdA16G1 from Greece, IIdA20G1 from Egypt, and IIdA21G1 from Spain) were used in PCR analysis of the two major indel targets. Each specimen was analysed in duplicate PCRs using 50 μl of PCR mixture which consisted of 1 × PCR buffer (Applied Biosystems, Foster City, CA, USA), 200 μM dNTP, 3.0 mM $MgCl_2$, 100 nM primers, 2.0 U of Taq polymerase (Promega, Madison, WI, USA), 400 ng/μL of non-acetylated BSA (Sigma-Aldrich, St. Louis, MO, USA), and 1 μL (~100 ng) of extracted DNA or 2 μL of primary PCR products (in secondary PCR). Both primary and secondary PCRs were performed in a GeneAmp 9700 (Applied Biosystems) for 35 cycles at 94 °C for 45 s, 55 °C for 45 s, and 72 °C for 60 s, with an initial denaturation (94 °C for 5 min) and a final extension (72 °C for 10 min). Positive PCR products were sequenced in both directions on an ABI3130 Genetic Analyzer (Applied Biosystems). Nucleotide sequences obtained were aligned with reference sequences of IIa and IId subtype families using ClustalX (http://www.clustal.org/).

2.6. Data deposition

Nucleotide sequences generated from the project, including all Sequence Read Achieve (SRA) data and assembled contigs, were submitted to the NCBI under the BioProject accession number PRJNA320419 and BioSample accession numbers SAMN04938568, SAMN04938569 and SAMN04938570. Raw SNV data from genome comparisons are presented in Supplementary Tables S1–S3. Full lists of list of indels in each of the three *C. parvum* specimens compared with the IOWA reference genome are presented in Supplementary Tables S4–S6.

3. Results

3.1. Genomes of sequenced C. parvum specimens

Approximately13.07–18.91 million 100 bp sequence reads were obtained from each *C. parvum* specimen sequenced. After trimming for poor quality scores and nucleotide ambiguity, 12.40–17.73 million of 65–91 bp reads remained when CLC Genomics Workbench was used as the trimming tool. In contrast, 8.62–13.27 million of 65–91 bp reads remained when BBDuk was used in sequence read trimming (Table 1). Improvements in the quality of sequence reads were achieved after the trimming, as indicated by the distribution of average Phred scores, nucleotide contribution by read base position, and quality distribution by read base position, which are illustrated in Supplementary Figs. S1 and S2. They produced 116.0-, 168.7- and 131.9-fold coverage of the IOWA reference genome for specimens 31727, 34902 and 35090, respectively, in sequence read mapping with CLC Genomics Workbench, and 78.5-, 126.5-, and 107.9-fold coverage in read mapping with BWA (Table 1). De novo assembly of read data using CLC Genomics Workbench generated assemblies of 9.06, 9.12 and 9.15 Mb in 3,269, 337 and 1,103 contigs for specimens 35090, 31727 and 34902, respectively (Table 1). Similar assemblies were obtained when SPAdes was used in the de novo assembly of genomes (assemblies of 9.23, 9.13, and 9.24 Mb in 1,390, 421, and 989 contigs for specimens 35090, 31727, and 34902, respectively). After the removal of contigs from contaminants, sequence data were obtained for 9.04–9.15 Mb of the genome. The genome from specimen 31727, sequenced by a 100 bp paired-end technique, was less fragmented than the two genomes sequenced by a 100 bp single-end technique, as reflected by the N50 contig values and the total number of contigs obtained (Table 1).

3.2. Gene gains and losses between sequenced genome and reference genome

The comparative genomic analysis of assemblies generated in this study and the reference IOWA genome identified some minor differences in gene contents (Table 2). At the 3′ end of chromosome 3, IId specimen 31727 had a 4,135 bp insert that had up to 81% sequence identity to gene cgd3_10, a member of the *Cryptosporidium* (conserved sequence motif) SKSR gene family, with signal peptide and SK and SR repeats. Gene prediction analysis identified a gene that encoded a 289 amino acid peptide that had 63% sequence identity to the protein encoded by cgd3_10. Similarly, IId specimen 34902 had a 4,158 bp insert that had up to 83% sequence identity to cgd3_10, which encoded a 292 amino acid peptide that had 64% sequence identity to the protein encoded by cgd3_10 (Fig. 1). The amino acid sequences obtained differed from the cgd3_10 sequence mainly in the nature of a repeat sequence in the encoded proteins, with the PSD/PSH repeat in the protein encoded by the

insertion and PSQ/PLQ repeat in the protein encoded by cgd3_10. The IIa specimen 35090 also appeared to have this insertion; the first 283 bp and another 2,368 bp fragment of the insertion were present in two contigs (contigs 2704 and 1254, respectively). In contrast, the reference IOWA genome does not have any of the sequences, and its chromosome 3 ends with telomeric repeats (Fig. 1).

Another major insertion of genomic fragments was seen at the 3′ end of chromosome 6 (Table 2). Compared with the reference genome of the *C. parvum* IOWA isolate, the three *C. parvum* specimens sequenced in this study, including the IIa specimen 35090 from Egypt, had a 5,273 bp insertion after cgd6-5490. This insertion had up to 66–71% nucleotide sequence identity to cgd3_4260 or cgd3_4270 (two insulinase-like peptidase genes similar to cgd6_5510 and cgd6_5520 downstream from the insertion). Gene prediction using geneid identified coding areas for the four M16 domains in typical insulinase-like peptidases. Upstream from the insertion, the three *C. parvum* specimens had only one of the two genes (cgd6_5480 and cgd6_5490) encoding the (conserved sequence motif) MEDLE family of secretory proteins (Fig. 2).

The presence of these gene gains and losses was confirmed by de novo assembly of genomes using alternative bioinformatics software and read mapping of published transcriptomic data from the *C. parvum* IOWA isolate to assemblies generated in this study. The assemblies generated using SPAdes showed the same deletions and insertions of genes at the 3′ end of chromosomes 3 and 6. In addition, a full 4,189 bp insertion was present at the 3′ end of chromosomes 3 in the genome from specimen 35090, and encoded for a 295 amino acid peptide that had 64% sequence identity to the protein encoded by cgd3_10, with the same PSD/PSH repeat seen in IId specimens (data not shown).

When sequence reads from a transcriptomic study of the *C. parvum* IOWA isolate in HCT-8 cultures (Isaza et al., 2015) were mapped to the assemblies of specimens 31727, 34902 and 35090, none of the reads were mapped to the 4,135–4,189 bp insertion in chromosome 3 and 5,273-bp insertion in chromosome 6 (Supplementary Fig. S3). In contrast, most other genes in the *C. parvum* IOWA genome were covered by these RNA sequencing reads, as reported previously (Isaza et al., 2015).

PCR analysis of sequences within the major indels further confirmed the differences in gene content between the reference IOWA genome and IId genomes. The three IId specimens analysed by PCR targeting the insertion at the 3′ end of chromosome 3 generated the expected product containing the partial insertion and upstream sequence. DNA sequence analysis produced one sequence identical to the target sequence from specimen 31727 and two sequences with two SNVs. In contrast, the three IIa specimens from the United States did not generate any product in PCR analysis of the 4,135–4,189 bp insertion (Fig. 3A). Similarly, PCR analysis of the sequence between cdg6_5490 and cgd6_5500 generated the expected PCR products for the three IIa specimens analysed. DNA sequence analysis of PCR products generated sequences identical to the target sequence from the three genomes from this study. Due to the presence of the 5,273-bp insertion between cdg6_5490 and cgd6_5500, the three IId specimens analysed did not generate any PCR product (Fig. 3B). We did not attempt to confirm the deletion of cgd6_5480 by PCR, as cgd6_5480 and

cgd6_5490 have very similar nucleotide sequences, which made the design of cgd6_5480-specific primers difficult.

There are many other smaller indels between *C. parvum* specimens sequenced in this study and the *C. parvum* IOWA reference genome throughout the eight chromosomes (Table 3).

3.3. Genes highly polymorphic between IIa and IId subtype families of C. parvum

Comparative genomic analysis revealed the presence of 5,191, 5,386 and 5,766 SNVs between specimens 35090 (IIaA15G1R1 from Egypt), 31727 (IIdA19G1 from China), or 34902 (IIdA20G1 from Egypt) and the reference IOWA genome, respectively, as indicated by sequence read mapping using BWA. Among them, 61.8–63.2% of the SNVs occurred in coding regions, although the latter accounted for 75.0% of the genome. Nearly 40% of the 3,805 predicted genes were polymorphic between the reference IOWA isolate and specimens sequenced in this study. Among the 3,210–3,630 SNVs in genes, 53.0–54.3% of the SNVs were non-synonymous (Table 4). Almost 3,000 core SNVs (>50% SNVs) identified in this study were shared among the three *C. parvum* genomes sequenced in this study, indicating that they were divergent from the reference IOWA genome. A similar percentage of non-synonymous SNVs and polymorphic genes were also shared by these *C. parvum* genomes (Supplementary Fig. S4). The highly divergent nature of the three sequenced *C. parvum* genomes obtained in this study was supported by phylogenetic analysis of the SNV data, in which they formed a cluster outside the IOWA genome (Fig. 4A).

Most of the SNVs detected in this study were in the subtelomeric regions of chromosomes 1, 4 and 6, with the exception of genes cgd1_470 and cgd6_1080, as shown in SNV rates along each of the eight chromosomes (Fig. 5). Between IIa and IId genomes from Egypt, most of the polymorphic genes encoded mucin proteins or were members of multigene families such as cgd1_120 (secreted protein with cysteine cluster with paralogs), cgd1_470 (mucin 8), cgd6_10 (secreted protein CP56 commonly used in subtyping), cgd6_40 (mucin MSC6-7 with paralogs, another common subtyping target), cgd6_1080 (mucin gp60, the most commonly used subtyping target), and cgd6_5490 (MEDLE secreted protein with paralogs) (Fig. 6A). A similar distribution of SNVs was present between the IIa genome from Egypt and IId genome from China, with cgd6_4570 also being highly polymorphic (Fig. 6B). In contrast, the genomes of the two IId specimens in this study differed mostly in cgd6_4570 and cgd6_5490 (Fig. 6C). Read mapping using BWA and SNV analyses using open source software generated the same pattern in the distribution of SNVs and identification of highly polymorphic genes (Supplementary Fig. S5). A phylogenetic tree constructed using concatenated sequences from these highly polymorphic genes showed a robust separation of IIa sequences and IId sequences (Fig. 4B).

4. Discussion

In this study, we sequenced the genomes of two IId specimens of *C. parvum* from China and Egypt. Sequence data were obtained for >99% of the expected genome. As expected, the de novo genome assembly obtained from paired-end sequencing was better than that from single-end sequencing. To assess the impact of geographic segregation on genetic differences between IIa and IId subtype families, we have also sequenced a IIa specimen

from Egypt, as IIa subtypes are rarely seen in China (Wang et al., 2014). As expected, significant genomic differences exist between genomes sequenced in this study and the reference IOWA genome. There were also numerous SNVs between IIa and IId subtype families across eight chromosomes. The highest number of SNVs observed between IIa and IId genomes, however, is approximately half of the SNVs observed previously between IIa and IIc genomes (Widmer et al., 2012). In the present studies, *C. parvum* genomes were sequenced after DNA was amplified using the REPLI-g Midi Kit. Previously, it was shown that *C. parvum* single cell genomes amplified using a REPLI-g Mini/Midi Kit showed no consistent biases in amplification of particular genomic regions and had 13–51 SNVs compared with the metagenome sequenced without whole genome amplification, with the majority of the SNVs belonging to the C > T and G > A transition (Troell et al., 2016). This amount to less than 1% of the magnitude of differences between the genomes sequenced in the present study and the reference IOWA genome, and thus should not affect significantly the outcome of comparative genomic analysis (Table 4).

Some differences were observed in the numbers of several genes that encode proteins thought to be involved in invasion. The specimens sequenced in this study have one more SKSR and insulinase genes and one less MEDLE gene than the reference IIa IOWA genome. Previously, *C. parvum* and *C. hominis* were shown to have different numbers of genes for these multigene families (Guo et al., 2015b; Isaza et al., 2015). As shown previously (Guo et al., 2015b; Isaza et al., 2015), none of the gene deletions occurred in the 10 sequence gaps present in the reference IOWA genome. Subtelomeric gene duplication or deletion of these genes was suggested to be involved in differences in host specificity between the two genetically related *Cryptosporidium* spp. (Guo et al., 2015b). Although the function of SKSR proteins is unknown, the reference *C. parvum* genome has eight such genes, and the only *C. hominis*-specific gene identified thus far, Chro.50011, also has SR repeats at its C terminus. Interestingly, a paralog of Chro.50011, Chro.50010, also has SR repeats at the C terminus. Thus, both Chro.50011 and Chro.50010 are probably members of the SKSR gene family. In *C. hominis*, Chro50011 is the last gene in chromosome 3 whereas Chro.50010 is the second-to-last gene in chromosome 5. The latter also has some sequence similarity to cgd2_4380, which is the last gene in chromosome 2 of *C. parvum* and was identified initially as a *C. parvum*-specific gene (Cops-1), but is known to have an orthologue in some *C. hominis* specimens (Bouzid et al., 2013). In *C. parvum*, an abridged, unannotated orthologue of Chro.50010 exists near the end of chromosome 5. Therefore, all genes involved in gene duplications and deletions among *Cryptosporidium* spp. are located in subtelomeric regions and are members of multigene families.

The biological significance of the gene gains and losses cannot be fully addressed in this study due to the nature of the specimens used. In this comparative study, the specimens analysed were all from calves and the gains and losses in three gene families were observed in both IIa and IId subtypes. However, results of PCR analysis of three *C. parvum* IIa specimens from the United Sates, where IId subtypes have not been reported, indicate the absence of the major insertions at the 3′ end of chromosomes 3 and 6. Indeed comparative genomic analysis of 30 additional *C. parvum* specimens has confirmed the absence of the two major insertions in IIa subtypes in the United States (unpublished data). Thus, there could be geographic differences in gene contents among *C. parvum* IIa isolates. In some

areas such as Spain, Greece and Australia, IId subtypes are commonly found in lambs and goats whereas IIa subtypes are found in calves (Quilez et al., 2008a, b; Tzanidakis et al., 2014; Yang et al., 2014). IIa subtypes, however, have been seen in lambs in some countries including Spain (Geurden et al., 2008; Caccio et al., 2013; Connelly et al., 2013; Imre et al., 2013; Diaz et al., 2015), and in some areas such as China, Egypt and Sweden, calves are commonly infected with IId subtypes (Amer et al., 2013a, b; Silverlas et al., 2013; Wang et al., 2014). Therefore, direct comparative genomic analysis of IIa and IId specimens from areas known to have distinct distribution of the two subtype families between calves and lambs is needed to infer the role of subtelomeric gene duplications in host adaptation by *C. parvum* subtype families.

The sequence differences between IIa and IId genomes involve mostly genes encoding mucin proteins such as cgd1_470 (encoding mucin 8), cgd6_40 (encoding MSC6-7), and cgd6_1080 (encoding gp60). As mucin proteins in *Cryptosporidium* spp. are best known for their roles in sporozoite invasion (O'Connor et al., 2009; Chatterjee et al., 2010; Ludington and Ward, 2016), the sequence polymorphism could contribute to the biological differences between IIa and IId subtype families of *C. parvum*. Mucins such as gp60 are also highly immunogenic, and thus are often immunodominant antigens. As such, genes encoding these proteins are naturally highly polymorphic (Strong et al., 2000; O'Connor et al., 2009). Nevertheless, only some of the mucin genes are divergent between IIa and IId *C. parvum*, as none of the seven mucin genes clustered in chromosome 2 have shown significant sequence differences.

Other highly polymorphic genes between the two subtype families include cgd1_120 (encoding a secretory protein with cysteine cluster), cgd6_10 (encoding the well-known secretory protein CP56), and cgd6_5490 (encoding a MEDLE secretory protein). These are also mostly multicopy genes in subtelomeric regions and encode secretory proteins. Higher sequence differences in genes encoding secretory proteins were previously observed between IIa and IIc subtype families of *C. parvum* (Widmer et al., 2012). The identity of genes with high sequence polymorphism, however, differed between the two studies. The earlier study identified high sequence divergence in genes encoding ABC transporters between IIa and IIc subtype families, whereas we have identified high sequence divergence in mucin protein genes. Further studies using comparative genomic analysis of more isolates with diverse phenotypic characteristics (including IIc isolates) and genetic aberrations are needed to examine the role of these proteins in host adaptation by *C. parvum*.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (31229005, 31425025, and 31110103901), Open Funding Project of the State Key Laboratory of Veterinary Etiological Biology, Lanzhou, China (SKLVEB2014KFKT008), and the US Centers for Disease Control and Prevention. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the US Centers for Disease Control and Prevention.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijpara.2016.12.002.

References

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, Buck GA, Xu P, Bankier AT, Dear PH, Konfortov BA, Spriggs HF, Iyer L, Anantharaman V, Aravind L, Kapur V. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. Science. 2004; 304:441–445. [PubMed: 15044751]

Amer S, Zidan S, Adamu H, Ye J, Roellig D, Xiao L, Feng Y. Prevalence and characterization of *Cryptosporidium* spp. in dairy cattle in Nile River delta provinces, Egypt. Exp Parasitol. 2013a; 135:518–523. [PubMed: 24036320]

Amer S, Zidan S, Feng Y, Adamu H, Li N, Xiao L. Identity and public health potential of *Cryptosporidium* spp. in water buffalo calves in Egypt. Vet Parasitol. 2013b; 191:123–127. [PubMed: 22963712]

Bouzid M, Hunter PR, McDonald V, Elwin K, Chalmers RM, Tyler KM. A new heterogeneous family of telomerically encoded *Cryptosporidium* proteins. Evol Appl. 2013; 6:207–217. [PubMed: 23467513]

Caccio SM, Sannella AR, Mariano V, Valentini S, Berti F, Tosini F, Pozio E. A rare *Cryptosporidium parvum* genotype associated with infection of lambs and zoonotic transmission in Italy. Vet Parasitol. 2013; 191:128–131. [PubMed: 22954678]

Chatterjee A, Banerjee S, Steffen M, O'Connor RM, Ward HD, Robbins PW, Samuelson J. Evidence for mucin-like glycoproteins that tether sporozoites of *Cryptosporidium parvum* to the inner surface of the oocyst wall. Eukaryot Cell. 2010; 9:84–96. [PubMed: 19949049]

Connelly L, Craig BH, Jones B, Alexander CL. Genetic diversity of *Cryptosporidium* spp. within a remote population of Soay Sheep on St Kilda Islands Scotland. Appl Environ Microbiol. 2013; 79:2240–2246. [PubMed: 23354707]

Diaz P, Quilez J, Prieto A, Navarro E, Perez-Creo A, Fernandez G, Panadero R, Lopez C, Diez-Banos P, Morrondo P. *Cryptosporidium* species and subtype analysis in diarrhoeic pre-weaned lambs and goat kids from northwestern Spain. Parasitol Res. 2015; 114:4099–4105. [PubMed: 26212102]

Geurden T, Thomas P, Casaert S, Vercruysse J, Claerebout E. Prevalence and molecular characterisation of *Cryptosporidium* and *Giardia* in lambs and goat kids in Belgium. Vet Parasitol. 2008; 155:142–145. [PubMed: 18565678]

Guo Y, Li N, Lysen C, Frace M, Tang K, Sammons S, Roellig DM, Feng Y, Xiao L. Isolation and enrichment of *Cryptosporidium* DNA and verification of DNA purity for whole-genome sequencing. J Clin Microbiol. 2015a; 53:641–647. [PubMed: 25520441]

Guo Y, Tang K, Rowe LA, Li N, Roellig DM, Knipe K, Frace M, Yang C, Feng Y, Xiao L. Comparative genomic analysis reveals occurrence of genetic recombination in virulent *Cryptosporidium hominis* subtypes and telomeric gene duplications in *Cryptosporidium parvum*. BMC Genomics. 2015b; 16:320. [PubMed: 25903370]

Imre K, Luca C, Costache M, Sala C, Morar A, Morariu S, Ilie MS, Imre M, Darabus G. Zoonotic *Cryptosporidium parvum* in Romanian newborn lambs (*Ovis aries*). Vet Parasitol. 2013; 191:119–122. [PubMed: 22995338]

Isaza JP, Galvan AL, Polanco V, Huang B, Matveyev AV, Serrano MG, Manque P, Buck GA, Alzate JF. Revisiting the reference genomes of human pathogenic *Cryptosporidium* species: reannotation of *C. parvum* Iowa and a new *C. hominis* reference. Sci Rep. 2015; 5:16324. [PubMed: 26549794]

Ludington JG, Ward HD. The *Cryptosporidium parvum* C-Type lectin CpClec mediates infection of intestinal epithelial cells via interactions with sulfated proteoglycans. Infect Immun. 2016; 84:1593–1602. [PubMed: 26975991]

O'Connor RM, Burns PB, Ha-Ngoc T, Scarpato K, Khan W, Kang G, Ward H. The polymorphic mucin antigens CpMuc4 and CpMuc5 are integral to *Cryptosporidium parvum* infection in vitro. Eukaryot Cell. 2009; 8:461–469. [PubMed: 19168754]

Quilez J, Torres E, Chalmers RM, Hadfield SJ, Del Cacho E, Sanchez-Acedo C. *Cryptosporidium* genotypes and subtypes in lambs and goat kids in Spain. Appl Environ Microbiol. 2008a; 74:6026–6031. [PubMed: 18621872]

Quilez J, Torres E, Chalmers RM, Robinson G, Del Cacho E, Sanchez-Acedo C. *Cryptosporidium* species and subtype analysis from dairy calves in Spain. Parasitology. 2008b; 135:1613–1620. [PubMed: 18980704]

Ryan U, Fayer R, Xiao L. *Cryptosporidium* species in humans and animals: current understanding and research needs. Parasitology. 2014; 141:1667–1685. [PubMed: 25111501]

Santin M. Clinical and subclinical infections with *Cryptosporidium* in animals. N Z Vet J. 2013; 61:1–10. [PubMed: 23134088]

Silverlas C, Bosaeus-Reineck H, Naslund K, Bjorkman C. Is there a need for improved *Cryptosporidium* diagnostics in Swedish calves? Int J Parasitol. 2013; 43:155–161. [PubMed: 23142404]

Strong WB, Gut J, Nelson RG. 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. 2000; 68:4117–4134. [PubMed: 10858229]

Troell K, Hallstrom B, Divne AM, Alsmark C, Arrighi R, Huss M, Beser J, Bertilsson S. *Cryptosporidium* as a testbed for single cell genome characterization of unicellular eukaryotes. BMC Genomics. 2016; 17:471. [PubMed: 27338614]

Tzanidakis N, Sotiraki S, Claerebout E, Ehsan A, Voutzourakis N, Kostopoulou D, Stijn C, Vercruysse J, Geurden T. Occurrence and molecular characterization of Giardia duodenalis and Cryptosporidium spp. in sheep and goats reared under dairy husbandry systems in Greece Parasite (Paris, France). 2014; 21:45.

Wang R, Zhang L, Axen C, Bjorkman C, Jian F, Amer S, Liu A, Feng Y, Li G, Lv C, Zhao Z, Qi M, Dong H, Wang H, Sun Y, Ning C, Xiao L. *Cryptosporidium parvum* IId family: clonal population and dispersal from Western Asia to other geographical regions. Sci Rep. 2014; 4:4208. [PubMed: 24572610]

Widmer G, Lee Y. Comparison of single- and multilocus genetic diversity in the protozoan parasites *Cryptosporidium parvum* and *C. hominis*. Appl Environ Microbiol. 2010; 76:6639–6644. [PubMed: 20709840]

Widmer G, Lee Y, Hunt P, Martinelli A, Tolkoff M, Bodi K. Comparative genome analysis of two *Cryptosporidium parvum* isolates with different host range. Infect Genet Evol. 2012; 12:1213–1221. [PubMed: 22522000]

Xiao L. Molecular epidemiology of cryptosporidiosis: an update. Exp Parasitol. 2010; 124:80–89. [PubMed: 19358845]

Xiao L, Hlavsa MC, Yoder J, Ewers C, Dearen T, Yang W, Nett R, Harris S, Brend SM, Harris M, Onischuk L, Valderrama AL, Cosgrove S, Xavier K, Hall N, Romero S, Young S, Johnston SP, Arrowood M, Roy S, Beach MJ. Subtype analysis of *Cryptosporidium* specimens from sporadic cases in Colorado, Idaho, New Mexico, and Iowa in 2007: widespread occurrence of one *Cryptosporidium hominis* subtype and case history of an infection with the *Cryptosporidium* horse genotype. J Clin Microbiol. 2009; 47:3017–3020. [PubMed: 19587303]

Yang R, Jacobson C, Gardner G, Carmichael I, Campbell AJ, Ng-Hublin J, Ryan U. Longitudinal prevalence, oocyst shedding and molecular characterisation of *Cryptosporidium* species in sheep across four states in Australia. Vet Parasitol. 2014; 200:50–58. [PubMed: 24332963]

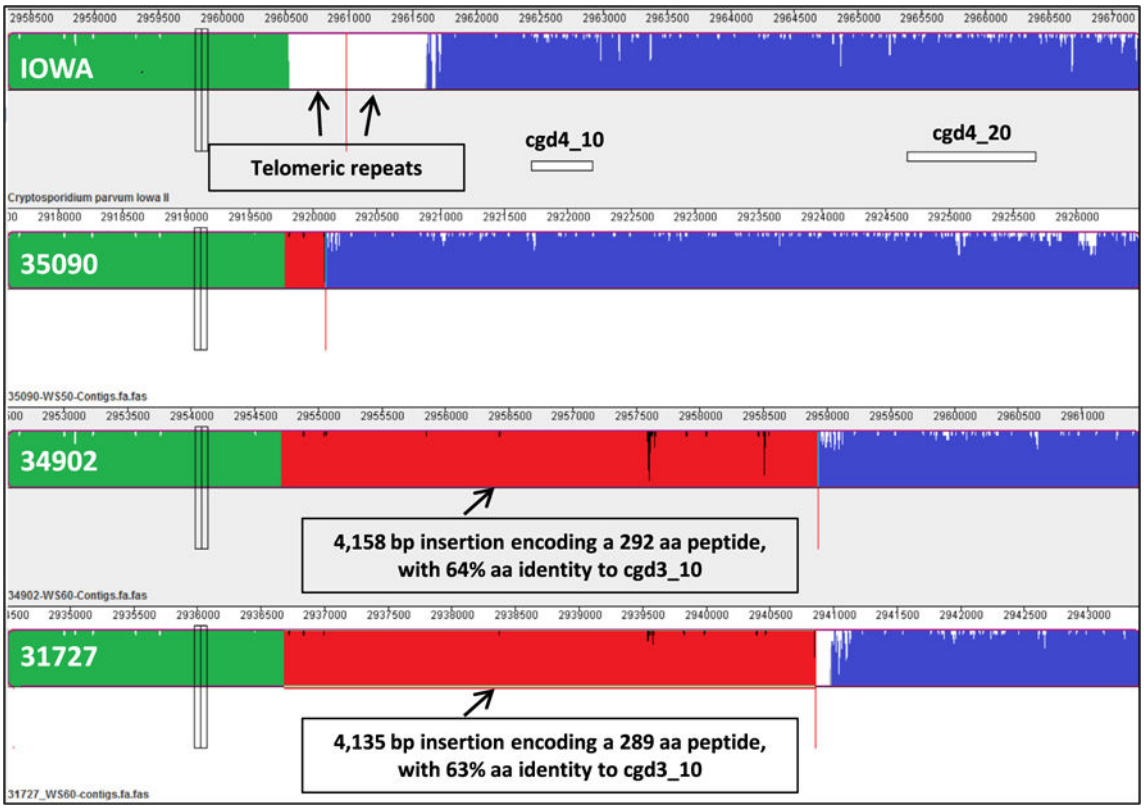

Fig. 1.

Insertion of one *Cryptosporidium*-specific (conserved sequence motif) SKSR gene at the 3′ end of chromosome 3 in genomes of *Cryptosporidium parvum* sequenced in this study. Compared with the published IIa reference IOWA isolate, IId specimens from China (31727) and Egypt (34902) had a 4,135 bp and 4,158 bp insertion at the 30 end of chromosome 3, which encodes a 289 and 292 amino acid (aa) peptide with up to 63% and 64% sequence identity to the *Cryptosporidium*-specific SKSR gene cgd3_10, respectively. IIa specimen 35090 appears to have the same insertion in this region. The vertical red (black) line in the IOWA reference isolate represents the border of the 3′ end of chromosome 3 and the 5′ end of chromosome 4, while the vertical black box (to the left) indicates the same sequence location in genomes under comparison. The coding regions of two genes at the 5′ end of chromosome 4, cgd4_10 and cgd4_20, are shown as horizontal bars. The sequence insertions in chromosome 3 are shown in red (grey), chromosome 3 sequences upstream from the insertion are shown in green (black) on the left, and chromosome 4 sequences are shown in blue (black) on the right. In the reference IOWA genome, the chromosomes 3 and 4 have numerous copies of the telomeric repeat sequences (TTTAGG at the 3′ end of chromosome 3 and CCTAAA at the 5′ end of chromosome 4), which are shown in white. Sizes of the full insertion in the 31727 and 34902 genomes are specified. Only the first 283 bp of the insertion is shown for the 35090 genome, but another 2,368 bp fragment of the insertion is present at the end of the genome sequence alignment. White peaks within each block are sequence divergence between the reference (IOWA) genome and genomes obtained in this study, while black peaks within the insertions are sequence differences

among the latter. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

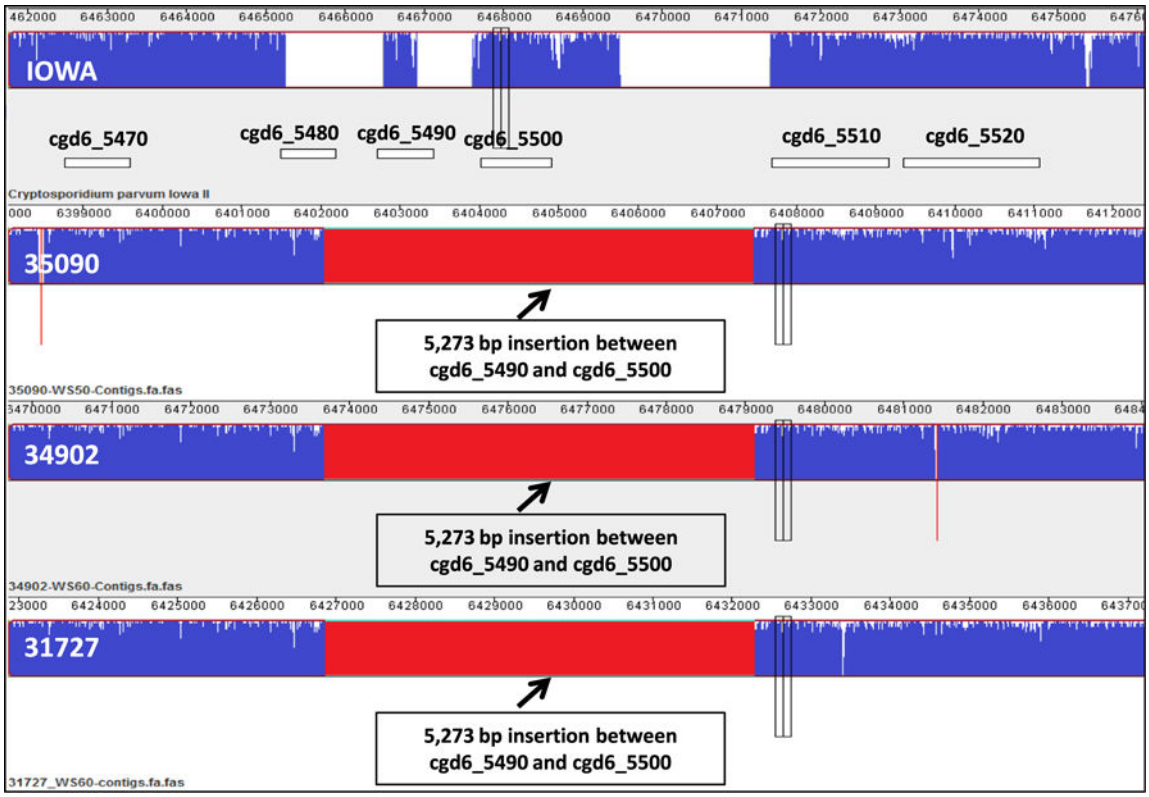

Fig. 2.

Deletion of one gene encoding the (conserved sequence motif) MEDLE family of secretory protein and insertion of one gene encoding the insulinase-like peptidase at the 3′ end of chromosome 6 in genomes of *Cryptosporidium parvum* sequenced in this study. Compared with the IIa reference IOWA gene, the three *C. parvum* specimens sequenced, including the IIa specimen from Egypt (35090) and IId specimens from and Egypt (34902) and China (31727), have a 5,273 bp insertion (shown in red (grey)) after gene cgd6-5490. The insertion has up to 66–71% nucleotide sequence identity to cgd3_4260 or cgd3_4270 (genes encoding insulinase-like peptidases similar to cgd6_5510 and cgd6_5520 downstream from the insertion). Upstream from the insertion, the reference IOWA genome has two genes encoding the MEDLE family of secreted proteins (cgd6_5480 and cgd6_5490), whereas genomes sequenced in this study have only one such gene (cgd6_5490) in this region (sequences shown in white blocks are not present in genomes sequenced in this study, including cgd6_5480 in the first white block in the IOWA reference isolate). Within the panels, the single vertical red (black) line represents the border of two contigs in the genome assembly, while the vertical black box indicates the relative location of the cgd6_5500 gene. White peaks within each block are sequence divergence between the reference (IOWA) genome and genomes obtained in this study. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

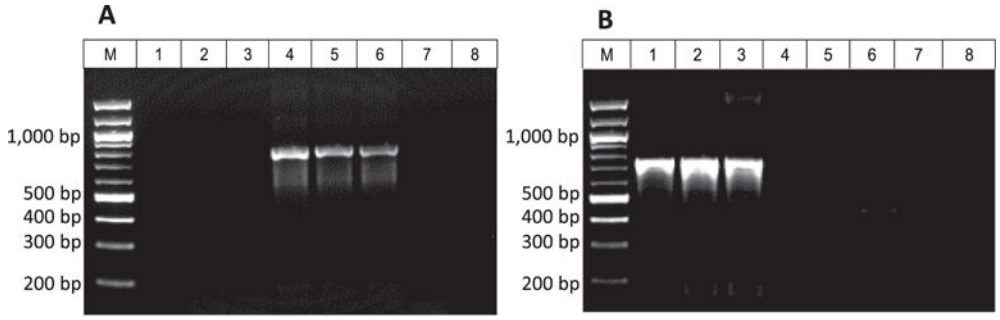

Fig. 3.

Confirmation of major insertions at the 3′ end of chromosomes 3 and 6 in *Cryptosporidium parvum* IId genomes by PCR. (A) Confirmation of the 4,135–4,189 bp insertion in chromosome 3 by PCR targeting on conserved sequences upstream of the telomeric repeats in the reference IOWA genome and 5′ end of the insertion in three genomes obtained from this study. Among the IIa and IId specimens analysed, only three IId specimens from Egypt, Spain and Greece produced the expected 781 bp PCR product. (B) Confirmation of the 5,273 bp insertion in chromosome 6 by PCR targeting conserved nucleotide sequences flanking the insertion (3′ end of cgd6_5490 and 5′ end of cgd6_5500). Among the IIa and IId specimens analysed, only three IIa specimens from the United States produced the expected 781 bp PCR product, as the large insertion in IId genomes between cgd6_5490 and cgd6_5500 had prevented the amplification of the target sequence. M, size marker in 100 bp; lane 1, IIaA15G2R2 from USA; lane 2, IIaA15G2R1 from USA; lane 3, IIaA18G3R1 from USA; lane 4, IIdA20G1 from Egypt; lane 5, IIdA21G1 from Spain; lane 6, IIdA16G1 from Greece; lane 7, negative control and primary PCR; and lane 8, negative control for secondary PCR.

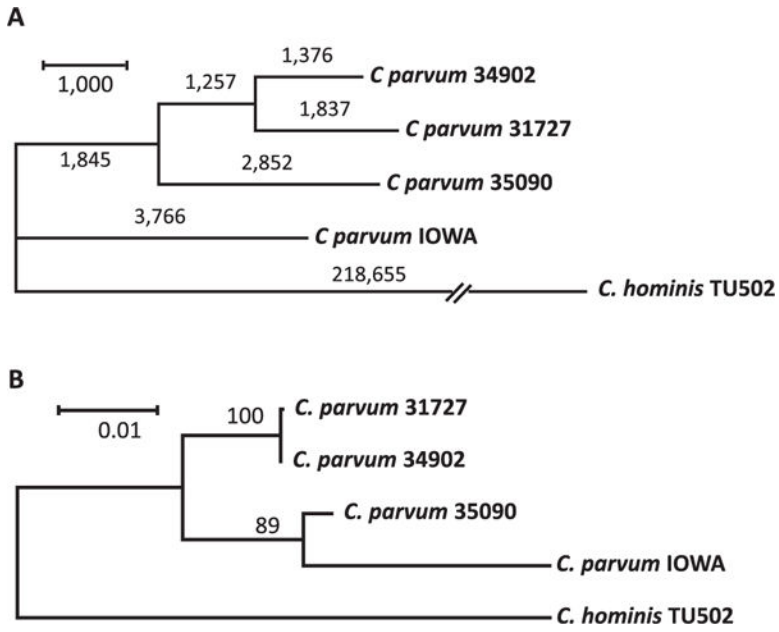

Fig. 4.

Phylogenetic relationship of *Cryptosporidium parvum* genomes characterised in this study as indicated by neighbour-joining analysis of whole genome single nucleotide variant data (A) and concatenated sequences (B) of 11 highly polymorphic genes (cgd1_120, cgd1_470, cgd4_10, cgd4_20, cgd4_30, cgd6_40, cgd6_1080, cgd6_5460, cgd6_5470, cgd6_5490, and cgd6_5500). In the former, genetic distances were computed using R package 'APE' and are in the number of single nucleotide variants, whereas in the latter, they were computed using the Kimura 2-parameter model and are in the units of the number of nucleotide substitutions per site. Numbers of single nucleotide variants and percentages of bootstrapping (1000 replicates) are shown on branches in A and B, respectively. Both trees are outgrouped with sequences from the *Cryptosporidium hominis* TU502 isolate.

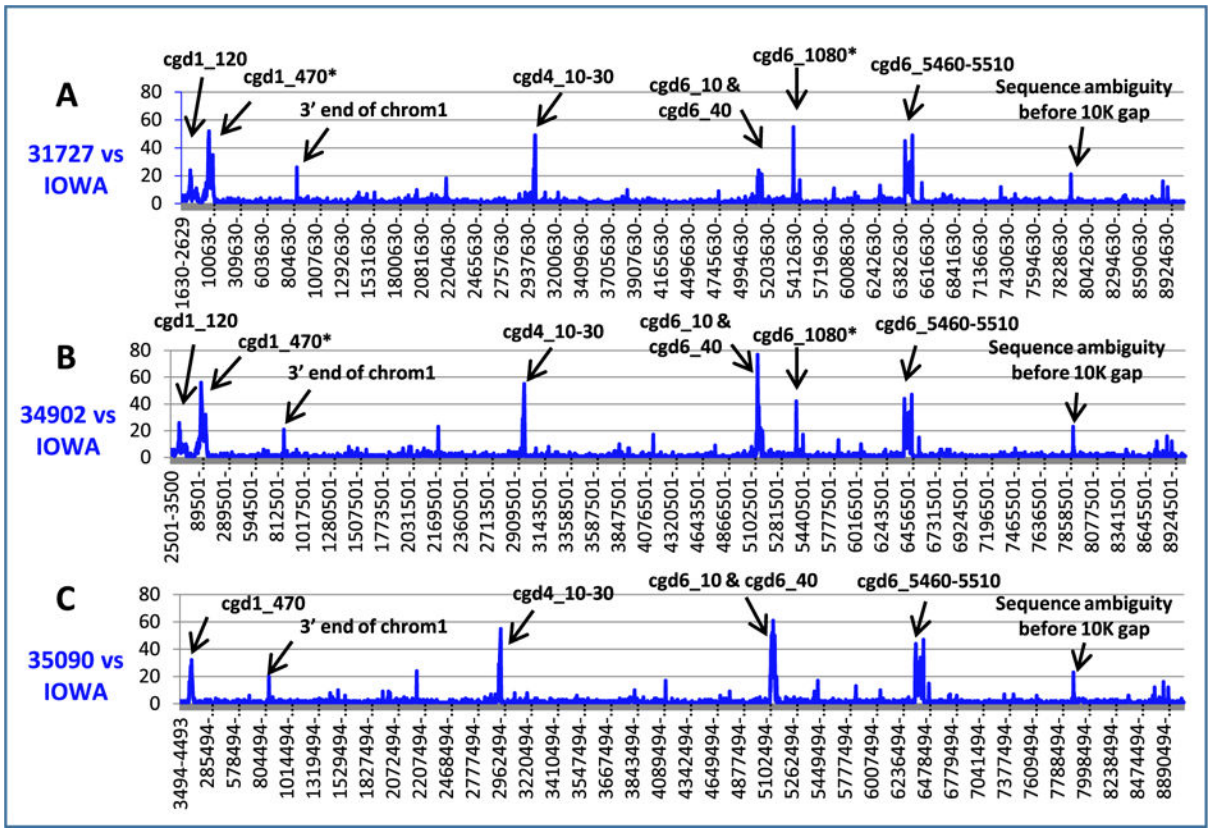

Fig. 5.

Distribution of single nucleotide variants in *Cryptosporidium parvum* genomes sequenced in this study in comparison with the published reference IOWA genome. To detect single nucleotide variants, sequence reads from each specimen were mapped to the 9.1 Mb reference IOWA genome and single nucleotide variants present were detected using the Basic Variant Detection tool in CLC Genomic Workbench 8.5. The *Y*-axis values are numbers of single nucleotide variants per 1,000 nucleotides and *X*-axis values are nucleotide positions of the single nucleotide variant peaks across the eight linked chromosomes (chrom). The *X*-axis labels for (A–C) are not drawn at the same scale, resulting in a slight shift of single nucleotide variant peaks. Highly polymorphic genes (single nucleotide variants values >20/1,000 nucleotides) are labelled in (A) 31727: IId specimen from China; (B) 34902: IId specimen from Egypt; (C) 35090: IIa specimen from Egypt. Genes only highly polymorphic between IId and the reference IOWA genomes are labelled with an asterisk.

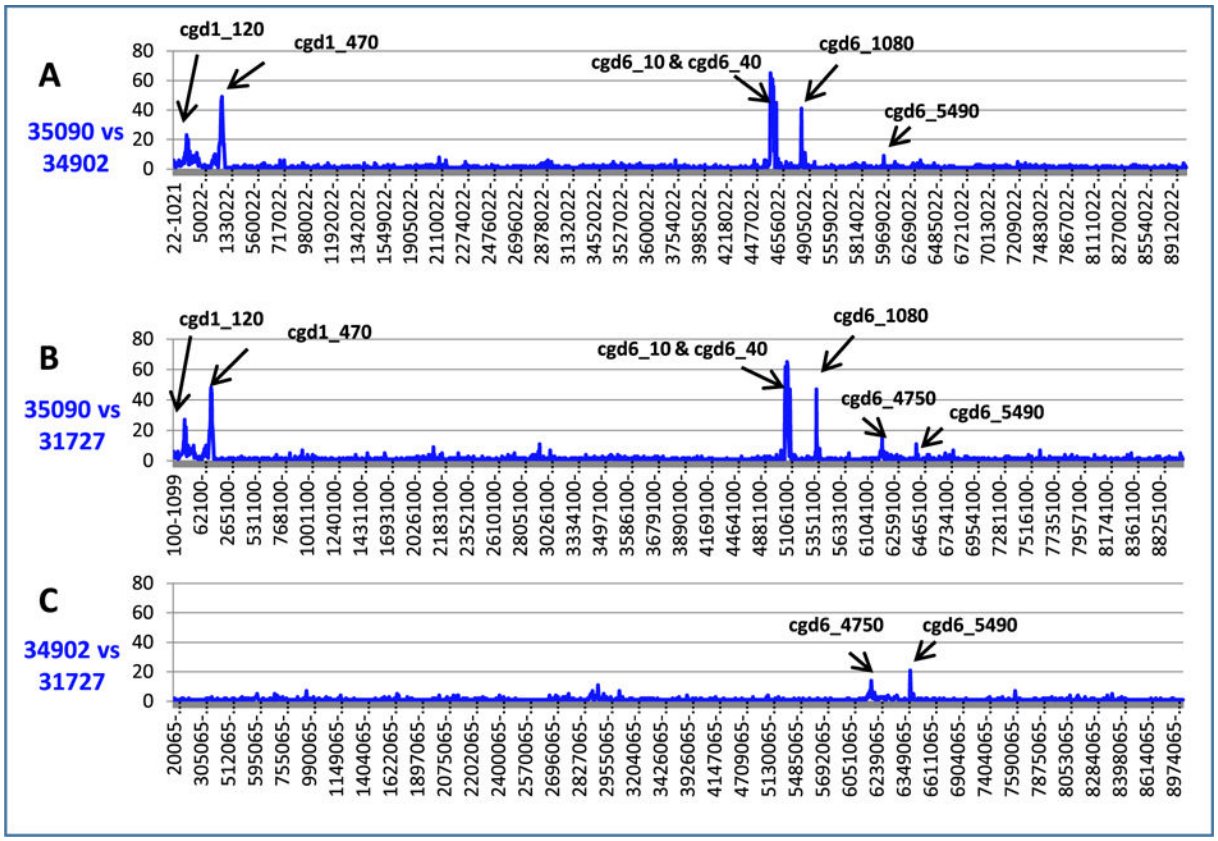

Fig. 6.

Distribution of single nucleotide variants among *Cryptosporidium parvum* genomes sequenced in this study. To detect single nucleotide variants, sequence reads from one specimen were mapped to the assembled genome of another specimen and single nucleotide variants present were detected using the Basic Variant Detection tool in CLC Genomic Workbench 8.5. The *Y*-axis values are numbers of single nucleotide variants per 1,000 nucleotides and *X*-axis values are nucleotide positions of the single nucleotide variant peaks across the eight linked chromosomes. Highly polymorphic genes (single nucleotide variants values > 20/1,000 nucleotides) are labelled. (A) 34902, IId specimen from Egypt versus 35090, IIa specimen from Egypt; (B) 31727, IId specimen from China versus 35090; (C) 34902 versus 31727.

Table 1

Summary of *Cryptosporidium parvum* genomes sequenced in this study.

Specimen ID	**31727**	**34902**	**35090**
Subtype	IIdA19G1	IIdA20G1	IIaA15G1R1
Host	Dairy cattle	Buffalo	Dairy cattle
Source location	Henan, China	Kafr El Sheikh, Egypt	El Beheira, Egypt
Sequencing technique[a]	100 bp PE	100 bp SE	100 bp SE
Total sequencing reads	13,074,496	18,907,631	14,188,762
Reads after cleaning	12,401,023/8,672,250	17,733,415/13,422,516	13,603,027/11,264,070
Reads mapped	11,802,700/8,619,306	17,450,035/13,270,834	13,574,206/11,233,559
Average coverage (fold) [b]	116.0/78.5	168.7/126.5	131.9/107.9
Total length of assembly (bp) [b]	9,119,494/9,131,157	9,146,840/9,237,554	9,062,877/9,229,217
N50 (bp) [b]	76,396/127,633	21,594/42,637	4,248/29,222
Maximum contig size (bp) [b]	232,359/429,383	85,438/170,615	48,893/116,545
No. of contigs [b]	337/421	1,103/989	3,269/1,390
Size of draft genome (bp) [b]	9,082,089/9,080,318	9,111,587/9,136,723	9,040,186/9,147,844
No. of contigs in draft genome [b]	309/242	1,076/567	3,256/841

N50, the number of contigs (sorted by length from longest to shortest) whose length when summed covers 50% or more of the genome assembly.

[a]Sequencing technique used as either paired-end sequencing (PE) or single-end sequencing (SE).

[b]Statistics from CLC Genomics Workbench (read mapping and de novo assembly)/BWA (Burrows Wheeler Alignment; read mapping) or SPAdes (de novo assembly).

Table 2

Gains and losses of genes in *Cryptosporidium parvum* IId subtype family compared with *Cryptosporidium hominis* and *Cryptosporidium parvum* IOWA reference isolate (subtype family IIa).

Chromosome	Gene family	*C. parvum* IOWA	*C. parvum* in this study	*C. hominis* TU502
3	SKSR[a]	–	cgd3_10 paralog [b]	Chro.50011
5	MEDLE family of secreted proteins	cgd5_4580	cgd5_4580	–
		cgd5_4590	cgd5_4590	–
		cgd5_4600	cgd5_4600	cgd5_4600
		cgd5_4610	cgd5_4610	–
6	MEDLE family of secreted proteins	cgd6_5480	–	–
		cgd6_5490	cgd6_5490 [b]	–
6	Insulinase-like proteases	cgd6_5510	cgd6_5510	–
		cgd6_5520	cgd6_5520	–
		–	cgd3_4260 paralog [b]	–
8	*Cryptosporidium*-specific paralogs	cgd8_660	cgd8_660	cgd8_660
		cgd8_680	cgd8_680	–
		cgd8_690	cgd8_690	–

[a] Chro.50011 is a paralog of Chro.50010, which encodes a secretory protein with SR motif repeats at the C terminus and is the second-to-last gene in chromosome 5 (the last gene is Chro.00007, which is a paralog of cgd8_10; *C. parvum* has orthologues of both Chro.50010 and Chro.00007). Both Chro.50010 and Chro.50011 are probably members of the SKSR gene family. The orthologues of Chro.50010 and Chro.00007 in *C. parvum* were not annotated in the IOWA genome.

[b] The gene was also present in *C. parvum* IIa specimen 35090.

Table 3

Summary of insertion and deletion (indel) events detected between *Cryptosporidium parvum* specimens sequenced in this study and the *Cryptosporidium parvum* IOWA reference genome (based on comparison of assemblies).[a]

Chromosome	No. of insertion events			No. of deletion events			Largest insertion (nt)			Largest deletion (nt)		
	35090	34902	31727	35090	34902	31727	35090	34902	31727	35090	34902	31727
1	104	143	162	85	142	150	18	88	88	18	26	28
2	116	107	99	91	117	145	256	175	315	256	36	318
3	103	103	129	143	140	148	4,189	4,158	4,135	742	1,914	329
4	193	187	211	193	204	202	36	210	210	276	146	231
5	140	165	127	226	232	172	42	90	37	95	156	156
6	260	219	246	364	330	342	5,273	5,273	5,273	1,877	1,877	1,877
7	71	81	92	84	113	119	29	219	119	64	84	84
8	142	158	157	239	256	244	78	54	54	150	150	144
Total	1,128	1,162	1,222	1,425	1,534	1,522	–	–	–	–	–	–

nt, nucleotide.

[a]Details on indels are presented in Supplementary Tables S4–S6.

Table 4

Number of single nucleotide variants in *Cryptosporidium parvum* genomes sequenced in this study compared with the *Cryptosporidium parvum* IOWA reference genome using Burrows Wheeler Alignment mapping.

Specimen ID.	Total SNVs	No. of genes with SNVs	No. of SNVs in genes	No. of non-synonymous SNVs
31727	5,386	1,505 (39.6%)	3,405 (63.2%)	1,821 (53.5%)
34902	5,766	1,525 (40.1%)	3,630 (63.0%)	1,971 (54.3%)
35090	5,191	1,459 (38.3%)	3,210 (61.8%)	1,702 (53.0%)