Current Research in Microbial Sciences 3 (2022) 100114

Contents lists available at ScienceDirect

Current Research in Microbial Sciences

journal homepage: www.sciencedirect.com/journal/current-research-in-microbial-sciences

A hybrid sequencing and assembly strategy for generating culture free *Giardia* genomes

Jenny G. Maloney [a], Aleksey Molokin [a], Gloria Solano-Aguilar [b], Jitender P. Dubey [c], Monica Santin [a,*]

[a] Environmental Microbial and Food Safety Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD 20705, USA
[b] Diet, Genomics and Immunology Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD 20705, USA
[c] Animal Parasitic Diseases Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD 20705, USA

ARTICLE INFO

Keywords:
Assemblage
Assembly
Genome
Giardia
Illumina MiSeq
Long-read sequencing
MinION

ABSTRACT

Giardia duodenalis is a pathogenic intestinal protozoan parasite of humans and many other animals. *Giardia duodenalis* is found throughout the world, and infection is known to have adverse health consequences for human and other mammalian hosts. Yet, many aspects of the biology of this ubiquitous parasite remain unresolved. Whole genome sequencing and comparative genomics can provide insight into the biology of *G. duodenalis* by helping to reveal traits that are shared by all *G. duodenalis* assemblages or unique to an individual assemblage or strain. However, these types of analyses are currently hindered by the lack of available *G. duodenalis* genomes, due, in part, to the difficulty in obtaining the genetic material needed to perform whole genome sequencing. In this study, a novel approach using a multistep cleaning procedure coupled with a hybrid sequencing and assembly strategy was assessed for use in producing high quality *G. duodenalis* genomes directly from cysts obtained from feces of two naturally infected hosts, a cat and dog infected with assemblage A and D, respectively. Cysts were cleaned and concentrated using cesium chloride gradient centrifugation followed by immunomagnetic separation. Whole genome sequencing was performed using both Illumina MiSeq and Oxford Nanopore MinION platforms. A hybrid assembly strategy was found to produce higher quality genomes than assemblies from either platform alone. The hybrid *G. duodenalis* genomes obtained from fecal isolates (cysts) in this study compare favorably for quality and completeness against reference genomes of *G. duodenalis* from cultured isolates. The whole genome assembly for assemblage D is the most contiguous genome available for this assemblage and is an important reference genome for future comparative studies. The data presented here support a hybrid sequencing and assembly strategy as a suitable method to produce whole genome sequences from DNA obtained from *G. duodenalis* cysts which can be used to produce novel reference genomes necessary to perform comparative genomics studies of this parasite.

1. Introduction

Giardia duodenalis (syn. *G. intestinalis, G. lamblia*) is a protozoan parasite which infects the intestinal tract of humans and a broad range of other mammals (Feng and Xiao, 2011). Infection with *G. duodenalis* is one of the most common causes of diarrhea in humans worldwide, and it has been estimated to cause over 180 million cases a year (Torgerson et al., 2015). Infection is usually self-limiting, causing diarrhea which can be severe, and chronic reinfection is considered a cause of failure to thrive in children from endemic areas (Allain and Buret, 2020). *Giardia* infection can have a lasting influence on host health as post-infection follow-up studies have found giardiasis to be a risk factor for the development of gastrointestinal disorders such as irritable bowel syndrome (Hanevik et al., 2014). Asymptomatic infection is also common in both humans and animal hosts, and undiagnosed infection may contribute to the spread of this parasite. *Giardia* is spread via the fecal-oral route and is transmitted through contact with infected hosts or ingestion of contaminated food or water (Dixon, 2021).

Giardia has two morphological stages, trophozoite and cyst. Trophozoites have two identical nuclei (binucleated), while fully differentiated cysts contain four nuclei (quadrinucleated) (Bernander et al., 2001). Each nucleus contains a diploid set of 5 chromosomes (Morrison et al., 2007). The trophozoite is the replicative stage of the parasite and survives only in the host intestines. The cyst, which is excreted in feces,

* Corresponding author.

https://doi.org/10.1016/j.crmicr.2022.100114
Received 5 January 2022; Received in revised form 11 February 2022; Accepted 14 February 2022
Available online 16 February 2022

is environmentally resistant and can survive for months in cold and wet conditions. *Giardia* also has a low infectious dose, and as few as 10 cysts have been shown to be sufficient for establishment of infection in humans (Rendtorff, 1954).

The taxonomy of *Giardia* remains a topic of debate. The genus is currently divided into eight species based to some extent on host specificity in a diverse group of hosts which include birds, amphibians, rodents, marsupials, and other mammals (Lyu et al., 2018). However, *G. duodenalis* is the only species that has been found in humans. *Giardia duodenalis* is considered a species complex and is further divided into eight assemblages named A through H based both on genetic differences and host specificity (Adam, 2021). Assemblages A and B are both found in humans as well as other mammals. Assemblages C and D are found in both wild and domestic canids, assemblage E in hoofed animals, assemblage F in cats, assemblage G in rodents, and assemblage H in marine animals. Genetic differences within and among assemblages are thought to play a role in determining differences in host specificity and pathogenicity, however, there remains a deficiency in the data needed to define these differences (Haque et al., 2009; Cai et al., 2021; Messa et al., 2021).

Whole genome sequencing has the potential to clarify many unresolved aspects of *Giardia's* biology and epidemiology. However, attaining high quality reference genomes has proven difficult, and the majority of published genomes describing *Giardia* whole genome sequences are from assemblages A and B (Table 1). This disparity is due in large part to the lack of cultured isolates for non-G. *duodenalis* species or *G. duodenalis* assemblages other than A and B, because attaining trophozoite cultures for other assemblages has proven difficult to impossible. Thus, data for assemblages C, D, and E are still limited, and to date, no whole genome sequences of assemblages F, G, or H have been published. The single assemblage E genome which has been published is the result of the only successful culture of an assemblage E isolate (P15) originally isolated from a pig (Jerlström-Hultqvist et al., 2010). Although three primary isolates of *G. duodenalis* from human feces representing two assemblage A isolates and one assemblage B isolate have been sequenced, they were not used to produce whole genome assemblies (Hanevik et al., 2015). Primary isolates of assemblages C and D have been sequenced from individual cysts or pooled DNA of 40 cysts from dog feces and used to produce whole genome assemblies which appear complete as compared to reference genomes but are highly fragmented (Kooyman et al., 2019). To date, no other reports of whole genome sequences for primary isolates have been published.

The lack of data on primary isolates from both multiple assemblages and multiple hosts impedes our ability to utilize whole genome sequences to understand genetic diversity as well as the host specificity and pathogenicity of different assemblages and strains of *Giardia*. Such data could also have important uses in identifying new genetic markers to improve detection, source tracking, and drug resistant strain identification (Capewell et al., 2021). Genetic difference between assemblages or strains could also have uses in clinical applications such as drug target determination and vaccine design. Whole genome sequencing of important food borne bacterial pathogens including *Escherichia coli, Listeria*, and *Salmonella* has produced significant health and economic benefits, however sequences from thousands of isolates are needed before such benefits can be attained (Brown et al., 2021). Clearly, methods to attain reference quality genomes from primary isolates are needed to bring the benefits of whole genome sequencing to the field of *Giardia*.

Whole genome sequencing strategies which employ long read sequencing platforms either alone or in combination with short reads have been demonstrated to produce superior genomes for virus, bacteria, fungi, and parasites (Díaz-Viraqué et al., 2019; Moolhuijzen et al., 2021; O'Donnell et al., 2020; Todd et al., 2018; Wick et al., 2017). A hybrid approach using short and long reads to produce *Giardia* genomes using isolates maintained in culture has also been recently used (Pollo et al., 2020; Xu et al., 2020a, b). However, the application of long read

Table 1
Summary of available *Giardia* spp. genomes including details of type of isolate (feces/cysts or culture/trophozoites), host, and sequencing platform. In bold are genomes obtained from DNA extracted from cysts.

Year	Assemblages sequenced (isolate ID)	Type of isolates	Host	Sequencing platform	Citations
2020	A1 (WB)[1] A (Beaver) B (GS)[2]	Culture	Human Beaver Human	Oxford Nanopore and Illumina	Pollo et al. 2020
2020	***G. muris* (Roberts-Thomson)**[3]	**Feces**	Mice	PacBio and Illumina	Xu et al. 2020b
2020	A1 (WB/C6)[4]	Culture	Human	PacBio and Illumina	Xu et al. 2020a
2019	**C (dog1/cyste1)**[5] **C (dog1/cyste3)**[5] **C (dog8/pool8)** [6] **D (dog1/cyste2)**[5] **D (dog1/cyste4)**[5] **D (dog5/pool5)**[6]	**Feces**	Dogs	Illumina	Kooyman et al. 2019
2019	A1 (ZX15)	Culture	Human	Illumina	Weisz et al. 2019
2018	A1 (18 isolates)[7] A1 (7 isolates)[7] A1 (1 isolates)[7] A1 (1 isolates)[7] A1 (2 isolates)[7] A1 (6 isolates)[7] A2 (6 isolates)[7] A2 (1 isolates)[7] A1/A2 (1 isolate)[7] B (6 isolates)[7] B (13 isolates)[7] B (1 isolates)[7] B (21 isolates)[7] A/E (1 isolates)[7] A/E (2 isolates)[7] A/E (2 isolates)[7]	Culture	Human Beaver Cat Dog Sheep Raw surface water Human Beaver Beaver Beaver Human Dog Raw surface water Human Beaver Raw surface water	Illumina	Tsui et al. 2018
2017	A2 (12 isolates)[8] B (8 isolates)[8]	Culture	Human	Illumina	Radunovic et al. 2017
2015	A1 (2 isolates)[9] A1 (1 isolate)[9] A2 (1 isolate)[9] B (5 isolates)[9] B (3 isolates)[9]	Culture	Human Beaver Human Human Drinking water	Illumina	Prystajecky et al. 2015
2015	B (BAH15c1)	Culture	Human	454 Life Science	Wielinga et al. 2015
2015	**A2 (sample 1)** **B (sample 2)** **B (sample 3)**	**Feces**	Human	SOLiD	Hanevik et al. 2015
2015	A2 (AS175) A2 (AS989)	Culture	Human	454 Life Science	Ankarklev et al. 2015

(*continued on next page*)

Table 1 (*continued*)

Year	Assemblages sequenced (isolate ID)	Type of isolates	Host	Sequencing platform	Citations
2013	A2 (DH) B (GS)	Culture	Human	454 Life Science	Adam et al. 2013
2010	E (P15)	Culture	Pig	454 Life Science	Jerlstrom-Hultqvist et al. 2010
2009	B (GS)[10]	Culture	Human	454 Life Science	Franzen et al. 2009
2007	A1 (WB/C6)[4]	Culture	Human	Sanger	Morrison et al.2007

[1] ATCC 30957.
[2] ATTCC 50580.
[3] Obtained from Waterborne.
[4] ATCC 50803.
[5] A single cyst was used.
[6] Pools of 40 cysts were used.
[7] Refer to Tsui et al. (2018) for specific isolate identification.
[8] No isolate identification provided.
[9] Refer to Prystajecky et al. (2015) for specific isolate identification.
[10] ATCC 50581.

sequencing or a hybrid approach to uncultured isolates obtained directly from feces to generate genomes has yet to be assessed. In this study, whole genome sequencing was performed using cysts cleaned and concentrated directly from feces. Two primary isolates from naturally infected hosts with loose stools containing high numbers of cysts were sequenced. The first isolate was obtained from a cat which received no anti-giardial treatment and the second from a dog which continued to excrete cysts following metronidazole treatment. The isolate from the cat is assemblage A making it useful for comparisons to well-accepted assemblage A references. While the isolate from the dog is assemblage D and represents a novel reference generated using these methods. Whole genome sequencing was performed using both long and short read sequencing platforms (Illumina MiSeq and Oxford Nanopore MinION), and assembly strategies using reads from both platforms either alone or in hybrid were compared to determine the best assembly for obtaining quality reference genomes from *Giardia* fecal isolates. Comparisons between the genomes produced in this study and reference genomes were also performed to assess quality and completeness.

2. Materials and methods

2.1. Source of isolates and cyst purification

Two primary isolates of *G. duodenalis* were obtained from naturally infected hosts. The first isolate came from a mixed breed domesticated cat that belonged to a closed cat colony that at the time of cyst collection had not received anti-*Giardia* treatment. The second isolate came from a privately-owned dog which after anti-*Giardia* treatment (metronidazole) continued to excrete cysts. Cat fecal sample collection was conducted under an animal use protocol approved by the Beltsville Area Animal Care and Use Committee (Protocol # 15–018). Dog fecal sample was a clinical sample submitted by dog owner without need for approval by the ethics committee.

Cysts from both fecal samples were sieved and then cleaned to concentrate parasite forms using cesium chloride density gradient centrifugation as previously described (Santín et al., 2004). Cysts were further cleaned and concentrated to obtain the cleanest possible starting material for whole genome sequencing using immunomagnetic beads (Dynabeads™ GC-Combo, Applied Biosystems, Carlsbad, CA) following manufacture's protocol with minor modifications. Changes to the original protocol include: scaling initial sample and buffer volumes down 10-fold to accommodate a 1.5 mL tube instead of the larger L10 tubes (1 mL cyst suspension + 100 μL 10X SL-Buffer A + 100 μL SL-Buffer B), addition of 1.2 mL 1X SL-Buffer A in a single step instead of 3 separate 400 μL additions, performance of the bead-cyst dissociation step twice, replacement of "Post-IMS" steps by a 10 min (1050 x G) centrifugation followed by removal of supernatant (leaving 10–20 μL in tube) and resuspension of purified cysts in 100 μL of PBS . Cysts were quantified before and after immunomagnetic bead purification by immunofluorescence microscopy using MerIFluor™ reagents (Meridian Biosciences, Cincinnati, OH) using a Zeiss Axioskop microscope equipped with epifluorescence and an FITC- Texas Red™ dual wavelength filter.

2.2. DNA preparation, PCR, and Sanger sequencing for assemblage identification

Total genomic DNA was extracted from cysts using the DNeasy Tissue Kit (Qiagen, Valencia, CA) following the manufacturer's instructions with minor modifications. Modifications were overnight incubation for proteinase K step and elution of DNA performed in 100 μl of AE buffer. The concentration of the extracted DNA was determined by Qubit™ (Invitrogen, Waltham, MA). DNA extraction yielded 18.39 ng and 77.84 ng of DNA from the cat and dog isolates, respectively.

Nested PCR amplification of fragments of the small subunit of the ribosomal RNA (*ssu*) and beta-giardin (*bg*) genes was performed as previously described (Hopkins et al., 1997; Lalle et al. 2005). PCR products were purified using Exonuclease I/Shrimp Alkaline Phosphatase (Exo-SAP-IT™ Express, Applied Biosystems, Foster City, CA) and sequenced in both directions using the same primers as the secondary PCRs in 10 μl reactions, Big Dye™ chemistries, and an ABI 3130xl Genetic Analyzer (Applied Biosystems, Foster City, CA). Sequence chromatograms of each strand were aligned and examined with Lasergene software (DNASTAR, Inc., Madison, WI).

2.3. Illumina library preparation and sequencing

Library preparation was performed using the Nextera DNA Flex kit (Illumina, San Diego, CA). Completed libraries were quantified using a Qubit™ (Invitrogen, Waltham, MA) and fragment size was estimated using a Bioanalyzer (Bio-Rad laboratories, Hercules, CA). Libraries were sequenced using an Illumina MiSeq (Illumina, San Diego, CA) and v2 300 cycle sequencing kit (2 × 150 bp) with paired end chemistry following the manufacturer's instructions.

2.4. Whole genome amplification and Nanopore library preparation and sequencing

Whole genome amplification (WGA) was performed to generate sufficient quantities of DNA for whole genome sequencing on the Nanopore MinION (Oxford Nanopre Technologies, Oxford, UK). The GenomiPhi V3 Ready-To-Go DNA Amplification Kit (Cytiva, Marlborough, MA) was used to amplify 10 ng of DNA from each isolate. Amplified DNA was quantified by Qubit™ (Invitrogen, Waltham, MA) and purity was measured by Nanodrop (ThermoFisher Scientific, Waltham, MA, USA). MinION library preparation was carried out using the Rapid Barcoding Sequencing kit (SQK-RBK004) (Oxford Nanopre Technologies, Oxford, UK) and 400 ng of DNA from each isolate. Following the fragmentation/barcoding step, XP bead clean-up was skipped based on recommendations in the protocol, and libraries were pooled using 5 μL of each sample. Libraries were then loaded onto an R9.4.1 flow cell (FLO-MIN106) and sequenced for approximately 20 hours. During the run, SQB buffer was added to the SpotON port at the 4-hour mark.

2.5. Read taxonomy and contaminant filtering

Raw Illumina and Nanopore reads were blasted against the NCBI nr protein database using DIAMOND v2.0.6 and the blastx command (Buchfink et al., 2021). Results were limited to the top hit and the

default e-value of 1e-3 was used. Blastx results were parsed using the taxonomizr v0.5.3 R package to map accessions to taxonomic IDs and ranks (Sherrill-Mix, 2019). Reads that did not have any protein hits were extracted and aligned to NCBI's nt nucleotide database using the blast+ (v2.11.0) megablast command (Camacho et al., 2009). Hit criteria included an e-value cutoff of 1e-3, max_hsps 1, and max_target_seqs 1. Megablast results were also parsed by taxonomizr and hit taxonomy was assigned. All reads that were classified as Bacterial via the superkingdom rank were filtered out from the original FASTQ files (hereafter referred to as "filtered reads").

2.6. Genome assembly

De novo assembly of filtered reads was performed using four different methods: short Illumina reads assembled with SPades v3.15.3, long Nanopore reads assembled with Canu v2.2, and a combination of short and long reads assembled with SPades or MaSuRCA v4.0.3 (Zimin et al., 2013; Antipov et al., 2016; Koren et al., 2017; Prjibelski et al., 2020). Prior to assembling Illumina reads with SPades, FASTQ read pairs were adapter trimmed, length filtered (minlength=75), and merged using bbduk and bbmerge from the bbtools software package v38.94 (Bushnell, 2014). Pairs were merged using the following options: rem, k=62, extend2=50, ecct, vstrict, mininsert=75. Both merged and unmerged reads were supplied to SPades for assembly with the following options: -k 127, –careful, –cov-cutoff 5. Canu correction, trimming, and assembly of Nanopore reads was performed using default parameters except for setting genome Size=12 m. SPades hybrid assembly used identical options to the Illumina-only assembly except for the addition of the the –nanopore flag with the path to the Nanopore FASTQ file. MaSuRCA assembler configuration was set based on recommendations provided in the example config file including LHE_COVERAGE=40 and FLYE_ASSEMBLY=0.

2.7. Mapping assembly to reference

Assembly contigs were mapped to reference genomes using minimap2 v2.22-r1101 (Li, 2018) with the option: -ax asm20. Samtools v1.13 was used to convert the minimap2 sam file to bam format and to extract mapping statistics (Li et al., 2009).

2.8. Gene/protein prediction and core protein alignment

Gene and protein predictions were performed using Prodigal v2.6.3 (Hyatt et al., 2010). A training file was first generated using the assemblage A1 WB isolate reference genome assembly (GenBank Assembly Accession GCA_000002435.2) and then supplied during the open reading frame (ORF) prediction step of the draft assembly. Predicted ORFs where then searched against using DIAMOND blastp and a list of homologous core proteins that are shared between the four reference genomes of assemblages A1, A2, GS_B, and P15_E (Adam et al., 2013). Blastp alignments were performed twice using query coverage and percent identity cutoffs of 50% or 90%.

2.9. Allelic sequence heterozygosity

Allelic sequence heterozygosity (ASH) was calculated within the draft assemblies using the bbtools package. First, filtered Illumina reads were mapped back to the MaSuRCA assembly contigs using bbmap.sh with the option: minid=95. Next, the sam file generated by bbmap along with the draft assembly FASTA file were used as input to callvariants.sh command with the options: ploidy=16, mincov=10, minallelefraction=0.15, calldel=f, and callins=f. ASH was then calculated as the percent of the number of substitutions divided by the total number of bases in the assembly.

2.10. BUSCO analysis

Estimation of genome completeness and redundancy of the draft genome assemblies was performed using BUSCO v5.2.1 which relies on OrthoDB v10 (Manni et al., 2021). The workflow was run in genome mode using the eukaryotic lineage.

2.11. Data availability and accession numbers

Genome assemblies and raw sequences are available at NCBI under the BioProject number PRJNA789594. The nucleotide sequences obtained by Sanger sequencing in this study for *ssu* and *bg* genes have been deposited in GenBank under the accession numbers OL965124–965125 and OL981636-OL981637, respectively.

3. Results

3.1. Impact of IMS purification on cyst counts

A multistep purification process which included CsCl density centrifugation and immunomagnetic separation was used to clean and concentrate *G. duodenalis* cysts from feces prior to DNA extraction and sequencing library preparation. Pre and post IMS purification cyst counts were performed to assess the impact of the IMS cleaning steps on cyst recovery. Prior to immunomagnetic separation there were 7.9×10^6 cysts from the cat sample. Following purification steps 1.6×10^6 cysts were available for DNA extraction. For the dog isolate, 10.5×10^6 cysts were observed prior to cleaning and 3.7×10^6 were available for DNA extraction.

3.2. Assemblage identification

Genotyping targeting *ssu* and *bg* genes was used to determine the assemblage of *G. duodenalis* for the two isolates used in this study. At both loci the *G. duodenalis* isolate obtained from the cat was identified as assemblage A, and the dog isolate was identified as assemblage D. At the *bg* locus, the sequence obtained from the cat isolate was identical to a sub-assemblage AI subtype A5 reference sequence (GQ329671) supporting its identification as sub-assemblage AI (Cai et al., 2021).

3.3. Assessment of read taxonomy

Both the assemblage A isolate obtained from the cat (abbreviated as CIA for Cat Isolate A) and the assemblage D isolate obtained from the dog (abbreviated as DID for Dog Isolate D) were sequenced via Illumina Miseq and Oxford Nanopore MinION platforms. Assessment of read taxonomy from CIA and DID identified that most reads generated using both Illumina and Nanopore sequencing platforms and which aligned to NCBI non redundant (nr/nt) database were from target sequence (80.8% and 84.2% for CIA and DID, respectively) (Table 2). Potential bacterial contamination in the reads was also assessed. Reads attributable to bacterial sequence varied by isolate and sequencing platform but generally represented a small proportion of total reads in any of the sequence pools (Table 2). Percentage of reads aligned to bacteria reference using MinION were 2.2% and 9.1% for CIA and DID, respectively. A higher percentage was found using Illumina, 15.0% for CIA and 18.9% for DID.

3.4. Features of de novo assemblies and comparison of assembly strategies

Comparisons between assembly methods utilizing Illumina reads, MinION reads, and a combination of the two were performed for both *G. duodenalis* isolates (Table 3). Using Illumina reads alone produced assemblies which were more fragmented compared to hybrid methods for both isolates. The assemblies comprised solely of MinION reads were more contiguous than Illumina-only assemblies but were still lower

Table 2
Assessment of read taxonomy for Cat Isolate A (CIA) and Dog Isolate D (DID) from reads obtained using Illumina MiSeq and Oxford Nanopore MinION sequencing platforms.

Isolate	Sequencing platform	Total read pairs	Percentage of reads aligned to NCBI non redundant (nr/nt) database	Percentage of reads aligned to *Giardia* reference [1]	Percentage of reads aligned to bacteria reference [1]
CIA	MiSeq	7,171,780	85.3%	80.0%	18.9%
	MinION	158,089	99.8%	96.1%	2.2%
DID	MiSeq	32,319,608	61.8%	84.1%	15.0%
	MinION	176,926	89.3%	90.2%	9.1%

[1] Includes only those reads which aligned to NCBI non redundant (nr/nt) database.s.

Table 3
Comparisons between assemblies from Illumina reads, MinION reads, and hybrids for Cat Isolate A (CIA) and Dog Isolate D (DID).

Isolate	Features	Assembly method SPAdes Illumina only	Canu MinION only	SPAdes hybrid	MaSuRCA hybrid
CIA	Number of scaffolds	526	201	273	93
	Total length (Mb)	10.4	10.1	10.7	10.7
	N50 (Kb)	58.3	83.5	132.7	149.1
	L50	63	43	28	18
	Max scaffold length (Kb)	192.5	206.3	340.8	644.7
	Scaffolds > 50 Kb	77	83	79	68
	Percentage of genome in scaffolds > 50 Kb	57.5	76.2	87.8	93.2
DID	Number of scaffolds	2821	824	2157	260
	Total length (Mb)	11.6	7.6	13.7	13.1
	N50 (Kb)	8.1	12.1	16.4	93.1
	L50	251	189	162	37
	Max scaffold length (Kb)	131.8	56.8	258.1	443.4
	Scaffolds > 50 Kb	26	2	46	80
	Percentage of genome in scaffolds > 50 Kb	15.8%	1.4%	25.7%	72.7%

quality compared to hybrid assemblies. Overall, the hybrid method that used both long and short reads produced higher quality *G. duodenalis* assemblies for both CIA and DID (Table 3). The CIA and DID hybrid assemblies generated using the MaSuRCA assembler had the best metrics of genome quality including, fewest number of contigs, the highest N50, lowest L50, and largest percentage of the genome in contigs greater than 50 KB. These assemblies were selected for use in downstream analysis of completeness and for comparison to available *G. duodenalis* reference sequences.

3.5. Comparison to reference genomes

The CIA genome consists of 93 contigs spanning 10.7 Mb. The DID genome consists of 260 contigs spanning 13.1 Mb. Mapping of CIA and DID hybrid assemblies to available reference assemblies, including assemblage A (A1 and A2), B, C, D, and E, was performed to assess coverage of reference assemblies, sequence similarity between genomes, and completeness of the hybrid assemblies (Table 4 and 5).

Mapping of CIA contigs to reference assemblies of the assemblage A1 WB isolate published in 2019 (GenBank Assembly Accession GCA_000002435.2) and 2020 (GenBank Assembly Accession GCA_011634545.1), yielded similar results. All CIA contigs and ≥97% of CIA bases mapped to WB assemblies (Table 4). Overall breadth of coverage was high with 85.2% and 88.6% of the A1 WB isolate published in 2019 and 2020 genome covered by CIA, respectively. A high degree of similarity and coverage was also observed when CIA was mapped to the assemblage A2 isolate DH reference assembly (GenBank Assembly Accession GCA_000498715.1). Mapping of CIA contigs to assemblage E (GenBank Assembly Accession GCA_000182665.1) and assemblage B (GenBank Assembly Accession GCA_000498735.1) reference genomes demonstrated less similarity and coverage between CIA and genomes from more distantly related assemblages.

The DID contigs were mapped against eight reference genomes including assemblage A1 WB isolate (GenBank Assembly Accession GCA_000002435.2), the assemblage A2 isolate DH (GenBank Assembly Accession GCA_000498715.1), the assemblage B isolate GS (GenBank Assembly Accession GCA_000498735.1), the assemblage C isolate Cyste1 (GenBank Assembly Accession GCA_902209425.1), the assemblage D isolates Cyste2 (GenBank Assembly Accession GCA_902221465.1), Cyste4 (GenBank Assembly Accession GCA_902221485.1), and Pool5 (GenBank Assembly Accession GCA_902221535.1), and the assemblage E isolate P15 (GenBank

Table 4
Mapping of Cat Isolate A (CIA) contigs to reference genomes (Assemblage A1, A2, B, and E).

	Reference genomes assemblage/isolate (GenBank Assembly Accession Number)				
	A1/WB (GCA_000002435.2)	A1/WB (GCA_011634545.1)	A2/DH (GCA_000498715.1)	B/GS (GCA_000498735.1)	E/P15 (GCA_000182665.1)
No. of scaffolds mapped	93	93	93	84	92
Scaffolds mapped (%)	100	100	100	90.3	98.9
No. of bases mapped	10,379,576	10,503,494	10,321,125	1,416,987	9,205,450
Bases mapped (%)	97.0	98.1	96.4	13.2	86.0
Base variation (%)	1.7	1.9	2.4	17.7	13.9
No. of reference scaffolds	35	37	239	543	820
No. reference bases	12,078,186	11,696,115	10,703,894	12,009,633	11,522,052
Reference scaffolds with any coverage (%)	42.8	89.2	78.2	32.0	26.9
Reference bases covered (%)	85.2	88.6	92.7	11.6	78.9

Table 5
Mapping of Dog Isolate D (DID) contigs to reference genomes (Assemblage A1, A2, B, C, D, and E).

	Reference genomes assemblage/isolate (GenBank Assembly Accession Number)								
	A1/WB (GCA_000002435.2)	A2/DH (GCA_000498715.1)	B/GS (GCA_000498735.1)	C/Cyste1 (GCA_902,209,425.1)	D/Cyste2 (GCA_902,221,465.1)	D/Cyste4 (GCA_902,221,485.1)	D/Pool5 (GCA_902,221,535.1)	E/P15 (GCA_000182665.1)
No. of scaffolds mapped	70	67	82	195	231	231	231	69
Scaffolds mapped (%)	26.9	25.8	31.5	75.0	88.8	88.8	88.8	26.5
No. of bases mapped	251,458	243,027	331,274	3,698,229	13,119,130	13,063,264	13,067,073	292,241
Bases mapped (%)	1.9	1.8	2.5	28.2	99.9	99.5	99.6	2.2
Base variation (%)	16.6	16.6	17.1	17.8	2.0	2.0	20	17.0
No. of reference scaffolds	35	239	543	3388	3269	2885	3489	820
No. reference bases	12,078,186	10,703,894	12,009,633	11,557,310	11,374,926	11,268,649	11,499,674	11,522,052
Reference scaffolds with any coverage (%)	20.0	22.6	17.3	9.9	43.1	44.6	41.1	9.5
Reference bases covered (%)	2.2	2.2	2.7	30.2	91.2	91.8	90.5	2.4

Assembly Accession GCA_000182665.1) (Table 5). Sequence similarity between DID and reference isolates for assemblage A (A1 and A2), B, and E was low with 25.8 to 31.5% of DID contigs mapping to any of these reference assemblies (Table 5). Similarly, coverage was low with 2.2 to 2.7% of total reference bases covered. To assess sequence similarity between more closely related isolates, DID contigs were also mapped against recently published assemblies obtained from *G. duodenalis* cysts from dogs which were isolated using flow cytometry and sequenced as individual cysts or as pooled cysts and identified as either assemblage C or assemblage D. Comparison with the three assemblage D isolates showed that 88.8% of contigs and 99.6 to 99.9% of bases from DID were successfully mapped, and over 90% of the bases in the assemblage D reference assemblies were covered by DID. Thus, a high degree of coverage and similarity was observed between the assemblage D isolate sequenced in this study and other published assemblage D genome assemblies. A lower degree of similarity was observed between DID and the assemblage C isolate, but there was more similarity between DID and assemblage C than observed between DID and sequences from assemblages A (A1 and A2), B, and E.

3.6. Allelic sequence heterozygosity

Allelic sequence heterozygosity (ASH) was assessed for both CIA and DID. There was far more ASH observed in DID than CIA with 0.65% and 0.08%, respectively.

3.7. ORF prediction and G. duodenalis core proteins present in assemblies

Prediction of open reading frames (ORFs) present in the hybrid assemblies was performed. There were 5815 and 7968 ORFs predicted within CIA and DID, respectively. A set of 4097 core genes was previously identified by Adam et al. (2013) from a four-way comparison of orthologs present in isolates WB, DH, GS, and P15 reference genomes. The presence of these core genes within the ORFs of CIA and DID was determined using two identification criteria. The first criterion used a cutoff of >90% identity and >90% coverage, and under these conditions, the percentage of *G. duodenalis* core genes present in CIA and DID was 97.9% and 13.0 %, respectively. The second and less restrictive criterion of >50% identity and >50% coverage greatly increased the percentage of core genes identified in DID to 94.6%.

3.8. BUSCO analysis for genome completeness

Benchmarking Universal Single Copy Orthologs (BUSCO) scores were assessed as an objective assessment of the completeness of the CIA and DID assemblies. Similar numbers of Eukaryota BUSCOs were observed in ORFs of both CIA and DID with total BUSCO scores of 29.8% and 30.3%, respectively (Table 6).

Table 6
Percentage of BUSCOs from Eukaryota dataset present in ORFs for genomes generated in this study, Cat Isolate A (CIA) and Dog Isolate D (DID), and four selected reference genomes for assemblages A and D.

	Assemblage (Isolate)					
	A (CIA)	D (DID)	A1 (WB/C6)[1]	A1 (WB/C6)[2]	D (Dog1/cyst4)[3]	D (Dog5/pool5)[3]
Complete	23.9	23.6	20.4	23.9	24.3	23.5
Fragmented	5.9	6.7	6.7	5.5	5.9	6.3
Total	29.8	30.3	27.1	29.4	30.2	29.8

[1] Morrison et al. (2007).
[2] Xu et al. (2020b).
[3] Kooyman et al. (2019).

4. Discussion

Giardia duodenalis is a species complex composed of eight assemblages with documented differences in host specificity (Cai et al., 2021). Furthermore, infection in the same host species with the same assemblage can present with different outcomes ranging from asymptomatic infection to severe gastrointestinal manifestations. Genetic factors both within and between these assemblages likely have key roles in determining many aspects of infection outcome as well as treatment resistance (Mørch and Hanevik, 2020). However, the drivers of assemblage and sub-assemblage level differences remain unknown. Whole genome sequencing and comparative genomics may help to explain the genetic determinants of parasite virulence, host specificity, and transmission. Yet, data from all known assemblages and from multiple isolates within each assemblage are needed to begin to parse out these relationships. A major hurdle in obtaining genomes from *Giardia* isolates is the inability to culture most *G. duodenalis* assemblages to obtain the DNA needed to produce complete genomes from primary isolates from feces or the environment.

In the present study, two primary isolates were sequenced, an assemblage A isolate obtained from a naturally infected domestic cat and an assemblage D isolate obtained from a naturally infected domestic dog. Sequencing using short read (Illumina MiSeq) and long read (Oxford Nanopore MinION) sequencing platforms was performed for both isolates. Assembly strategies using short reads, long reads, and a hybrid approach using reads from both sequencing platforms were compared to determine the best method for producing complete whole genome assemblies of *G. duodenalis*.

Working with fecal samples requires contending with potential sample contaminants which may include host DNA, DNA from host diet, bacterial DNA, and DNA from other eukaryotes which might be present in a fecal sample. Cleaning and concentrating parasite forms is necessary to obtain quality DNA for use in library preparation and to limit contaminant sequence from other artifacts present in the starting material. A multistep cyst cleaning process which included CsCl density centrifugation followed by immunomagnetic separation of cysts was used prior to DNA extraction, and cyst loss was observed following cleaning. Pre and post-purification IFA counts indicated that a large proportion of cysts were lost from both samples during cleaning. There was an 80% reduction in cysts for the cat isolate, and a 65% reduction in cysts available from the dog isolate. The number of cysts recovered from the isolates used in this study provided ample DNA material for Illumina library preparation, however WGA was needed to produce enough DNA for long read sequencing on the MinION platform which requires a relatively large starting concentration of DNA for library preparation. Limited starting material presents a unique challenge for producing whole genomes from organisms which cannot be cultured. WGA can be used to turn nanogram quantities of DNA into microgram quantities of amplified products. WGA has been demonstrated as a suitable method for obtaining the quantity and quality of DNA needed for whole genome sequencing when working with organisms for which DNA concentrations may be limited. In a recent study on the impact of WGA on *G. duodenalis* mutation identification, it was determined that WGA had no significant impact on mutation identification and that whole genome sequences produced from WGA material are suitable for use in comparative genomics studies (Weisz et al., 2019). WGA has also been employed for the production of WGS from individual *Giardia* cysts (Kooyman et al., 2019). Thus, WGA currently has an important role in aiding in the production of *Giardia* genomes from fecal or environmental samples as it can be used to amplify DNA from isolates which contain few parasite forms.

When working with primary *Giardia* isolates obtained from fecal samples, bacterial contamination in the reads can be limited through cleaning of cysts prior to DNA extraction. However, some degree of bacterial contamination is still likely to be present in the sequence pool. In this study, bacterial sequence contamination varied widely by isolate and sequencing platform (Table 2). A larger percentage of Illumina reads aligned to the bacterial reference database for both CIA and DID than reads obtained from MinION sequencing. Illumina sequences aligned to the bacterial reference database for both CIA and DID represented 18.9% and 15.0% of Illumina reads, respectively. Whereas the proportion for MinION reads was 2.2% and 9.1% for CIA and DID, respectively. This difference may be a function of the average read length being quite different between the two sequencing platforms. Illumina MiSeq reads generated in this study have a maximum read length of 150 bp, while MinION reads were nearly 20X as long with a mean read length of almost 3 Kbp and maximum read length of over 65 Kbp for CIA and a mean and maximum read length of over 2 Kbp and 43 Kbp for DID. Longer reads may be less likely to falsely aligned to similar sequences from another organism which may explain the difference in bacterial sequences detected from the two sequencing platforms used in this study. As 100% of the contigs and 97% of the bases from CIA mapped to other assemblage A isolates which were sequenced using axenic cultures, the cyst cleaning and data filtering steps used in this study appear to effectively limit bacterial sequence contamination in assemblies produced from fecal isolates.

Whole genome sequencing of CIA and DID was performed using short read and long read sequencing platforms. Reads from both platforms were assembled individually and as hybrid assemblies to assess the optimal method for producing the highest quality genomes from primary isolates of *G. duodenalis.* We found hybrid assemblies generated using the MaSuRCA assembler were the most contiguous with fewer contigs and higher N50s than any of the other assembly methods tested (Tables 3). A similar conclusion was drawn from a recent study which also found a hybrid assembly strategy combing Illumina and Nanopore reads produced optimal assemblies from cultured isolates of *G. duodenalis* (Pollo et al., 2020). Long-read sequencing technologies are better equipped to handle regions which are challenging to assemble with short reads such as repetitive elements or duplicated gene regions, as the short reads align equally well at more than one genomic location and cannot be unambiguously aligned to a reference, thereby offering an opportunity to enhance fragmented genome assemblies derived from only short reads (Lu et al., 2016). Thus, a hybrid strategy combining short but highly accurate Illumina reads with the long reads generated using the MinION can aid in producing reference quality genomes from difficult or unculturable isolates allowing for whole genomes of novel isolates and assemblages of *G. duodenalis* to be produced.

The hybrid assemblies of CIA and DID were mapped against reference genomes of *G. duodenalis* to assess sequence similarity between genomes from fecal and cultured isolates. The CIA assembly was very similar to reference assemblies of assemblage A from culture with all contigs and 97% of bases from CIA mapping to the assemblage A isolate WB reference assemblies (Table 4). The CIA assembly covered 79.5 to 88.6% of the bases present in the assemblage A isolate WB reference assemblies, thus it is relatively complete as compared to a closely related isolate obtained from culture. The DID assembly shared little similarity with reference assemblies of assemblages A, B, or E (Table 5). Although this observation was not surprising given that previous phylogenetic analysis of *G. duodenalis* genomes that included assemblages A (2 isolates), B (2 isolates), C (3 isolates), D (3 isolates) and E (1 isolate) demonstrated that assemblages C and D formed a separate clade from other assemblages (Kooyman et al., 2019). A similar phylogenetic relationship among assemblages have been long supported by analysis of individual loci commonly used for species differentiation and genotyping *Giardia* spp. such as triphosphate isomerase, glutamate dehydrogenase, or *bg* (Feng and Xiao, 2011). Therefore, the DID assembly was also mapped against recently published assemblies of assemblage C and assemblage D generated using single cysts or cysts pools obtained from dog fecal samples by flow cytometry (Kooyman et al., 2019). When comparing DID to an assemblage C assembly generated from a single cyst, 75.0% of contigs but only 28.2% of bases mapped (Table 5). Thus, assemblage C and assemblage D are more similar to each other than

assemblage D is to other *G. duodenalis* assemblages A, B, and E. Comparing DID to recently published assemblage D assemblies demonstrated 88.8% of contigs and 99.5 to 99.7% of bases from the assemblage D assembly produced in this study mapped to the three assemblage D assemblies from dogs (Table 5). The DID assembly also covered over 90% of the bases present in the assemblage D assemblies from the dog isolates obtained from The Netherlands indicating a high degree of similarity between the assembly from this study and those recently published for assemblage D (Kooyman et al., 2019).

The number of ORFs observed in CIA are similar to the number of genes previously reported in assemblage A with 5901 and 4963 genes reported for WB isolates (Morrison et al., 2007; Xu et al., 2020a). The number of ORFs observed in DID are higher than gene counts reported in other assemblage D assemblies (Kooyman et al., 2019). However, given that the assemblies reported for assemblage D draft genomes generated using cell sorting are far more fragmented than the DID hybrid assembly from this study, this difference could be attributed to the DID assembly being more contiguous. Indeed, a recent study which employed a hybrid assembly strategy for cultured *G. duodenalis* isolates reported highly contiguous assemblies using this strategy with 9639 gene models present in assemblage A isolate WB and 7234 gene models present in assemblage B isolate GS (Pollo et al., 2020). Future comparative studies between genomes produced using different sequencing and assembly strategies may help to reveal the source and importance of these differences.

Comparing ORFs between different assemblages and isolates can be used to define the core genes of the *G. duodenalis* species complex. Such comparisons can also provide a useful snapshot of the completeness of whole genome sequences from primary isolates. The ORFs in both CIA and DID were assessed for the presence of a previously described core set of orthologs thought to have the same function in genomes from assemblages A, B, and E (Adam et al., 2013). Of the 5815 predicted ORFs in CIA, 98% of core proteins were observed using 50% coverage and identity criteria, and 97.9% of core proteins were observed using a more stringent criteria of 90% coverage and identity. These findings support both the completeness of the CIA assembly and that this set of ORFs represent core proteins shared by assemblages A, B, and E. In contrast, of the 7968 predicted ORFS in DID, 94.6% of core proteins were observed using 50% coverage and identity criteria but only 13% of core proteins were observed using the more stringent criteria of 90% coverage and identity. The observation of the majority of core proteins in DID using less stringent coverage and identity supports the completeness of the assembly produced in this study. However, the striking differences in sequence identity between core proteins in DID and the assemblages used to generate the core protein list highlight the divergence of assemblage D from currently available reference genomes. These differences could help to explain biological differences between assemblages A and B which are zoonotic and have a wider host range and assemblage D which demonstrates stricter host specificity and is observed almost exclusively in canine hosts. Future comparative genomic studies including additional isolates may help to elucidate the genetic basis of such differences.

All *G. duodenalis* assemblages are tetraploid in the trophozoite form, and the cyst form of the parasite contains four tetraploid nuclei (Bernander et al., 2001). Yet, levels of ASH have been reported to vary widely by assemblage from less than 0.0023% in the assemblage E isolate P15 genome up to 0.53% in the assemblage B isolate GS genome (Franzén et al., 2009; Jerlström-Hultqvist et al., 2010). Variation within assemblages and even within individual isolates has also been reported, with ASH shown to be present between cells of the same isolate (Ankarklev et al., 2012). In this study, ASH of 0.08% was observed in CIA which is slightly higher that previous reports of ASH in assemblage A, A1 isolate WB (0.01 to 0.03%) (Morrison et al., 2007; Xu et al., 2020a) and A2 isolate DH (0.037%) (Adam et al. 2013) but lower than ASH reported for assemblage A2 isolates AS98 (0.35%) and AS175 (0.25%) (Ankarklev et al. 2015). While ASH of 0.65% was observed for DID which is similar to ASH reported for assemblage B (0.53%) (Franzén et al., 2009), DID had lower ASH than reported in assemblage C and assemblage D dog isolates from The Netherlands (Kooyman et al., 2019). The average ASH reported for both assemblage C and assemblage D in that study was 0.89% and 0.74 %, respectively (Kooyman et al., 2019). This difference between assemblage D isolates from different studies could be attributable to differences in sequencing and assembly methodologies. The observed ASH for CIA further supports assemblage A being less heterozygous than assemblage B, although data from more primary isolates are needed to fully characterize this relationship. The ASH observed in DID suggests assemblage D, like assemblage B and C, is far more heterozygous than assemblages A and E, although this observation would also be strengthened by further investigation.

Moderate to high levels of heterozygosity (greater than 1%) also present a challenge for genome assembly. Highly polymorphic regions containing multiple alleles can be erroneously split into alternative contigs and lead to more fragmented genomes with higher-than-expected sizes (Asalone et al., 2020). Hybrid de novo assemblers, like MaSuRCA, leverage long reads corrected by short accurate reads and are able to assemble polymorphic regions up to a certain level of heterozygosity. Beyond such thresholds post-assembly processing steps are needed to separate out different haplotypes (Zimin et al., 2013). This issue is not unique to *Giardia*, but the impact of ASH should be considered in future analyses as assembly algorithms become increasingly optimized for heterozygous genomes.

The significance of the differences in ASH between assemblages and within assemblages and isolates remains to be defined. However, it has recently been observed that regions around variant-specific surface protein (VSP) which may have roles in virulence and host specificity tend to be gene poor, but higher allelic sequence variation is observed in these regions in assemblage A isolate WB (Xu et al., 2020a). In another recent study, it was observed that VSPs were enriched in structural variants, defined as polymorphisms greater than 100 bp, in both assemblage A and assemblage B hybrid assemblies with more structural variants present in assemblage B overall (Pollo et al., 2020). These observations could potentially indicate a role for ASH in assemblage or isolate level differences in VSP repertoires which could impact resistance to treatment or clinical presentation.

BUSCO scores are intended to provide a quantitative assessment of genome assembly completeness and can be used as an objective indicator of genome quality (Manni et al., 2021). An objective measure of quality and completeness is useful in assessing whole genome sequences for organisms like *G. duodenalis* where significant genetic differences may be present between and within assemblages. Given the existences of such differences, comparisons between reference genomes and novel isolates may not be the ideal assessments of assembly quality for *G. duodenalis*. As such, ORFs present in the CIA and DID assemblies were scored for the presence of Eukaryotic BUSCOs alongside the previously published reference assemblies for assemblage A1 (isolate WB) and two assemblage D isolates representing assemblies from a single cyst and pooled cysts (Table 6). A similar percentage of Eukaryotic BUSCOs was observed in all the assemblies (Table 6). However, CIA and DID both had slightly higher scores than previously published assemblies. Using the methods described in this study, we obtained whole genome sequences from fecal isolates with the same level of genome completeness observed in cultured isolates, further supporting the use of this strategy to generate *G. duodenalis* genomes using genetic material obtained from cysts isolated from fecal isolates.

Whole genome sequencing and comparative genomics studies have the potential to provide important details about the genetic basis of virulence, resistance to treatment, transmission, and host specificity of *Giardia* spp. which have remained elusive. Yet, relatively few whole genome sequences have been produced for *G. duodenalis*, and currently, most genomes have been obtained from cultured isolates (Table 1). The lack of primary isolate sequences means that not all assemblages have been sequenced, and comparisons between strains from different

geographical regions, host species, symptom presentations, and treatment responses cannot and have not been performed. A major hurdle to obtaining whole genome sequences from primary isolates from feces or the environment is a lack of methodology both for obtaining DNA and for producing high quality assemblies. In this study, we demonstrate that a multistep cleaning process coupled with a hybrid sequencing and assembly strategy produce high quality, relatively complete whole genome sequences using primary isolates obtained from fecal samples of naturally infected hosts. Comparisons between the assemblies obtained in this study and other published assemblies from both cultured and primary isolates support the suitability of these methods for use with *G. duodenalis*. Moreover, we have generated reference quality genomes for two novel isolates of *G. duodenalis* including the most contiguous assembly currently available for assemblage D. The field of *G. duodenalis* whole genome sequencing will benefit from studies which use multiple sequencing strategies and assembly methods to produce whole genome sequences from a variety of isolates. Only through such experimentation and subsequent comparison can we begin to understand the ideal approach for producing *G. duodenalis* genomes and begin to fill the knowledge gaps in both assemblage and sub-assemblage level genetic data which is needed to understand the complex biology and epidemiology of *G. duodenalis*.

CRediT authorship contribution statement

Jenny G. Maloney: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Aleksey Molokin:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – review & editing, Visualization. **Gloria Solano-Aguilar:** Resources, Writing – review & editing. **Jitender P. Dubey:** Resources, Writing – review & editing. **Monica Santin:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

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.

References

Adam, R.D., 2021. *Giardia duodenalis*: biology and pathogenesis. Clin. Microbiol. Rev. 34 https://doi.org/10.1128/cmr.00024-19 e00024-19.

Adam, R.D., Dahlstrom, E.W., Martens, C.A., Bruno, D.P., Barbian, K.D., Ricklefs, S.M., Hernandez, M.M., Narla, N.P., Patel, R.B., Porcella, S.F., Nash, T.E., 2013. Genome sequencing of *Giardia lamblia* genotypes A2 and B isolates (DH and GS) and comparative analysis with the genomes of Genotypes A1 and E (WB and pig). Genome Biol. Evol. 5, 2498–2511. https://doi.org/10.1093/gbe/evt197.

Allain, T., Buret, A.G., 2020. Pathogenesis and post-infectious complications in giardiasis. Adv. Parasitol. 107, 173–199. https://doi.org/10.1016/bs.apar.2019.12.001.

Ankarklev, J., Svärd, S.G., Lebbad, M., 2012. Allelic sequence heterozygosity in single *Giardia* parasites. BMC Microbiol. 12, 65. https://doi.org/10.1186/1471-2180-12-65.

Ankarklev, J., Franzén, O., Peirasmaki, D., Jerlström-Hultqvist, J., Lebbad, M., Andersson, J., Andersson, B., Svärd, S.G., 2015. Comparative genomic analyses of freshly isolated *Giardia intestinalis* assemblage A isolates. BMC Genomics 16 (1), 697. https://doi.org/10.1186/S12864-015-1893-6.

Antipov, D., Korobeynikov, A., McLean, J.S., Pevzner, P.A., 2016. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics 32, 1009–1015. https://doi.org/10.1093/bioinformatics/btv688.

Asalone, K.C., Ryan, K.M., Yamadi, M., Cohen, A.L., Farmer, W.G., George, D.J., Joppert, C., Kim, K., Mughal, M.F., Said, R., Toksoz-Exley, M., 2020. Regional sequence expansion or collapse in heterozygous genome assemblies. PLoS Comput. Biol. 16 (7), e1008104 https://doi.org/10.1371/journal.pcbi.1008104.

Bernander, R., Palm, J.E.D., Svärd, S.G., 2001. Genome ploidy in different stages of the *Giardia lamblia* life cycle. Cell. Microbiol. 3, 55–62. https://doi.org/10.1046/J.1462-5822.2001.00094.X.

Brown, B., Allard, M., Bazaco, M.C., Blankenship, J., Minor, T., 2021. An economic evaluation of the Whole Genome Sequencing source tracking program in the U.S. PLoS One 16, e0258262. https://doi.org/10.1371/JOURNAL.PONE.0258262.

Buchfink, B., Reuter, K., Drost, H.G., 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods 18, 366–368. https://doi.org/10.1038/s41592-021-01101-x.

Bushnell, B., 2014. BBTools Software Package. http://bbtools.jgi.doe.gov (accessed 12/17/2021).

Cai, W., Ryan, U., Xiao, L., Feng, Y., 2021. Zoonotic giardiasis: an update. Parasitol. Res. 1, 1–20. https://doi.org/10.1007/s00436-021-07325-2.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinform. 10, 1–9. https://doi.org/10.1186/1471-2105-10-421.

Capewell, P., Krumrie, S., Katzer, F., Alexander, C.L., Weir, W., 2021. Molecular epidemiology of *Giardia* infections in the genomic era. Trends Parasitol 37, 142–153. https://doi.org/10.1016/j.pt.2020.09.013.

Díaz-Viraqué, F., Pita, S., Greif, G., de Souza, R.de C.M., Iraola, G., Robello, C., 2019. Nanopore sequencing significantly improves genome assembly of the protozoan parasite *Trypanosoma cruzi*. Genome Biol. Evol. 11, 1952–1957. https://doi.org/10.1093/GBE/EVZ129.

Dixon, B.R., 2021. *Giardia duodenalis* in humans and animals – Transmission and disease. Res. Vet. Sci. 135, 283–289. https://doi.org/10.1016/j.rvsc.2020.09.034.

Feng, Y., Xiao, L., 2011. Zoonotic potential and molecular epidemiology of *Giardia* species and giardiasis. Clin. Microbiol. Rev. 24, 110–140. https://doi.org/10.1128/CMR.00033-10.

Franzén, O., Jerlström-Hultqvist, J., Castro, E., Sherwood, E., Ankarklev, J., Reiner, D.S., Palm, D., Andersson, J.O., Andersson, B., Svärd, S.G., 2009. Draft genome sequencing of *Giardia intestinalis* assemblage B isolate GS: Is human giardiasis caused by two different species? PLoS Pathog. 5, e1000560 https://doi.org/10.1371/journal.ppat.1000560.

Hanevik, K., Bakken, R., Brattbakk, H.R., Saghaug, C.S., Langeland, N., 2015. Whole genome sequencing of clinical isolates of *Giardia lamblia*. Clin. Microbiol. Infect. 21, 192.e1–192.e3. https://doi.org/10.1016/J.CMI.2014.08.014.

Hanevik, K., Wensaas, K.A., Rortveit, G., Eide, G.E., Mørch, K., Langeland, N., 2014. Irritable bowel syndrome and chronic fatigue 6 years after Giardia infection: A controlled prospective cohort study. Clin. Infect. Dis. 59, 1394–1400. https://doi.org/10.1093/cid/ciu629.

Haque, R., Mondal, D., Karim, A., Molla, I.H., Rahim, A., Faruque, A.S., Ahmad, N., Kirkpatrick, B.D., Houpt, E., Snider, C., Petri Jr., W.A., 2009. Prospective case-control study of the association between common enteric protozoal parasites and diarrhea in Bangladesh. Clin. Infect. Dis. 48, 1191–1197. https://doi.org/10.1086/597580.

Hopkins, R.M., Meloni, B.P., Groth, D.M., Wetherall, J.D., Reynoldson, J.A., Thompson, R.C., 1997. Ribosomal RNA sequencing reveals differences between the genotypes of *Giardia* isolates recovered from humans and dogs living in the same locality. J. Parasitol. 83, 44–51.

Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. 11, 1–11. https://doi.org/10.1186/1471-2105-11-119.

Jerlström-Hultqvist, J., Franzén, O., Ankarklev, J., Xu, F., Nohýnková, E., Andersson, J. O., Svärd, S.G., Andersson, B., 2010. Genome analysis and comparative genomics of a *Giardia intestinalis* assemblage E isolate. BMC Genomics 11, 543. https://doi.org/10.1186/1471-2164-11-543.

Kooyman, F.N.J., Wagenaar, J.A., Zomer, A., 2019. Whole-genome sequencing of dog-specific assemblages C and D of *Giardia duodenalis* from single and pooled cysts indicates host-associated genes. Microb. Genom. 5, e000302 https://doi.org/10.1099/mgen.0.000302.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27, 722–736. https://doi.org/10.1101/gr.215087.116.

Lalle, M., Pozio, E., Capelli, G., Bruschi, F., Crotti, D., Cacciò, S.M., 2005. Genetic heterogeneity at the β-giardin locus among human and animal isolates of *Giardia duodenalis* and identification of potentially zoonotic subgenotypes. Int. J. Parasitol. 35, 207–213. https://doi.org/10.1016/j.ijpara.2004.10.022.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Lu, H., Giordano, F., Ning, Z., 2016. Oxford Nanopore MinION sequencing and genome assembly. Genom. Proteom. Bioinform. 14, 265–279. https://doi.org/10.1016/j.gpb.2016.05.004.

Lyu, Z., Shao, J., Xue, M., Ye, Q., Chen, B., Qin, Y., Wen, J., 2018. A new species of *Giardia* Künstler, 1882 (Sarcomastigophora: Hexamitidae) in hamsters. Parasit. Vectors 11, 202. https://doi.org/10.1186/s13071-018-2786-8.

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., Zdobnov, E.M., 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol. Biol. Evol. 38, 4647–4654. https://doi.org/10.1093/MOLBEV/MSAB199.

Messa Jr., A., Köster, P.C., Garrine, M., Gilchrist, C., Bartelt, L.A., Nhampossa, T., Massora, S., Kotloff, K., Levine, M.M., Alonso, P.L., Carmena, D., Mandomando, I., 2021. Molecular diversity of *Giardia duodenalis* in children under 5 years from the Manhiça district, Southern Mozambique enrolled in a matched case-control study on the aetiology of diarrhoea. PLoS Negl. Trop. Dis. 15, e0008987 https://doi.org/10.1371/journal.pntd.0008987.

Moolhuijzen, P., See, P.T., Moffat, C.S., 2021. The first genome assembly of fungal pathogen *Pyrenophora tritici-repentis* race 1 isolate using Oxford Nanopore MinION sequencing. BMC Res. Notes 14, 334. https://doi.org/10.1186/S13104-021-05751-0.

Mørch, K., Hanevik, K., 2020. Giardiasis treatment: an update with a focus on refractory disease. Curr. Opin. Infect. Dis. 33, 355–364. https://doi.org/10.1097/QCO.0000000000000668.

Morrison, H.G., McArthur, A.G., Gillin, F.D., Aley, S.B., Adam, R.D., Olsen, G.J., Best, A. A., Cande, W.Z., Chen, F., Cipriano, M.J., Davids, B.J., Dawson, S.C., Elmendorf, H. G., Hehl, A.B., Holder, M.E., Huse, S.M., Kim, U.U., Lasek-Nesselquist, E., Manning, G., Nigam, A., Nixon, J.E.J., Palm, D., Passamaneck, N.E., Prabhu, A., Reich, C.I., Reiner, D.S., Samuelson, J., Svard, S.G., Sogin, M.L., 2007. Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. Science 317, 1921–1926. https://doi.org/10.1126/SCIENCE.1143837.

O'Donnell, V.K., Mayr, G.A., Sturgill Samayoa, T.L., Dodd, K.A., Barrette, R.W., 2020. Rapid sequence-based characterization of African swine fever virus by use of the Oxford Nanopore MiniOn sequence sensing device and a companion analysis software tool. J. Clin. Microbiol. 58, e01104–e01119. https://doi.org/10.1128/JCM.01104-19.

Pollo, S.M.J., Reiling, S.J., Wit, J., Workentine, M.L., Guy, R.A., Batoff, G.W., Yee, J., Dixon, B.R., Wasmuth, J.D., 2020. Benchmarking hybrid assemblies of *Giardia* and prediction of widespread intra-isolate structural variation. Parasit. Vectors 13, 108. https://doi.org/10.1186/s13071-020-3968-8.

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., Korobeynikov, A., 2020. Using SPAdes de novo assembler. Curr. Protoc. Bioinform. 70, e102. https://doi.org/10.1002/cpbi.102.

Prystajecky, N., Tsui, C.K.M., Hsiao, W.W.L., Uyaguari-Diaz, M.I., Ho, J., Tang, P., Isaac-Renton, J., 2015. *Giardia* spp. are commonly found in mixed assemblages in surface water, as revealed by molecular and whole-genome characterization. Appl. Environ. Microbiol. 81, 4827–4834. https://doi.org/10.1128/AEM.00524-15.

Radunovic, M., Klotz, C., Saghaug, C.S., Brattbakk, H.R., Aebischer, T., Langeland, N., Hanevik, K., 2017. Genetic variation in potential *Giardia* vaccine candidates cyst wall protein 2 and α1-giardin. Parasitol. Res. 116, 2151–2158. https://doi.org/10.1007/S00436-017-5516-9.

Rendtorff, R.C., 1954. The experimental transmission of human intestinal protozoan parasites: II. *Giardia lamblia* cysts give in capsules. Am. J. Epidemiol. 59, 209–220. https://doi.org/10.1093/oxfordjournals.aje.a119634.

Santín, M., Trout, J.M., Xiao, L., Zhou, L., Greiner, E., Fayer, R., 2004. Prevalence and age-related variation of *Cryptosporidium* species and genotypes in dairy calves. Vet. Parasitol. 122, 103–117. https://doi.org/10.1016/j.vetpar.2004.03.020.

Sherrill-Mix, S., 2019. Taxonomizr: Functions to work with NCBI accessions and taxonomy (Version 0.5.3.). https://cran.r-project.org/package=taxonomizr (accessed 12/14/2021).

Todd, S.M., Settlage, R.E., Lahmers, K.K., Slade, D.J., 2018. *Fusobacterium* genomics using MinION and Illumina sequencing enables genome completion and correction. mSphere 3. https://doi.org/10.1128/MSPHERE.00269-18 e00269-18.

Torgerson, P.R., Devleesschauwer, B., Praet, N., Speybroeck, N., Willingham, A.L., Kasuga, F., Rokni, M.B., Zhou, X.-N., Fèvre, E.M., Sripa, B., Gargouri, N., Fürst, T., Budke, C.M., Carabin, H., Kirk, M.D., Angulo, F.J., Havelaar, A., Silva, N.de, 2015. World Health Organization estimates of the global and regional disease burden of 11 foodborne parasitic diseases, 2010: a data synthesis. PLOS Med 12. https://doi.org/10.1371/journal.pmed.1001920 e1001920.

Tsui, C.K.-M., Miller, R., Uyaguari-Diaz, M., Tang, P., Chauve, C., Hsiao, W., Isaac-Renton, J., Prystajecky, N., 2018. Beaver fever: whole-genome characterization of waterborne outbreak and sporadic isolates to study the zoonotic transmission of giardiasis. mSphere 3. https://doi.org/10.1128/MSPHERE.00090-18 e00090-18.

Weisz, F., Lalle, M., Nohynkova, E., Sannella, A.R., Dluhošová, J., Cacciò, S.M., 2019. Testing the impact of whole genome amplification on genome comparison using the polyploid flagellated *Giardia duodenalis* as a model. Exp. Parasitol. 207, 107776 https://doi.org/10.1016/j.exppara.2019.107776.

Wick, R.R., Judd, L.M., Gorrie, C.L., Holt, K.E., 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb. Genomics 3. https://doi.org/10.1099/MGEN.0.000132 e000132.

Wielinga, C., Thompson, R.C.A., Monis, P., Ryan, U., 2015. Identification of polymorphic genes for use in assemblage B genotyping assays through comparative genomics of multiple assemblage B *Giardia duodenalis* isolates. Mol. Biochem. Parasitol. 201, 1–4. https://doi.org/10.1016/J.MOLBIOPARA.2015.05.002.

Xu, F., Jex, A., Svärd, S.G., 2020a. A chromosome-scale reference genome for *Giardia intestinalis* WB. Sci. Data 7, 1–8. https://doi.org/10.1038/s41597-020-0377-y.

Xu, F., Jiménez-González, A., Einarsson, E., Ástvaldsson, Á., Peirasmaki, D., Eckmann, L., Andersson, J.O., Svärd, S.G., Jerlström-Hultqvist, J., 2020b. The compact genome of *Giardia muris* reveals important steps in the evolution of intestinal protozoan parasites. Microb. Genomics 6, 1–15. https://doi.org/10.1099/mgen.0.000402.

Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., Yorke, J.A., 2013. The MaSuRCA genome assembler. Bioinformatics 29, 2669–2677. https://doi.org/10.1093/bioinformatics/btt476.