Published in final edited form as:

Water Res. 2024 May 01; 254: 121333. doi:10.1016/j.watres.2024.121333.

Sequence introgression from exogenous lineages underlies genomic and biological differences among *Cryptosporidium parvum* IOWA lines

Wanyi Huang[a,1], **Kevin Tang**[b,1], **Chengyi Chen**[a], **Michael J. Arrowood**[c], **Ming Chen**[a], **Yaqiong Guo**[a], **Na Li**[a], **Dawn M. Roellig**[c,*], **Yaoyu Feng**[a,*], **Lihua Xiao**[a,*]

[a] State Key Laboratory for Animal Disease Control and Prevention, South China Agricultural University, Guangzhou 510642, China

[b] Division of Scientific Resources, Centers for Disease Control and Prevention, Atlanta, GA 30341, USA

[c] Division of Foodborne, Waterborne, and Environmental Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia 30341, USA

Abstract

The IOWA strain of *Cryptosporidium parvum* is widely used in studies of the biology and detection of the waterborne pathogens *Cryptosporidium* spp. While several lines of the strain have been sequenced, IOWA-II, the only reference of the original subtype (IIaA15G2R1), exhibits significant assembly errors. Here we generated a fully assembled genome of IOWA-CDC of this subtype using PacBio and Illumina technologies. In comparative analyses of seven IOWA lines maintained in different laboratories (including two sequenced in this study) and 56 field isolates, IOWA lines (IIaA17G2R1) with less virulence had mixed genomes closely related to IOWA-CDC but with multiple sequence introgressions from IOWA-II and unknown lineages. In addition, the IOWA-IIaA17G2R1 lines showed unique nucleotide substitutions and loss of a gene associated with host infectivity, which were not observed in other isolates analyzed. These genomic differences among IOWA lines could be the genetic determinants of phenotypic traits in *C. parvum*. These data provide a new reference for comparative genomic analyses of *Cryptosporidium* spp. and rich targets for the development of advanced source tracking tools.

[*] Corresponding authors. iyd4@cdc.gov (D.M. Roellig), yyfeng@scau.edu.cn (Y. Feng), lxiao@scau.edu.cn (L. Xiao).
[1] These authors contributed equally: Wanyi Huang, Kevin Tang.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Wanyi Huang: Writing – review & editing, Writing – original draft, Methodology, Formal analysis. **Kevin Tang:** Writing – review & editing, Software, Methodology, Formal analysis. **Chengyi Chen:** Writing – review & editing, Software, Methodology, Formal analysis. **Michael J. Arrowood:** Writing – review & editing, Resources, Data curation. **Ming Chen:** Writing – review & editing, Software, Formal analysis. **Yaqiong Guo:** Writing – review & editing, Resources, Data curation. **Na Li:** Writing – review & editing, Resources, Data curation. **Dawn M. Roellig:** Writing – review & editing, Resources, Conceptualization. **Yaoyu Feng:** Writing – review & editing, Resources, Conceptualization. **Lihua Xiao:** Writing – review & editing, Writing – original draft, Supervision, Data curation, Conceptualization.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2024.121333.

Keywords

Cryptosporidium parvum ; IOWA strain; Genome; Evolution; Virulence determinants

1. Introduction

Cryptosporidium spp. are the leading cause of diarrhea-related deaths in humans and various animals, and are among the most important waterborne pathogens (Checkley et al., 2015; Efstratiou et al., 2017). Currently, 46 species and more than 120 genotypes of *Cryptosporidium* are recognized, and most of them have preferred hosts and different pathogenicity (Ryan et al., 2021). Among them, *C. parvum* and *C. hominis* are two dominant species in humans. In addition, *C. parvum* is found in various animals, making it a major zoonotic species (Feng et al., 2018). *C. parvum* is often used in studies of the biology of *Cryptosporidium* spp. and in the development of detection tools for oocysts in water and other environmental samples (Dumaine et al., 2020). The two species also differ in human infectivity and virulence (Feng et al., 2018).

Cryptosporidium spp. have genomes of approximately 9 Mb in eight chromosomes, encoding approximately 4000 protein-coding genes (Abrahamsen et al., 2004; Baptista et al., 2021). With the development of next-generation sequencing (NGS) technologies, whole-genome sequence (WGS) data from *Cryptosporidium* spp. have increased rapidly in recent years. To date, 15 species have assembled genomes, of which eight are annotated (Baptista et al., 2021). However, except for three *C. parvum* isolates and one *C. hominis* isolate with fully assembled genomes (Abrahamsen et al., 2004; Isaza et al., 2015; Baptista et al., 2021; Menon et al., 2022), the genomes of most *Cryptosporidium* spp. have been assembled in varying numbers of contigs. Because WGS data are important in studies of the genetic basis of phenotypic traits in *Cryptosporidium* spp. (Fan et al., 2019), it is crucial to have fully assembled and annotated genomes for comparisons of gene content and characterization of copy number variation among species or subtypes.

For more than 40 years, the *C. parvum* IOWA strain has been the most widely used isolate for biological studies of *Cryptosporidium* spp., evaluation of therapeutic agents and vaccine candidates, and development of detection and diagnostic tools (Moon and Bemrick 1981; Cama et al., 2006). It was also the first *Cryptosporidium* isolate to have its entire genome sequenced (Abrahamsen et al., 2004). For nearly 20 years, the first fully assembled *C. parvum* genome, designated as IOWA-II, was the reference genome for comparative genomics analyses. In recent years, taking advantage of long-read sequencing technologies, *C. parvum* IOWA genomes from two additional sources have been sequenced and assembled into 8 chromosomes (IOWA-ATCC and IOWA-KWI52) (Baptista et al., 2021; Menon et al., 2022). Although these lines are derived from the original IOWA strain, their subtype in the gene encoding the 60 kDa glycoprotein (GP60) has changed from IIaA15G2R1 to IIaA17G2R1, possibly due to contamination by an exogenous isolate during calf passage (Zhang and Zhu 2020).

The IOWA strain was first isolated in the late 1970s. It has been maintained and propagated for research in neonatal calf or mouse infection models at several laboratories, including the

U.S. Centers for Disease Control and Prevention (CDC), Sterling Parasitology Laboratory, Waterborne, Inc., and Bunch Grass Farm (Cama et al., 2006). It is also the source of oocysts used for performance evaluation and quality assurance of the detection tool used in the official Method 1623 for the detection of *Cryptosporidium* oocysts in water samples (USEPA 2012). Interestingly, studies in mouse models have shown variation in infectivity and virulence when using IOWA from different sources, with IOWA-IIaA15G2R1-CDC being lethal and commercial IOWA-IIaA17G2R1 lines being avirulent and less infectious in IFN-knockout (GKO) mice (Ndao et al., 2013; Sonzogni-Desautels et al., 2015; Audebert et al., 2020; He et al., 2022; Jia et al., 2022). However, the genetic determinants of these biological differences are largely unknown.

In this study, we sequenced the *C. parvum* IOWA strain maintained at the U.S. CDC (IOWA-CDC of subtype IIaA15G2R1) using long-read (PacBio) and short-read (Illumina) sequencing technologies. We performed *de novo* genome assembly and annotation to generate a new complete reference genome of the *C. parvum* IOWA strain. In addition, we acquired WGS data from two other IOWA lines from Waterborne, Inc. and Bunch Grass Farm. Together with published data from three other IOWA lines and 56 field isolates, we performed comparative genomic analyses to assess the genetic differences between different lines of the IOWA strain. The knowledge gained from the analyses will deepen our understanding of the evolution of the IOWA strain and facilitate the development of molecular tools for advanced tracing of sources of *Cryptosporidium* oocyst contamination in source and drinking water.

2. Materials and methods

2.1. Cryptosporidium parvum samples

The *C. parvum* IOWA-CDC line (38,783: 43th passage since its initial arrival at CDC) belongs to the IIaA15G2R1 subtypes and was obtained from CDC, Atlanta, Georgia. It was originally obtained from the University of Arizona in 1989 and has been maintained at CDC through regular calf passages ever since. The IOWA-Waterborne line, referred to by the supplier as the IOWA isolate but belonging to the IIaA17G2R1 subtype, was purchased from Waterborne, Inc. (New Orleans, LA). The IOWA-Bunchgrass line (also IIaA17G2R1 subtype) was supplied by Prof. Guan Zhu of the Texas A&M University, but was originally obtained from Bunch Grass Farm (Deary, ID). All oocyst preparations were stored in 2.5 % potassium dichromate solution or antibiotics at 4 °C.

2.2. Whole-genome sequencing and sources of other WGS data

DNA was extracted from the purified oocysts using the Qiagen Blood and Tissue Kit after five cycles of freezing and thawing in liquid nitrogen and a 37 °C water bath. DNA was sequenced on an Illumina HiSeq 2500 (Illumina, San Diego, CA, United States) using the 250-bp paired-end approach as described (Guo et al., 2015). DNA was also extracted from IOWA-CDC oocysts using the traditional phenol-chloroform method and whole-genome sequenced using standard PacBio procedures.

Since the subtype identity of the IOWA strain has changed from IIaA15G2R1 to IIaA17G2R1 due to possible contamination by an exogenous IIa isolate during passage, we retrieved a total of 59 sets of whole genome sequence (WGS) data of *C. parvum* IIa subtypes from around the world to track the evolution of IOWA strain. All data were downloaded from the Sequence Read Archive (SRA) database of the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/sra/). They were derived from published studies (Hadfield et al., 2015; Troell et al., 2016; Nash et al., 2018; Audebert et al., 2020; Baptista et al., 2021; Corsi et al., 2022; Menon et al., 2022; Wang et al., 2022) and obtained through BioProject numbers PRJNA253836, PRJNA253840, PRJNA253843, PRJNA253845, PRJNA253846, PRJNA253847, PRJNA308172, PRJNA439211, PRJNA573722, PRJNA633764, PRJNA634014, PRJNA744539, PRJNA759721, PRJNA810562, PRJNA818164. Three fully assembled *C. parvum* genomes (IOWA-II, IOWA-ATCC, and IOWA-KWI52) and the gene annotation files (.gff) of the IOWA-II and IOWA-ATCC genomes were downloaded from the *Cryptosporidium* Information Resources (CryptoDB v.64). Some basic information about the WGS data is provided in Supplemental Table 1.

2.3. Genome assembly

The Illumina reads for IOWA-CDC were initially assembled using ABySS version 2.0.2 with Kmer values of 31,49, 57, 67, 77, 80, and 87 (Jackman et al., 2017). Based on the N50 size and the similarities to the published IOWA-II genome, the best assembly was obtained by running ABySS with a Kmer value of 57. A hybrid *de novo* assembly was then performed with SPAdes, version 3.8.0, using the PacBio reads, the Illumina reads and the AbySS contigs as trusted contigs (Antipov et al., 2016). The PacBio reads were also assembled using the hierarchical genome assembly process 3 (HGAP3) implemented in smartportal 2.3.0 (Chin et al., 2013). The final assembly was polished by manual inspection of the long PacBio reads aligned to the contigs, PCR and Sanger sequencing.

The Illumina reads for two other IOWA lines from Waterborne, Inc. (IOWA-Waterborne) and Bunch Grass Farm (IOWA-Bunchgrass) were assembled *de novo* using CLC Genomics Workbench with a word size of 63 and bubble size of 400. In addition, all WGS data from SRA database were assembled using SPAdes 3.1 (http://cab.spbu.ru/software/spades/) with Kmer of 63 and the careful mode.

2.4. Molecular characterization

The sequences of the *18S rRNA* and *gp60* genes were extracted from genomes using Blastn v2.15.0 (https://blast.ncbi.nlm.nih.gov/Blast.cgi). These isolates were identified to the subtype level using the established nomenclature (Xiao and Feng 2017). In addition, these sequences were aligned using MUSCLE v5.1 (https://www.ebi.ac.uk/). A maximum likelihood (ML) tree of the *gp60* sequences was reconstructed using RAxML-NG v1.0.0 (Kozlov et al., 2019), with the GTR + I model and 1000 bootstrap replicates.

2.5. Gene prediction and annotation

Gene prediction for the IOWA-CDC genome was performed using AUGUSTUS v2.5.5 and GeneMark-ES v4.32 (Lomsadze et al., 2005). Annotations of the *C. hominis* genome

(UdeA01) were used to train AUGUSTUS for the prediction (Isaza et al., 2015). The Rapid Annotation Transfer Tool was also used to transfer annotations from the reference *C. hominis* genome to a IOWA-CDC genome based on conserved synteny (Otto et al., 2011). The uncharacteristic genes were annotated by searching the NCBI non-redundant protein database using Blastp v2.10.1 with e-value of 1×10^{-6}. The gene nomenclatures used are similar to those of IOWA-II (e.g., CPCDC_6g1080 for the *gp60* gene in IOWA-CDC versus cgd6_1080 for the ortholog in IOWA-II), which have become familiar to *Cryptosporidium* researchers.

2.6. Comparative genomic analysis

Structural variations (SVs) between genomes were detected using Mauve v20150226 and Mummer v3. Major SVs between the two well-assembled IIaA15G2R1 genomes (IOWA-CDC and IOWA-II) were verified by PCR. Copy number variations (CNVs) between the genomes, especially in subtelomeric genes, were determined using Orthofinder v2.5.2. Motifs in genes were searched using Tandem Repeats Finder (https://tandem.bu.edu/trf/) and MEME (Bailey et al., 2009).

2.7. Variant analysis

Reads were trimmed and mapped to the IOWA-CDC and IOWA-II genomes using the BWA-MEM v0.7.17 (Li and Durbin 2009) and called wgSNPs using BCFtools v1.12 (https://samtools.github.io/bcftools/) following procedures described previously (Feng et al., 2017). Genome coverage and sequence depth were estimated using the mpileup algorithm of SAMtools v1.7 (http://samtools.sourceforge.net/).

The SNPs identified above were annotated for variant types and affected genes using SnpEff (https://pcingola.github.io/SnpEff/). SNP distributions were calculated in sliding windows using Vcftools 0.1.16 (https://vcftools.github.io/index.html) and visualized using ‘ggplot2’ in the R package (https://www.rdocumentation.org/packages/ggplot2/versions/3.3.5).

2.8. Phylogenetic analysis

To investigate the relationships among *C. parvum* IOWA-related isolates, a maximum likelihood (ML) tree was constructed from the wgSNPs using RAxML-NG v1.0.0 (https://github.com/amkozlov/raxml-ng). The TVM substitution model used in the ML tree construction was selected using jModelTest v2.1.10 based on Akaike Information Criterion values (Posada 2008). The robustness of clustering was assessed by bootstrapping with 1000 replicates.

2.9. Assessment of gene flow among populations

The wgSNPs were further used phylogenetic network and absolute divergence (*dxy*) analyses. Phylogenetic networks were generated using the neighbor-net algorithm of SplitsTree5 (Huson and Bryant 2006). The *dxy* value between two populations was calculated in 10-kb sliding windows using popgenWindows.py in Genomics_general (https://github.com/simonhmartin/genomics_general) and visualized in line plots using ‘ggplot2’ in the R package.

3. Results

3.1. Genome assemblies for C. parvum IOWA lines maintained at different facilities

We sequenced the IOWA-CDC genome (No. 38,783 from passage 43IA8) using PacBio and obtained a total of 1.24 Gb sequences with 155-fold coverage of the *C. parvum* genome (Supplemental Table 1). After filtering out contigs from contaminants in the *de novo* assembly, we obtained a *Cryptosporidium* genome of 9153,992 bp in 18 contigs with an N50 of 1020,670 bp (Supplemental Table 1). In addition, we sequenced the same genome using the Illumina technology with 264-fold coverage of the genome in 250 bp paired-end reads. The quality of the PacBio-generated assembly was further improved using the short paired-end reads, resulting in the final assembly (named as IOWA-CDC) of eight contigs. Chromosomal identity of the contigs was established by alignment to the published *C. parvum* reference (IOWA-II).

In addition to the IOWA-CDC line, two other lines of IOWA maintained at Waterborne, Inc. and Bunch Grass farm were sequenced using Illumina, and the resulting data were assembled into contigs. Approximately 361-fold coverage of paired-end reads was obtained from each line, resulting in genome assemblies of 9.09 Mb in 91 and 47 contigs, respectively (Supplemental Table 1).

3.2. A new annotated reference genome of the C. parvum IOWA strain

The new IOWA-CDC genome had no gaps or ambiguous bases and contains 13 telomeric ends with the AAACCT or AGGTTT repeats (Supplemental Table 2). The telomeric sequences were present in most chromosomes (2, 3, 4, 5, and 6), except for the 5′ end of chromosome 1 and the 3′ end of chromosomes 7 and 8. Compared to the IOWA-II and IOWA-ATCC reference genomes, no coding sequences were present in the missing telomeric regions.

Both the *de novo* and homology-based approaches were used to annotate the IOWA-CDC genome. A total of 3914 protein-coding genes were predicted in IOWA-CDC. This represented a reduction of 30 genes from IOWA-II, which was the first fully predicted and annotated genome for *C. parvum* IOWA strain at the time of our initial data analyses. Differences between the two genomes included newly predicted, lost, merged and separated genes as well as relocated genes (Supplemental Table 3).

3.3. Subtype diversity among C. parvum IOWA lines maintained at different facilities

At the major subtyping locus of the *gp60* gene, the IOWA lines maintained at different facilities showed two subtypes: IIaA15G2R1 and IIaA17G2R1. The *gp60* sequences of IOWA-II and IOWA-CDC were identical to each other and belonged to the IIaA15G2R1 subtype (Fig. 1). In contrast, the IOWA lines from Waterborne, Inc (IOWA-Waterborne-O16), University of Arizona (IOWA-ATCC), and Bunch Grass Farm (IOWA-Bunchgrass and IOWA-KWI52) produced sequences of the IIaA17G2R1 subtype, with 2–3 SNPs in the nonrepeat region of the gene compared to IOWA-II and IOWA-CDC. Among them, the IOWA-Bunchgrass and IOWA-KWI52 sequences had one more SNP than IOWA-ATCC and IOWA-Waterborne (Fig. 1a). However, the extra SNP was an artifact of the genome assembly, because all IIaA17G2R1 lines had both nucleotides at this site upon close inspection of the read-mapping results (Fig. 1b). In addition, the three unique SNPs in the IOWA-IIaA17G2R1 lines were not detected in any of the field isolates analyzed in the study (Supplemental Table 4), and the ML tree of the *gp60* sequences classified the IIa subtypes of *C. parvum* into two clades with a long branch formed by the IOWA-IIaA17G2R1 lines (Supplemental Fig. 1). The data indicated that the IOWA lines maintained in different laboratories were not genetically identical, and that the *gp60* sequences of IOWA-Waterborne, IOWA-Bunchgrass, IOWA-KWI52, and IOWA-ATCC were unique.

3.4. Structural differences among long-read assemblies of C. parvum IOWA lines

To identify the structural differences among the *C. parvum* IOWA lines maintained at different facilities, we aligned the fully assembled genomes of IOWA-II, IOWA-CDC, IOWA-ATCC, and IOWA-KWI52) using MAUVE and MUMmer. The results indicated that the IOWA-CDC, IOWA-ATCC, and IOWA-KWI52 genome assemblies were largely collinear. In contrast, the genomic structure of IOWA-II differed from the others, with multiple rearrangements, inversions, and deletions of segments of various sizes (Fig. 2a–c).

Of the four fully assembled genomes, IOWA-ATCC and IOWA-KWI52 had an 8.7 kb deletion in the 5′ region of chromosome 5. IOWA-ATCC and IOWA-II had additional deletions of ~9 kb at the 5′ end of chromosome 7 and of ~6 kb at the 3′ end of both chromosomes 7 and 8, and IOWA-II had a deletion of 5.4 kb at the 5′ end of chromosome 1. These deletions contained the rRNA units (Fig. 2b and c). In the IOWA-II assembly, another deletion of 5.7 kb was present in chromosome 3 (Fig. 2b, c, and Supplemental Fig. 2a), although this deletion was not seen in other isolates analyzed in the study except Uppsala1499 (Supplemental Table 4). In contrast, IOWA-KWI52 had large insertions of 13.0 kb, 21.4 kb, and 10.0 kb at the 5′ end of chromosome 1 and the 3′ ends of chromosomes 7 and 8, respectively (Fig. 2b). These regions contained rRNA units and showed high identity to the 5′ region of chromosome 7, which was conserved among the four fully assembled genomes (Fig. 2d and e). In addition, IOWA-II had a 10.3 kb sequence ambiguity (sequencing gap) in chromosome 8 (Fig. 2a and b).

Using existing PacBio long reads, we identified five rRNA units in chromosomes 1, 2, 7, and 8 (Fig. 2e). In the new fully assembled genome, the 5′ regions of chromosomes 2 and 7 began with telomeric repeats, followed by subtelomeric regions and then the rRNA unit, while the telomeric region of the 5′ end of chromosome 1 and the 3′ end of chromosomes 7 and 8 began with rRNA units (Fig. 2e). We were unable to obtain the missing telomeric regions by re-assembly of the region using PacBio reads. In contrast, there were only two rRNA units in the IOWA-ATCC and IOWA-II genomes, while IOWA-KWI52 had five rRNA units after a large region of conserved sequences except in chromosome 2, which was the only non-telomeric unit in all fully assembled IOWA genomes (Fig. 2e). Similar to the IOWA-CDC genome, IOWA-KWI52 also lacked the telomeric repeat at the 5′ end of chromosome 1 and the 3′ end of chromosomes 7 and 8 (Supplemental Table 2).

In the rRNA units, the 18 s, 5.8 s, and 28 s rRNA genes were conserved in length and sequence between copies except for the genes on chromosome 2 of IOWA-CDC, which had the B-type RNA sequence described previously (Le Blancq et al. 1997). In particular, this rRNA unit had very different ITS1 and ITS2 sequences (Supplemental Fig. 3 and Supplemental Table 5). The results of blast analysis and read mapping indicated that the B-type ITS1 sequence was present in 36 *C. parvum* isolates examined in the study, while it was absent in other 27 isolates, including the IOWA-IIaA17G2R1 lines (Supplemental Table 4).

Among the newly assembled IOWA-ATCC, IOWA-CDC and IOWA-KWI52 genomes, the IOWA-ATCC genome had the translocation of a fragment of chromosome 7 to the 5′ end of chromosome 1, which was not found in other isolates analyzed in the study (Fig. 2b and Supplemental Table 4). In addition, the IOWA-II genome had several rearrangements in chromosomes 4, 5, and 6 compared to the others (Supplemental Fig. 2). For example, a large fragment in chromosome 4 was inversely assembled (Supplemental Fig. 2b). Three fragments in the 3′ region of chromosome 5 were translocated, and two of the three were also reversely assembled. One fragment in the 3′ region of chromosome 6 was translocated from chromosome 5 and reverse assembled (Supplemental Fig. 2c). These were likely due to assembly errors, as most Illumina assemblies of isolates analyzed in the study were collinear with IOWA-CDC and IOWA-KWI52 with no rearrangements observed in IOWA-II (Supplemental Table 4). To confirm the correct assembly of the IOWA-CDC genome, PCR primers were designed based on sequences in the transition regions of the rearrangements. The results of these PCR analyses supported the validity of the new genome assembly (Supplemental Fig. 2d-h).

3.5. Genomic differences between C. parvum IOWA lines

The genomes of the IOWA lines differed in the copy number of subtelomeric genes encoding several protein families. Compared to IOWA-CDC, the new IOWA lines (IOWA-ATCC, IOWA-KWI52, IOWA-Waterborne, and IOWA-Bunchgrass) had lost the CPCDC_5g5511 gene in the 5′ region of chromosome 5, while the gene previously renamed as cgd6_5510–5520 in IOWA-II was translocated to the 3′ end of chromosome 6 and was known as CPCDC_6g5511 in IOWA-CDC (Fig. 3). The presence of this additional gene on chromosome 5 of the IOWA-CDC genome was confirmed by PCR analysis (Supplemental Fig. 2f and S2g). Notably, the CPCDC_5g5511 and its paralog CPCDC_6g5511 encoded insulinases and had identical sequences in over 2000 bp (Supplemental Fig. 4a). The CPCDC_5g5511 gene was also present in the Illumina assemblies of all other IIa isolates analyzed in this study (Fig. 3 and Supplemental Table 4). In contrast, the CPCDC_6g5511 gene was absent from the assemblies of IOWA-II and three isolates from Europe (Supplemental Table 4). Two other genes encoding insulinases, CPCDC_3g4260 and its paralog CPCDC_3g4270, had high sequence identity in the middle of the genes (Supplemental Fig. 4b). This partial sequence identity in two genes often led to breaks in contig assembly in these regions when isolates were sequenced using the Illumina short-read technique (Supplemental Fig. 4c).

Pairwise comparisons of the average nucleotide identity (ANI) of IOWA genomes revealed that IOWA-IIaA17G2R1 lines were homologous to each other but distinct from IOWA-II and IOWA-CDC, with the latter having more sequence identity to IOWA-IIaA17G2R1 (Fig. 4a). This finding was supported by the maximum likelihood (ML) tree based on

14,230 whole-genome SNPs (wgSNPs) among 63 *C. parvum* IIa isolates (Fig. 4b). The IOWA-IIaA17G2R1 lines formed a subclade and had nearly identical sequences. IOWA-II formed a long branch outside the subclade, while IOWA-CDC and some other IIaA15G2R1 and IIaA17G2R1 genomes from the United States clustered with them. IIa isolates from other countries formed clusters outside the cluster formed by the US isolates.

3.6. Gene flow between C. parvum IOWA lines

The phylogenetic network constructed using the wgSNP data indicated the presence of gene flow among IOWA-II, IOWA-CDC, and IOWA-IIaA17G2R1, with edges connecting clades in the main cluster (Fig. 4c). In addition, the *dxy* values between the IOWA-CDC and IOWA-IIaA17G2R1 lines were the lowest in most regions across the eight chromosomes, suggesting that they likely had similar origins (Fig. 4a, d). However, lower *dxy* values were present in 820 kb to 1000 kb of chromosome 8 between IOWA-IIaA17G2R1 and IOWA-II (Fig. 4e and Supplemental Fig. 5a). In contrast, IOWA-IIaA17G2R1 lines had substantial *dxy* values compared to other isolates analyzed in the study, such as the 660–700 kb region of chromosome 1 (Supplemental Fig. 5a). This indicated the presence of introgression of exogenous sequences into the IOWA-IIaA17G2R1 lines (Fig. 4f). This was supported by the read-mapping results indicating the presence of two types of nucleotides at some polymorphic loci in IOWA-IIaA17G2R1 (Fig. 1b, Supplemental Fig. 5b).

3.7. Genomic differences between IOWA-CDC and IOWA-IIaA17G2R1

Because the IOWA-CDC and IOWA-IIaA17G2R1 lines differ from each other in virulence in murine models, we investigated the genomic differences between them in more detail. The 306 SNPs between the two types of genomes were unevenly distributed across the eight chromosomes, with much higher numbers on chromosomes 2, 3 and 7 (Fig. 5a). A total of nine highly polymorphic protein-coding genes (HPPGs, diversity > standard deviation from the mean) were identified between the two IOWA lines, including three genes (CPCDC_2g450, CPCDC_2g420, and CPCDC_6g1080, which are named as cgd2_450, cgd2_420, and cgd6_1080 in the IOWA-II genome, respectively) encoding mucin-like glycoproteins and one gene (CPCDC_3g4270 or cgd3_4270 in IOWA-II) encoding an insulinase-like protein (Table 1). On closer inspection of the mapping result of short Illumina reads at the CPCDC_3g4270 locus, the sequence polymorphism here was mostly due to mis-mapping of CPCDC_3g4260 reads with high sequence identity (Supplemental Fig. 4d). However, this gene was identified as a HPPG between IOWA-CDC and IOWA-ATCC by comparisons of PacBio reads.

Although not a HPPG, the CPCDC_8g661 (cgd8_660_670 in the IOWA-II genome) gene in the IOWA-IIaA17G2R1 genomes had a stop codon (position #685), two non-synonymous variations (position #1047 and #1438), and 24-bp deletions (Fig. 5b and Supplemental Table 4). This stop codon in IIaA17G2R1 caused protein truncation and split the gene into two coding sequences. This sequence feature was also present in IOWA-II but only in four of the 56 field isolates examined in this study (Fig. 5C and Supplemental Table 4). Looking closely at the read mapping results, the IOWA-IIaA17G2R1 lines all had two types of nucleotides at most SNP sites (281/306) and both with and without the 24-bp deletion in CPCDC_8g661 (Fig. 5c). In total, 48 IOWA-IIaA17G2R1 specific SNPs were detected, including the 3 unique SNPs in the *gp60* gene described above (Fig. 5c). These comparative data indicated that IOWA-IIaA17G2R1 lines were distinct among *C. parvum* isolates.

4. Discussion

Data from combined PacBio and Illumina sequencing have led to the generation of a new fully assembled and annotated reference genome of the *C. parvum* IOWA strain (IOWA-CDC). Comparative genomic analyses of IOWA lines maintained in different laboratories and other published *C. parvum* IIa isolates revealed that recent IOWA lines of subtype IIaA17G2R1 have more sequence similarity to IOWA-CDC than the originally published IOWA-II but have undergone genetic recombination with unknown strains. In addition, these sequence data have provided evidence for genetic determinants of virulence differences between IOWA-CDC and IOWA-IIaA17G2R1 lines. Their low infectivity and virulence in mouse models, unique sequence features at key genetic loci, and the presence of mixed nucleotides at most polymorphic sites indicate that the IOWA-IIaA17G2R1 lines are poor representatives of field isolates of *C. parvum*. This has important implications for the continued usage of IOWA-IIaA17G2R1 lines in studies of *Cryptosporidium* biology and regulatory testing for *Cryptosporidium* oocysts in water samples.

The use of third-generation long-read sequencing in combination with short-read sequencing has improved the quality of genome assembly for *Cryptosporidium* spp. Using this approach, we have provided a gapless reference with improved gene prediction and annotation for the IIaA15G2R1 subtype, which is the dominant subtype of *C. parvum* (Feng et al., 2018). We have resolved 13 of the 16 telomeres in the present assembly. The completeness of the new genome assembly is similar to that of the two recently reported IOWA genomes. Since the three missing chromosomal ends in the IOWA-CDC genome are also the only three missing chromosomal ends in the IOWA-ATCC and IOWA-KWI52 genomes (Baptista et al., 2021; Menon et al., 2022), there may be intrinsic issues that have prevented the acquisition of these chromosomal ends during genome sequencing. These chromosome-level genomes facilitate studies of copy number variation and sequence divergence of subtelomeric genes, which could contribute to the biological differences between *Cryptosporidium* species (Xu et al., 2019b). Therefore, the resolution of telomeres and subtelomere regions across chromosomes may lead to a better understanding of the relationship between biological traits and genomic features. There are minor structural differences among the three recently sequenced IOWA genomes. Because they all contain chromosomal ends with the large rRNA unit, these differences are likely due to differences in the approaches used to resolve sequence gaps generated by large repetitive sequences.

C. parvum IOWA lines maintained in different laboratories have divergent genomes and evolutionary histories. Over the years, at least seven lines known as the IOWA strain and maintained at five facilities have been used in research. Three of them, have been sequenced and fully assembled, designated as IOWA-ATCC, IOWA-KWI52, and IOWA-II (Abrahamsen et al., 2004; Audebert et al., 2020; Baptista et al., 2021; Menon et al., 2022). Based on the *gp60* genes, these IOWA lines exhibit two subtypes with SNPs in the non-repeat region of the *gp60* gene. As shown in the present study, recent IOWA lines of the IIaA17G2R1 subtype (IOWA-ATCC, IOWA-Waterborne, IOWA-Bunchgrass, and IOWA-KWI52) have nearly identical genomes, suggesting that they may share the same origin. However, the genomes of the two IOWA lines of the IIaA15G2R1 subtype (IOWA-II and IOWA-CDC) show heterogeneity. In addition, copy number variations of subtelomeric genes are detected among these genomes. These data suggest that these IOWA lines have heterogeneous genomes of different origins. These results are consistent with previous multilocus characterizations of IOWA lines of subtype IIaA15G2R1 maintained in different laboratories (Cama et al., 2006).

The IOWA-IIaA17G2R1 lines appear to be derived from the original IOWA strain (IIaA15G2R1 subtype), but have undergone genetic recombination with unknown strains. In the present study, IOWA-IIaA17G2R1 lines from the University of Arizona, Waterborne, and Bunch Grass Farm show high genomic similarity to IOWA-CDC in most genomic regions. However, gene flow from another IIaA15G2R1 line (IOWA-II) into these IOWA-IIaA17G2R1 genomes was observed, for example on chromosome 8. A previous multilocus sequence type analysis of IOWA lines kept in different laboratories suggests that IOWA-CDC may represent the original IOWA isolate, as multiple passages of this line produced sequences identical to those obtained from IOWA DNA achieved in 1989 (Cama et al., 2006). The other IOWA lines, however, go by different names but are either produced by Bunch Grass Farm or produced in-house from oocysts originally provided by Bunch Grass Farm. IOWA-II was generated by the Pleasant Hill Farm (Abrahamsen et al., 2004), which was subsequently renamed as Bunch Grass Farm (Cama et al., 2006). Therefore, these data suggest that IOWA-IIaA17G2R1 was generated from the original IOWA strain, but it has undergone recombination with IOWA-II. In addition, we observed gene flow from unsampled lines into the IOWA-IIaA17G2R1 genomes, and we showed that the unique sequence features of the *gp60* and CPCDC_8g661 genes and the absence of the CPCDC_5g5511 gene in IOWA-IIaA17G2R1 lines are not seen in any of the other isolates analyzed in this study. These data suggest that a unique exogenous strain was introduced and recombined with the original IOWA strain to generate the IOWA-II, which evolved further as IOWA-IIaA17G2R1 with more sequence introgression and backcrossing.

The presence of two nucleotide types at most polymorphic loci within the IOWA-IIaA17G2R1 lines supports the existence of exogenous contamination. It also suggests that the IOWA-IIaA17G2R1 lines contain mixed *C. parvum* genomes, which in turn facilitates that the occurrence of genetic recombination. These findings are not surprising given that the calf infection model is used to propagate and maintain IOWA, and *C. parvum* infection of newborn calves is very common in the United States. Therefore, there is a significant risk of exogenous parasite contamination of the IOWA isolate during calf passage. This finding is consistent with the recent conclusion that genetic recombination plays an important role in shaping the genetic and biological characteristics of *Cryptosporidium* isolates (Nader et al., 2019; Corsi et al., 2022; Wang et al., 2022; Huang et al., 2023). The occurrence of mixed *C. parvum* populations in IOWA-IIaA17G2R1 lines also suggests that these isolates may be biologically and genetically unstable due to the likely occurrence of additional genetic recombination and natural selection of the progeny.

Copy number variation and sequence polymorphism of invasion-associated genes may underlie the differences in virulence between different *C. parvum* IOWA lines. Data from several recent studies indicate that the IOWA-IIaA17G2R1 lines derived from Waterborne and Bunchgrass differ from the IOWA-CDC line. While IOWA-CDC is highly virulent in GKO and other immunocompromised mice, IOWA-IIaA17G2R1 lines produce only a mild infection without clinical signs in these animals (Ndao et al., 2013; Sonzogni-Desautels et al., 2015; He et al., 2022; Jia et al., 2022). Compared to the IOWA-IIaA17G2R1 genomes, IOWA-CDC had an additional subtelomeric insulinase gene (CPCDC_5g5511) and significant sequence differences in genes encoding mucin-like proteins, such as cgd2_420, cgd2_450, and cgd6_1080 (*gp60*). Insulinase genes have been implicated in *C. parvum* invasion and development of (Xu et al., 2019a; Zhang et al., 2019; Cui et al., 2022), while mucin glycoproteins play a critical role in sporozoite invasion (Ludington and Ward 2016). In addition, the CPCDC_8g661 gene in IOWA-IIaA17G2R1 lines has a stop codon that truncates the coding sequence. Therefore, these gene gains and sequence polymorphisms may contribute to the differences in virulence among *C. parvum* isolates.

5. Conclusions

We have generated a complete assembly with improved gene prediction and annotation as a new reference genome of the *C. parvum* IOWA strain. The data indicate that the IOWA lines maintained in different laboratories have different genomes. The commercial IOWA-IIaA17G2R1 lines all contain mixed populations of *C. parvum* and have sequence polymorphisms and gene deletions that are not observed in field isolates of *C. parvum*. These unique genomic characteristics appear to have resulted from contamination of the original IOWA isolate with an exogenous strain during passages in calves. The presence of mixed parasite populations in IOWA-IIaA17G2R1 lines underscores their use in regulatory testing of *Cryptosporidium* oocysts in water samples. The polymorphic genetic loci identified in this study may facilitate the development of advanced molecular tools for precisely tracking contamination sources of *Cryptosporidium* oocysts in both source and drinking water. The virulence-associated genes identified in this study should be further investigated using advanced biological and genetic tools.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (32030109, 32150710530, U1901208), the Guangdong Major Project of Basic and Applied Basic Research (2020B0301030007), 111 Project (D20008), Double First-Class Discipline Promotion Project (2023B10564003), and the Advanced Molecular Detection Program of 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 official position of the Centers for Disease Control and Prevention. We thank Prof. Guan Zhu for providing IOWA-Bunchgrass oocysts.

Data availability

The sequence data and annotations generated in this study were submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA446067. The final genome assembly of IOWA-CDC, along with annotations, was submitted to GenBank in March 2018 under accession numbers CP029780-CP029787. To maintain the objectivity of the work, no changes were made to the genome assembly and annotations (except for the recent change in the locus_tag prefix requested by NCBI staff) after the publication of the recent IOWA-IIaA17G2R1 genomes. Sequences of rRNA units generated from this study were submitted to GenBank under accession numbers OR419798, OR419799, OR419800, OR421304, and OR421305.

References

Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, et al. , 2004. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. Science 304, 441–445. 10.1126/science.1094786. [PubMed: 15044751]

Antipov D, Korobeynikov A, McLean JS, Pevzner PA, 2016. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics 32, 1009–1015. 10.1093/bioinformatics/btv688. [PubMed: 26589280]

Audebert C, Bonardi F, Caboche S, Guyot K, Touzet H, Merlin S, Gantois N, Creusy C, Meloni D, Mouray A, et al. , 2020. Genetic basis for virulence differences of various *Cryptosporidium parvum* carcinogenic isolates. Sci. Rep. 10, 7316. 10.1038/s41598-020-64370-0. [PubMed: 32355272]

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS, 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37, W202–W208. 10.1093/nar/gkp335. [PubMed: 19458158]

Baptista RP, Li Y, Sateriale A, Brooks KL, Tracey A, Sanders MJ, Ansell BRE, Jex AR, Cooper GW, Smith ED, et al. , 2021. Long-read assembly and comparative evidence-based reanalysis of *Cryptosporidium* genome sequences reveals expanded transporter repertoire and duplication of entire chromosome ends including subtelomeric regions. Genome Res. 32, 203–213. 10.1101/gr.275325.121. [PubMed: 34764149]

Cama VA, Arrowood MJ, Ortega YR, Xiao L, 2006. Molecular characterization of the *Cryptosporidium parvum* IOWA isolate kept in different laboratories. J. Eukaryot Microbiol. 53 (Suppl 1), S40–S42. 10.1111/j.1550-7408.2006.00168.x. [PubMed: 17169063]

Checkley W, White AC Jr., Jaganath D, Arrowood MJ, Chalmers RM, Chen XM, Fayer R, Griffiths JK, Guerrant RL, Hedstrom L, et al. , 2015. A review of the global burden, novel diagnostics, therapeutics, and vaccine targets for *Cryptosporidium*. Lancet Infect. Dis. 15, 85–94. 10.1016/S1473-3099(14)70772-8. [PubMed: 25278220]

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. , 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods 10, 563–569. 10.1038/nmeth.2474. [PubMed: 23644548]

Corsi GI, Tichkule S, Sannella AR, Vatta P, Asnicar F, Segata N, Jex AR, van Oosterhout C, Caccio SM, 2022. Recent genetic exchanges and admixture shape the genome and population structure of the zoonotic pathogen *Cryptosporidium parvum*. Mol. Ecol. 0, 1–13. 10.1111/mec.16556.

Cui H, Xu R, Li Y, Guo Y, Zhang Z, Xiao L, Feng Y, Li N, 2022. Characterization of dense granule metalloproteinase INS-16 in *Cryptosporidium parvum*. Int. J. Mol. Sci. 23, 7617. 10.3390/ijms23147617. [PubMed: 35886965]

Dumaine JE, Tandel J, Striepen B, 2020. Cryptosporidium parvum. Trends Parasitol. 36, 485–486. 10.1016/j.pt.2019.11.003. [PubMed: 31836286]

Efstratiou A, Ongerth JE, Karanis P, 2017. Waterborne transmission of protozoan parasites: review of worldwide outbreaks - An update 2011–2016. Water Res. 114, 14–22. 10.1016/j.watres.2017.01.036. [PubMed: 28214721]

Fan Y, Feng Y, Xiao L, 2019. Comparative genomics: how has it advanced our knowledge of cryptosporidiosis epidemiology? Parasitol. Res. 118, 3195–3204. 10.1007/s00436-019-06537-x. [PubMed: 31724068]

Feng Y, Li N, Roellig DM, Kelley A, Liu G, Amer S, Tang K, Zhang L, Xiao L, 2017. Comparative genomic analysis of the IId subtype family of *Cryptosporidium parvum*. Int. J. Parasitol. 47, 281–290. 10.1016/j.ijpara.2016.12.002. [PubMed: 28192123]

Feng Y, Ryan UM, Xiao L, 2018. Genetic diversity and population structure of *Cryptosporidium*. Trends Parasitol. 34, 997–1011. 10.1016/j.pt.2018.07.009. [PubMed: 30108020]

Guo Y, Tang K, Rowe LA, Li N, Roellig DM, Knipe K, Frace M, Yang C, Feng Y, Xiao L, 2015. Comparative genomic analysis reveals occurrence of genetic recombination in virulent *Cryptosporidium hominis* subtypes and telomeric gene duplications in *Cryptosporidium parvum*. BMC Genomics 16, 320. 10.1186/s12864-015-1517-1. [PubMed: 25903370]

Hadfield SJ, Pachebat JA, Swain MT, Robinson G, Cameron SJ, Alexander J, Hegarty MJ, Elwin K, Chalmers RM, 2015. Generation of whole genome sequences of new *Cryptosporidium hominis* and *Cryptosporidium parvum* isolates directly from stool samples. BMC Genomics 16, 650. 10.1186/s12864-015-1805-9. [PubMed: 26318339]

He X, Huang W, Sun L, Hou T, Wan Z, Li N, Guo Y, Kvac M, Xiao L, Feng Y, 2022. A productive immunocompetent mouse model of cryptosporidiosis with long oocyst shedding duration for immunological studies. J. Infect. 84, 710–721. 10.1016/j.jinf.2022.02.019. [PubMed: 35192895]

Huang W, Guo Y, Lysen C, Wang Y, Tang K, Seabolt MH, Yang F, Cebelinski E, Gonzalez-Moreno O, Hou T, et al. , 2023. Multiple introductions and recombination events underlie the emergence of a hyper-transmissible *Cryptosporidium hominis* subtype in the USA. Cell Host Microbe 31, 112–123. 10.1016/j.chom.2022.11.013 e114. [PubMed: 36521488]

Huson DH, Bryant D, 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267. 10.1093/molbev/msj030. [PubMed: 16221896]

Isaza JP, Galván AL, Polanco V, Huang B, Matveyev AV, Serrano MG, Manque P, Buck GA, Alzate JF, 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. 10.1038/srep16324. [PubMed: 26549794]

Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, et al. , 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res. 27, 768–777. 10.1101/gr.214346.116. [PubMed: 28232478]

Jia R, Huang W, Huang N, Yu Z, Li N, Xiao L, Feng Y, Guo Y, 2022. High infectivity and unique genomic sequence characteristics of *Cryptosporidium parvum* in China. PLoS Negl. Trop. Dis. 16, e0010714 10.1371/journal.pntd.0010714. [PubMed: 35994488]

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A, 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35, 4453–4455. 10.1093/bioinformatics/btz305. [PubMed: 31070718]

Le Blancq SM, Khramtsov NV, Zamani F, Upton SJ, Wu TW, 1997. Ribosomal RNA gene organization in *Cryptosporidium parvum*. Mol. Biochem. Parasitol. 90, 463–478. 10.1016/s0166-6851(97)00181-3. [PubMed: 9476794]

Li H, Durbin R, 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. 10.1093/bioinformatics/btp324. [PubMed: 19451168]

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M, 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33, 6494–6506. 10.1093/nar/gki937. [PubMed: 16314312]

Ludington JG, Ward HD, 2016. The *Cryptosporidium parvum* c-type lectin CpClec mediates infection of intestinal epithelial cells via interactions with sulfated proteoglycans. Infect. Immun. 84, 1593–1602. 10.1128/iai.01410-15. [PubMed: 26975991]

Menon VK, Okhuysen PC, Chappell CL, Mahmoud M, Mahmoud M, Meng Q, Doddapaneni H, Vee V, Han Y, Salvi S, et al. , 2022. Fully resolved assembly of *Cryptosporidium parvum*. Gigascience 11, 1–8. 10.1093/gigascience/giac010.

Moon HW, Bemrick WJ, 1981. Fecal transmission of calf cryptosporidia between calves and pigs. Vet. Pathol. 18, 248–255. 10.1177/030098588101800213. [PubMed: 7467084]

Nader JL, Mathers TC, Ward BJ, Pachebat JA, Swain MT, Robinson G, Chalmers RM, Hunter PR, van Oosterhout C, Tyler KM, 2019. Evolutionary genomics of anthroponosis in *Cryptosporidium*. Nat. Microbiol. 4, 826–836. 10.1038/s41564-019-0377-x. [PubMed: 30833731]

Nash JHE, Robertson J, Elwin K, Chalmers RA, Kropinski AM, Guy RA, 2018. Draft genome assembly of a potentially zoonotic *Cryptosporidium parvum* isolate, UKP1. Microbiol. Resour. Announc. 7 10.1128/MRA.01291-18 e01291–01218.

Ndao M, Nath-Chowdhury M, Sajid M, Marcus V, Mashiyama ST, Sakanari J, Chow E, Mackey Z, Land KM, Jacobson MP, et al. , 2013. A cysteine protease inhibitor rescues mice from a lethal *Cryptosporidium parvum* infection. Antimicrob Agents Chemother 57, 6063–6073. 10.1128/AAC.00734-13. [PubMed: 24060869]

Otto TD, Dillon GP, Degrave WS, Berriman M, 2011. RATT: rapid annotation transfer tool. Nucleic Acids Res. 39, e57. 10.1093/nar/gkq1268. [PubMed: 21306991]

Posada D, 2008. jModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25, 1253–1256. 10.1093/molbev/msn083. [PubMed: 18397919]

Ryan UM, Feng Y, Fayer R, Xiao L, 2021. Taxonomy and molecular epidemiology of *Cryptosporidium* and *Giardia* - a 50 year perspective (1971–2021). Int. J. Parasitol. 51, 1099–1119. 10.1016/j.ijpara.2021.08.007. [PubMed: 34715087]

Sonzogni-Desautels K, Renteria AE, Camargo FV, Di Lenardo TZ, Mikhail A, Arrowood MJ, Fortin A, Ndao M, 2015. Oleylphosphocholine (OlPC) arrests *Cryptosporidium parvum* growth in vitro and prevents lethal infection in interferon gamma receptor knock-out mice. Front. Microbiol. 6, 973. 10.3389/fmicb.2015.00973. [PubMed: 26441906]

Troell K, Hallström B, Divne AM, Alsmark C, Arrighi R, Huss M, Beser J, Bertilsson S, 2016. *Cryptosporidium* as a testbed for single cell genome characterization of unicellular eukaryotes. BMC Genomics 17, 471. 10.1186/s12864-016-2815-y. [PubMed: 27338614]

Wang T, Guo Y, Roellig DM, Li N, Santin M, Lombard J, Kvac M, Naguib D, Zhang Z, Feng Y, et al. , 2022. Sympatric recombination in zoonotic *Cryptosporidium* leads to emergence of populations with modified host preference. Mol. Biol. Evol. 39, msac150. 10.1093/molbev/msac150. [PubMed: 35776423]

Xiao L, Feng Y, 2017. Molecular epidemiologic tools for waterborne pathogens *Cryptosporidium* spp. and Giardia duodenalis. Food Waterborne Parasitol. 8–9, 14–32. 10.1016/j.fawpar.2017.09.002.

Xu R, Guo Y, Li N, Zhang Q, Wu H, Ryan U, Feng Y, Xiao L, 2019a. Characterization of INS-15, A metalloprotease potentially involved in the invasion of *Cryptosporidium parvum*. Microorganisms 7, 452. 10.3390/microorganisms7100452. [PubMed: 31615118]

Xu Z, Guo Y, Roellig DM, Feng Y, Xiao L, 2019b. Comparative analysis reveals conservation in genome organization among intestinal *Cryptosporidium* species and sequence divergence in potential secreted pathogenesis determinants among major human-infecting species. BMC Genomics 20, 406. 10.1186/s12864-019-5788-9. [PubMed: 31117941]

Zhang H, Zhu G, 2020. High-throughput screening of drugs against the growth of *Cryptosporidium parvum* in vitro by qRT-PCR. Methods Mol. Biol. 2052, 319–334. 10.1007/978-1-4939-9748-0_18. [PubMed: 31452170]

Zhang S, Wang Y, Wu H, Li N, Jiang J, Guo Y, Feng Y, Xiao L, 2019. Characterization of a species-specific insulinase-like protease in *Cryptosporidium parvum*. Front. Microbiol. 10, 354. 10.3389/fmicb.2019.00354. [PubMed: 30894838]

Office of Water, U.S. Environmental Protection Agency 2012. Method 1623.1: *Cryptosporidium* and Giardia in water by filtration/IMS/FA EPA 816-R-12–001.

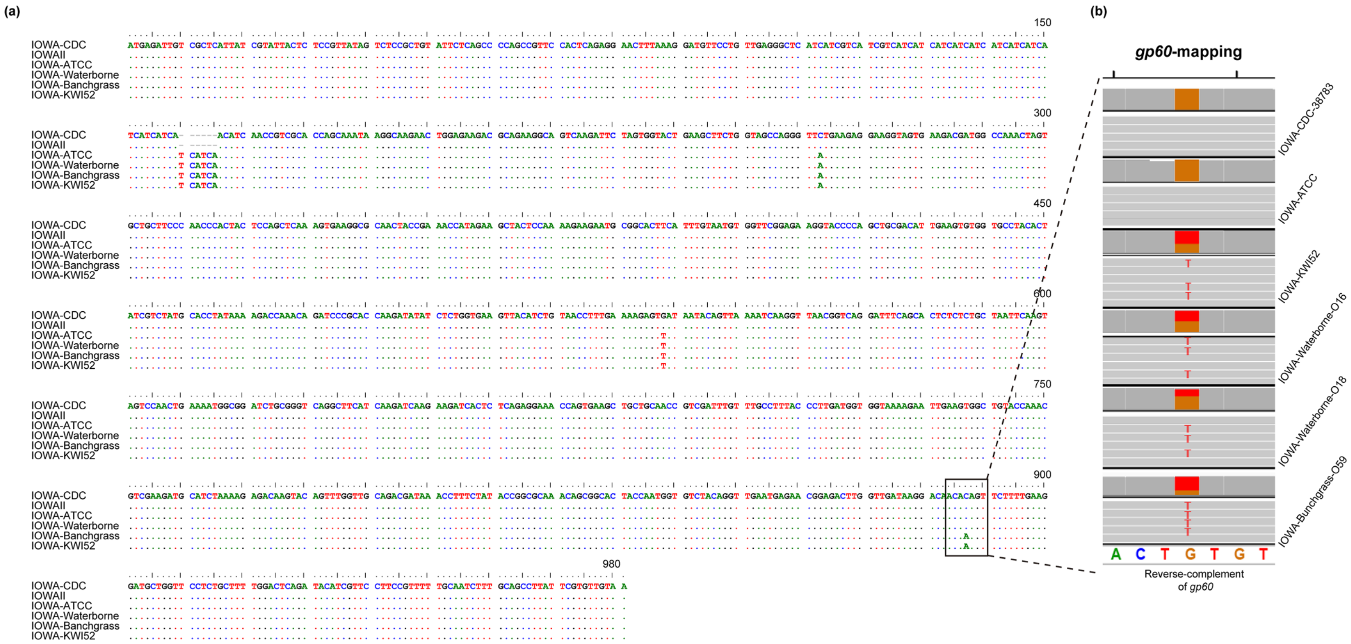

Fig. 1.

Differences in *gp60* gene sequences between IOWA lines maintained in different laboratories. (a) Aligned *gp60* sequences of IOWA lines, with dots representing nucleotides identical to those in the IOWA-CDC reference. (b) Confirmation of the nucleotide substitution at nucleotide 887 of the *gp60* gene using read mapping. Vertical bars indicate the percentage of nucleotides at the locus.

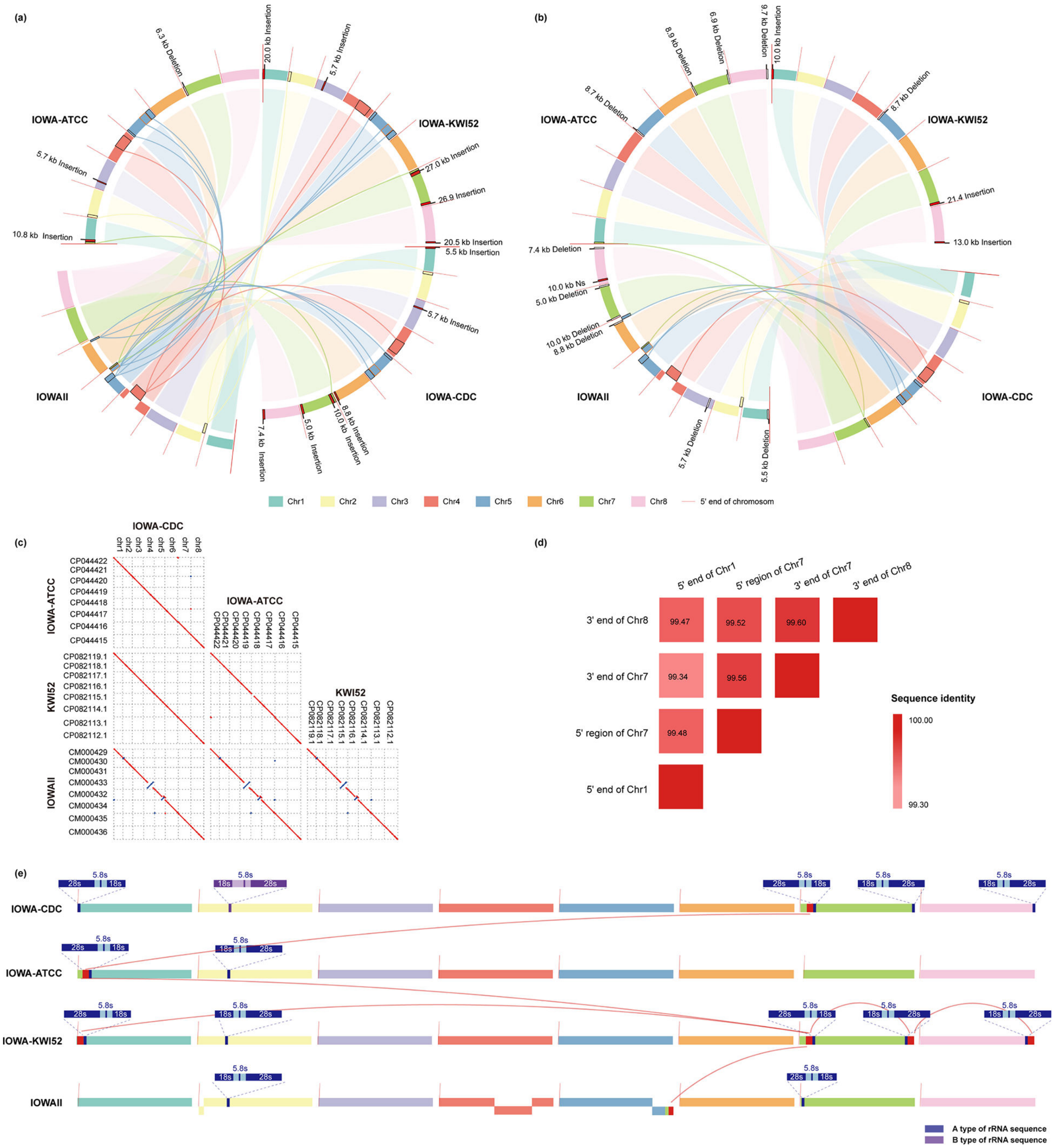

Fig. 2.

Comparisons of chromosome-level assemblies from four IOWA lines. (a,b) Syntenic relationship in genomic structure between the genomes of IOWA-CDC, IOWA-II, IOWA-ATCC, and IOWA-KWI52 using the IOWA-II (a) or the IOWA-CDC (b) as reference. Colors represent eight chromosomes of *C. parvum*. Sequence rearrangements are connected by lines. Inversions and deletions are shown as red and white blocks, respectively. (c) Dot plots showing the collinear relationship between the IOWA-CDC, IOWA-ATCC, and IOWA-KWI52 genome assemblies. Clear assembly differences are visible between the IOWA-II assembly and the others. (d) Pairwise comparisons of sequence identity between several large insertions and the 5′ region of chromosome 7 in the IOWA-KWI52 assembly. (e) Relationship of the insertions in the IOWA-KWI52 genome to the 5′ regions of chromosome 7 in other fully assembled genomes. This figure also shows the rRNA units in these genomes.

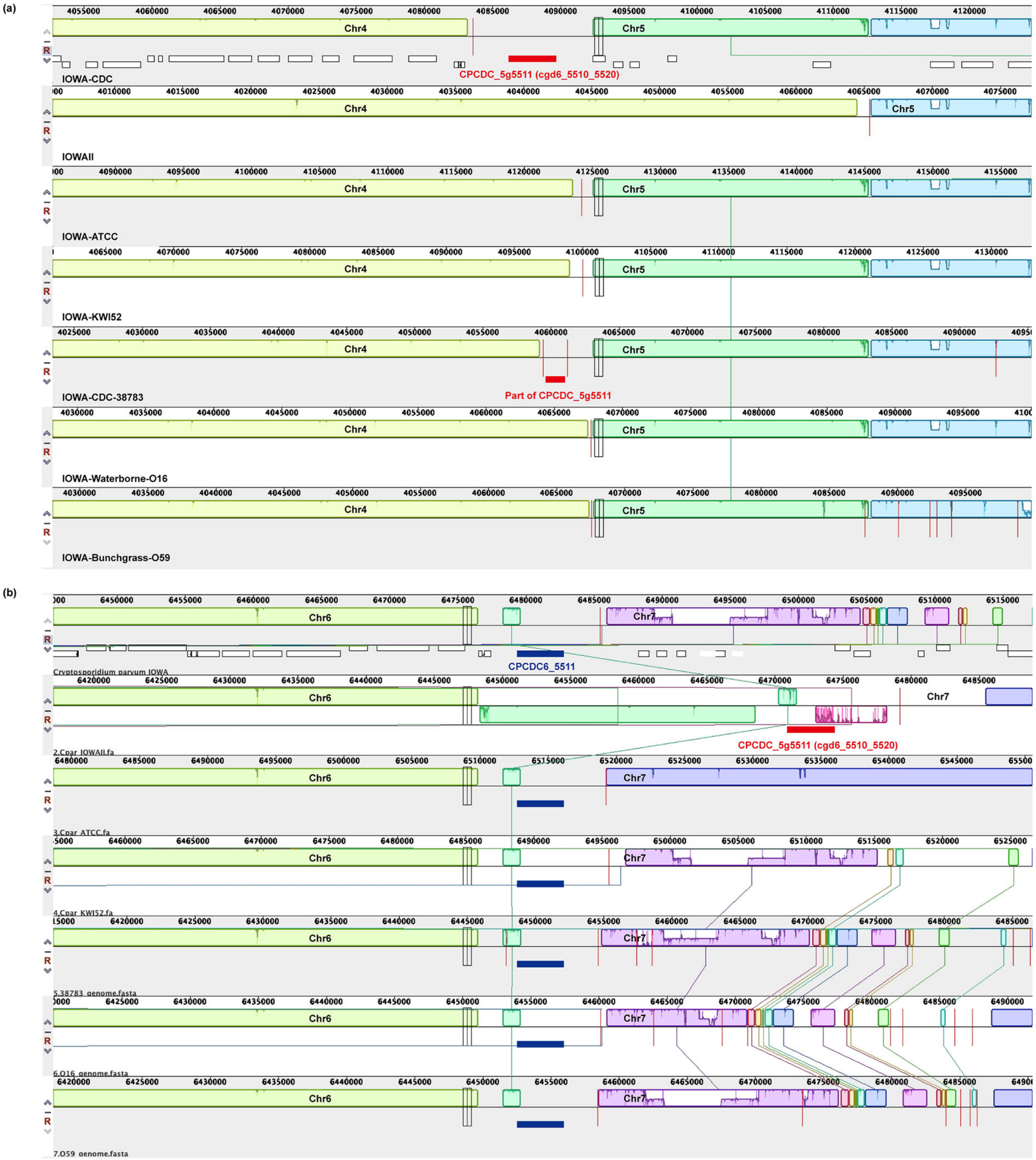

Fig. 3.

Deletion of genes between genomes of IOWA lines maintained in different laboratories in comparison with IOWA-CDC. The colored blocks (known as locally collinear blocks) are conserved segments of sequence that are internally free from genome rearrangements, while the inverted white peaks within each block are sequence divergence between the reference genome and other genomes. Assembled chromosomes or contigs are indicated by red vertical lines. (a) Deletions at the 5′ end of chromosome 5 compared with the IOWA-CDC (indicated by small red blocks). (b) Rearrangements and deletions at the 3′ end of chromosome 6 compared to the IOWA-CDC (indicated by colored blocks and small blue boxes, respectively).

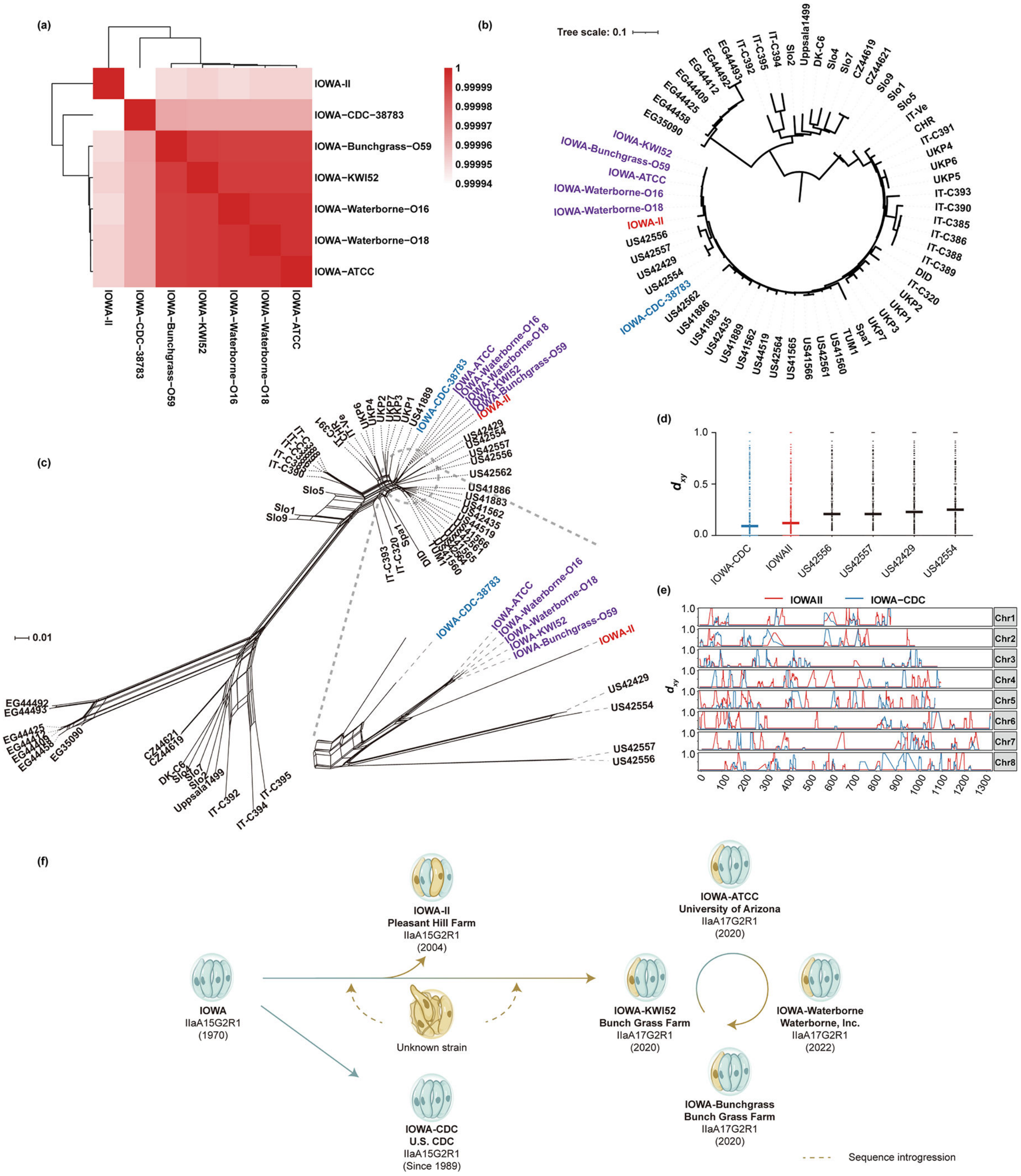

Fig. 4.

Population substructures of IOWA lines and origin of *C. parvum* IOWA IIaA17G2R1. (a) Pairwise comparisons of the average nucleotide identity of genomes from IOWA lines. (b) Phylogenetic relationships among *C. parvum* genomes inferred by ML using 14,230 wgSNPs. (c) Phylogenetic network of 63 isolates based on wgSNPs. The parallel edges in the network indicates gene flow between isolates. (d) Absolute divergence (*dxy*) values between IOWA IIaA17G2R1 and others in a 10-kb sliding window. (e) Results of *dxy* comparisons between IOWA IIaA17G2R1 and IOWA-II or IOWA-CDC across eight chromosomes in a 10-kb sliding window. (f) Summary of the likely evolutionary history of IOWA lines maintained in different laboratories. The formation of different IOWA lines is indicated by the solid arrow, while the putative genetic recombination is indicated by the dashed arrow.

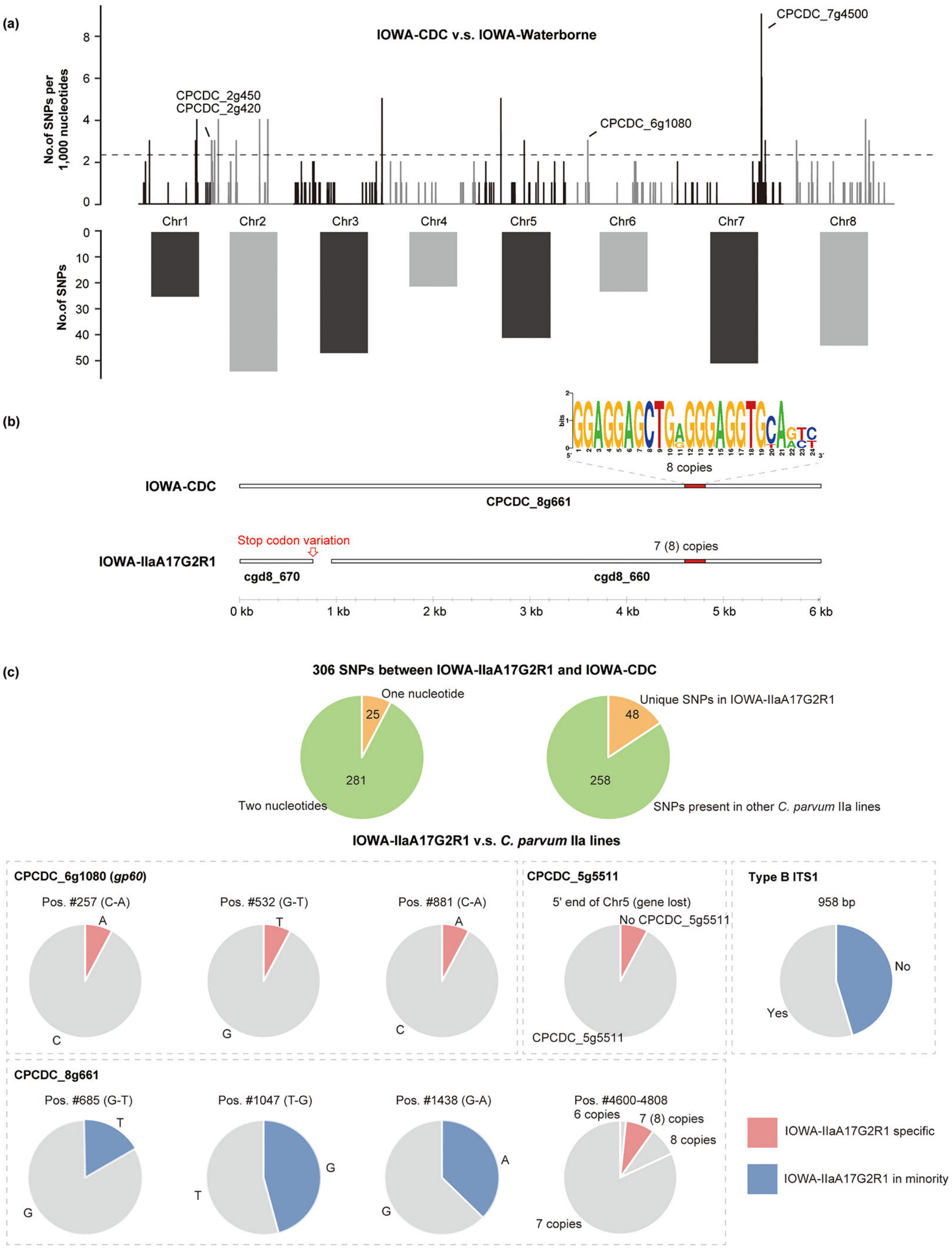

Fig. 5.

Summary of major genomic differences between IOWA-CDC and IOWA-IIaA17G2R1 lines. (a) Distribution of single nucleotide variants in the IOWA-IIaA17G2R1 genomes compared to the IOWA-CDC genome. (b) Sequence characteristics of the CPCDC_8g661 gene and structural differences of the gene between IOWA-CDC and IOWA-IIaA17G2R1. (c) Genetic uniqueness of the IOWA-IIaA17G2R1 lines. Among the 306 SNPs between IOWA-IIaA17G2R1 and IOWA-CDC, 281 sites have two nucleotides in the genomes of IOWA-IIaA17G2R1 lines, and there are 48 IOWA-IIaA17G2R1-specific sites that are absent in other *C. parvum* IIa lines. Among the nine highly polymorphic protein-coding genes shown in Table 1, the *gp60* gene of the IOWA-IIaA17G2R1 lines has three SNPs in the non-repeat region, which were not found in any other isolates analyzed in the study. Only IOWA-IIaA17G2R1 lines lack the CPCDC_5g5511 gene and have both types of CPCDC_8g661 gene sequences with and without the 24 bp deletion. Furthermore, the absence of type B ITS1 sequences and the presence of three SNPs in the CPCDC_8g661 gene in IOWA-IIaA17G2R1 lines, including the stop codon variation (position #685), are rarely found in other *C. parvum* IIa isolates.

Table 1

Highly polymorphic genes between IOWA-CDC and IOWA-Waterborne genomes.

Chromosomes	Gene ID in IOWA- CDC	Ortholog gene in IOWA- II	No. of SNPs	Non synonymous	Synonymous	Gene length	Annotation
2	CPCDC_2g450	cgd2_450	2	2	0	549	Mucin-like protein
	CPCDC_2g420	cgd2_420	2	2	0	618	Mucin-like protein
	CPCDC_2g2780	cgd2_2780	4	2	2	1122	Dcd1p-like dCMP deaminase
3	CPCDC_3g4270	cgd3_4270	9	8	1	3531	Insulinase-like peptidase
5	CPCDC_5g2550	cgd5_2550	3	0	3	690	Polyubiquitin
6	CPCDC_6g1080	cgd6_1080	3	3	0	975	GP60 (mucin)
7	CPCDC_7g4500	cgd7_4500	7	6	1	2493	Signal peptide-containing secreted protein
	CPCDC_7g4680	cgd7_4680	3	2	1	1167	Hypothetical protein
8	CPCDC_8g4020	cgd8_4020	6	1	5	2352	Hypothetical protein
	CPCDC_8g4181	cgd8_4181	6	4	2	2433	Hypothetical protein