scientific data

OPEN

DATA DESCRIPTOR

A chromosome-scale genome assembly of *Giardia duodenalis* by long-read sequencing of ten trophozoites

Jingjing Sun[1,2,3], Yuancai Chen[1,2,3], Yizhuo Wei[1,2,3], Kaihui Zhang[1,2,3], Yin Fu[1,2,3], Zi Yan[1,2,3], Xingyu Zhu[1,2,3], Sumei Zhang[1,2,3], Longxian Zhang[1,2,3] ✉ & Junqiang Li[1,2,3] ✉

***Giardia duodenalis*, a protozoan parasite of important zoonotic concern, is estimated to cause approximately 280 million human infections annually worldwide. Currently, genome sequencing of *G. duodenalis* mainly relies on *in vitro* axenic clones; however, for non-culturable or hard-to-cultivate microorganisms, obtaining sufficient DNA for whole-genome sequencing poses a great challenge. In this study, we isolated ten *G. duodenalis* trophozoites using single-cell selection technology, followed by the extraction of whole-genome DNA and its amplification via multiple displacement amplification (MDA). The *G. duodenalis* DNA was sequenced by long-read sequencing (Oxford Nanopore Technologies and Pacific Biosciences), and three main assembly tools (Canu, MECAT2, and RagTag) were used to assemble the sequenced data. As a result, a chromosome-scale genome of *G. duodenalis* was successfully assembled (assemblage A1 isolate g12a2), with a total genome size of 11.1 Mbp, five contigs, and an N50 value of 3.1 Mbp. This study achieved a chromosome-scale *G. duodenalis* genome sequencing and assembly from groups of 10 trophozoites, which facilitates protozoan single cell genomics research.**

Background & Summary

Giardia duodenalis (synonyms *Giardia lamblia* and *Giardia intestinalis*) is a unicellular flagellated parasite that infects the gastrointestinal tract of a wide range of mammalian animals, including humans[1]. It is estimated that *G. duodenalis* causes infection in approximately 280 million people annually worldwide[2]. Investigations into DNA polymorphisms indicate that *G. duodenalis*, despite being morphologically indistinguishable, should be recognized as a species complex[3]. *G. duodenalis* can be classified into eight assemblages (A-H) commonly identified by nucleic acid polymorphisms of the beta-giardin (*bg*) gene, glutamate dehydrogenase (*gdh*) and triosephosphate isomerase (*tpi*)[4].

The first sequenced genome of *G. duodenalis* (WB isolate) was reported in 2007, which yielded a genome of 11.7 Mbp in size and was composed of 306 contigs (92 scaffolds)[5]. So far, at least 38 *G. duodenalis* genomes have been sequenced and reported (assemblage A to E)[6], which were represented by assemblage A (WB/C6[5], Be-2[7]), assemblage B (GS[8], BAH15c1[9]), assemblage C (cyste1, cyste3, pool8)[10], assemblage D (cyste2, cyste4, pool5)[10], and assemblage E (P15), and some other isolates[11,12]. However, the majority of the *Giardia* genomes had been assembled only at the contig level[6]. The fragmented assembly of *G. duodenalis* genomes imposes great constraints on our understanding of its chromosomal structure and the evolutionary dynamics at the chromosomal scale. With the rapid development of genome sequencing technology, long-read sequencing using Oxford Nanopore Technologies (ONT) and Pacific Biosciences (Pacbio) have been used to generate genome for *G. duodenalis*[13,14]. Among them, the genomes of the *G. duodenalis* reference isolate WB-C6 and the Be2 isolate have been assembled to the chromosomal scale by means of sequencing millions of trophozoites or cysts[7,14].

[1]College of Veterinary Medicine, Henan Agricultural University, Zhengzhou, 450046, China. [2]International Joint Research Laboratory for Zoonotic Diseases of Henan, Zhengzhou, 450046, China. [3]Key Laboratory of Quality and Safety Control of Poultry Products, Ministry of Agriculture and Rural Affairs, Zhengzhou, 450046, China. ✉e-mail: zhanglx8999@henau.edu.cn; lijunqiangcool@126.com

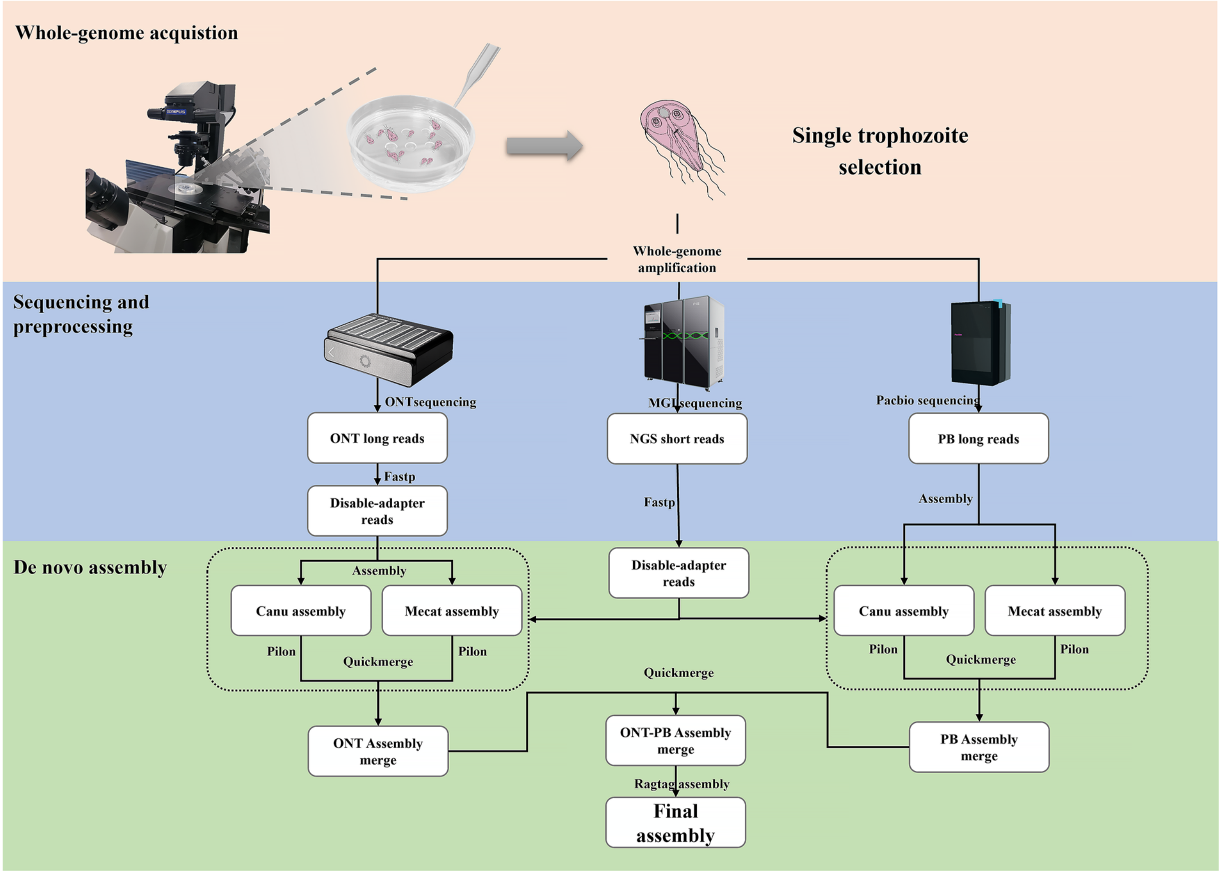

Fig. 1 Workflow for the generation of *Giardia duodenalis* genome assembly in this study.

For most microorganisms, the sample quantity needed for genomic sequencing is generally acquired through *in vitro* axenic clones, which is both time-consuming and labor-intensive. However, for microorganisms that are either non-culturable or hard to cultivate *in vitro*, securing an adequate quantity for whole-genome sequencing poses a great challenge. In more recent years, a technology that combines multiple displacement amplification (MDA) with single-cell genomic sequencing could be a breakthrough for this problem, where only a small number of homogeneous cells are needed. Now, single-cell sequencing is widely used for human cells[15,16], ruminant livestock[17], and some microorganisms, such as *Plasmodium vivax*[18], *Eimeria tella*[19], *Cryptosporidium*[20], etc. Furthermore, Kooyman *et al*. first isolated single and pooled cysts of *G. duodenalis*, and then performed genome sequencing using MDA technology and Illumina paired-end sequencing, which led to the generation of a draft genome[11]. Therefore, employing the MDA technology to amplify the genomic DNA of single microbial cells enables the acquisition of a substantial quantity of DNA, which is suitable for subsequent genome sequencing.

In this study, we performed genome assembly of *G. duodenalis* using ten trophozoites sequenced on both the ONT and PacBio platforms and assembled a contiguous, chromosome-scale genome of *G. duodenalis* (g12a2). The total length of the assembled *G. duodenalis* genome was 11.1 Mbp, with five contigs. The genome results indicated that the longest contig was 3,279,057 bp, and the N50 and N90 of the contigs were 3,062,379 bp and 1,457,149 bp, respectively. The GC% and BUSCO completeness score of the newly assembled genome were 49.23% and 21.6%, respectively. A high-resolution genetic map of *G. duodenalis* can enhance our understanding of the diversity and evolution of *G. duodenalis*, promote research in comparative genomics of *G. duodenalis*, and open up new avenues for drug and vaccine development.

Methods

Single trophozoite selection.

The *G. duodenalis* trophozoites used in this study were preserved in the Henan International Joint Laboratory for Zoonoses (Zhengzhou, China). They were axenically cultured in TYI-S-33 medium, and these trophozoites belong to sub-assemblage AI (named isolate g12a2 in this study). The axenically cultured trophozoites were aliquoted into 2 mL centrifuge tubes, and 30 μL of the trophozoite suspension was placed as droplets into the culture medium, which was then sealed with paraffin oil. The selected trophozoites were washed with 0.1 mol/L phosphate-buffered saline (PBS, pH 7.0) and then isolated under an inverted microscope (OLYMPUS-BX53, Japan) using a micromanipulator (World Precision Instruments Inc., USA). The isolated 10 trophozoites of *G. duodenalis* were preserved together in one PCR tube containing 4 microliters of phosphate buffered saline (PBS) for subsequent DNA extraction (Fig. 1).

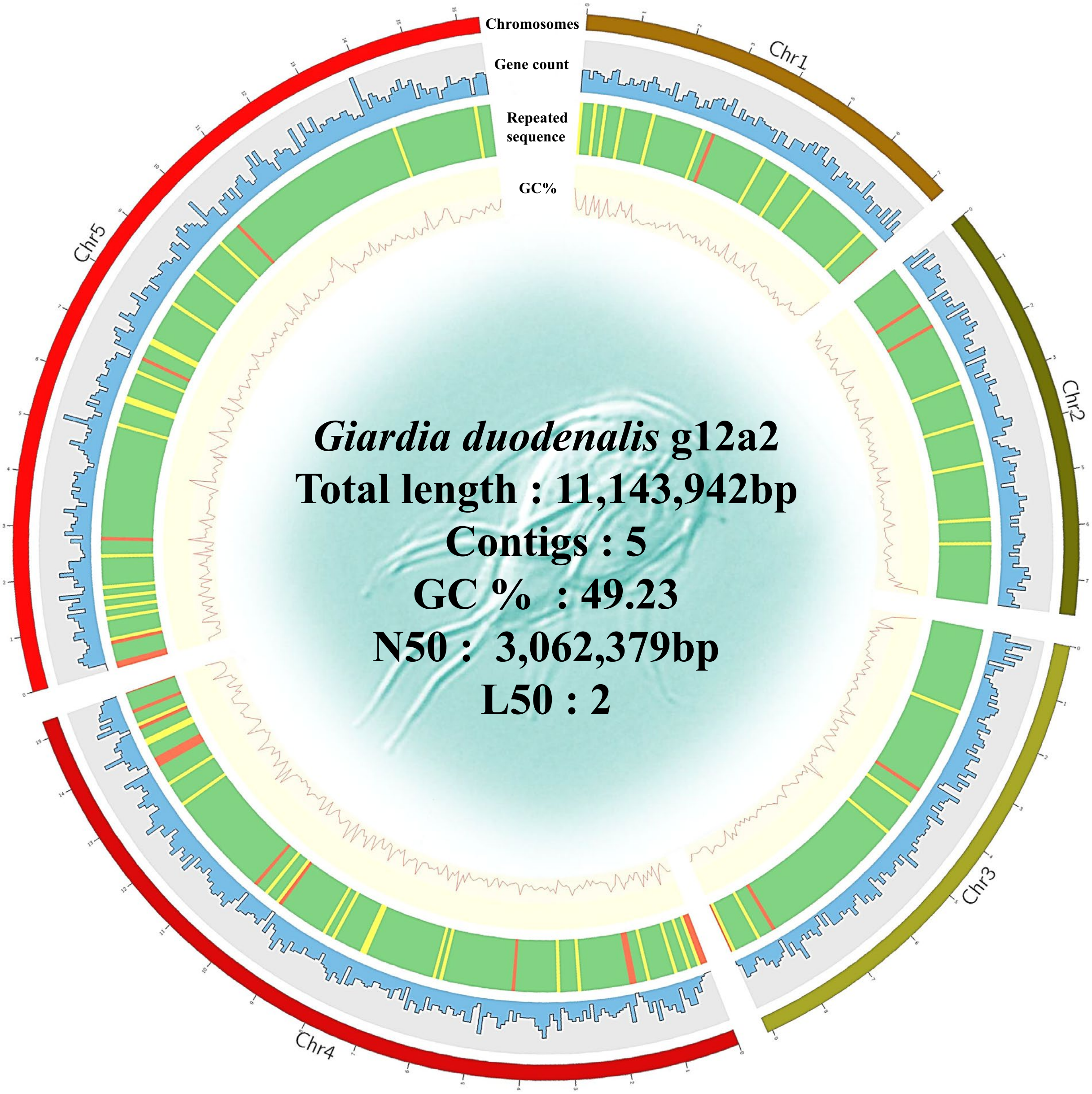

Fig. 2 Composition of the *Giardia duodenalis* genome of isolate g12a2 in this study. The outermost layer represents chromosomes, with chr1, chr2, chr3, chr4, and chr5 represented by five different colors. Gene content information is represented by a histogram. Repeat sequences are represented by a heat map. The GC content is represented by a line graph.

Genomic DNA preparation. Ten *G. duodenalis* trophozoites samples were lysed and subsequently subjected to whole-genome amplification using the REPLI-g single cell kit (QIAGEN, Germany), which was based on a MDA method. The whole-genome amplification (WGA) products were purified using Agencourt AMPure XP beads (BECKMAN, USA) at a ratio of 1:0.9. This process effectively removed dNTP, primers, primer dimers, salt ions, and other impurities from the amplified products, so as to achieve the selective retention of DNA fragments.

Quality assessment of the total genomic DNA. The concentration of the WGA product of *G. duodenalis* was determined to be 1094.0 ng/μL using a NanoDrop One spectrophotometer (Thermo Fisher Scientific, USA). The ratios of A260/A280 and A260/A230 were 1.85 and 2.29, respectively. These values indicate that the DNA was highly pure and free from contamination by impurities, such as polysaccharides, phenol, and salts (Table 1).

Fifteen pairs of specific primer sequences were designed based on the reference genome sequence of *G. duodenalis* in the National Center for Biotechnology Information (NCBI) database (Table 2). Among the 15 pairs of designed primers, 14 successfully amplified the target amplicons through PCR amplification. Furthermore, nested PCR amplification of the *bg*, *gdh*, and *tpi* loci of *G. duodenalis* yielded target bands at

Statistic	WB	WB-C6	Be-2	g12a2*	
Report year	2007[5]	2019[14]	2022[7]	2024 (this study)	
Life-cycle stage	Trophozoites	Trophozoites	Cysts	Trophozoites	
No of trophozoites/cycts	—	2.3×10^8	1.0×10^8	10	
Sequencing instrument	LI-COR, ABI3700	PacBio R SII	MinIONR9.4.1	Nanopore PromethION 48	PacBio Revio
Reads	224,000	411,835	148,144	2,966,909	1,618,041
Bases	—	3.6 billion	—	11.5 billon	12.1 billion
Assembler	ARACHNE2.0	HGAP3	Flye	Canu, MECAT2, RagTag	
Genome size (Mbp)	11.7	12.1	11.4	11.1	
Total ungapped length (Mbp)	—	11.7	11.4	11.1	
Gaps	137	35	1	13	
Gap size (Mbp)	1.6	4	0.01	0.4	
Number of chromosomes	5	5	5	5	
Number of scaffolds	306	35	5	5	
Number of contigs	92	38	6	5	
Contig N50 (Mbp)	—	2.8	3.0	3.1	
Contig L50	—	2	2	2	
GC percent	49.0%	46.3%	49.5%	49.2%	
Genome coverage	11x	217x	84x	1027x	1086x
Assembly level	Contig	Chromosome	Chromosome	Chromosome	

Table 1. Comparison of the *Giardia duodenalis* genome. Note: *g12a2 refers to the genome name obtained by sequencing in this study.

the corresponding positions (Table 2). This result suggests that the extracted DNA has a relatively complete coverage. Quantitative polymerase chain reaction (q-PCR) primers were designed and used for detection. The upstream primer F was "CTCCCTCCTTTGTGCACCTTTCTACAGCTA", and the downstream primer R was "CTCGCCCATGATTCTACGTCTTTCAGAGTG". We then performed the q-PCR detection on the *G. duodenalis* genomic DNA to assess its purity. The obtained CT values ranged from 21.39~23.00, indicating that the DNA content was within the normal range.

Short-read sequencing.

The qualified whole genome DNA of *G. duodenalis* was sent to a commercial genome sequencing company (Wuhan Nextomics Biosciences Co., Ltd., China) for genome sequencing. For MGI Tech Sequencing (MGI), 1 μg of this genomic DNA was used to prepare a second-generation sequencing library with the MGIEasy Universal DNA Library Prep Kit V1.0 (CAT#1000005250, MGI). The library was then sequenced on the MGISEQ-2000 platform (BGI, Shenzhen, China), generating second-generation sequencing data. After that, the reads were demultiplexed, and the files were exported in fastq format (Fig. 1). Approximately 2.2 Gb of 150 bp paired-end reads (clean data) with 198 × coverage were generated using the MGI sequencing platform.

Long-read sequencing.

For Oxford Nanopore Technologies sequencing (ONT), approximately 8–10 μg of genomic DNA was repaired using the NEBNext FFPE DNA Repair Mix (Cat #M6630, NEB, USA). Subsequently, end-repair and dA-tailing were carried out with NEBNext Ultra II reagents (NEB, USA). The dA-tailed insert molecules were further ligated with an ONT adaptor via ligation kit SQK-LSK114 (Oxford Nanopore Technologies, UK). The concentrations of library fragments were quantified with the Qubit® 3.0 Fluorometer (Invitrogen, USA). Based on the quantified concentrations, an appropriate amount of the DNA library was selected. The DNA library was then loaded into the primed PromethION 48 sequencer (ONT, UK) flow cell and the ONT sequencing data were collected in fastq format.

For Pacific Biosciences (PacBio) sequencing, the quality control of at least 15 μg of genomic DNA was conducted using the Femto Pulse system (Agilent, USA), followed by shearing of the DNA with the Megaruptor 3 system (Diagenode). The dA-tailed insert fragments were further ligated and subjected to nuclease treatment using the SMRTbell® prep kit 3.0 (PacBio, USA). Library size selection and purification were performed using the PippinHT system (Sage Science, USA). Ultimately, the DNA library was constructed using the Revio enzyme binding kit (Pacific Biosciences, USA), and sequenced on the PacBio platform. The resulting PacBio sequencing data were collected in BAM format. Specifically, the ONT and PacBio sequencing generated approximately 11.5 Gb and 12.1 Gb of data, respectively, achieving 1029 × coverage for the former and 1088 × coverage for the latter against the *G. duodenalis* genome (estimated at ~12 Mb).

Pre-assembly data processing.

The sequencing data for the g12a2 isolate from the three platforms of MGI, ONT, and PacBio were preprocessed. This involved the use of Fastp (v 0.23.4)[21] to remove adapter contamination, filter out low-quality reads, and perform a thorough quality control analysis. Following, the BWA alignment tool (v 0.7.17)[22], using the BWA-MEM algorithm, was employed to align the sequencing data from MGI, ONT, and PacBio with the reference genome Be2, which yielded alignment rates for each dataset. The calculated alignment rate for each sequencing dataset served as a metric to assess the completeness and integrity of the sequencing data.

Primer sequence	**Primers (5′–3′)**	**Size (bp)**	**Annealing temperature (°C)**
1	F: TCAGACACAGCAGATAGTAAG	821	50
	R: TCACCTCAAGCAGCAGTA		
2	F: ACGAAGACGAAGCGATAC	752	50
	R: CTACAACGAAGACGAACCT		
3	F: TCCATTCGTGACCATATCG	520	50
	R: ATACTGATGTGACCGTTAGG		
4	F: AGCCTCCTCCAACACTATA	957	50
	R: TGACAGCGTATGCGATAG		
5	F: CATTGTGTAGGTAGGCAGAA	991	50
	R: TTAGAGGCGTGGATGACT		
6	F: GCATTATCTGGCGGTATCA	799	50
	R: GAGGAGCATAGGATTCACAA		
7	F: GGACCTCCTCTCAACAAC	508	50
	R: AGTGTAATGCCAGAGAATGA		
8	F: GCTCCTGACTGAAGATTGT	596	50
	R: GCGATTGATGACGATATACC		
9	F: CTCAAGCATCGGCAGTAA	609	50
	R: AAGCAGTCTACAGCAAGTAA		
10	F: GAAGAAGGTGAGCGTGTT	624	50
	R: CTGCCGATTGATGGATGT		
11	F: GATAGCATCCGCACGAAT	968	50
	R: TACTTCCTGTCTCTTCCTCT		
12	F: TGATACAGGTGGCAAGGT	558	50
	R: TAGGAAGTGAAGAGGAAGGA		
13	F: CATCACATCCAGAGTAGAAGA	860	50
	R: CACATAAGCACGCAGAGA		
14	F: TGAGATAGTTGATGGCAGAC	682	50
	R: TGACCGATTGGATGGAGTA		
15	F: TAACGGTGGATGACTATGAC	562	50
	R: CGACGGAAGTGGATTGAA		
bg	F1: AAGCCCGACGACCTCACCCGCAGTGC	511	65
	R1: GAGGCCGCCCTGGATCTTCGAGACGAC		
	F2: BG3 GAACGAACGAGATCGAGGTCCG		55
	R2: BG4 CTCGACGAGCTTCGTGTT		
gdh	F1: TTCCGTRTYCAGTACAACTC	520	50
	R1: ACCTCGTTCTGRGTGGCGCA		
	F2: ATGACYGAGCTYCAGAGGCACGT		50
	R2: GTGGCGCARGGCATGATGCA		
tpi	F1: AAATIATGCCTGCTCGTCG	530	50
	R1: CAAACCTTITCCGCAAACC		
	F2: CCCTTCATCGGIGGTAACTT		50
	R2: GTGGCCACCACICCCGTGCC		

Table 2. Primer sequences for *Giardia duodenalis* genome coverage detection.

Genome assembly. For the ONT sequencing data, the Canu (v 2.0)[23] software was used to generate an assembly result based on the Overlap-Layout-Consensus (OLC) algorithm with default parameters. Meanwhile, the MECAT2 assembly tool[24] produced a mecat assembly result by exploiting the overlap relationships among the reads, using default parameters. Both the Canu and MECAT2 assembly results were processed to remove contigs with abnormal GC content. Subsequently, the assemblies were aligned to the reference genome using NUCmer (v 3.1)[25], and contigs with extremely low alignment rates were removed. To further improve the assembly results, MGI sequencing data were used for error correction. The MGI sequencing data were aligned to the assembly results using BWA (v 0.7.17), and the aligned sequences were sorted using Samtools (v 1.9)[26] to generate index files. These BAM files were then polished in two rounds using Pilon (v 1.24)[27] (with the parameters:–changes–vcf–diploid), which iteratively corrected the contigs in the original assembly results. Ultimately, the corrected Canu and MECAT2 assembly results were obtained.

To improve the quality of the genome assembly, we compared the overlapping regions and sequence features of the assembly results generated by Canu and MECAT2. By integrating the information from both assembly results, we employed the Quickmerge tool (v 0.3)[28] (with the parameters -hco 5.0 -c 1.0 -l 10000 -ml 100) to

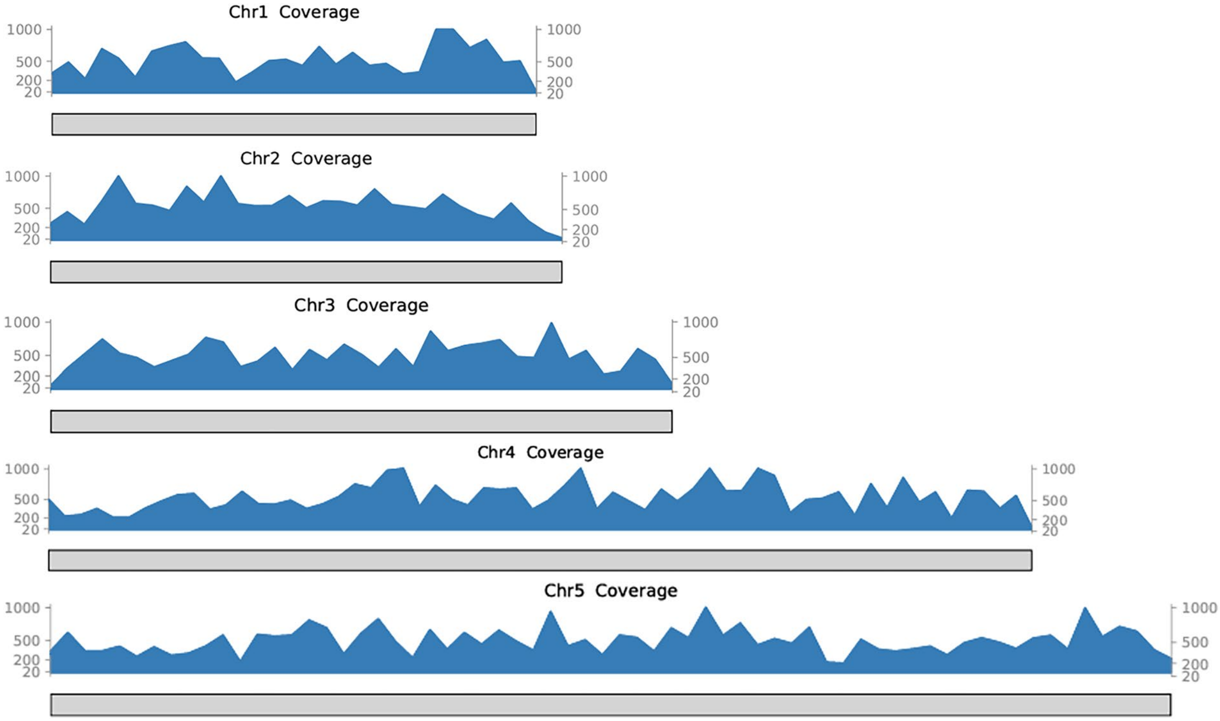

Fig. 3 Mapping the Oxford Nanopore Technologies sequencing data after Fastp to the reads of the g12a2 assembly. The vertical axis represents the amount of Oxford Nanopore Technologies sequencing data mapped to this region. The horizontal axis represents the new genome.

merge the outputs of Canu and MECAT2. This process yielded a higher-quality genome assembly, namely the ONT assembly merge.

For the PacBio sequencing data, we performed the same processing as for the ONT sequencing data, which generated a PacBio merged assembly result, referred to as the PacBio assembly merge. We used the Quickmerge tool (v 0.3) with default parameters to integrate the ONT and PacBio merged assembly results, thereby obtaining an ONT-PacBio merged assembly result.

With the default parameters of the RagTag tool (v 2.1.0)[29], we conducted error correction, scaffolding, and gap filling between scaffolds. We used the Oxford ONT-PacBio merged assembly result as the input file for a secondary assembly. Ultimately, we obtained the final assembly result for *G. duodenalis* (isolate g12a2), which was 11.1 Mbp in size and had a contig N50 of 3.1 Mbp (Fig. 2).

Data Records

The raw sequencing data for *G. duodenalis* (isolate g12a2), including MGI short reads (accession CRA021161)[30], PacBio (accession CRA021160)[31] and ONT long reads (accession CRA021157)[32], and the whole-genome assembly (accession GWHFIGV00000000.1)[33] of the *G. duodenalis* isolate can be accessed through National Genomics Data Center, China National Centre for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (PRJCA033502)[34]. The genome assembly for *G. duodenalis* generated in this study has been submitted to the NCBI GenBank database with the assembly accession GCA_049639855.1[35]. Moreover, the genomic annotation results have been deposited in the Figshare database[36].

Technical Validation

The ONT sequencing data for the *G. duodenalis* genome, which were processed by Fastp (v 0.23.4), were aligned to the assembled g12a2 genome, and each position in the assembly result of g12a2 was supported by at least 171 × reads. Subsequently, Python was used to plot and present the results (Fig. 3).

NUCmer (v 3.1) with the parameters of "–maxgap = 50000" and "–mincluster = 100" was employed to align the final genome assembly result (isolate g12a2) with the two reference genomes, Be2 (GCA_026248805.1)[37] (Fig. 4a) and WB-C6 (GCA_000002435.2)[38] (Fig. 4b). Subsequently, the dot-plot function of MUMmerplot (v 3.5)[39] was used to assess and visualize the generated delta files. Additionally, GenomeSyn (v 1.41)[40] was used to identify the synteny between g12a2 and the two reference genomes, and the findings were displayed in a graphical format (Fig. 4c). The alignment analysis confirmed that these five contigs generated in this study (isolate g12a2) exhibited a high degree of congruence with reference genome of GCA_026248805.1 (Be2), and also demonstrated a strong similarity to the initial five contigs of reference GCA_000002435.2 (WB-C6), which correspond to the previously identified chromosomes.

The N50 in genome assembly is an important indicator for evaluating the quality of genome assembly. For a genome, a larger N50 indicates less fragmentation and can reduce the disruption of gene structure caused by

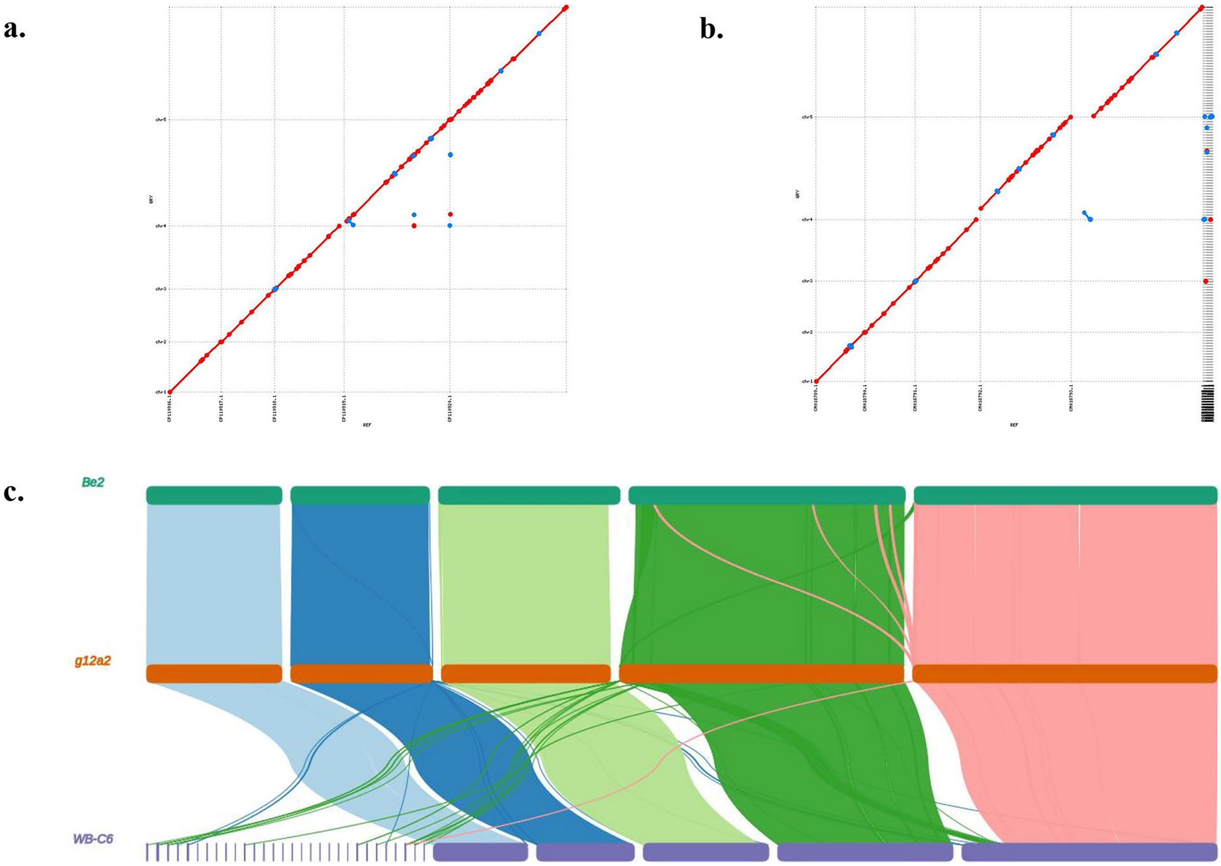

Fig. 4 The comparison between the g12a2 assembly and the previously published Be2, WB-C6 isolate assembly. (**a**) The g12a2 assembly shows a high concordance with the previously published Be2 isolate assembly. (**b**) The g12a2 assembly shows a high concordance with the first five contigs of the previously published WB-C6 isolate assembly. (**c**) Genome collinearity analysis of g12a2 with Be2 and WB-C6. The collinear gene blocks were determined by NGenomeSyn between genomes for three *Giardia duodenalis* isolates. Syntenic relationship in gene organization across the genomes of g12a2, Be2 and WB-C6. Syntenic blocks, which consist of regions containing orthologous genes, are connected by lines.

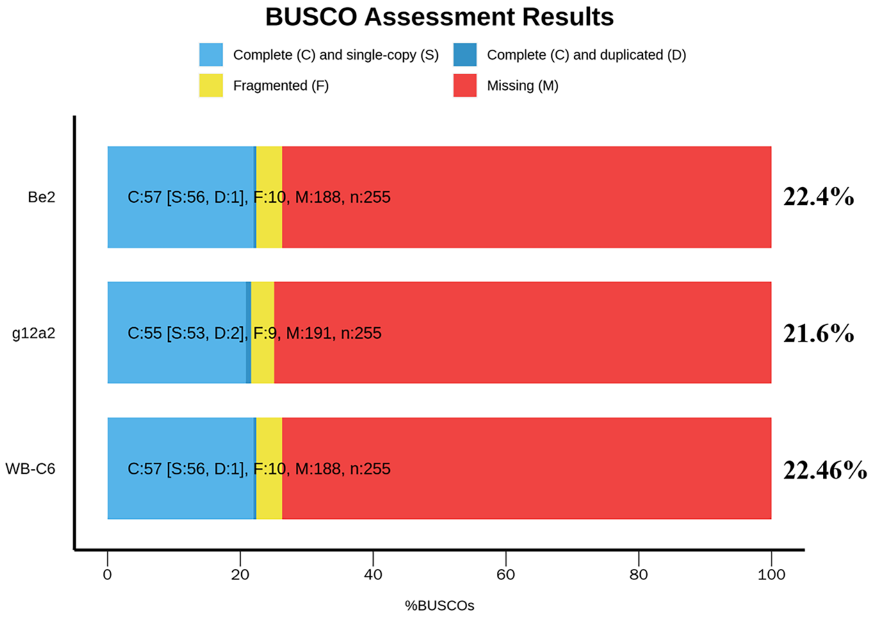

Fig. 5 Comparison of Benchmarking Universal Single-Copy Orthologs (BUSCO) evaluation results of g12a2 (this study) with reference results of Be2 and WB-C6.

fragmentation[41]. In this study, the N50 of contigs for the novel assembly of isolate g12a2 was 3.1 Mbp, which was higher than the N50 of the reference genome of WB-C6 (2.8 Mbp) and that of the Be2 isolate (3.0 Mbp).

Benchmarking Universal Single-Copy Orthologs (BUSCO, v 5.2.2)[42] was used to evaluate the integrity of genomes g12a2, Be2, and WB-C6 with the parameters "-m geno -l eukaryota_odb10 -offline -i" based on the eukaryotic database updated on January 8, 2024. It was reported that although BUSCO was not an optimal tool in diplomonads since many of the conserved proteins used in the BUSCO analyses have been lost[43], the analysis was conducted for isolate g12a2 in this study. A BUSCO analysis was carried out on the genome assemblies of g12a2, WB-C6, and Be2, and the completeness percentages were 21.6%, 22.4%, and 22.4%, respectively. The BUSCO output scores were plotted and visually presented using the ggplot2 library available via the R Tidyverse package (Fig. 5).

In this study, a chromosome-scale genome assembly of *G. duodenalis* was obtained through long-read sequencing of ten trophozoites. The quality of novel genome sequencing and assembly was comprehensively evaluated using multiple methods, including genome coverage detection, synteny analysis with the reference genome, and BUSCO evaluation. This study provides alternative opportunities for high-quality genome sequencing of free-living microorganisms, as well as those that are non-culturable or hard to cultivate *in vitro* in the short term.

Code availability

No custom code was used in this study. The data analyses used standard bioinformatic tools specified in the methods.

Data availability

All of the data generated or analyzed during this study are included in this manuscript.

Received: 20 December 2024; Accepted: 16 June 2025;
Published online: 01 July 2025

References

1. Alja'fari, J. *et al*. Assessing human-source microbial contamination of stormwater in the U.S. *Water Res* **268**, 122640, https://doi.org/10.1016/j.watres.2024.122640 (2024).
2. Jingjing, S. *et al*. Assessment of potential zoonotic transmission of *Giardia duodenalis* from dogs and cats. *One health* **17**, 100651, https://doi.org/10.1016/j.onehlt.2023.100651 (2023).
3. Natali, L. *et al*. The exosome-like vesicles of *Giardia* assemblages A, B, and E are involved in the delivering of distinct small RNA from parasite to parasite. *Int J Mol Sci* **24**, 9559, https://doi.org/10.3390/ijms24119559 (2023).
4. Fusaro, C. *et al*. Burden and epidemiology of human intestinal *Giardia duodenalis* infection in Colombia: a systematic review. *Trop Med Infect Dis* **7**, 325, https://doi.org/10.3390/tropicalmed7100325 (2022).
5. Morrison, H. G. *et al*. Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. *Science* **317**, 1921–1926, https://doi.org/10.1126/science.1143837 (2007).
6. Wei, Y. *et al*. Advances in *Giardia duodenalis* genomics. *Chin J Zoonoses* **40**, 977–983, https://doi.org/10.3969/i.issn.1002-2694.2024.00.140 (2024).
7. de Paula Baptista, R. *et al*. Comparative genomics of *Giardia duodenalis* sub-assemblage AI beaver (Be-2) and human (WB-C6) strains show remarkable homozygosity, sequence similarity, and conservation of VSP genes. *Sci Rep* **4**, 13582, https://doi.org/10.1038/s41598-024-63783-5 (2024).
8. Franzén, O. *et al*. 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 (2009).
9. Maloney, J. G. *et al*. A hybrid sequencing and assembly strategy for generating culture free *Giardia* genomes. *Curr Res Microb Sci* **3**, 100114, https://doi.org/10.1016/j.crmicr.2022.100114 (2022).
10. Kooyman, F. N. J., Wagenaar, J. A. & Zomer, A. 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 (2019).
11. Klotz, C. *et al*. Highly contiguous genomes of human clinical isolates of *Giardia duodenalis* reveal assemblage- and sub-assemblage-specific presence-absence variation in protein-coding genes. *Microb. Genom* **9**, mgen000963, https://doi.org/10.1099/mgen.0.000963 (2023).
12. Seabolt, M. H., Roellig, D. M. & Konstantinidis, K. T. Genomic comparisons confirm *Giardia duodenalis* sub-assemblage AII as a unique species. *Front Cell Infect Microbiol* **12**, 1010244, https://doi.org/10.3389/fcimb.2022.1010244 (2022).
13. Pollo, S. M. J. *et al*. 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 (2020).
14. Xu, F., Jex, A. & Svärd, S. G. A chromosome-scale reference genome for *Giardia intestinalis* WB. *Sci Data* **7**, 38, https://doi.org/10.1038/s41597-020-0377-y (2020).
15. Pan, H. *et al*. Single-cell genomics reveals a novel cell state during smooth muscle cell phenotypic switching and potential therapeutic targets for atherosclerosis in mouse and human. *Circulation* **142**, 2060–2075, https://doi.org/10.1161/CIRCULATIONAHA.120.048378 (2020).
16. Lloréns-Rico, V. *et al*. Single-cell approaches in human microbiome research. *Cell* **185**, 2725–2738, https://doi.org/10.1016/j.cell.2022.06.040 (2022).
17. Lyons, A., Brown, J. & Davenport, K. M. Single-cell sequencing technology in ruminant livestock: challenges and opportunities. *Curr Issues Mol Biol* **46**, 5291–5306, https://doi.org/10.3390/cimb46060316 (2024).
18. Hazzard, B. *et al*. Single-cell analyses of polyclonal *Plasmodium vivax* infections and their consequences on parasite transmission. *Nat Commun* **15**, 7625, https://doi.org/10.1038/s41467-024-51949-8 (2024).
19. Blake, D. P. *et al*. Population, genetic, and antigenic diversity of the apicomplexan *Eimeria tenella* and their relevance to vaccine development. *Proc Natl Acad Sci USA* **112**, E5343–50, https://doi.org/10.1073/pnas.1506468112 (2015).
20. Agyabeng-Dadzie, F., Xiao, R. & Kissinger, J. C. *Cryptosporidium* genomics- current understanding, advances, and applications. *Curr Trop Med Rep* **11**, 92–103, https://doi.org/10.1007/s40475-024-00318-y (2024).
21. Wang, P. & Wang, F. A proposed metric set for evaluation of genome assembly quality. *Trends Genet* **39**, 175–186, https://doi.org/10.1016/j.tig.2022.10.005 (2023).
22. Aalam, A. A. *et al*. Use of advanced platelet-rich fibrin for the treatment of multiple adjacent mucogingival recessions: a technical report of the fibrin-assisted soft-tissue promotion protocol. *J Esthet Restor Dent* **2024**, 19, https://doi.org/10.1111/jerd.13341 (2024).

23. Glotov, O. S. *et al*. The benefits of whole-exome sequencing in the differential diagnosis of hypophosphatasia. *Int J Mol Sci* **25**, 11728, https://doi.org/10.3390/ijms252111728 (2024).
24. Koren, S. *et al*. 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 (2017).
25. Xiao, C. L. *et al*. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat Methods* **14**, 1072–1074, https://doi.org/10.1038/nmeth.4432 (2017).
26. Kurtz, S. *et al*. Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12, https://doi.org/10.1186/gb-2004-5-2-r12 (2004).
27. Makino, J. *et al*. Fast and accurate short-read alignment with hybrid hash-tree data structure. *Genomics Inform* **22**, 19, https://doi.org/10.1186/s44342-024-00012-5 (2024).
28. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963, https://doi.org/10.1371/journal.pone.0112963 (2014).
29. Alonge, M. *et al*. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol* **23**, 258, https://doi.org/10.1186/s13059-022-02823-7 (2022).
30. *NGDC/CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA021161 (2025).
31. *NGDC/CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA021160 (2025).
32. *NGDC/CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA021157 (2025).
33. *NGDC/CNCB* https://ngdc.cncb.ac.cn/gwh/Assembly/88048/show (2025).
34. *NGDC/CNCB* https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA033502 (2025).
35. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_049639855.1 (2025).
36. Sun, J. Genome annotation data for the *Giardia duodenalis* g12a2, figshare. *Dataset* https://doi.org/10.6084/m9.figshare.28935134.v1 (2025).
37. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_026248805.1 (2022).
38. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_000002435.2 (2019).
39. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14**, e1005944, https://doi.org/10.1371/journal.pcbi.1005944 (2018).
40. Zhou, Z. W. *et al*. GenomeSyn: a bioinformatics tool for visualizing genome synteny and structural variations. *J Genet Genomics* **49**, 1174–1176, https://doi.org/10.1016/j.jgg.2022.03.013 (2022).
41. Jung, H. *et al*. Comparative evaluation of genome assemblers from long-read sequencing for plants and crops. *J Agric Food Chem* **68**, 7670–7677, https://doi.org/10.1021/acs.jafc.0c01647 (2020).
42. Manni, M. *et al*. BUSCO: Assessing genomic data quality and beyond. *Curr Protoc* **1**, e323, https://doi.org/10.1002/cpz1.323 (2021).
43. Akdeniz, Z. *et al*. The expanded genome of *Hexamita inflata*, a free-living diplomonad. *Sci Data* **12**, 192, https://doi.org/10.1038/s41597-025-04514-x (2025).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32102698), National Key Research and Development Program of China (2023YFD1801200, 2023YFD18010000), Henan Provincial Key Research and Development Program, China (231111111500). We sincerely acknowledge the contributions of the Protist 10,000 Genomes Project (P10K) consortium members, whose expert suggestions refined our methodological approach and data interpretation.

Author contributions

Conceived and designed: J.Q.L. and L.X.Z. Manuscript: J.J.S., Y.C.C. and J.Q.L. Analysis: J.J.S., Y.C.C., Y.Z.W., K.H.Z., F.Y., Z.Y., and X.Y.Z. Reagents/materials: J.J.S., Y.C.C., J.Q.L. and S.M.Z. Supervision: J.Q.L. and L.X.Z. All of the authors have read and approved the fnal manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.Z. or J.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.