



OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of *Trichomonas vaginalis* strain IR-78 (ATCC 50138)

Mostafa Y. Abdel-Glil^{1,2}✉, Johannes Solle¹, Heinrich Neubauer¹ & Lisa D. Sprague¹✉

Trichomonas vaginalis is a flagellated protozoan parasite and the causative agent of trichomoniasis, the most prevalent non-viral sexually transmitted infection, with significant impact on public health and economy. Here, we present a chromosome-level genome of *T. vaginalis* strain IR-78 (ATCC 50138), assembled using short and long-read sequencing technologies as well as chromatin conformation capture sequencing (Hi-C). The assembled genome is 173.3 Mb in size, with an N50 scaffold length of 26.4 Mb. The assembly is anchored in six super-scaffolds, corresponding to the six chromosomes of *T. vaginalis*, covering 93.5% of the genome. A total of 43,326 protein-coding genes were predicted, of which 95.4% were annotated using multiple databases, including Gene Ontology, eggNOG, and KEGG. Among these, 16,656 genes encoded hypothetical proteins. This high-quality genome assembly provides an additional resource for ongoing research on *T. vaginalis* pathogenesis, drug resistance, and host-parasite interactions. It furthermore enables comparative genomic studies to assess genomic variety across different *Trichomonas* species.

Background & Summary

Trichomonas vaginalis is a protozoan parasite that causes trichomoniasis, the most common non-viral sexually transmitted infection (STI), affecting approximately 156 million people worldwide annually (WHO, 2025)¹. This parasite mainly infects the urogenital tract and symptoms range from mild irritation to severe inflammation. However, a considerable proportion of infections are asymptomatic, which contributes to its widespread transmission^{2,3}. *T. vaginalis* can cause serious health complications in women if left untreated, with complications including infertility, premature birth, pelvic inflammatory disease, cervical cancer and increased vulnerability to HIV. In men, the infection is frequently asymptomatic, prostatitis and urethritis may occur¹. The public health implications of *T. vaginalis* are particularly severe in low- and middle-income countries in which access to healthcare and programmes for preventing sexually transmitted diseases is limited.

The true global burden of *T. vaginalis* is underestimated for numerous reasons among which are: inadequate understanding of its epidemiology and public health implications, asymptomatic carriers, inadequate diagnostic capabilities in low-resource settings and because *T. vaginalis* infection is not a notifiable disease^{4,5}. Efforts to contain its spread face challenges, such as the lack of routine screening in many healthcare systems, the stigma associated with sexually transmitted diseases, and the parasite's ability to persist in asymptomatic carriers. The problem is exacerbated by the emergence of drug-resistant *T. vaginalis* strains, which complicates treatment and prolongs transmission and infection cycles. Addressing these issues requires improved diagnostics, routine screening programmes and the development of new therapeutic options to fight resistance⁴.

The genome of *T. vaginalis* is characterised by its size and complexity⁶; it is approximately 160 Mb in size and consists to a significant extent of repetitive sequences and transposable elements. This genomic diversity may play a role in the parasite's ability to evade the host's immune system and to develop drug resistance⁷. However, only a limited number of *T. vaginalis* genome sequences are currently available in public repositories. Genome sequencing of further strains will enable comparative analyses and could reveal strain-specific patterns of pathogenicity.

This work describes the construction of a chromosome-level genome assembly for *T. vaginalis* IR-78 (ATCC 50138) using Illumina short-read, Nanopore long-read, and Hi-C sequencing techniques. The final genome

¹Friedrich-Loeffler-Institut, Institut für Bakterielle Infektionen und Zoonosen (IBIZ), Naumburger Str. 96a, 07743, Jena, Germany. ²Jena University Hospital – Friedrich Schiller University, Institute for Infectious Diseases and Infection Control, Jena, Germany. ✉e-mail: mostafa.abdelglil@fli.de; lisa.sprague@fli.de

assembly of strain IR-78 is 173.3 Mb in size with a N50 scaffold length of 26.4 Mb (Fig. 1, Table 1). The six super-scaffolds, representing the genome's six chromosomes, account for 162 Mb of the total genome size. The analysis revealed 43,326 predicted protein-coding genes.

Methods

***T. vaginalis* IR-78 sequencing.** Genomic DNA was extracted from *T. vaginalis* IR-78 cultured in InPouch® TF-Bovine (Megacor, Austria) at a concentration of 1.5×10^6 cells/mL. The extraction process utilized UNSET lysis buffer, followed by phenol-chloroform purification as described by Horner *et al.* (1996). DNA quality was evaluated using a NanoDrop Spectrophotometer (Thermo Fisher, USA), while quantification was performed with a Qubit Fluorometer (Invitrogen, USA).

Long-read sequencing was conducted with the R9.4.1 chemistry from Oxford Nanopore Technologies (ONT) on a GridION platform, using the SQK-LSK109 ligation kit. Data acquisition and basecalling were conducted via Guppy 6.5.7 using the sup-accurate model, yielding 3.65 million reads totalling 5.21 billion bases. Trimming and filtering of these ONT reads were performed with Porechop (default mode) and Filtrlong (-min_length 1000), producing a refined dataset of 850,194 reads with an average length of 4.7 kb and a cumulative base count of 4 billion bases.

Short-read sequencing was carried out on an Illumina MiSeq platform using the XT DNA Library kit, generating 300 bp paired-end reads. This resulted in 52.7 million raw reads, corresponding to 13.9 billion bases. Adapter trimming and quality control using fastp v0.23.2⁸ improved the data quality, reducing the dataset to 51.4 million high-quality reads with a total of 13.2 billion bases and a Q30 score of 86.3%.

***De novo* genome assembly.** The genome assembly was performed using a hybrid approach based on long-read assembly with subsequent polishing using short reads to enhance both contiguity and accuracy. The assembly utilized the tricho-workflow pipeline (available at https://gitlab.com/FLI_Bioinfo/tricho-workflow, accessed on 12.11.2024), which has previously been successfully applied to *Trichomonas foetus*.

Nanopore long reads, filtered to include only those with a minimum length of 1000 bases, were used for the initial assembly with Flye v2.9⁹. The draft assembly underwent multiple rounds of polishing. Two iterations of Racon v1.5.0¹⁰ with Minimap2 v2.24¹¹ were applied to refine the assembly using the long-read data. This was followed by two additional polishing rounds with Polypolish with bwa v0.50¹² using high-quality, fastp-processed Illumina reads. Illumina data corrected 116,679 substitution errors in the draft assembly, achieving a final consensus quality of 99.87%. Nanopore reads were then mapped back to the polished contigs using Minimap2, and mapping quality was assessed with Qualimap v2.2.2a¹³. Contigs with an average read depth below 10-fold were excluded, resulting in a final set of 423 contigs and a total genome size of 173.27 Mb, which were subsequently prepared for scaffolding.

Hi-C scaffolding. To generate a chromosome-level genome assembly, Hi-C proximity ligation data were utilized. This was achieved through *in-situ* Hi-C sequencing, performed with the EpiTect Hi-C Kit (Qiagen). Sequencing was conducted using the Illumina MiSeq platform, yielding 43.49 million raw reads, equivalent to 13.09 billion bases. After filtration with fastp (v0.23.2), a refined dataset of 27.4 million clean reads, representing 7.4 billion bases, was retained for downstream analysis. These reads were critical for identifying spatially interacting genomic regions, enabling the reconstruction of chromosomal structures. The ARIMA Genomics Hi-C Mapping Pipeline (https://github.com/ArimaGenomics/mapping_pipeline; accessed February 2024) was employed for read alignment and interaction identification. Hi-C reads were aligned using BWA-MEM v0.7.17¹⁴ in single-end mode, and chimeric or ligation-spanning reads were trimmed to ensure accurate mapping. Mapped reads were filtered based on quality thresholds using Samtools v1.19.2¹⁵, and PCR duplicates were removed with Picard v3.5.3¹⁶.

To assemble the genome into candidate chromosomes, YAHS v1.2a.2¹⁷ was utilized. This assembly process transformed the draft genome into an organized, chromosome-scale structure. The resulting Hi-C maps, generated during assembly, were visualized using HapHic Plot v1.0.3¹⁸ (Fig. 1).

The final genome assembly of *T. vaginalis* IR-78 was 173.3 Mb in total. Hi-C analysis confirmed the presence of six super-scaffolds of 14, 17, 21, 26, 35, and 45 Mb in size, corresponding to the six chromosomes previously identified in *T. vaginalis*¹⁹ (Fig. 1). These super-scaffolds collectively anchored 93.5% of the total genome assembly (162.1 Mb). Key metrics of the assembly included a scaffold N50 of 26.4 Mb and a contig N50 of 1.13 Mb, highlighting the assembly's high contiguity. The average GC content across the genome was calculated to be 32.79% (Table 1, Fig. 2), consistent with previous findings for *T. vaginalis* genomes. This assembly is comparable in size to the *T. vaginalis* G3 strain (assembly name NYU_TvagG3_2, accession GCF_026262505.1)²⁰ (Table S1). Despite differences in assembly techniques and sequencing technologies, the IR-78 genome exhibits contiguity comparable to the G3 strain²⁰, thus providing a further reliable resource for future comparative genomic studies investigating genomic variation and functionality across *T. vaginalis* and other *Trichomonas* spp. strains.

Repeat annotation. Tandem repeats in the *T. vaginalis* IR-78 genome were predicted using the Tandem Repeats Finder tool v4.09²¹, which identifies and classifies repeat motifs based on sequence similarity. To annotate other repetitive elements, we used RepeatModeler2 v2.0.5²², a *de novo* automated tool designed to construct species-specific transposable element (TE) libraries. These libraries were then used by RepeatMasker v4.1.6 (<https://www.repeatmasker.org/>) to identify both known and novel TEs in the *T. vaginalis* IR-78 genome²³. The *T. vaginalis* IR-78 genome is highly enriched for repetitive elements, with a total of 115.3 Mb (66.57% of the genome) classified as repeat sequences. RepeatMasker classified the predicted repetitive sequences as follows: DNA transposons were the most prevalent, comprising approximately 35% of the total genome. Retroelements

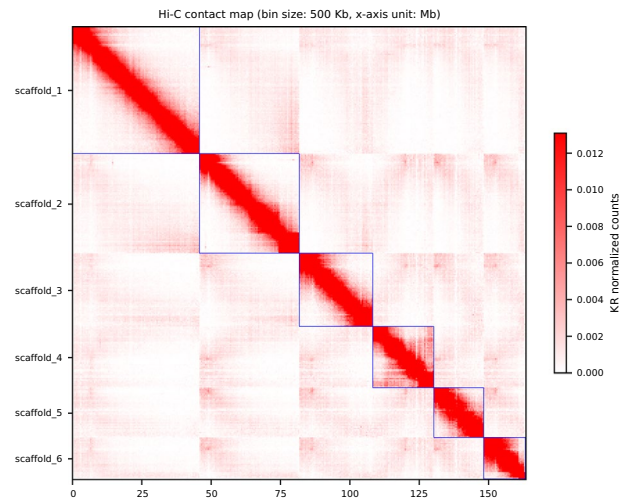


Fig. 1 Hi-C contact map of the *Trichomonas vaginalis* IR-78 genome assembly. Chromosomal scaffolds are depicted. Interaction frequencies of Hi-C data are indicated by the color gradient bar on the right.

Metric	Value
Total length of assembly	173,3 Mb
Total length of chromosomes (Mb)	162,17 Mb (93,5%)
Total number of contigs	204
G + C content (%)	32,79%
Average coverage (with long reads)	30,2-fold
N50 scaffold/contig length	26,4/1,13 Mb
n. protein-coding genes	43326
Average gene length (bp)	1204
n. tRNA genes	460

Table 1. Genome statistics of *Trichomonas vaginalis* strain IR-78.

and long terminal repeat (LTR) elements each represented 0.97%, while simple repeats comprised 0.36%. A significant proportion of the repetitive elements (29.89%) remained unclassified, likely due to the limited research on repetitive sequences in Parabasalia species. Minor proportions of repetitive elements were attributed to rolling-circle elements and short interspersed nuclear elements (SINEs).

As these annotations were generated using automated tools and current repeat databases, which are known to be of limited quality for *T. vaginalis*, the presence of some predicted repeat categories—such as retroelements, LTR elements, rolling-circle elements, and SINEs—should be interpreted with caution. Manual annotation and validation of these sequences are required to confirm their presence and classification, and to rule out potential artifacts arising from automated prediction. As tools for annotation and prediction continue to advance, the classification and robust identification of repetitive elements in *T. vaginalis* genomes are expected to improve, potentially supported by extensive manual curation.

Annotation of non-coding RNA genes. Homologues of structural RNAs in the *T. vaginalis* IR-78 genome were identified using Infernal v1.1.5²⁴ to perform alignments with the Rfam RNA database v14.9²⁵ (<http://rfam.xfam.org/>). The search was carried out with cmscan v1.1.5²⁴, revealing 681 high-scoring hits corresponding to structural RNAs, with 460 transfer RNAs (tRNAs) and 165 ribosomal RNAs (rRNAs) (Table 1, Table S2).

Gene prediction. Protein-coding genes in the *T. vaginalis* IR-78 genome were predicted *de novo* using BRAKER v3.0.8^{26–36}. Prior to gene prediction, the assembled genome was soft-masked for repetitive elements using RepeatMasker^{22,23} with the Dfam libraries, in order to reduce spurious gene calls from transposable elements. BRAKER was run in genome-annotation mode using protein homology evidence from the OrthoDB v11 database³⁷, which provided conserved orthologous sequences across eukaryotes to guide gene structure inference. Gene models were trained iteratively within BRAKER, and parameters were optimized based on both *ab initio* prediction and external protein alignments. Using this approach, a total of 43,326 protein-coding genes were predicted in the *T. vaginalis* IR-78 genome, with an average gene length of 1,204 base pairs. The genome annotation features are summarized in Table 1 and plotted across the six chromosomes of the *T. vaginalis* IR-78 genome in Fig. 2.

It should be noted that the *de novo* gene prediction workflow applied in this study was not supported by experimental transcriptomic or proteomic evidence, which may render the resulting annotation susceptible

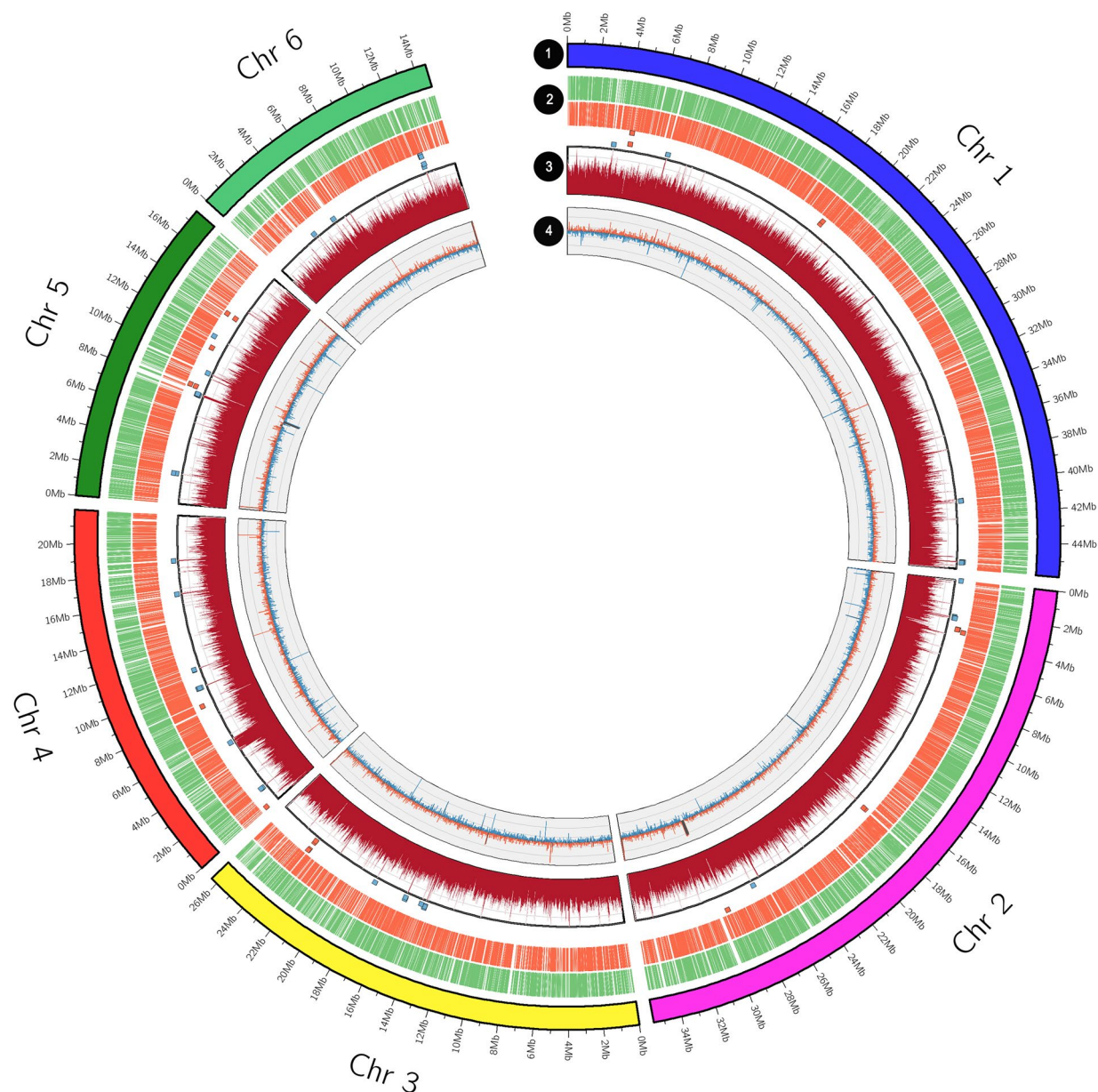


Fig. 2 Circos plot illustrating the chromosomal architecture of the *Trichomonas vaginalis* IR-78 genome. Circle 1 shows the six major scaffolds with chromosome lengths indicated in megabases. Circle 2 displays annotated gene sequences in clockwise and anticlockwise orientations. Circle 3 presents the mapping profile of Nanopore long-read data across the chromosomes, with regions of deviated coverage variation highlighted. Circle 4 represents the GC content along the genome.

to artifacts. Consequently, additional validation will be required to confirm the accuracy of the predicted gene structures. Notably, the automated annotation workflow identified 3961 introns (average length = 232 bp), whereas previous studies have characterized *T. vaginalis* as a largely intron-poor organism^{38–40}.

Functional genome annotation. The functional annotation of the *T. vaginalis* IR-78 genome was carried out using several databases to obtain insights into gene function and biological pathways. The predicted protein sequences of 43,326 genes were analysed with DIAMOND Blastp³⁵ (e-value cutoff: 1.0e-3) within the Functional Analysis Module of OmicsBox v3.0.30⁴¹. Comparisons were made with the databases NR protein database, SWISS-PROT⁴², Gene Ontology (GO)⁴³, eggNOG⁴⁴ and KEGG⁴⁵. To complement the BLAST-based functional annotation, InterProScan v5.69-101.0⁴⁶ was used to identify conserved protein domains and motifs using the EMBL-EBI version of InterPro⁴⁷. This allowed a detailed analysis of predicted protein function (Table S3).

The annotation results showed good matches in the above-mentioned functional databases. 42,776 predicted protein sequences had matches in the NR database, 42,663 in RefSeq, and 12,748 in SWISS-PROT. InterProScan identified functional domains in 31,664 predicted protein sequences, which corresponds to 73.1%

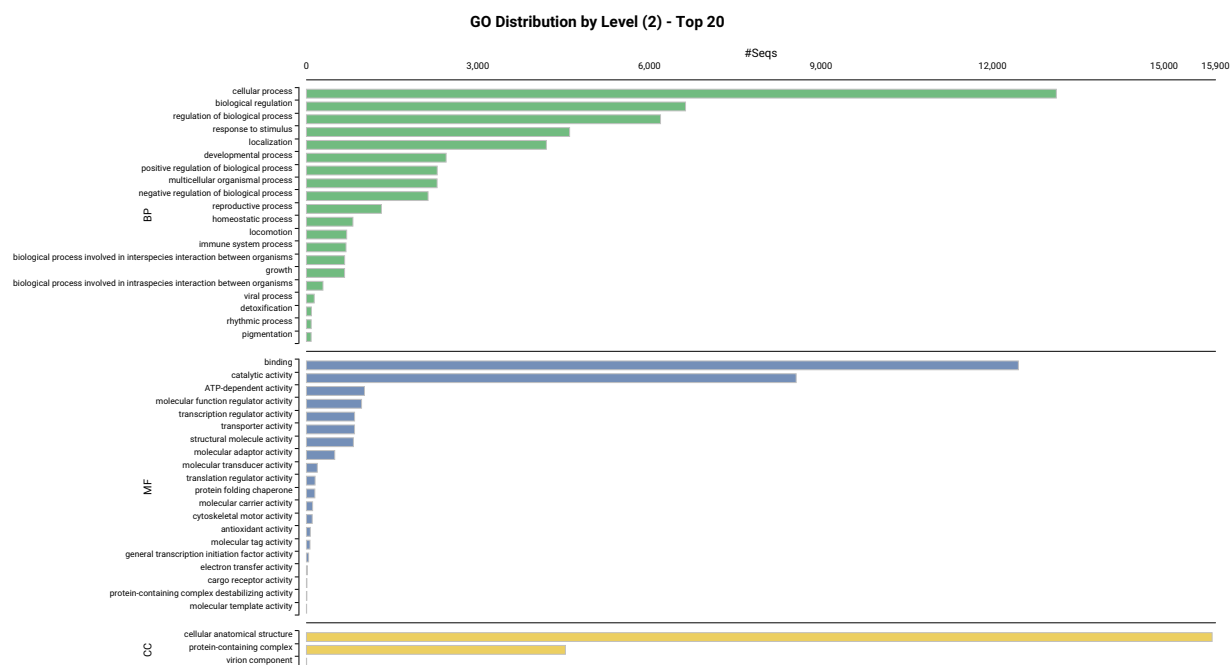


Fig. 3 Gene Ontology (GO) categorization of functionally annotated genes in the *Trichomonas vaginalis* IR-78 genome. The chart summarizes the distribution of gene functions across biological process (green), molecular function (blue), and cellular component (yellow) domains, based on GO term assignments.

of the predicted proteome. Assignment to Gene Ontology (GO) terms revealed that 21,038 predicted protein sequences (48.5%) could be assigned to specific biological processes, molecular functions and cellular components. 10,015 sequences (23.1%) were associated with enzyme code (EC) annotations, linking these predicted proteins to specific enzymatic activities. Further pathway analysis using eggNOG revealed that 22,149 sequences (51.1%) were associated with orthologous groups (OG) and functional categories. The KEGG database identified 414 metabolic and signalling pathways, encompassing 9,488 sequences associated with these pathways. Reactome analysis⁴⁸ identified a further 2,274 metabolic pathways, encompassing 1,472 protein sequences.

Gene Ontology classification of the functionally annotated genes revealed that they are distributed across three primary GO domains: biological process, molecular function and cellular component. Of the annotated genes, 32% (n = 13,899) were associated with 'biological process', 40.6% (n = 17,612) with 'Molecular Function' and 41% (n = 17,783) with 'cellular component' (Fig. 3). In the 'Biological Process' category, the GO terms at level 2 'cellular process' (39%, n = 16,064) and 'metabolic process' (30%, n = 13,115) were the most common classes. In the 'Cellular Component', the predominant GO terms were 'cellular anatomical entity' (36.5%, n = 15,840), which represents genes associated with structural and organelle functions, and 'protein complex' (10.4%, n = 4,529). The most abundant GO classes in the 'molecular function' class included 'binding' (28.7%, n = 12,451) and 'catalytic activity' (19.7%, n = 8,565), indicating a significant presence of enzymes that drive the organism's biochemical processes (Fig. 3 and Tables S3, S4).

Data Records

All raw data of the whole genome have been deposited at the National Center for Biotechnology Information (NCBI) under the SRA accession number SRP589323⁴⁹. The completed genome assembly is available in GenBank under the accession number GCA_052324655.1⁵⁰. The corresponding annotation data have been deposited on Zenodo and can be accessed via the following link: <https://zenodo.org/records/16924891>⁵¹.

Technical Validation

Quality assessment of the genome assembly. To validate the genome assembly, QUAST⁵² was utilised to report assembly contiguity metrics. Additionally, we calculated the percentage of long reads mapping to the final assembly. For this, nanopore data were mapped using minimap2, and mapping quality was assessed with Qualimap¹³. Genome completeness was evaluated using two approaches. First, we used BUSCO v5.5.0⁵³ with the eukaryota odb10 dataset, which assesses genome completeness by searching for conserved single-copy orthologs across eukaryotic species. The analysis identified 127 complete BUSCOs, corresponding to a completeness score of 49.8%, consistent with ranges reported for other Parabasalia genomes^{54,55}. BUSCO's focus on evolutionarily conserved orthologs makes it a reliable benchmark for cross-species comparisons for genome integrity. However, the relatively low BUSCO score reflects the absence of representative genes for highly divergent protists in the reference set, rather than a true deficiency of the assembly. Currently, no universally appropriate benchmark exists for genome completeness in such evolutionary distant lineages.

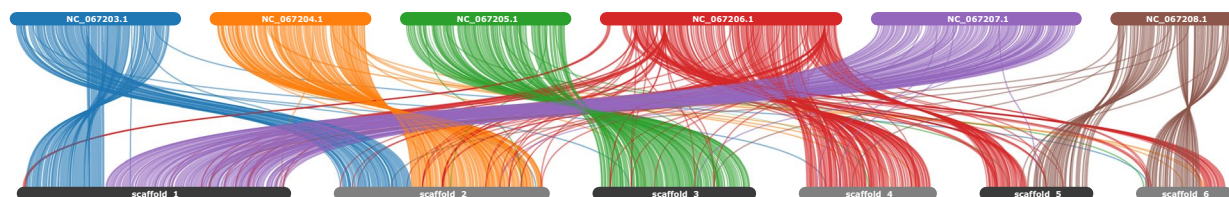


Fig. 4 Synteny analysis between the *Trichomonas vaginalis* IR-78 sequenced in this study (bottom) and the reference genome of *T. vaginalis* G3 strain (top). Conserved genomic regions are displayed as colored links, illustrating collinear blocks and highlighting the extent of genome conservation and rearrangement between the two strains.

Additionally, we employed OMARK v0.3.0⁵⁶, a hierarchical orthology analysis tool designed to detect conserved Hierarchical Orthologous Groups (HOGs) across eukaryotic lineages. Using the Eukaryota ancestral clade, OMARK estimated a completeness score of 71.26%, identifying 709 of the 995 conserved HOGs. Among these, 32.96% were classified as single-copy, 38.29% as duplicate, and 28.74% as missing. Of the 43,325 predicted proteins, 14,897 (34.38%) showed consistent lineage placement, reflecting a substantial alignment with established phylogenetic expectations.

Chromosome synteny. To evaluate the structural completeness and accuracy of the genome assembly of *T. vaginalis* strain IR-78 (ATCC 50138), we conducted a synteny analysis against the reference genome of strain G3 (NC_067203.1 – NC_067208.1)^{20,57}. Chromosome synteny was generated using MCScanX v1.0.0⁵⁸ with default parameters and visualized with SynVisio (no version)⁵⁹. The resulting plot (Fig. 4) shows a high degree of collinearity and conservation between the two strains, as expected for intraspecies comparisons. However, several ribbons span across different chromosomes may suggest either genuine structural variations (e.g., inversions or rearrangements) or assembly-related fragmentation in the IR-78 genome.

Data availability

The datasets generated and analyzed in this study are publicly available. The raw sequencing data have been deposited at NCBI under the SRA accession number SRP589323, BioProject accession number PRJNA1271105, and BioSample accession number SAMN48852979. The completed genome assembly is available in GenBank under the accession number GCA_052324655.1. The corresponding genome annotation is accessible via Zenodo at <https://zenodo.org/records/16924891>.

Code availability

All software and tools in this study were used with their default parameters, unless otherwise stated.

Received: 20 June 2025; Accepted: 23 October 2025;

Published online: 05 November 2025

References

1. WHO. World Health Organization (WHO) <https://www.who.int/news-room/fact-sheets/detail/trichomoniasis>, Accessed 02.09.2025 (2025).
2. Tian, W. *et al.* Systematic review and meta-analysis of the global prevalence and infection risk factors of *Trichomonas vaginalis*. *Parasite* **32**, 56, <https://doi.org/10.1051/parasite/2025051> (2025).
3. Ozdemir, E., Keleştemur, N. & Kaplan, M. *Trichomonas vaginalis* as a rare cause of male factor infertility at a hospital in East Anatolia. *Andrologia* **43**, 283–285, <https://doi.org/10.1111/j.1439-0272.2010.01119.x> (2011).
4. Ibáñez-Escribano, A. & Nogal-Ruiz, J. J. The Past, Present, and Future in the Diagnosis of a Neglected Sexually Transmitted Infection: Trichomoniasis. *Pathogens* **13**, <https://doi.org/10.3390/pathogens13020126> (2024).
5. Harfouche, M. *et al.* Epidemiology of *Trichomonas vaginalis* infection in the Middle East and North Africa: systematic review, meta-analyses, and meta-regressions. *EBioMedicine* **106**, 105250, <https://doi.org/10.1016/j.ebiom.2024.105250> (2024).
6. Carlton, J. M. *et al.* Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science (New York, N.Y.)* **315**, 207–212, <https://doi.org/10.1126/science.1132894> (2007).
7. Mohammadi, K. *et al.* Unexpected intron plasticity and trans-splicing capability suggest spliceosome diversification in the evolutionarily divergent protozoan parasite *Trichomonas vaginalis*. *Int J Parasitol.* <https://doi.org/10.1016/j.ijpara.2025.08.011> (2025).
8. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, <https://doi.org/10.1093/bioinformatics/bty560> (2018).
9. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology* **37**, 540–546, <https://doi.org/10.1038/s41587-019-0072-8> (2019).
10. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research* **27**, 737–746, <https://doi.org/10.1101/gr.214270.116> (2017).
11. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, <https://doi.org/10.1093/bioinformatics/bty191> (2018).
12. Wick, R. R. & Holt, K. E. Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLOS Computational Biology* **18**, e1009802, <https://doi.org/10.1371/journal.pcbi.1009802> (2022).
13. García-Alcalde, F. *et al.* Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28**, 2678–2679, <https://doi.org/10.1093/bioinformatics/bts503> (2012).
14. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, <https://doi.org/10.1093/bioinformatics/btp324> (2009).

15. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, <https://doi.org/10.1093/bioinformatics/btp352> (2009).
16. Broad Institute of MIT and Harvard. Picard: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. <https://broadinstitute.github.io/picard> (2014).
17. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, <https://doi.org/10.1093/bioinformatics/btac808> (2022).
18. Zeng, X. *et al.* Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nature Plants* **10**, 1184–1200, <https://doi.org/10.1038/s41477-024-01755-3> (2024).
19. Zubáková, Z., Cimbůrek, Z. & Tachezy, J. Comparative analysis of trichomonad genome sizes and karyotypes. *Mol Biochem Parasitol* **161**, 49–54, <https://doi.org/10.1016/j.molbiopara.2008.06.004> (2008).
20. Sullivan, S. A., Carlton, J. M. & Schatz, M. C. A new chromosome-level assembly of the sexually transmitted parasite *Trichomonas vaginalis*. G3 genome assembly NYU_TvagG3_2. Whole genome shotgun sequencing project. *GenBank* https://identifiers.org/ncbi/insdc:gca:GCA_026262505.1 (2023).
21. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**, 573–580, <https://doi.org/10.1093/nar/27.2.573> (1999).
22. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457, <https://doi.org/10.1073/pnas.1921046117> (2020).
23. Smith, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. *RepeatMasker Open-4.0* (2013).
24. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337, <https://doi.org/10.1093/bioinformatics/btp157> (2009).
25. Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic acids research* **49**, D192–D200, <https://doi.org/10.1093/nar/gkaa1047> (2020).
26. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769, <https://doi.org/10.1093/bioinformatics/btv661> (2016).
27. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform* **3**, lqaa108, <https://doi.org/10.1093/nargab/lqaa108> (2021).
28. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-Genome Annotation with BRAKER. *Methods Mol Biol* **1962**, 65–95, https://doi.org/10.1007/978-1-4939-9173-0_5 (2019).
29. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform* **2**, lqaa026, <https://doi.org/10.1093/nargab/lqaa026> (2020).
30. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566, <https://doi.org/10.1186/s12859-021-04482-0> (2021).
31. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62, <https://doi.org/10.1186/1471-2105-7-62> (2006).
32. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644, <https://doi.org/10.1093/bioinformatics/btn013> (2008).
33. Iwata, H. & Gotoh, O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic acids research* **40**, e161, <https://doi.org/10.1093/nar/gks708> (2012).
34. Gotoh, O. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic acids research* **36**, 2630–2638, <https://doi.org/10.1093/nar/gkn105> (2008).
35. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60, <https://doi.org/10.1038/nmeth.3176> (2015).
36. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic acids research* **33**, 6494–6506, <https://doi.org/10.1093/nar/gki937> (2005).
37. Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic acids research* **51**, D445–D451, <https://doi.org/10.1093/nar/gkac998> (2023).
38. Callejas-Hernández, F., Shiratori, M., Sullivan, S. A., Blow, F. & Carlton, J. M. Redefining the spliceosomal introns of the sexually transmitted parasite *Trichomonas vaginalis* and its close relative in columbid birds. *PLOS Pathogens* **21**, e1013282, <https://doi.org/10.1371/journal.ppat.1013282> (2025).
39. Deng, X. L. *et al.* A 25-bp ancient spliceosomal intron in the TvRab1a gene of *Trichomonas vaginalis*. *Int J Biochem Cell Biol* **41**, 417–423, <https://doi.org/10.1016/j.biocel.2008.09.014> (2009).
40. Vanáčová, S., Yan, W., Carlton, J. M. & Johnson, P. J. Spliceosomal introns in the deep-branching eukaryote *Trichomonas vaginalis*. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 4430–4435, <https://doi.org/10.1073/pnas.0407500102> (2005).
41. Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research* **36**, 3420–3435, <https://doi.org/10.1093/nar/gkn176> (2008).
42. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research* **28**, 45–48, <https://doi.org/10.1093/nar/28.1.45> (2000).
43. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature genetics* **25**, 25–29, <https://doi.org/10.1038/75556> (2000).
44. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research* **47**, D309–D314, <https://doi.org/10.1093/nar/gky1085> (2019).
45. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27–30, <https://doi.org/10.1093/nar/28.1.27> (2000).
46. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, <https://doi.org/10.1093/bioinformatics/btu031> (2014).
47. Paysan-Lafosse, T. *et al.* InterPro in 2022. *Nucleic acids research* **51**, D418–D427, <https://doi.org/10.1093/nar/gkac993> (2022).
48. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic acids research* **48**, D498–D503, <https://doi.org/10.1093/nar/gkz1031> (2020).
49. NCBI Sequence Read Archive. <https://identifiers.org/ncbi/insdc.sra:SRP589323> (2025).
50. Abdel-Glil, M., Solle, J., Neubauer, H. & Sprague, L. Chromosome-level genome assembly of *Trichomonas vaginalis* strain IR-78 (ATCC 50138), assembly FLI_Tvag_1.1. Whole genome shotgun sequencing project. *GenBank* https://identifiers.org/ncbi/insdc:gca:GCA_052324655.1 (2025).
51. Abdel-Glil, M. Chromosome-level genome assembly of *Trichomonas vaginalis* strain IR-78 (ATCC 50138). *Zenodo* <https://doi.org/10.5281/zenodo.16924891> (2025).
52. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUASt: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075, <https://doi.org/10.1093/bioinformatics/btt086> (2013).
53. Manni, M., Berkeley, M. R., Seppely, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **38**, 4647–4654, <https://doi.org/10.1093/molbev/msab199> (2021).

54. Alves-Ferreira, E. V. C. *et al.* Whole Genome Sequence of the gut commensal protist *Tritrichomonas musculus* isolated from laboratory mice. *Scientific Data* **12**, 590, <https://doi.org/10.1038/s41597-025-04921-0> (2025).
55. Abdel-Glil, M. Y., Solle, J., Wibberg, D., Neubauer, H. & Sprague, L. D. Chromosome-level genome assembly of *Tritrichomonas foetus*, the causative agent of Bovine Trichomonosis. *Scientific Data* **11**, 1030, <https://doi.org/10.1038/s41597-024-03818-8> (2024).
56. Nevers, Y. *et al.* Quality assessment of gene repertoire annotations with OMArk. *Nature biotechnology* <https://doi.org/10.1038/s41587-024-02147-w> (2024).
57. Sullivan, S. A. *et al.* Comparative genomics of the parasite *Trichomonas vaginalis* reveals genes involved in spillover from birds to humans. *Nature Communications* **16**, 6487, <https://doi.org/10.1038/s41467-025-61483-w> (2025).
58. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene syntenic and collinearity. *Nucleic acids research* **40**, e49–e49, <https://doi.org/10.1093/nar/gkr1293> (2012).
59. Bandi, V. & Gutwin, C. Interactive exploration of genomic conservation. *Proceedings of the 46th Graphics Interface Conference on Proceedings of Graphics Interface 2020 (GI'20)*. Canadian Human-Computer Communications Society, Waterloo, CAN. (2020).

Author contributions

M.Y.A.-G.: Next Generation Sequencing, bioinformatics, data interpretation and writing of the manuscript, J.S.: Next Generation sequencing, Hi-C sequencing and data analysis; H.N.: data interpretation and manuscript revision; L.D.S.: Conceptualisation, project supervision, cultivation of *T. vaginalis* and DNA preparation; data interpretation, and writing of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-025-06182-3>.

Correspondence and requests for materials should be addressed to M.Y.A.-G. or L.D.S.

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.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025