Parasites & Vectors

https://doi.org/10.1186/s13071-026-07378-1

Article in Press

De novo genome assembly and annotation of *Gnathostoma spinigerum*

Received: 19 Dec 2025

Accepted: 16 Mar 2026

Published online: 11 April 2026

Cite this article as: González-Bertolín, B., Monzón, S., Zaballos, Á. *et al.* De novo genome assembly and annotation of *Gnathostoma spinigerum* . *Parasites Vectors* (2026). https://doi.org/10.1186/s13071-026-07378-1

Belén González-Bertolín, Sara Monzón, Ángel Zaballos, Pilar Jiménez, Paron Dekumyoy, Poom Adisakwattana, Sarai Varona, Isabel Cuesta, Sunita Sumanam, Javier Sotillo, Ana Hernández-González, Iñigo Marcos-Alcalde, Paulino Gómez-Puertas, Neil Young & Maria Perteguer

De novo genome assembly and annotation of *Gnathostoma spinigerum*

Belén González-Bertolín[1,2,3*], Sara Monzón[4], Ángel Zaballos[5], Pilar Jiménez[5], Paron Dekumyoy[6], Poom Adisakwattana[6], Sarai Varona[4], Isabel Cuesta[4], Sunita B. Sumanam[7], Javier Sotillo[1], Ana Hernández-González[1], Iñigo Marcos-Alcalde[8], Paulino Gómez-Puertas[8], Neil D. Young[7‡], Maria J. Perteguer[1,9*‡]

[1]Helminths Unit, Parasitology Reference and Research Laboratory, National Centre for Microbiology, Instituto de Salud Carlos III, Majadahonda, Madrid, Spain.

[2]Eicosanoids Research Division, Institute of Biomedicine and Molecular Genetics (IBGM), Consejo Superior de Investigaciones Científicas (CSIC), Valladolid, Spain.

[3] CIBER de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Instituto de Salud Carlos III, Madrid, Spain.

[4]Bioinformatics Unit, Core Scientific and Technical Units, Instituto de Salud Carlos III, Madrid, Spain.

[5]Genomics Unit, Core Scientific and Technical Units, Instituto de Salud Carlos III, Madrid, Spain.

[6]Department of Helminthology, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand.

[7]Department of Veterinary Biosciences, Melbourne Veterinary School, Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, Victoria, Australia.

[8]Molecular Modeling Group, Centro de Biologia Molecular Severo Ochoa (CBM, CSIC-UAM), E-28049 Madrid, Spain

[9]CIBER de Enfermedades Infecciosas (CIBERINFEC), Institute of Health Carlos III, Madrid, Spain.

*Correspondence: belengbertolin@gmail.com and chus.perteguer@isciii.es

[‡]Neil D Young and Maria J. Perteguer contributed equally to this work

Abstract

Background

Gnathostoma spinigerum is a parasitic nematode implicated in human cases of eosinophilic meningitis. This species is mainly endemic to Thailand and frequently occurs in other geographical regions as imported cases. Despite the fact this parasite poses a significant pathogenic risk, its genome has not yet been assembled nor annotated. The aim of our study is to generate the first genome assembly of *G. spinigerum*.

Methods

Whole-genome sequence libraries were generated from genomic DNA extracted from a pooled sample of advanced stage 3 larvae. After sequencing, the assembly of the genome was produced using a combination of second and third generation sequencing technologies. Multiple draft assemblies were generated and evaluated, and the absence of contamination was determined. An identification and modelling of new *G. spinigerum* metalloproteinases and tissue inhibitors of metalloproteinases was performed, and molecular dynamics simulations were used to analyze their potential interactions. The final assembly was annotated and made publicly available via the NCBI genome database.

Results

The hybrid assembly approach (using short and long reads) using the SPAdes assembler and with post-assembly polishing (Pilon/Picard) yielded the most complete genome (222 Mb genome size, N50=14,149 bp, 69.6% BUSCO assembly). A total of 14,451 protein-

coding genes were predicted in the *G. spinigerum* genome with 62.3% BUSCO annotation. 3D computational modelling and molecular dynamics of six new metalloproteinases and two new tissue inhibitors of metalloproteinases are presented.

Conclusions

We provide the first sequence assembly and annotation for the nematode *G. spinigerum*. This draft genome will be an essential resource for future scientific and applied investigations of diseases caused by this parasite.

Keywords:

Gnathostoma; *G. spinigerum*; Genome; *de novo* assembly, annotation; M10 metalloproteinases; tissue inhibitors of metalloproteinases; TIMPs.

Background

Parasitic infection by *Gnathostoma spinigerum* (gnathostomiasis) can lead to a severe condition of cerebral involvement, potentially triggering eosinophilic meningitis or meningoencephalitis caused by advanced third-stage larvae (L3) bp. Efforts to develop effective diagnostics and therapeutics for gnathostomiasis have been limited, partly, by the lack of genomic information available for this parasite.

In 1836, Richard Owen described the genus *Gnathostoma* spp. following the discovery of adult parasites in the stomach of a Bengal tiger (*Panthera tigris*), naming the species *G. spinigerum* [3]. Humans are incidental hosts who become infected mainly by eating raw or undercooked freshwater fish containing L3 larvae [4]. In humans, the larvae are unable to develop into adults and can migrate through various tissues and organs [5,6]. The most common symptoms are skin-related, but the most severe form of visceral involvement is invasion of the central nervous system (CNS), including brain and

spinal cord, which can lead to death [7]. Members of the genus *Gnathostoma* are distributed worldwide, but *G. spinigerum* is the only species currently linked with cases of cerebral involvement. Gnathostomiasis has been reported in Japan, India, China, and Southeast Asia, including Thailand, Laos, Myanmar, Indonesia, Malaysia, and the Philippines [3]. The common incidence of *G. spinigerum* infection underscores the significant public health concern posed by gnathostomiasis, with a notable concentration of human cases reported in Thailand [8]. Moreover, gnathostomiasis has emerged as a sporadic yet alarming phenomenon in non-endemic regions, fueling concerns about its potential for global dissemination. Travel-related cases have been documented in Europe, Australia, and the Americas, reflecting the disease's ability to transcend geographical boundaries [2,9].

The absence of expedited therapeutic intervention in gnathostomiasis can result in severe and prolonged morbidity. The parasite, with a longevity in the human host that may exceed a decade, is responsible for asymptomatic intervals followed by symptomatic recurrences. The clinical sequelae are significant, and involvement of the CNS confers a risk of fatal outcome. The treatments of choice (albendazole and/or ivermectin with corticosteroids) do not ensure a complete cure, with several case relapses documented. No definitive mechanisms of resistance to available drugs have been described for *G. spinigerum* so far [10].

A reference genome for *G. spinigerum* will provide a much-needed molecular resource for gnathostomiasis diagnostic and therapeutic advancements. Our study aims to fill this knowledge gap by achieving a draft genome assembly of *G. spinigerum*.

Methods

DNA extraction of parasitic larvae

Gnathostoma spinigerum larvae were collected from the livers of naturally infected eels (*Monopterus albus*) in Thailand. Briefly, livers of freshwater eels were extracted and digested with 1% hydrochloric acid-pepsin at 37°C water bath for 3 hours. Larvae were identified by microscopy and washed several times with 85% saline and distilled water. A total of 250 advanced third-stage larvae (L3), corresponding to human infective forms, were kept at -80°C.

DNA extraction was conducted, employing automated methodologies (QIAcube), with the QIAamp DNA Mini Kit combined with the ATL buffer (QIAgen) for 48-hour at 56ºC.

Each *G. spinigerum* larva was taxonomically verified by sequencing mitochondrial (COI) and ribosomal (ITS) genes using a conventional polymerase chain reaction (PCR) protocol previously published [11]. Both PCR amplicons were sequenced using the Sanger technique on a 3730XL DNA Analyzer sequencer (Applied Biosystems).

Library preparation and sequencing

Two sequencing technologies were employed: Illumina and Oxford Nanopore Technologies (ONT). In the case of Illumina, a total of four samples were prepared and two different library preparation kits were utilised: the Nextera DNA Flex Library Prep kit (Illumina) with eight amplification cycles, in accordance with the manufacturer's instructions. The samples used with this library were ILGs-1 and ILGs-2 (n = 10 pooled larvae in each library). Subsequently, the libraries were quantified with the Promega QuantiFluor® ONE dsDNA System and their average size determined with the 2100 Bioanalyzer High Sensitivity DNA Analysis (Agilent). These first two samples were sequenced in a NextSeq 550 (Illumina), using the sequencing kit NextSeq 500/550 High

Output Kit v2.5 (300 cycles) (Illumina). The second sequencing kit employed was DNA PCR-Free Prep, Tagmentation (Illumina), which did not require amplification. The samples used for this second Illumina modality were ILGs-3 and ILGs-4 (n = 51 larvae per library). The second set of samples were sequenced using a NovaSeq sequencer (Illumina) with the NovaSeq 6000 SP Reagent Kit v1.5 (300 cycles) (Illumina) sequencing kit. In both cases, the reads obtained with the Illumina technology were paired-end and 150 bp in length.

Prior to sequencing the long reads, the libraries were prepared using the Ligation Sequencing Kit SQK-LSK110 (Oxford Nanopore Technologies) and the NEBNext Companion Module for Oxford Nanopore Technologies Ligation Sequencing (New England Biolabs), in accordance with the manufacturer's instructions. ONTGs-5 sample consisted of n = 162 pooled larvae. In the case of the long read approach, the sequencer employed was the MinION MK1C (Oxford Nanopore Technologies), while the flow cell was (R9.4.1) FLO-MIN106D (Oxford Nanopore Technologies). Basecalling of raw FAST5 files was performed using the program Guppy v5.1.12 and reads were stored in FASTQ format.

Genome assembly approaches

All bioinformatics analyses were performed on a high-performance computing (HPC) cluster from the Instituto de Salud Carlos III.

The FastQC software (v.0.11.9) [12] was employed to assess the quality of the FASTQ files and any necessary read filtering steps were performed using the fastp program (v.0.20.0) [13].

The *de novo* assembly of the genome was performed individually using: i) short reads from the Illumina (NextSeq) technology with the following assembler: SGA

(v.0.10.15) [14], SPAdes (v.3.15.4) [15], Soap de novo 2 (v.2.04) [16], Platanus (v.1.2.4) [17], ABySS (v.2.3.5) [18], and Unicycler (v.0.5.0) [19], ii) short reads from the Illumina (NovaSEQ) technology with SPAdes (v.3.15.4) and iii) long reads from the ONT technology with Flye (v.2.9) [20] and NECAT (v.0.0.1) [21]. The quality evaluation of the different assemblies was performed with QUAST(v. 5.0.2) [22] and BUSCO (v. 5.3.2) in genome mode. [23] The workflow of short reads approach is summarized in Figure 1.

After obtaining draft genome assemblies using each sequencing technology individually, hybrid assemblies were then constructed. Specifically, short reads were integrated with long reads to create hybrid assemblies. Two distinct approaches were employed for this purpose. A workflow of the whole process is shown in Figure 2. Two assembly approaches were used: one a hybrid SPAdes assembly of both Illumina and ONT reads, and the other a Flye assembly of long reads that was then polished with Illumina reads as follows. Reads were mapped to the genome using Bowtie2 (v.2.4.2) [24]. Alignment files were converted from SAM to BAM format using SAMtools (v.1.12) [25]. Duplicates were removed using the Picard program (v.2.25.1) [26]. The contigs were then corrected using final BAM, the Flye genome assembly and Pilon (v.1.23) [27]. The same evaluation procedure was used for the generated draft assemblies, with BUSCO and QUAST.

Further information related to the HPC, genome assembly software and approaches is included in Additional file 1 (text 1).

Structural and functional annotation

Repeat element annotation was carried out using RepeatModeler (v.2.0.5) [28], RepeatMasker (v. 4.1.6) and the DFAM repeat database [29].

Protein-coding gene annotation of the final draft genome assembly was performed using the Braker pipeline (v3.0.8) [30]. Gene evidence for Braker was obtained by genome mapping metazoan protein sequences and RNAseq data of *G. spinigerum* previously published (NCBI accession number SRR8137628) [31]. The gene finders selected were GeneMark-ET and AUGUSTUS, and the gene sets were combined with TSEBRA. Functional annotation of inferred proteins was performed using eggNOG-mapper (v 2.1.12) based on eggNOG data [32] and InterproScan (v. 5.56) [33]. BUSCO (protein mode) and OMArk software [34] (v.0.3.0) were used to measure proteome completeness. The latter software characterizes the consistency of the protein coding genes after annotation with the genome assembly and transcriptome, identifying also the presence of contamination from other species.

Mitochondrial genome assembly comparison

A comparison was conducted with the previously published mitochondrial genome of *G. spinigerum* [35] to assess the quality of the assembled genome. This evaluation involved aligning the draft assembly exhibiting the best metrics with the corresponding mitochondrial genome sequence using the command-line BLAST tool (v.2.11.0). To identify gene families associated with host-parasite interactions, we performed targeted searches for matrix metalloproteinases (MMPs) and SCP/TAPS proteins. The latter were identified by searching the predicted proteome against the Pfam CAP domain profile (PF00188) using hmmsearch (HMMER v3.4).

Structural modeling of putative metalloproteinases M10 and tissue inhibitors of metalloproteinases (TIMPs) identified in the genome. Molecular Dynamics simulation of M10-TIMP interaction.

Identification of the proteins corresponding to sequences in the *Gnathostoma spinigerum* genome predicted as putative metalloproteinases M10 and/or tissue inhibitors of metalloproteinases (TIMPs) was performed searching for conserved domains. We have performed genome-wide identification of M10 metalloproteinases and TIMPs in the *G. spinigerum* protein-coding gene set using Hidden Markov Models (HMMER v3.3.2) against the Pfam-A database (PF00413 for M10 and PF00965 for TIMPs). Hits with less than 60% domain coverage were excluded. Structural modeling of the newly identified metalloproteinases and TIMPS, as well as the only *G. spinigrum* metalloproteinase M10 described so far in Genbank (AAF82802), was performed using as template the crystal structure of the human proMMP-9 catalytic domain (Protein Data Bank id: 7OGT) [36], the human pro-matrix metalloproteinase-2 (Protein Data Bank id: 1CK7) [37] or the human tissue inhibitor of metalloproteinases-2 (Protein Data Bank id: 1BR9) [38]. The structures were modeled by combining residue positions obtained using Phyre 2.2 [39] and SwissModel [40].

The interaction complexes between the seven M10 metalloproteinases and the two TIMPs (in total, 14 models) were obtained using the structure of the complex between the human tissue inhibitor of metalloproteinases-1 and the matrix metalloproteinase-3 catalytic domain (Protein Data Bank id: 6N9D) [41] as a template. Structure complexes were subjected to 200 ns of unrestrained Molecular Dynamics (MD) simulation using the Amber18 package (https://ambermd.org; University of California-San Francisco, CA), essentially as previously described [42]. Briefly, after solvation, initial wild-type and variant model structures were subjected to 10,000 cycles of energy minimization, followed by a 1 ns restrained equilibration phase in which the temperature was smoothly raised to 297 K, after which the restraints were gradually removed over 10 ns. Each system was then subjected to a 200 ns free MD production phase.

Trajectories were analyzed using cpptraj [43] and VMD [44]. The NAMD Energy Plugging of VMD was used to evaluate nonbonding energy contributions of the surface interaction between M10 metalloproteinases and TIMPs along the simulations using NAMD [45]. Plots were generated using Pymol (https://pymol.org).

Detailed arguments for programs described in this section are given in Additional file 1, Text 1.

Results

Microscopical and molecular classification of *Gnathostoma* larvae

Microscopic observation confirmed that the advanced 3 larvae (AL3) belonged to the *Gnathostoma spinigerum* species (Fig 3). The *G. spinigerum* characteristics follow the criteria of a cephalic bulb bearing four rows of hooklets, a long muscular esophagus and four cervical sacs (Fig.3A). A higher magnification of the anterior end of the body is displayed in Figure 3B with the two lips located at the extreme. The cephalic bulb hooklets have an oblong shape, and a sharp-pointed end at the base. The number of hooklets are 40 or more in each row.

The ribosomal ITS2 and mitochondrial (COI) PCR amplified DNA products were 650 and 250 bp, respectively. Upon sequencing, the BLAST results indicated 100% identity with the species *G. spinigerum* (GenBank accession number MK033974.1).

DNA extraction and library evaluation

The DNA concentrations of the different samples after extraction were as follows: 0.539 ng/µL (ILGs-1), 0.882 ng/µL (ILGs-2), 0.7 ng/µL (ILGs-3), 2.4 ng/µL (ILGs-4) and 5.5 ng/µL (ONTGs-5). After construction of the short-read libraries, the concentrations were: 9.5 ng/µL (ILGs-1), 2.8 ng/µL (ILGs-2), 0.243 ng/µL (ILGs-3) and 0.975 ng/µL (ILGs-

4). For the long-read sample (ONTGs-5), 258.5 ng of genomic DNA was used to make the library.

Sequencing and pre-processing

Illumina short-read DNA libraries were constructed from four out of the five samples processed, resulting in the sequencing of paired-end 2x150 bp reads for samples ILGs-1, ILGs-2, ILGs-3, and ILGs-4. Raw and filtered reads for each library, including the platform used to sequence the data, are detailed in supplementary material (Table S1).

The Illumina-NextSeq sequenced ILGs-1 and ILGs-2 samples had a guanine-cytosine content of 37%. Samples ILGs-3 and ILGs-4, sequenced using Illumina NovaSeq technology, yielded a content of guanine-cytosine in the range of 37-39%. The length range of the sequenced reads was 35-151 bp. Sample ONTGs-5, sequenced by MinION (ONT), bases passed sequencer inner filters were 2.19 Gb and estimated 6-fold coverage of the predicted genome size. The size length range was 89 - 55,853 bp. The percentage of guanine-cytosine content was 35%.

Following pre-processing of the Illumina NextSeq samples with the fastp programme, the percentage of guanine-cytosine content was found to be identical to that observed prior to processing, at 37%. The length range of the sequences was 100-151 bp, resulting in a reduction in the number of reads (Additional file 2: Table S1). The processed readings of the Illumina NovaSeq samples exhibited a range of guanine-cytosine content, with values ranging from 37 to 38%. Following pre-processing, the total number of reads was 58 million and 253 million (Additional file 2: Table S1).

Assembly evaluation

Using Illumina-NextSeq short-read data only, different assemblers generated distinct draft assemblies (Table 1).

Table 1. Metrics of the Illumina-NextSeq sample assembly drafts from a selection of QUAST and BUSCO data

Sample	**Assembler**	**Contigs**	**Total length** (pb)	**N50** (pb)	**GC** (%)	**BUSCO*** (%)
ILGs-1	SPAdes	73,799	254,431,934	6,747	37.91	54.1
ILGs-1	SoapDeNovo 2	35,667	25,836,549	687	38.87	1.2
ILGs-1	Platanus	52,833	84,533,917	2,118	38.6	20.6
ILGs-1	SGA	104,106	77,372,582	695	38.87	2.7
ILGs-1	AbySS	155,591	138,089,355	851	38.48	8.3
ILGs-1	Unicycler	180,012	257,638,835	1,752	38	19.3
ILGs-2	SGA	36935	24,302,369	624	38.57	0.6
ILGs-2	SPAdes	188,774	320,734,975	2,477	38.02	31,1
ILGs-2	SoapDeNovo 2	31,720	23,225,428	691	38.9	1.2

*Complete BUSCOs genome mode.

The SPAdes assembler was inferred to be the best, as evidenced by the higher N50 value observed in sample ILGs-1 (6,457) and sample ILGs-2 (2,477). Furthermore, the SPAdes assembly yielded the most accurate BUSCO results, with 54.1% observed in sample ILGs-1 and 31.1% in sample ILGs-2.

Next, Illumina-NovaSeq short-read data only was assembled using SPAdes (Table 2).

Table 2. Metrics of draft assemblies obtained with SPAdes and Illumina sequencing samples

Sample	**Sequencer**	**Contigs**	**Total length** (pb)	**N50** (pb)	**GC** (%)	**BUSCO*** (%)
ILGs-3 SPAdes	NovaSeq	72,841	225,114,680	5,733	37.5	50.2

ILGs-4 SPAdes	NovaSeq	120,494	316,541,976	5,442	37.62	54.8
ILGs-1 SPAdes	NextSeq	73,799	254,431,934	6,747	37.91	54.1
ILGs-2 SPAdes	NextSeq	188,774	320,734,975	2,477	38.02	31.1

*Complete BUSCOs genome mode, nematoda db.

Hybrid approach

The optimal assemblies for the read and long read sequencing using the SPAdes assembler in the hybrid mode (based on short reads) were sample ILGs-3 (Illumina NovaSeq) and sample ONTGs-5 (ONT) with the fewest contigs assembled (72,960), the highest N50 value (9,834), the highest percentage of single BUSCOs (58.7%) (Additional file 2: Table S2) and the lowest percentage of duplicated, fragmented and missing BUSCO genes (Additional file 2: Table S3). The hybrid approach has improved the metrics of the individual assemblies of both short and long read based assemblies (Additional file 2: Table S2). In the long-read based approach, the best result is obtained for samples ONTGs-5 and ILGs-1 with Flye, Bowtie2, Samtools, Pilon and Picard, although the metrics of both BUSCO and QUAST are very similar between the different assemblies.

Table 3. Assembly draft metrics, after combining different assembly approaches, obtained with QUAST and BUSCO

Samples	**Assembler**	**Contigs**	**Total length** (pb)	**N50** (pb)	**GC** (%)	**BUSCO*** (%)
ILGs-1 *Pilon/Picard ONTGs-5-ILGs-1*	SPAdes *Untrusted*	105,154	288,742,554	6,608	37.77	58.1
ILGs-3 *Pilon/Picard ONTGs-5-ILGs-1*	SPAdes *Untrusted*	80,021	269,148,826	8,645	37.67	58.3

ILGs-3 *Pilon ONTGs-5- ILGs-1*	SPAdes *Untrusted*	80,014	269,151,703	8,646	37.67	59.1
ILGs-1 *Pilon ONTGs-5-ILGs-1*	SPAdes *Trusted*	100,998	299,435,360	7,492	37.8	57.5
(V1) ILGs-3 - ONTGs-5 *Pilon/Picard ONTGs-5-ILGs-1*	SPAdes hybrid *Untrusted*	70,253	264,481,023	11,416	37.64	61.4
(V1.2) ILGs-3 - ONTGs-5 *Pilon/Picard ONTGs-5-ILGs-1*	SPAdes hybrid *Untrusted* Soft-masked purged	32,589	222,261,784	14,149	37.34	69.6

*Complete BUSCO genome mode.

The samples with the best statistics were ILGs-3 and ILGs-1 (Illumina) with the SPAdes assembler. Combining with the long read sample (ONTGs-5, ONT) improves the assembly metrics, as does using the untrusted parameter with assemblies from the hybrid approach (combination of long and short reads, based on long reads) and subsequent hardening with other programs. A resume of assemblies metrics is displayed in Additional file 2: Figures S1 and S2.

The draft ILGs-3/ONTGs-5 SPAdes hybrid Pilon/Picard ONTGs-5/ILGs-1 assembly was referred to as version 1 (V1). After soft-masking and purging, it was renamed version 1.2 (V1.2).

Mitochondrial genome and structural and functional annotation

The published mitochondrial genome is 14,079 bp in size. A total of 17 scaffolds of the version assembly were aligned against the published mitochondrial genome, with almost 90% of the reference mitochondrial genome being assembled (Additional file 2: Figure S3).

The *G. spinigerum* V1.1 draft genome was 264.48 Mb in size (NCBI BioProject accession number PRJNA1141240) and assembled into 70,253 scaffolds (N50: 11,416 bp). We assembled with RNAseq data; the number of genes was 27,592. Genome annotation produced 30,047 predicted protein coding sequences. After purging repetitive elements and duplicates, a total of 11,866 genes.

The *G. spinigerum* V1.2 draft genome was 222.26 Mb in size ((NCBI BioProject accession number PRJNA1141240) and assembled into 32,589 scaffolds (N50: 14,149 bp).

The number of proteins was 11,865, the total consistent lineage placement was 7,249 (61.10%), the number total contaminants was 0 (table 5 and figure 4). The clade most consistent with the taxonomic distribution of gene families was Spirurina.

Table 4. Summary features of genome assemblies for *Gnathostoma spinigerum*

Assembly version	Assembly size (Mb)	Scaffolds	N50	GC (%)	BUSCO genome (%)	Genes	BUSCO annotation (%)
v1.1	264.48	70,253	11,416	37.64	61.4	27,592	62.3
v1.2	222.26	32,589	14,149	37.34	69.6	11,866	62.6

The percentage of proteome and conserved HOGs of *G. spinigerum* annotated assembly is displayed in figure 4.

Table 5. Features of the genome of *Gnathostoma spinigerum* V1.2

Number of genes/mRNA	11,866
Gene length[a]	2,786 ± 6,238
mRNA length[a] **Coding domain length[a]**	921 ± 841 921 ± 841
Number of exons	70,087

Exon length[a]	921 ± 841
Protein length[a]	307 ± 280
Completeness:	
Complete BUSCOs[b]	2,179 (69.6%)
Complete single-copy BUSCOs	2,141 (68.4%)
Complete and duplicated BUSCOs	38 (1.2%)
Fragmented BUSCOs	306 (9.8%)
Missing BUSCOs	646 (20.6%)
OMArk completeness (conserved HOGs):	
Single	3,551 (67.27%)
Duplicated	390 (7.39%)
Missing	1,338 (25.35%)
OMArk consistency[c]	
Total consistent lineage placements	7,249 (61.10%)
Total inconsistent lineage placements	748 (6.30%)
Contaminants	0
Unknown	3,868 (32.60%)

[a]Lengths (bp); mean ± standard deviation.

[b]Number of BUSCOs identified (in genome mode) using nematoda_odb10 dataset (3131 genes).

[c]Proportion of annotated protein-coding genes in the metazoan proteome.

EggNOG-mapper annotated 11,071 (93.30%) of predicted proteins, there was a predominance of genes involved in biological process and cellular components. The most abundant COG categories were “Function unknown (S)” (22.7%), “Signal transduction mechanisms (T)” (10.7%) and “Post-translational, modification, protein turnover, chaperones (O)” (8.8%).

Table 6. Metrics of the annotated assembly with EggNOG

Metric	Value
Total proteins	11,071 (93.30%)
Annotated proteins	10,282 (92.00%)
Proteins with GO terms	7,662 (74.00%)

Proteins with KEGG terms	7,398 (71.00%)

*GO: gene ontology

A total of 10 gene models containing the CAP domain (PF00188) were identified, representing the SCP/TAPS family in *G. spinigerum.*

Functional annotation of the *G. spinigerum* protein-coding set revealed a diverse repertoire of M10 matrix metalloproteinases (MMPs). Based on their domain architecture and comparative analysis with host and other nematode orthologs, these enzymes were classified into three distinct structural groups: one containing those with Signal Peptide, Propeptide, Catalytic, Hinge, and Hemopexin-like domains; another containing sequences with only Signal Peptide, Propeptide, and Catalytic domains; and finally, another group with a single sequence containing only an apparently incomplete catalytic domain (Table 7).

Table 7. Domain architecture and predicted localization of *G. spinigerum* Matrix Metalloproteinases (MMPs)

Sequence ID (*G.spinigerum* genome)	**Key Domains Present**	**Key Features**	**MMP type**	**Secretion**
G9182.T2	Signal Peptide + Propeptide + Catalytic + Hinge + Hemopexin-like	Resembles human MMP-2/9	Typical Nematode MMP	Secreted
G1280.T1	Signal Peptide + Propeptide + Catalytic + Hinge + Hemopexin-like	Resembles human MMP-2/9	Typical Nematode MMP	Secreted
G9438.T1	Signal Peptide + Propeptide + Catalytic + Hinge + Hemopexin-like	Resembles human MMP-2/9	Typical Nematode MMP	Secreted
G12031.T2	Signal Peptide + Propeptide + Catalytic	Similar to *Anisakis* matrilysins, lacks the hemopexin-like and the hinge domain	MMP-7-like	Secreted

AAF82802	Signal Peptide + Propeptide + Catalytic	Similar to *Anisakis* matrilysins, lacks the hemopexin-like and the hinge domain	MMP-7-like	Secreted
G8944.T1	Signal Peptide + Propeptide + Catalytic	Similar to *Anisakis* matrilysins, lacks the hemopexin-like and the hinge domain	MMP-7-like	Secreted
G7562.T1	Catalytic	Possible intracellular form or fragmented gene model	"Minimal" MMP	Non-secreted

Computational simulation of the interaction between putative metalloproteinases M10 and tissue inhibitors of metalloproteinases (TIMPs).

After identifying the sequences of six new M10 metalloproteinases (G12031.T2, G9182.T2, G1280.T1, G9438.T1, G8944.T1, and G7562.T1) and two tissue inhibitors of metalloproteinases (TIMPs) (G7222.T1 and G1596.T1), molecular modeling and molecular dynamics techniques were used to study the possible interactions between the M10 metalloproteinases and the TIMPs. The previously described M10 metalloproteinase (AAF82802) was also included in the computational simulation. First, structural models of the M10 metalloproteinases (Figure 5A) and TIMPs (Figure 5B) were generated. The M10 metalloproteinases can be divided into two groups depending on the length of the sequence found in the genome. The sequences of the proteins AAF82802, G7562.T1, G12031.T2, and G8944.T1 mainly correspond to the catalytic domain (in magenta in Figure 5A). The sequences of the proteins G9438.T1 and G1280.T1, and G9182.T2 contain both the catalytic domain (in magenta in Figure 5A) and the propeptide and the Hemopexin domains (in gray and ochre, respectively, in Figure 5A). The G7562.T1

protein does not appear to have a complete catalytic domain based on its sequence length. However, both possible Zn^{2+} and Ca^{2+} binding sites appear to be present. Moreover, after 200 ns of unrestricted molecular dynamics simulation, the overall structure of the domain did not change, suggesting that it may function as a catalytically active protein.

After generating the individual models, we proceeded to generate models of possible interaction complexes between M10 metalloproteinases and TIMPs (for a total of 14 models), using the structure of a known homologous complex [41] as a template. Figure 5B illustrates the interaction model between the metalloproteinase AAF82802 and the two TIMPs, for example. We then subjected these models to 200 ns of Molecular Dynamics simulation to evaluate the stability of each metalloproteinase's interaction with the two TIMPs. The interaction energy was obtained along the trajectories using NAMD [45], as shown in Figure 5C. Finally, the energy values of each model during the first 100 ns of the simulation were discarded, assuming stabilization of the complexes during this time. The values from the remaining 100 ns were then averaged to obtain mean values +/- STD (Figure 5D).

The energy analysis appears to categorize the metalloproteinase-TIMP pairs into three groups. The first group consists of the metalloproteinases G8944.T1 and G12031.T2, which show modest interaction energy values between their active centers and both TIMPs. The second group consists of the metalloproteinases G1280.T1, G7562.T1, G9182.T2, and G9438.T1. This group shows a clear preference for TIMP G1596.T1 (blue bars in Figure 5D) over TIMP G7222.T1 (green bars). The third group consists of a single member: the metalloproteinase AAF82802. This metalloproteinase shows the opposite behavior, with lower interaction energy (indicative of higher affinity) with G7222.T1 than with G1596.T1.

Discussion

The present study reports the first complete genome assembly and annotation of *G. spinigerum*, a nematode parasite that causes severe or even fatal eosinophilic meningitis in humans. The genome was sequenced, assembled and annotated using a hybrid assembly approach and will be a valuable resource for the characterization of the genetic basis of the parasite and the identification of potential drug and diagnostic candidates.

A pool of larvae was employed instead of a single individual, as the requisite concentration could not be achieved with a single larva. For the Illumina sequencing, the number of individuals was 51 and 10 and with ONT, 162 larvae were utilized. Other *de novo* genome assemblies have used a pool of larvae with successful results, such as *Wuchereria bancrofti*, *Anisakis simplex*, *Enterobius vermicularis* and *Brugia pahangi* [46]. However, it can be reasonably assumed that this factor contributes to elevated rates of heterozygosity and a high number of SNPs. These factors—assembly fragmentation and high genetic diversity—likely influence the total number of genes identified in our study. Because we sequenced a pool of many larvae, the assembly software may have occasionally struggled to tell the difference between two versions of the same gene (alleles) and instead reported them as two separate genes. This can lead to an overestimation of the size of certain gene families. Conversely, because the assembly is still fragmented (as indicated by the 69.6% BUSCO score), some genes may be uncompleted or missing from our data, which could lead to an underestimation of other gene groups. While these are common challenges when working with wild, non-inbred parasites, our current gene counts should be viewed as a preliminary draft. Future work using single-worm sequencing and higher-level scaffolding will be necessary to confirm the exact copy number of these genes. In future, obtaining adult specimens would

facilitate whole genome sequencing from a single worm and permit the construction of a contact map using Hi-C technology to assemble contigs into chromosomes.

Mitochondrial genomes should be included in the evaluation of whole genome assemblies, as they are indispensable components of an organism's complete genome [47]. In our case, despite the challenges associated with performing mitochondrial assembly without organelle-specific purification, the results obtained in the first assembly version were encouraging. Practically the entire mitochondrial genome was covered. This aspect is reinforced by the absence of contamination of the samples, whether from human, host (eel), bacterial or viral organisms.

G. spinigerum invades the central nervous system by migrating along the nerve roots. The larva has cuticular spines and apical rows of hooks, that allow it to burrow through soft tissues, resulting in the necrosis in the brain and spinal cord and cerebral haemorrhages directly associated with its mortality [48,49]. A hypothetical role in the degradation of extracellular matrix macromolecules of host tissues has been assigned to a 24 kDa secreted a matrix metalloproteinase (MMP) like protein from *G. spinigerum* L3 (AAF82802), identified as matrilysin [50]. Matrilysins have a simpler structure compared to other MMPs, characterised by the absence of a C-terminal hemopexin domain [51]. Besides, *Gnathostoma* L3 MMP has been proven to be antigenic and useful in human diagnosis [52,53]. The immunogenic proteins of *G. spinigerum*, are known to be located in the 24 kDa region and efforts have been done to isolate and characterize immunodiagnostic candidates and drug targets by proteomics and RNA seq experiments [53–55], always with the constraint of the absence of a reference genome. It is known that inflammatory and parasite MMPs might represent suitable therapeutic targets to prevent the blood-brain-barrier disruption [56]. Indeed, one of the most abundant metalloproteinase inhibitors expressed in *G. spinigerum*, homologous to the putative *C.*

elegans inhibitor cri-2, without sequence identification so far, has been postulated as a potential target for both anthelmintics and vaccines [55]. In this work we have identified six new M10 metalloproteinases and two tissue inhibitors of metalloproteinases (TIMPs). Our analysis revealed that these MMPs possess two distinct architectures: "short-form" matrilysin-like enzymes and "long-form" typical nematode MMPs. As previously described, the matrilysin-like proteins (e.g. AAF82802 and G12031.T2) lack the C-terminal hemopexin-like domain, a feature shared with invasive stages of *Anisakis simplex*. Their simplified structure likely facilitates rapid diffusion through host tissues during the larva's characteristic burrowing migration. Conversely, the typical MMPs (e.g., G9182.T2) contain a hemopexin domain, resembling human MMP-2/9 (gelatinases). These "long-form" MMPs are common in other clade III and IV nematodes like *Strongyloides*, where they are utilized for the targeted degradation of host collagen and basement membranes. The presence of signal peptides in nearly all identified MMPs confirms their role as secreted factors, directly facilitating the necrosis and hemorrhaging observed along larval migration pathways in the central nervous system[56].

The computational analysis of interactions homologous to the putative *C. elegans* inhibitor cri-2 between these new M10 metalloproteinases, the one previously described and two TIMPs detected in the genome suggests different behaviours for each. This relatively low selectivity for different metalloproteinases shown by TIMP G1596.T1 is frequent in TIMPs, usually forming tight 1:1 complexes and also participating in pro-MMP activation and in suppression of different biological functions tumour-growth; matrix binding; inhibition of angiogenesis; or induction of apoptosis [57]. The metalloproteinases G8944.T1 and G12031.T2, the later a matrilysin structurally similar to AAF82802, are not predicted to exhibit strong affinity for either TIMP. This could imply a distinct mode of action, the presence of additional TIMPs in the genome, that

could not be detected in this first draft, or the existence of other proteins with equivalent functions. Future comprehensive studies would give enlightenment regarding utility of the identified TIMPs as therapeutic target as well as the potential role of novel MMP in pathogenies or diagnosis. In addition to matrilysins, we performed a survey for other gene families classically associated with nematode parasitism. Specifically, we identified 10 members of the SCP/TAPS family (Pfam: PF00188) within the *G. spinigerum* genome. These proteins, frequently identified in the secretomes of parasitic nematodes, are implicated in the transition to parasitism, larval migration, and the modulation of the host immune response. While the identification of ten candidates highlights a significant repertoire of these virulence-associated factors, this count should be considered a preliminary survey. Given the high heterozygosity and fragmented nature of the current draft assembly, future work is needed to distinguish between true paralogous expansions and uncollapsed allelic variants.

The application of next-generation sequencing (NGS) techniques to diagnosis of infectious viral, bacterial and parasite meningitis and eosinophilic encephalitis in a single step, using cerebrospinal fluid as a sample, has been the subject of recent reports in the scientific literature. One such study, conducted across eight hospitals in the United States, revealed the presence of *Angiostrongylus cantonensis* in two of the cases [58]. This parasite, along with *G. spinigerum*, is the primary etiological agent responsible for parasitic eosinophilic meningitis. Although the incidence of *A. cantonensis* meningoencephalitis is higher than that of *G. spinigerum*, cases of neurognathostomiosis are more severe, with greater sequelae and mortality (less than 1% versus 7-25%) [48,59]. The availability of the *G. spinigerum* genome would allow for the detection of the worm in these diagnostic/analysis panels. For example, in regions endemic to parasitic eosinophilic meningitis, such as Southeast Asia, the availability of the *A. cantonensis*

genome has already facilitated the diagnosis of numerous cases, which are particularly relevant in the pediatric population due to the immaturity of their immune systems [58,60,61]. The same could be applied to cerebral *Gnathostoma* cases, where early diagnosis is essential due to the severity of pathology and sequelae [62]. Continued development of genomic technologies will reduce the costs of sequencing and library preparation enabling the routine genomic sequencing of clinical samples. However, this approach requires the availability of a genome, like the one that was generated in this article.

Conclusions

The first genome assembly of *G. spinigerum* marks a major advance in parasitic genomics, yet this study represents only an initial step toward full genomic characterization. To resolve structural and functional complexities, future work should prioritize high-resolution sequencing of a single individual. Chromosome-scale assembly will further improve accuracy, enabling deeper insights into the parasite's genetic architecture. Despite current limitations, this genome provides a critical resource for advancing diagnostics, drug discovery, and basic research on this species. In addition, we present novel molecules that could lead to new research opportunities into their potential role in pathogenesis / diagnosis or as a therapeutic target in the case of MMP and TIMPs respectively.

List of abbreviations

BUSCO: Benchmarking Universal Single-Copy Orthologs; GC: Guanine Cytosine; HPC: High-performance computing cluster; IL: Illumina, short read sequencing; L3: Third-stage larvae; MMP: matrix metalloproteases; N50: Sequence length of the shortest contig

at 50% of the total assembly length; ONT: Oxford Nanopore Technologies; bp: base pair; PCR: Conventional polymerase chain reaction; QUAST: Quality Assessment Tool for Genome Assemblies; TIMPs: Tissue inhibitors of metalloproteinases.

Supplementary Information

Additional file 1: **Text S1.** Bioinformatics software, assemblers used and procedures an additional analysis.

Additional file 2: **Table S1.** Sequencing results pre- and post-processing on the two platforms. **Table S2.** Metrics, of the draft assemblies hybrid approach, with different software, obtained using QUAST and BUSCO. **Table S3.** Busco results of assemblies for *Gnathostoma spinigerum*. **Figure S1.** Parallel chart of hybrid assemblies metrics. **Figure S2.** Spider chart of hybrid assemblies metrics.

Acknowledgments

We would like to thank all the people who have contributed to this project: the staff at Mahidol University, especially Tippayarat Yoonuan, as well as members of the Gasser Laboratory, Tanapan Sukee, Joe Byrne, Aya Taki, Anson V Koehler and Ross Hall and also the team at the Instituto de Salud Carlos III, especially Cristina García Amil, Guillermo Gorines, Alberto Lema and Erika Kvalem. We also thank Lourdes Castro and Angela Ceballos-Caro, from the ISCIII, for their help with the figures. The computational support of the "Centro de Computación Científica CCC-UAM" is gratefully recognized. Finally, we would like to express our gratitude to Geoffry Gobbert for facilitating connections between researchers from the University of Melbourne, Australia and the Instituto de Salud Carlos III, Spain.

Funding

This study was funded by Instituto de Salud Carlos III (AESI-PI22CIII/00010) and by CIBERINFEC-UE-(CB21/13/00120) to MJP and AEI (PID2021-126625OB-I00) to PGP. BGB received a Predoctoral Fellowship award for career development by Fundación Rafael Folch (2021/E02) and a mobility grant from the Spanish Society of Infectious Diseases and Clinical Microbiology (SEIMC). The funders had no role in study design, data collection and analysis, decision or publish, or preparation of the manuscript.

Availability of data and materials

Following the FAIR principles, all datasets generated and analyzed during the *Gnathostoma spinigerum* Genome Project are openly available at the following addresses:

BioProject PRJNA1141240 *Gnathostoma spinigerum* isolate:aL3 Genome sequencing https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_043882145.1

Author´s contributions

Conceptualization: BGB, NDY, MJP. Methodology: BGB, SM, AZ, PJ, SV, IC, JS, AH, SS, NDY, MJP, IMA, PGP. Programming and Data curation: BGB, SM, SV, IC, NDY, IMA, PGP. Formal analysis: BGB, SM, AZ, SV, IC, SS, NDY, MJP, IMA, PGP. Investigation: BGB, SM, AZ, PJ, SV, IC, SS, NDY, IMA, PGP. Resources: AZ, PD, IC, PA, RPG, NDY, MJP, IMA, PGP. Writing-original draft preparation: BGB, NDY, MJP. Review and editing: BGB, SM, AZ, PJ, PD, SV, IC, SS, PA, JS, AHG, NDY, MJP, IMA, PGP. Supervision: NDY, MJP. Funding acquisition: MJP and PGP. All authors have read and agreed to the published version of the manuscript.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare no competing interests.

References

1. Hung M-N, Huang H-W, Dekumyoy P, Pakdee W, Lee Y-S, Ji D-D. First case of neurognathostomiasis in Taiwan—A Thai laborer presenting with eosinophilic meningitis and intracranial hemorrhage. J Formos Med Assoc. 2015;114:1280–4. https://doi.org/10.1016/j.jfma.2013.07.006

2. Sawanyawisuth K, Chlebicki MP, Pratt E, Kanpittaya J, Intapan PM. Sequential imaging studies of cerebral gnathostomiasis with subdural hemorrhage as its complication. Trans R Soc Trop Med Hyg. 2009;103:102–4. https://doi.org/10.1016/j.trstmh.2008.09.011

3. Herman JS, Chiodini PL. Gnathostomiasis, Another Emerging Imported Disease. Clin Microbiol Rev. 2009;22:484–92. https://doi.org/10.1128/CMR.00003-09

4. Janwan P, Intapan PM, Sanpool O, Sadaow L, Thanchomnang T, Maleewong W. Growth and development of Gnathostoma spinigerum (Nematoda: Gnathostomatidae) larvae in Mesocyclops aspericornis (Cyclopoida: Cyclopidae). Parasit Vectors. 2011;4:93. https://doi.org/10.1186/1756-3305-4-93

5. Sharma C, Piyaphanee W, Watthanakulpanich D. Case Report: Clinical Features of Intermittent Migratory Swelling Caused by Gnathostomiasis with Complete Follow-up. Am J Trop Med Hyg. 2017;97:1611–5. https://doi.org/10.4269/ajtmh.17-0239

6. Sivakorn C, Promthong K, Dekumyoy P, Viriyavejakul P, Ampawong S, Pakdee W, et al. Case Report: The First Direct Evidence of Gnathostoma spinigerum Migration through Human Lung. Am J Trop Med Hyg. 2020;103:1129–34. https://doi.org/10.4269/ajtmh.20-0236

7. Liu G-H, Sun M-M, Elsheikha HM, Fu Y-T, Sugiyama H, Ando K, et al. Human gnathostomiasis: a neglected food-borne zoonosis. Parasit Vectors. 2020;13:616. https://doi.org/10.1186/s13071-020-04494-4

8. Katchanov J, Sawanyawisuth K, Chotmongkol V, Nawa Y. Neurognathostomiasis, a Neglected Parasitosis of the Central Nervous System. Emerg Infect Dis J - CDC [Internet]. 2011 [cited 2021 Aug 27];17. https://doi.org/10.3201/eid1707.101433

9. Leroy J, Cornu M, Deleplancque AS, Loridant S, Dutoit E, Sendid B. Sushi, ceviche and gnathostomiasis - A case report and review of imported infections. Travel Med Infect Dis. 2017;20:26–30. https://doi.org/10.1016/j.tmaid.2017.10.010

10. Nograd K, Adisakwattana P, Reamtong O. Human gnathostomiasis: A review on the biology of the parasite with special reference on the current therapeutic management. Food Waterborne Parasitol. 2023;33:e00207. https://doi.org/10.1016/j.fawpar.2023.e00207

11. Jongthawin J, Intapan PM, Sanpool O, Sadaow L, Janwan P, Thanchomnang T, et al. Three Human Gnathostomiasis Cases in Thailand with Molecular Identification of Causative Parasite Species. Am J Trop Med Hyg. 2015;93:615–8. https://doi.org/10.4269/ajtmh.15-0284

12. Babraham Bioinformatics. FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. 2019 [cited 2023 Apr 26]. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 26 Apr 2023

13. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90. https://doi.org/10.1093/bioinformatics/bty560

14. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012;22:549–56. https://doi.org/10.1101/gr.126953.111

15. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes De Novo Assembler. Curr Protoc Bioinforma. 2020;70:e102. https://doi.org/10.1002/cpbi.102

16. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012;1:18. https://doi.org/10.1186/2047-217X-1-18

17. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 2014;24:1384–95. https://doi.org/10.1101/gr.170720.113

18. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol İ. ABySS: A parallel assembler for short read sequence data. Genome Res. 2009;19:1117–23. https://doi.org/10.1101/gr.089532.108

19. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017;13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595

20. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–6. https://doi.org/10.1038/s41587-019-0072-8

21. Chen Y, Nie F, Xie S-Q, Zheng Y-F, Dai Q, Bray T, et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat Commun. 2021;12:60. https://doi.org/10.1038/s41467-020-20236-7

22. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinforma Oxf Engl. 2013;29:1072–5. https://doi.org/10.1093/bioinformatics/btt086

23. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinforma Oxf Engl. 2015;31:3210–2. https://doi.org/10.1093/bioinformatics/btv351

24. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. https://doi.org/10.1038/nmeth.1923

25. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinforma Oxf Engl. 2009;25:2078–9. https://doi.org/10.1093/bioinformatics/btp352

26. Broad Institute. Picard Toolkit (https://broadinstitute.github.io/picard/) [Internet]. 2019 [cited 2023 Apr 28]. https://github.com/broadinstitute/picard/wiki/Home. Accessed 28 Apr 2023

27. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One. 2014;9:e112963. https://doi.org/10.1371/journal.pone.0112963

28. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117:9451–7. https://doi.org/10.1073/pnas.1921046117

29. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. Nucleic Acids Res. 2016;44:D81-89. https://doi.org/10.1093/nar/gkv1272

30. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics Bioinforma. 2021;3:lqaa108. https://doi.org/10.1093/nargab/lqaa108

31. Nuamtanong S, Reamtong O, Phuphisut O, Chotsiri P, Malaithong P, Dekumyoy P, et al. Transcriptome and excretory-secretory proteome of infective-stage larvae of the nematode Gnathostoma spinigerum reveal potential immunodiagnostic targets for development. Parasite Paris Fr. 2019;26:34. https://doi.org/10.1051/parasite/2019033

32. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated

orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47:D309–14. https://doi.org/10.1093/nar/gky1085

33. Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 2021;49:D344–54. https://doi.org/10.1093/nar/gkaa977

34. Nevers Y, Warwick Vesztrocy A, Rossier V, Train C-M, Altenhoff A, Dessimoz C, et al. Quality assessment of gene repertoire annotations with OMArk. Nat Biotechnol. 2025;43:124–33. https://doi.org/10.1038/s41587-024-02147-w

35. Liu G-H, Shao R, Cai X-Q, Li W-W, Zhu X-Q. Gnathostoma spinigerum Mitochondrial Genome Sequence: a Novel Gene Arrangement and its Phylogenetic Position within the Class Chromadorea. Sci Rep. 2015;5:12691. https://doi.org/10.1038/srep12691

36. Nishikawa-Shimono R, Kuwabara M, Fujisaki S, Matsuda D, Endo M, Kamitani M, et al. Discovery of novel indole derivatives as potent and selective inhibitors of proMMP-9 activation. Bioorg Med Chem Lett. 2024;97:129541. https://doi.org/10.1016/j.bmcl.2023.129541

37. Morgunova E, Tuuttila A, Bergmann U, Isupov M, Lindqvist Y, Schneider G, et al. Structure of human pro-matrix metalloproteinase-2: activation mechanism revealed. Science. 1999;284:1667–70. https://doi.org/10.1126/science.284.5420.1667

38. Tuuttila A, Morgunova E, Bergmann U, Lindqvist Y, Maskos K, Fernandez-Catalan C, et al. Three-dimensional structure of human tissue inhibitor of metalloproteinases-2 at 2.1 A resolution. J Mol Biol. 1998;284:1133–40. https://doi.org/10.1006/jmbi.1998.2223

39. Powell HR, Islam SA, David A, Sternberg MJE. Phyre2.2: A Community Resource for Template-based Protein Structure Prediction. J Mol Biol. 2025;437:168960. https://doi.org/10.1016/j.jmb.2025.168960

40. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46:W296–303. https://doi.org/10.1093/nar/gky427

41. Raeeszadeh-Sarmazdeh M, Greene KA, Sankaran B, Downey GP, Radisky DC, Radisky ES. Directed evolution of the metalloproteinase inhibitor TIMP-1 reveals that its N- and C-terminal domains cooperate in matrix metalloproteinase recognition. J Biol Chem. 2019;294:9476–88. https://doi.org/10.1074/jbc.RA119.008321

42. Ros-Pardo D, Gómez-Puertas P, Marcos-Alcalde Í. STAG2: Computational Analysis of Missense Variants Involved in Disease. Int J Mol Sci. 2024;25:1280. https://doi.org/10.3390/ijms25021280

43. Roe DR, Cheatham TE. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J Chem Theory Comput. 2013;9:3084–95. https://doi.org/10.1021/ct400341p

44. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph. 1996;14:33–8, 27–8. https://doi.org/10.1016/0263-7855(96)00018-5

45. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. J Comput Chem. 2005;26:1781–802. https://doi.org/10.1002/jcc.20289

46. International Helminth Genomes Consortium. Comparative genomics of the major parasitic worms. Nat Genet. 2019;51:163–74. https://doi.org/10.1038/s41588-018-0262-1

47. Wang P, Wang F. A proposed metric set for evaluation of genome assembly quality. Trends Genet TIG. 2023;39:175–86. https://doi.org/10.1016/j.tig.2022.10.005

48. Lo Re V, Gluckman SJ. Eosinophilic meningitis. Am J Med. 2003;114:217–23. https://doi.org/10.1016/s0002-9343(02)01495-x

49. Chayangsu C, Ampawong S, Reamtong O, Viriyavejakul P, Kanjanapruthipong T, Fongsodsri K, et al. Detection of Gnathostoma spinigerum larva in the brain with complete follow-up after surgical treatment of human neurognathostomiasis. Food Waterborne Parasitol. 2024;35:e00229. https://doi.org/10.1016/j.fawpar.2024.e00229

50. Uparanukraw P, Morakote N, Harnnoi T, Dantrakool A. Molecular cloning of a gene encoding matrix metalloproteinase-like protein from Gnathostoma spinigerum. Parasitol Res. 2001;87:751–7. https://doi.org/10.1007/s004360100440

51. Sreesada P, Vandana null, Krishnan B, Amrutha R, Chavan Y, Alfia H, et al. Matrix metalloproteinases: Master regulators of tissue morphogenesis. Gene. 2025;933:148990. https://doi.org/10.1016/j.gene.2024.148990

52. Janwan P, Intapan PM, Yamasaki H, Laummaunwai P, Sawanyawisuth K, Wongkham C, et al. Application of Recombinant Gnathostoma spinigerum Matrix Metalloproteinase-Like Protein for Serodiagnosis of Human Gnathostomiasis by Immunoblotting. Am J Trop Med Hyg. 2013;89:63–7. https://doi.org/10.4269/ajtmh.12-0617

53. Saenseeha S, Penchom J, Yamasaki H, Laummaunwai P, Tayapiwatana C, Kitkhuandee A, et al. A dot-ELISA test using a Gnathostoma spinigerum recombinant matrix metalloproteinase protein for the serodiagnosis of human gnathostomiasis. Southeast Asian J Trop Med Public Health. 2014;45:990–6.

54. Thiangtrongjit T, Nograd K, Ketboonlue T, Malaitong P, Adisakwattana P, Reamtong O. Proteomics of Gnathostomiasis: A Way Forward for Diagnosis and Treatment Development. Pathogens. 2021;10:1080. https://doi.org/10.3390/pathogens10091080

55. Nograd K, Thiangtrongjit T, Adisakwattana P, Dekumyoy P, Muangnoicharoen S, Thawornkuno C, et al. Protein and antigen profiles of third-stage larvae of Gnathostoma spinigerum assessed with next-generation sequencing transcriptomic information. Sci Rep. 2022;12:6915. https://doi.org/10.1038/s41598-022-10826-4

56. Bruschi F, Pinto B. The significance of matrix metalloproteinases in parasitic infections involving the central nervous system. Pathogens. Basel, Switzerland; 2013;2:105–29. https://doi.org/10.3390/pathogens2010105

57. Tallant C, Marrero A, Gomis-Rüth FX. Matrix metalloproteinases: fold and function of their catalytic domains. Biochim Biophys Acta. 2010;1803:20–8. https://doi.org/10.1016/j.bbamcr.2009.04.003

58. Wilson MR, Sample HA, Zorn KC, Arevalo S, Yu G, Neuhaus J, et al. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. N Engl J Med. 2019;380:2327–40. https://doi.org/10.1056/NEJMoa1803396

59. Ramirez-Avila L, Slome S, Schuster FL, Gavali S, Schantz PM, Sejvar J, et al. Eosinophilic meningitis due to Angiostrongylus and Gnathostoma species. Clin Infect Dis Off Publ Infect Dis Soc Am. 2009;48:322–7. https://doi.org/10.1086/595852

60. Liu J, Tao J, Chen W, Wang T, Chen X, Shen M, et al. The application of metagenomic next-generation sequencing for Angiostrongylus eosinophilic meningitis in a pediatric patient: A case report. Front Public Health. 2022;10:1003013. https://doi.org/10.3389/fpubh.2022.1003013

61. Liu D, Li N, Zhu Y, Chen Q, Fan X, Feng J. Diagnosis of human angiostrongyliasis in a case of hydrocephalus using next-generation sequencing: a case report and literature review. BMC Neurol. 2024;24:281. https://doi.org/10.1186/s12883-024-03663-7

62. Bunyaratavej K, Pongpunlert W, Jongwutiwes S, Likitnukul S. Spinal gnathostomiasis resembling an intrinsic cord tumor/myelitis in a 4-year-old boy. Southeast Asian J Trop Med Public Health. 2008;39:800–3.

Figure 1. Short reads genome assembly workflow. **A, B** and **C** boxes represent the main processes in the short reads bioinformatic analysis, after getting fastq files from Illumina sequencing. **A)** Preprocessing was performed by three steps: first, FastQC analysis to get the quality evaluation of raw reads **(A.1);** second, Fastp, main preprocessing step, it consisted of filtering the reads that were longer than 100 base pairs (bp) and those with a Phred score greater than 15, finally the trimming of homopolymeric and adapters regions (detailed in supplementary material); third, FastQC **(A.2)** again to get the quality evaluation of selected and trimmed reads. **B)** Assembly process using different de novo assemblers. **C)** Evaluation process of the several draft assemblies, each one was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) and Quality

Assessment Tool for Genome Assemblies (QUAST) software to obtain metrics for comparison and selection of the best draft assembly.

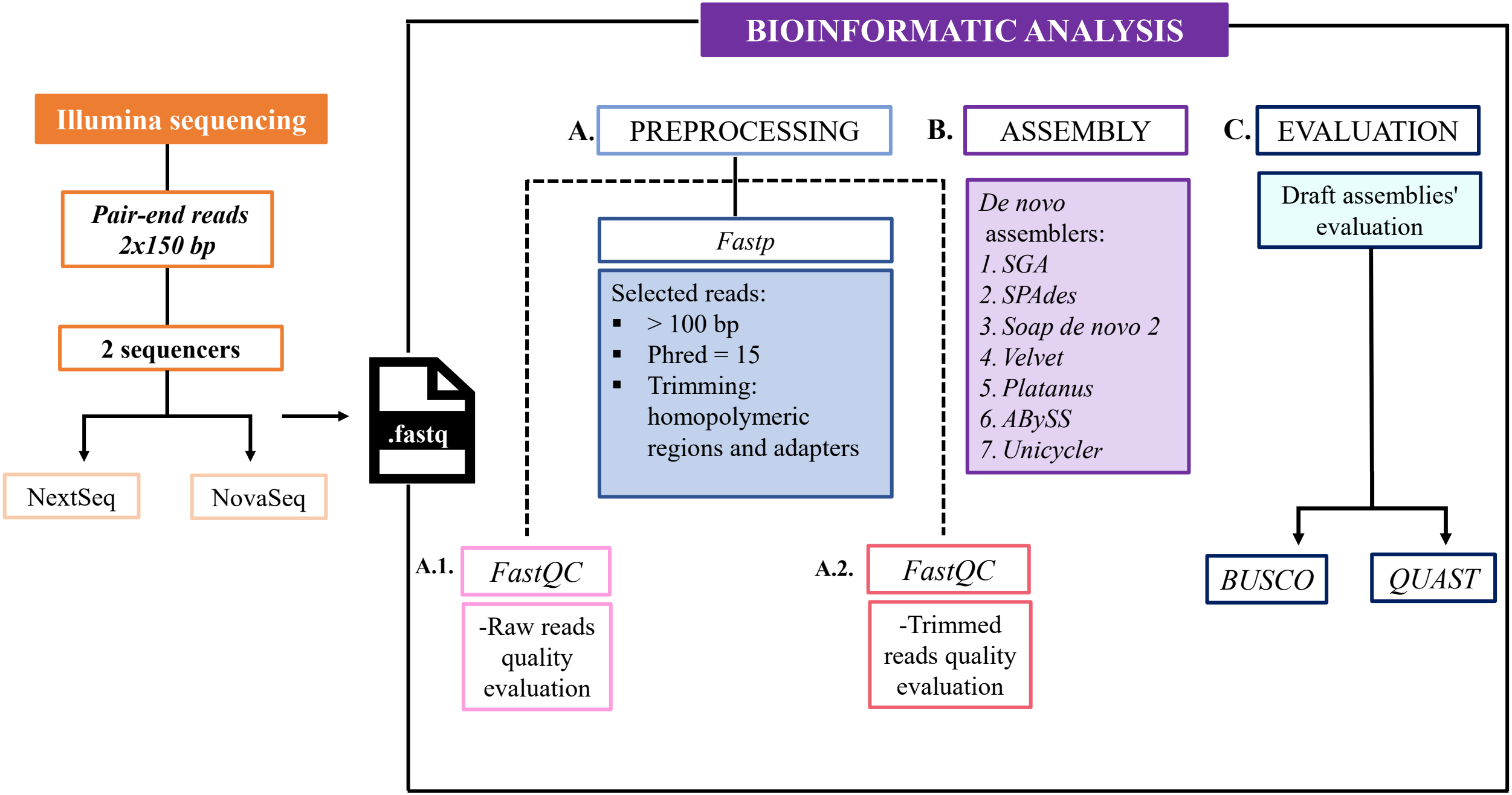
Figure 2. Hybrid genome assembly workflow. **A)** Based on short reads: the SPAdes assembler was employed using the hybrid mode, incorporating short reads (Illumina) and long reads from ONT technology with the nanopore option. **B)** Based on long reads: it used contigs assembled using Flye and the long reads to align short reads (Illumina) employing Bowtie2.

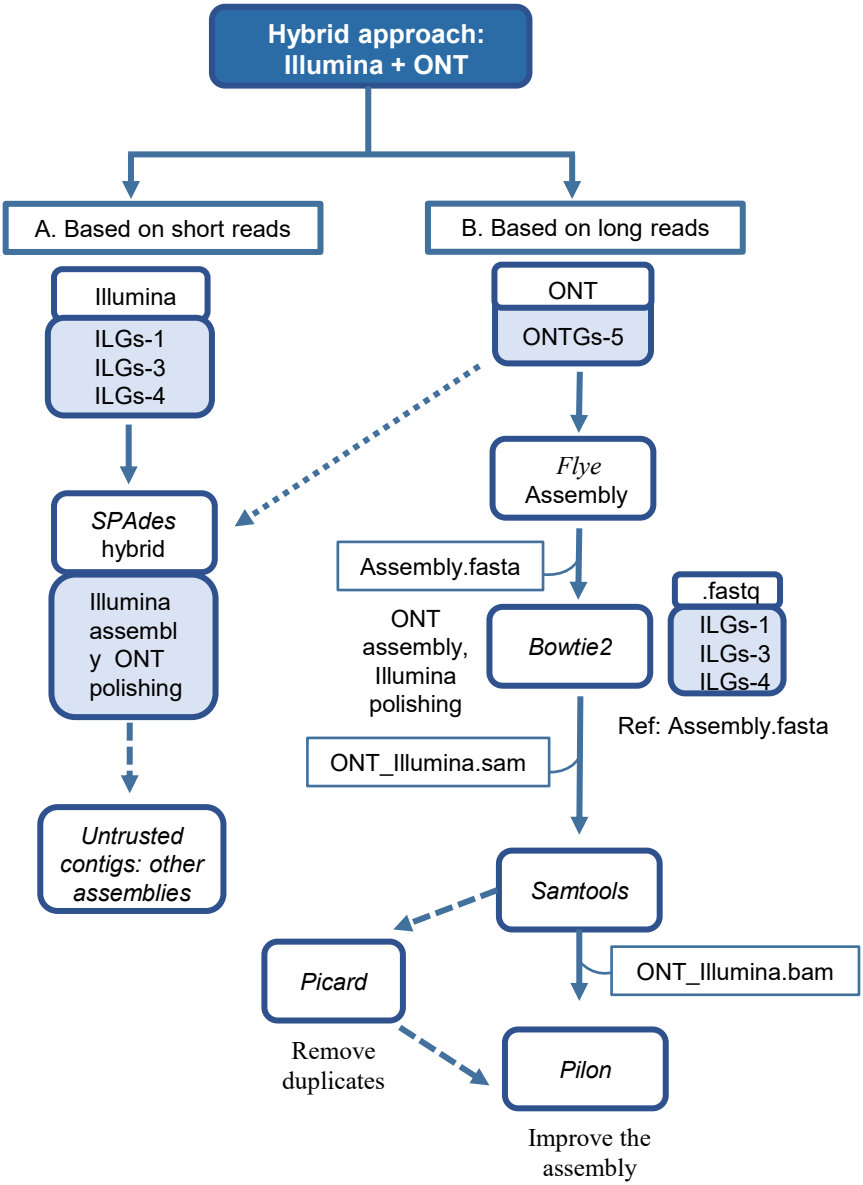
Figure 3. A) Larva of Gnathostoma spinigerum advanced stage 3. Cervical sacs (black arrows) and cephalic bulb. **B)** Magnification of the head bulb, with four rows of hooklets and lips, details (black arrows) of the hooklets.

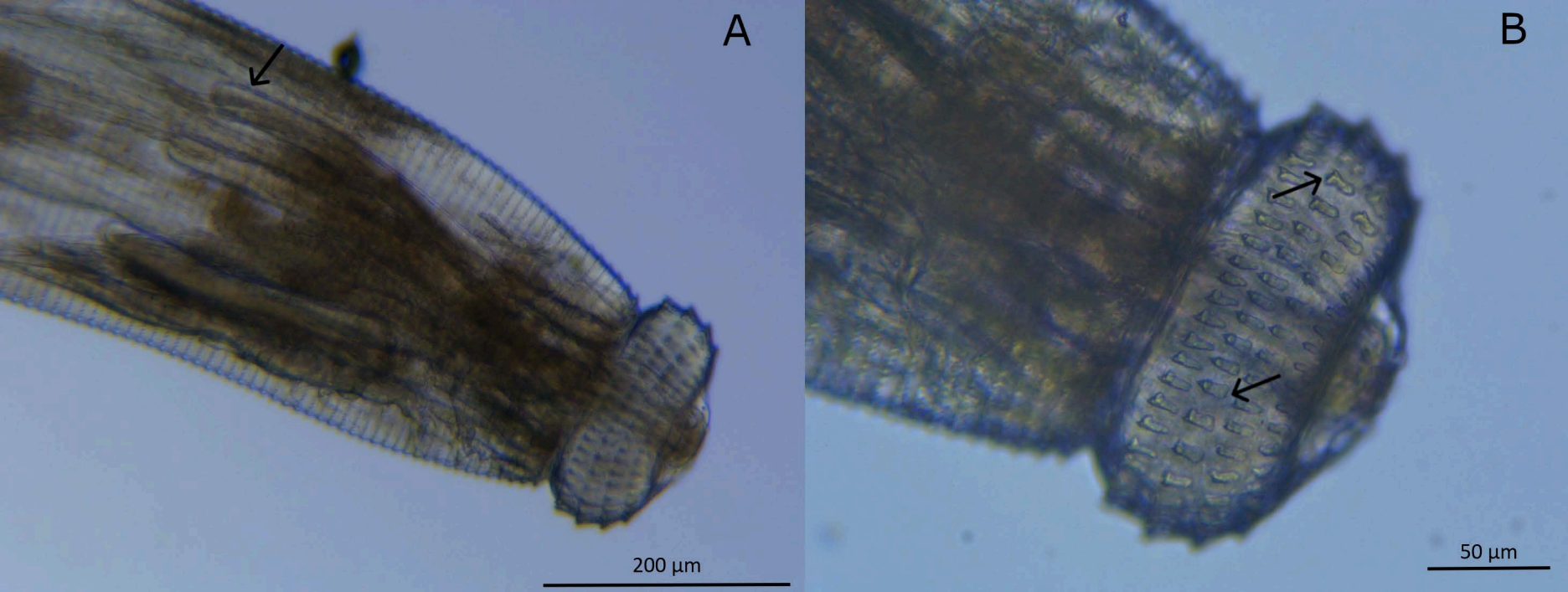
Figure 4. Percentage of proteome and conserved HOGs of *Gnathostoma spinigerum* annotated assembly. **HOGs**: Hierarchical Orthologous Groups. The graph categorized the HOGs as single, duplicated or missing, and proteome in consistent, contaminant, inconsistent, unknown, partial mapping and fragments.

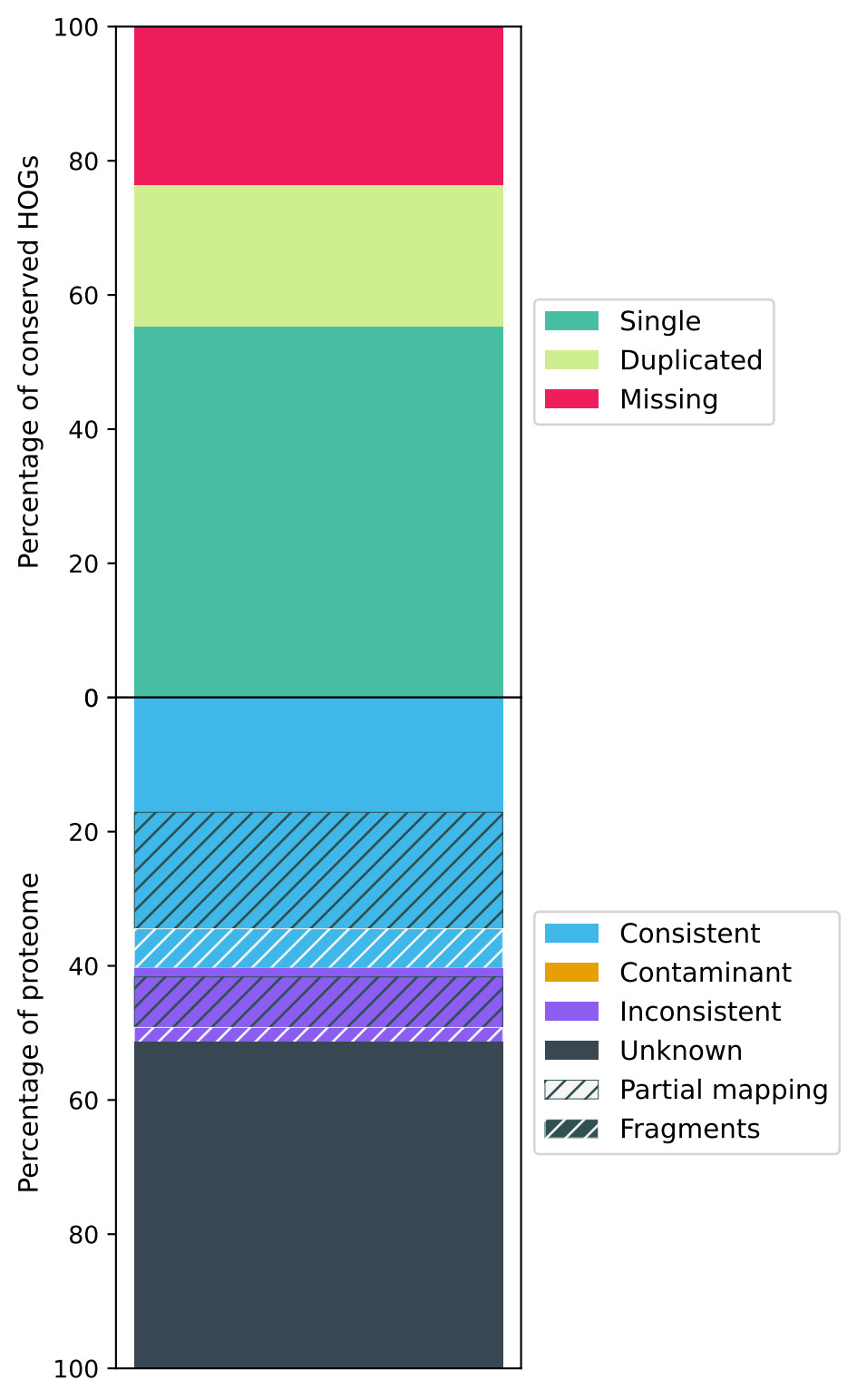
Figure 5. Structural models and computational analysis of the interaction between metalloproteinases M10 and TIMPs. **A)** Structural models of the seven metalloproteinases M10 (AAF82802, G12031.T2, G9182.T2, G1280.T1, G9438.T1, G8944.T1 and G7562.T1). Catalytic domains were depicted in magenta. Putative positions of Ca++ (green) and Zn++ ions (light green) is indicated. **B)** Structural model of the interaction between the catalytic domain of the metalloproteinase AAF82802 (showing the electrostatically charged surface) and the model of tissue inhibitors of

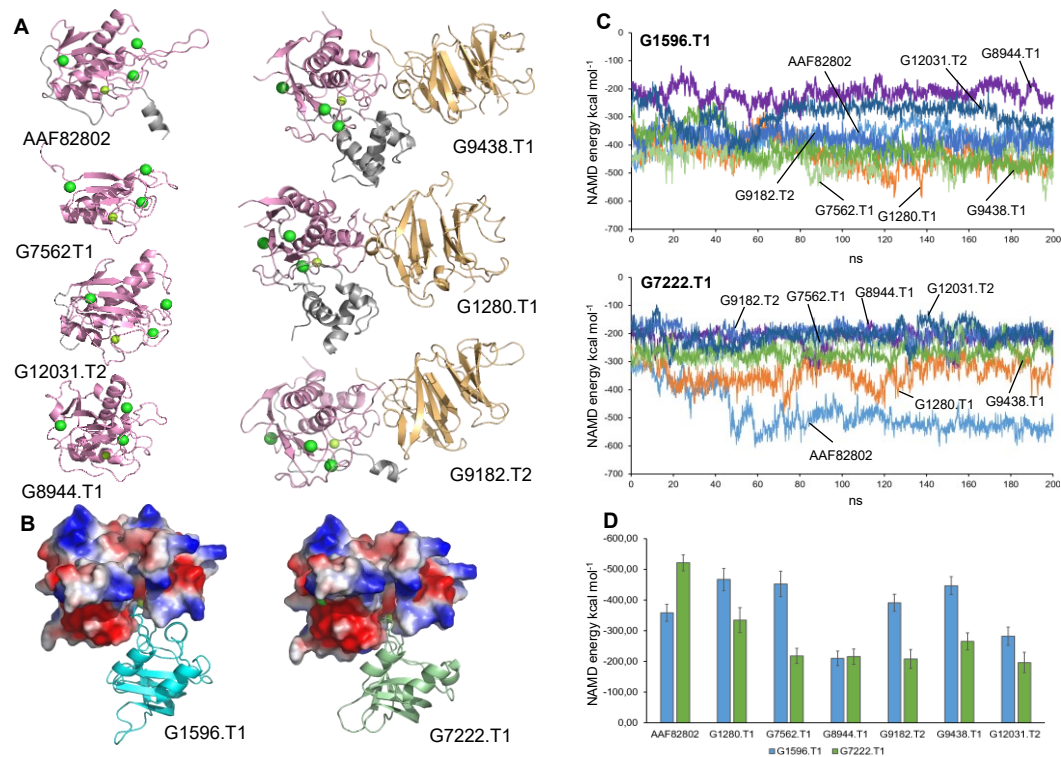
metalloproteinases -TIMPs- G7222.T1 (left) and G1596.T1 (right), represented as secondary structure. **C)** Interaction energy measured, using NAMD, over 200 ns of Molecular Dynamics between inhibitor G1596.T1 (upper panel) or G7222.T1 (lower panel) and the seven metalloproteinases M10. D) Plot showing the mean +/- STD values of energy shown in B and C and corresponding to the last 100 ns of the Molecular Dynamics trajectory. Blue bars: G1596.T1; green bars: G7222.T1.


BIOINFORMATIC ANALYSIS
Illumina sequencing
Pair-end reads 2x150 bp
2 sequencers
NextSeq
NovaSeq
.fastq
A. PREPROCESSING
Fastp
Selected reads:
> 100 bp
Phred = 15
Trimming: homopolymeric regions and adapters
A.1. FastQC
-Raw reads quality evaluation
A.2. FastQC
-Trimmed reads quality evaluation
B. ASSEMBLY
De novo assemblers:
1. SGA
2. SPAdes
3. Soap de novo 2
4. Velvet
5. Platanus
6. ABySS
7. Unicycler
C. EVALUATION
Draft assemblies' evaluation
BUSCO
QUAST


Hybrid approach:
Illumina + ONT
A. Based on short reads
B. Based on long reads
Illumina
ILGs-1
ILGs-3
ILGs-4
ONT
ONTGs-5
SPAdes
hybrid
Illumina
assembl
y ONT
polishing
Untrusted
contigs: other
assemblies
Flye
Assembly
Assembly.fasta
ONT
assembly,
Illumina
polishing
Bowtie2
.fastq
ILGs-1
ILGs-3
ILGs-4
Ref: Assembly.fasta
ONT_Illumina.sam
Samtools
Picard
ONT_Illumina.bam
Remove
duplicates
Pilon
Improve the
assembly


A
200 µm
B
50 µm


Percentage of conserved HOGs
100
80
60
40
20
0
Single
Duplicated
Missing
Percentage of proteome
20
40
60
80
100
Consistent
Contaminant
Inconsistent
Unknown
Partial mapping
Fragments


A
AAF82802
G9438.T1
G7562T1
G1280.T1
G12031.T2
G9182.T2
G8944.T1
B
G1596.T1
G7222.T1
C
G1596.T1
AAF82802
G12031.T2
G8944.T1
G9182.T2
G7562.T1
G1280.T1
G9438.T1
NAMD energy kcal mol-1
ns
G7222.T1
G9182.T2
G7562.T1
G8944.T1
G12031.T2
G1280.T1
G9438.T1
AAF82802
NAMD energy kcal mol-1
ns
D
NAMD energy kcal mol-1
AAF82802
G1280.T1
G7562.T1
G8944.T1
G9182.T2
G9438.T1
G12031.T2
G1596.T1
G7222.T1

De novo genome assembly and annotation of *Gnathostoma spinigerum*

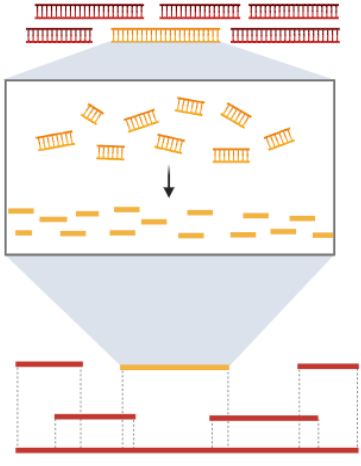
DNA
G.spinigerum

Long and short reads sequencing

Bioinformatics

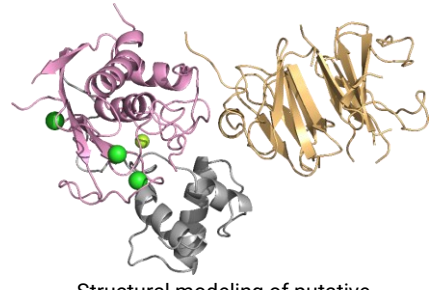
Assembly and Analysis

Sequenced Genome



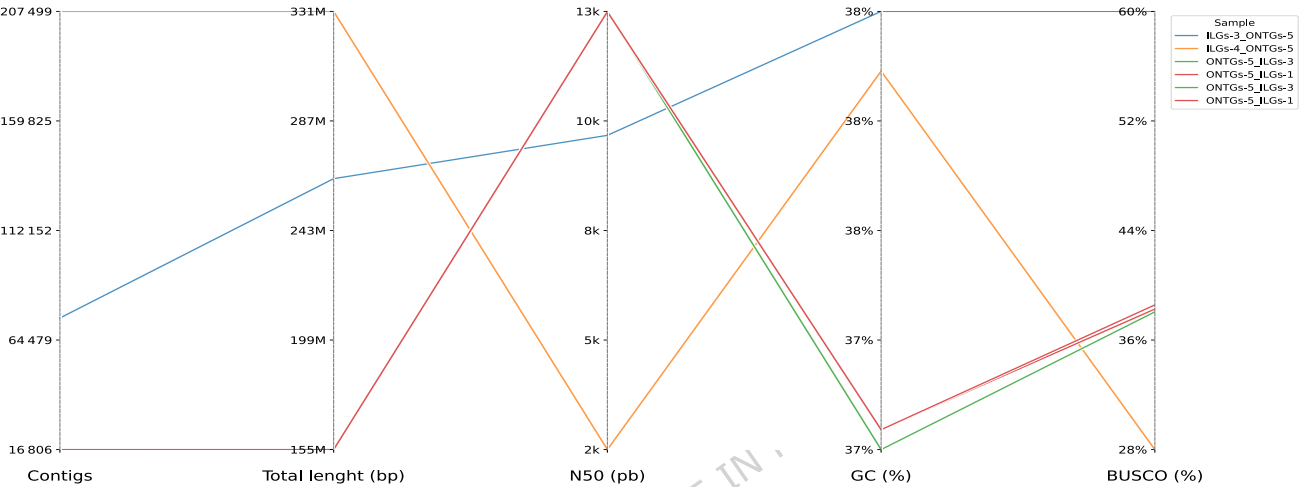
De novo genome assembly of *Gnathostoma spinigerum*

Genome annotation

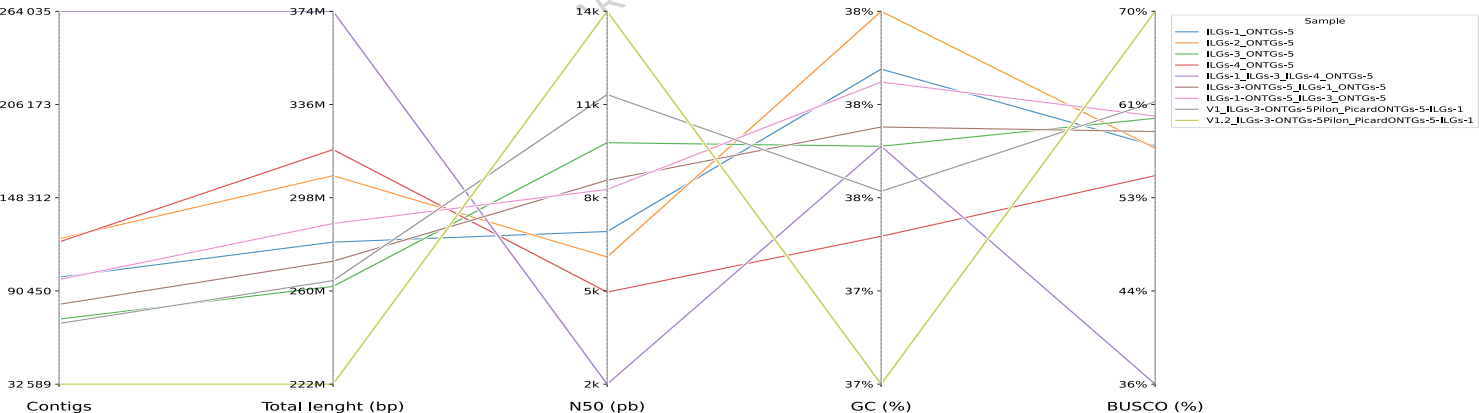


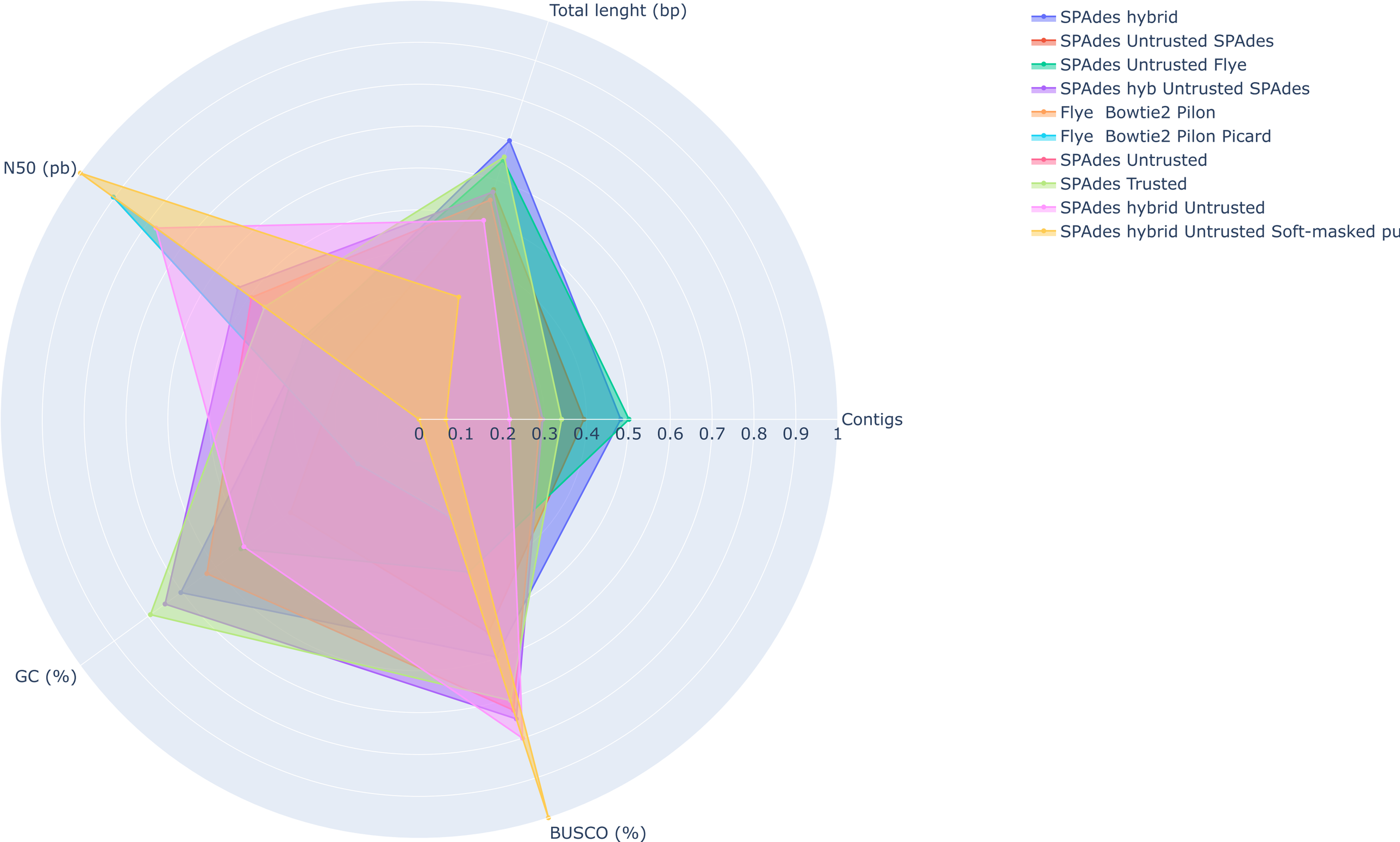
Structural modeling of putative metalloproteinases & inhibitors

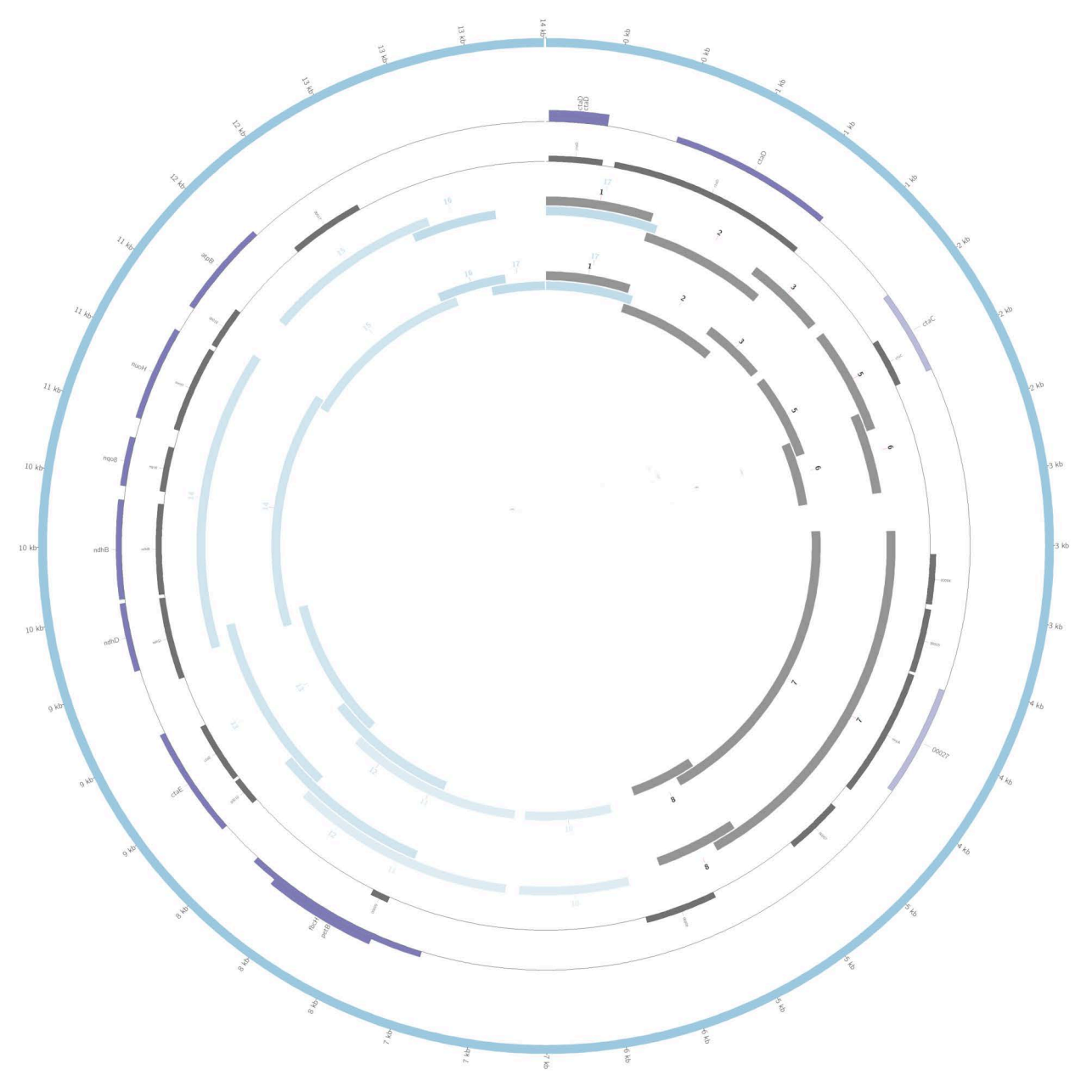
A



B




Total lenght (bp)
N50 (pb)
Contigs
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
GC (%)
BUSCO (%)
SPAdes hybrid
SPAdes Untrusted SPAdes
SPAdes Untrusted Flye
SPAdes hyb Untrusted SPAdes
Flye Bowtie2 Pilon
Flye Bowtie2 Pilon Picard
SPAdes Untrusted
SPAdes Trusted
SPAdes hybrid Untrusted
SPAdes hybrid Untrusted Soft-masked pu


14 kb
0 kb
1 kb
2 kb
3 kb
4 kb
5 kb
6 kb
7 kb
8 kb
9 kb
10 kb
11 kb
12 kb
13 kb
cpcD
cpcG
cpcC
00027
ftsH
petB
petA
ctaE
ndhD
ndhB
bbpu
hcnu
atpe