First chromosome-level genome assembly of a ribbon worm from the Hoplonemertea clade, *Emplectonema gracile*, and its structural annotation

Alberto Valero-Gracia[1]*, Nickellaus G. Roberts[2], Meghan Yap-Chiongco[2], Ana Teresa Capucho[1], Kevin M. Kocot[2, 3], Michael Matschiner[1], Torsten H. Struck[1].

[1] Natural History Museum, University of Oslo, Blindern, P.O. Box 1172, 0318 Oslo, Norway

[2] Department of Biological Sciences, University of Alabama, Tuscaloosa, AL 35487, USA

[3] Alabama Museum of Natural History, University of Alabama, Tuscaloosa, AL 35487, USA

* Correspondence: Alberto Valero-Gracia, Natural History Museum, University of Oslo, Norway, alberto.valero-gracia@nhm.uio.no

Abstract

Genome-wide information has so far been unavailable for ribbon worms of the clade Hoplonemertea, the most species-rich class within the phylum Nemertea. While species within Pilidiophora, the sister clade of Hoplonemertea, possess a pilidium larval stage and lack stylets on their proboscis, Hoplonemertea species have a planuliform larva and are armed with stylets employed for the injection of toxins into their prey. To further compare these developmental, physiological, and behavioral differences from a genomic perspective, the availability of a reference genome for a Hoplonemertea species is crucial. Such data will be highly useful for future investigations towards a better understanding of molecular ecology, venom evolution, and

regeneration not only in Nemertea, but also in other marine invertebrate phyla. To this end, we herein present the annotated chromosome-level genome assembly for *Emplectonema gracile* (Nemertea; Hoplonemertea; Monostilifera; Emplectonematidae), an easily collected nemertean well-suited for laboratory experimentation. The genome has an assembly size of 157.9 Mbp. Hi-C scaffolding yielded chromosome level scaffolds, with a scaffold N50 of 10.0 Mbp and a score of 95.1% for complete BUSCO genes found as a single copy. Annotation predicted 20,684 protein-coding genes. The high-quality reference genome reaches an Earth BioGenome standard level of 7.C.Q50.

Keywords: de novo assembly, genome sequence, 3D genomics, Hi-C, HiFi

Significance

The genome of *Emplectonema gracile* is highly contiguous, well annotated, and shorter than those of the other ribbon worm species sequenced to date. This genome is a valuable resource for studies on molecular ecology, venom evolution, and regeneration in marine invertebrates.

Introduction

Nemerteans, commonly known as ribbon worms, are a phylum of about 1,200 species of predatory worms that exhibit spiral cleavage and a variety of life histories, typically including pelagic and benthic stages (Gibson 1994, Maslakova and Hiebert 2015). While nemerteans are mainly marine, some species have entered freshwater habitats, and a few have colonized moist, terrestrial habitats (Gibson 1994). Nemerteans are a venomous animals known for their remarkable regeneration capacities and enigmatic phylogenetic position (Stricker and Cloney 1983; Zattara et al. 2019).

Phylogenetically, Nemertea is nested within Spiralia (*sensu* Giribet and Edgecombe 2020). However, their exact phylogenetic position is not well established (Struck and Fisse 2008; Struck et al. 2014; Andrade et al. 2014; Laumer et al. 2015; Kocot et al. 2017; Bleidorn 2019; Marlètaz et al. 2019; Dràbková et al. 2022). Nemertea is divided into three main clades: Paleonemertea, Pilidiophora, and Hoplonemertea (Figure 1A) (Andrade et al. 2014; Kvist et al. 2014).

Six nemertean nuclear genomes, belonging either to Paleonemertea or Pilidiophora species, have been sequenced to date. However, of these six genomes, only the assembly of *Lineus longissimus* meets the reference standards set by the Earth BioGenome Project (Kwiatkowski et al. 2021). The remaining five genomes, from *Cephalotrix simula* (GeneBank ID GCA_035591015.1), *Cephalothrix spiralis* (GCA_032353305.1), *Notospermus geniculatus* (GCA_002633025.1, Luo et al. 2018), *Tubulanus punctatus* (GCA_036785005.1), and *Tubulanus ruber* (GCA_036873915.1) were sequenced using short-read technologies. Except for *Notospermus geniculatus*, these assemblies are extremely fragmented with more than 200,000 scaffolds and a scaffold N50 of around 2 kb, and even the one of *Notospermus geniculatus* has over 10,000 scaffolds and a scaffold N50 of only 239.2 kb.

In this study, we aimed to enrich the available genomic resources for Nemertea. To this end, we have sequenced, assembled, and annotated the genome of one representative of the Hoplonemertea clade, *Emplectonema gracile* (Figure 1A). *Emplectonema gracile* inhabits the rocky shores of the North Atlantic Ocean and the Mediterranean Sea. This species has been selected for ease of collection, culturing, and spawning in the lab. Its slender, bi-toned body, armored with a venomous stylet used for capturing prey, can reach lengths of up to about 50 cm. The genome of *Emplectonema gracile* is the first hoplonemertean genome sequenced. Its assembly is highly contiguous, well annotated, and smaller than those of the other ribbon worm species sequenced to date. Given these features, the newly sequenced genome stands to be a valuable resource for studies

on molecular ecology, venom evolution, and regeneration in marine invertebrates, and will ultimately contribute to clarifying the phylogenetic position of nemerteans within the animal Tree of Life.

Results and Discussion

HiFi sequencing yielded 26.6 Gbp of information contained in a total of 1,779,646 reads and a kmer coverage of 70.9x. Analysis of the genomic data with GenomeScope (Vurture et al. 2017) inferred a genome size of 153.7 Mbp with a heterozygosity of 1.5%, a uniqueness of 76.8%, and an error rate of 0.1% (Figure S1A). As no genomic information is yet available for this species, we also checked if the assumed diploidy is given. K-mer analysis indicated 92% of all heterozygous k-mer pairs as diploid, and only a small percentage as triploid or tetraploid (4%); these values therefore confirm the assumed diploidy (Figure S1B).

Assembly, purging redundant haplotigs, and filtering out contamination resulted in a final genome assembly of 157.9 Mbp, consisting of 22 contigs with an N50 of 10.0 Mbp (Figure 1B, Table 1). Next, we assessed the completeness and quality of the genome by comparing the k-mers of the de novo assembly to those of the unassembled reads using Merqury (Rhie et al. 2020). The base pair quality of the *E. gracile* genome was established as 66.0, i.e., an error rate of less than 1/5,000,000. As many as 99.8% of the k-mers mapped to the combined primary and alternative assembly, while 79.1% mapped to the assembly using only the primary one. Moreover, 99.1% of the assembly mapped to the putative chromosomes as determined by Hi-C scaffolding. The 0.7% of difference between the k-mer completeness and the mapped assemblies is probably due to remaining contigs resulting from contamination (e.g., gut content). The assembly, with a gene count of 20,684, is rather entire, with 95.3% out of 954 BUSCO markers detected as complete genes (including 0.2% duplicated markers), plus another 1.0% detected as fragmented. The selected

values for the different assembly parameters for each step (i.e., the unpurged primary genome assembly, the primary assembly after purging haplotypic duplications, the decontaminated primary genome assembly, and the HiC scaffolded genome) are shown in Table 1.

For Hi-C, 52,305,505 reads were obtained, corresponding to a coverage of 49.7x. After scaffolding with Hi-C reads, the longest scaffold was 17.8 Mbp, and the scaffold L50 was 6 (for further details, please see Table 1). Although a karyotype is so far unavailable for this species, our Hi-C assembly and structural annotation allow us to determine that the *E. gracile* genome contains between 17 and 19 chromosomes (Fig. 1C). This chromosome number estimation is supported by an analysis of telomere-associated motifs using the telomere identification kit tidk (Brown et al. 2023) (see Fig. S2). After annotation, the resulting BUSCO protein score was 89.6% complete (84.6% single copy, 5.0% duplicated, 3.1% fragmented). Our genome annotation predicted 20,684 protein-coding genes.

With a genome size estimate ranging from 153.7 to 161.8 Mbp, the *Emplectonema gracile* genome is shorter than the size of 210 ± 5 Mbp previously estimated based on flow cytometry (Paule et al. 2021). Moreover, the *E. gracile* genome is substantially smaller than those of the species previously sequenced: *Lineus longissimus* (391.2 Mbp in 19 putative chromosomes) (Kwiatkowski et al. 2021), *Cephalotrix simula* (427.7 Mbp in 262,498 scaffolds), *Cephalothrix spiralis* (650.6 Mbp in 427,796 scaffolds), *Notospermus geniculatus* (558.6 Mbp in 11,108 scaffolds) (Luo et al. 2018), *Tubulanus punctatus* (367.7 Mbp in 230,562 scaffolds), and *Tubulanus ruber* (312.5 Mbp in 205,164 scaffolds). As reported by Paul et al. 2021, such a genome size variation among nemertean genomes may be related to evolutionary processes of genome expansion and reduction. These genome size differences can be ascribed to the reproductive and developmental strategies of these species, among others (Paule et al. 2021).

Despite being the *E. gracile* genome shorter than the other nemertea species previously sequenced, the congruence in gene counts between our Hoplonemertea species (20,684 annotated genes), *N. geniculatus* (Pilidiophora sp.; 20,473 annotated genes; Luo et al. 2018), and *L. longissimus* (Pilidiophora sp.; 21,203 annotated genes; NCBI RefSeq GCF_910592395.1) reinforces the accuracy of our genome assembly.

The repetitive content of the *E. gracile* genome (30.5%) is slightly lower than that reported for the genomes of the nemertean *Notospermus geniculatus* (37.5%) and of the phoronid *Phoronis australis* (39.4%), and higher than the one reported for the brachiopod *Lingula anatina* genome (23.3%) (Luo et al. 2018). A list of the repetitive elements found on the *E. gracile* genome grouped in their respective families, as well as their relative occurrence, can be seen on Table S3.

Materials and Methods

Sampling

For this work, two specimens of *Emplectonema gracile* of unknown sex were selected. The specimens, approximately 10 cm long, were bisected. The anterior body part of N53 was preserved in 4% formaldehyde as a voucher (Natural History Museum, University of Oslo, Norway; catalog number NHMO C7190). Remaining pieces of N53, and all N59, were flash frozen in liquid nitrogen. One specimen was used for HiFi sequencing (specimen ID N59), while the other one was used for Hi-C sequencing (specimen ID N53) (Lieberman-Aiden et al. 2009; Hu et al. 2021). Both individuals were collected on October 16th 2020 from a beach on Jeløya (Moss, Viken, Norway; N 59º25’23.6”, E 010º37’05.9”; WGS84; ±2 m).

Identification

Morphological identification and DNA barcoding were performed for both specimens. Specimen N59 was barcoded using the DNA extracted for HiFi sequencing, while the DNA of N53 was extracted using the Qiagen DNeasy Blood and Tissue Kit following manufacturer's instructions. 16S and 18S rRNA gene sequences of N59 were PCR amplified using the following primers: forward (16S: 5' CCGGTCTGAACTCAGATCACGT 3'; 18S: 5' CCCCGTAATTGGAATGAGTACA 3'), reverse (16S: 5' CGCCTGTTTATCAAAAACAT 3'; 18S: 5' AGCTCTCAATCTTGTCAATCCT 3'); and the following settings: 1x 2 minutes at 94 ºC, 40x [30 seconds at 94 ºC, 60 seconds at 51 ºC, 60 seconds at 72 ºC], 1x 2 minutes at 72 ºC. For N53, CO1 was PCR amplified using forward primer LCO1490-JJ (5' CHACWAAYCATAAAGATARYGG 3'), and reverse primer HCO2198 JJ (5' AWACTTCVGGRTGVCCAAARAARCA 3') (Astrin and Stüben 2008). Resulting PCR products were Sanger sequenced by Macrogen (Amsterdam). The 16S, 18S, and COI sequences confirmed the morphological identification (complete sequences can be found associated with BioSamples SAMN39983511 and SAMN39991962 in NCBI). The COI sequence for N53 was identical to the COI sequence of the publicly available mitochondrial genome of *E. gracile* (NC_016952.1). The mitochondrial genome of the *Emplectonema gracile* specimen N59 determined by BLAST had a >99% similarity to the previously published mitochondrial genome for the same species (NCBI accession number JF727825).

Genome Sequencing

For HiFi sequencing, DNA was extracted from posterior parts of N59. Samples were weighed and minced on dry ice followed by tissue disruption using a TissueRuptor II (QIAGEN, Germany) on its maximum settings for 10 seconds. High molecular weight (HMW) DNA was

extracted using the Nanobind Tissue Big DNA kit (Circulomics Inc, USA). DNA quality and concentration were determined with a Nanodrop UV/Vis spectrophotometer (Thermo Fisher Scientific, USA), a Qubit BR dsDNA assay (Thermo Fisher Scientific), a pulsed field gel, and a Fragment Analyzer (Agilent, USA) run. Low molecular weight DNA smaller than 15 kb was removed using the BluePippin (Sage Science, USA) system with High Pass Plus Gel cassettes. DNA was further purified and concentrated using the AMpure XP purification kit (Beckman Coulter, USA). A final concentration of 143 ng/μL in a volume of 75 μL was obtained. The library for HiFi circular consensus sequencing was constructed and sequenced on a SEQUEL II (Pacific Biosciences) platform at the Norwegian Sequencing Centre (Oslo, Norway).

For Hi-C sequencing, a library was prepared using the Arima High Coverage Hi-C+ Kit (Arima Genomics, USA). Specifically, the restriction enzymes used for Arima Hi-C 2.0 cut at the following recognition sites: ^GATC, G^ANTC, C^TNAG, T^TAA. For this reaction, ~40 mg of disrupted tissue was used. Subsequently, a library was generated following the manufacturer's instructions. The quality of extracted HWM DNA is crucial for obtaining accurate sequencing results; therefore, concentrations and fragment length distributions were assessed using Qubit (Thermo Fisher Scientific, USA) and Fragment Analyzer (Agilent, USA). Additionally, the GC content of the Illumina libraries was measured using a Kapa library quantification kit (Roche, Switzerland). The final barcoded library was pooled on a quarter S4 flow cell in 2x150 bp paired-end mode on an Illumina NovaSeq sequencer (Illumina, USA). Hi-C library preparation and sequencing were done at the Norwegian Sequencing Centre (Oslo, Norway).

Genome Profiling

Genome profiling steps were followed to assess k-mer frequencies within raw sequencing reads, and to estimate major genome characteristics such as genome size, heterozygosity, and

repetitiveness. The k-mer distribution, with a k-mer size of 21, was generated using Jellyfish 2.3.0 using default settings. Based on this k-mer distribution, SmudgePlot 0.2.4 was run to test the ploidy of the genome (Marcais et al. 2012; Ranallo-Benavidez et al. 2020). GenomeScope 2.0 (Ranallo-Benavidez et al. 2020) with a k-mer size of 21, diploid level and a high-bound value of 1 million, was used to calculate the genome size, repetitiveness, and heterozygosity for a diploid genome using a combinatorial approach fitting a mathematical model to the k-mer distribution.

De Novo Genome Assembly

Assembly of HiFi reads was carried out with Hifiasm 0.18.2 (Cheng et al. 2021), using default settings. Haplotypic duplications were purged with Purge_dups 1.4 (Guan et al. 2020). The primary assembly was checked for contamination and corrected using the BlobToolKit 3.1.4 software (Challis et al. 2020), leading to the removal of 0.97% of the primary assembly content. In this removal step, all contigs not matching to a metazoan phylum, but to Proteobacteria, Ascomycota, Verrucomicrobia, or to any other non-metazoan clade were excluded (Fig. S1C).

Computational Scaffolding

The Arima Mapping Pipeline (https://github.com/ArimaGenomics/mapping_pipeline) was used for mapping raw Hi-C reads to the purged and decontaminated assembly outlined above. Briefly, Hi-C paired reads are first aligned to the reference independently using BWA-MEM to identify potential chimeric reads to be filtered out. Filtered single-end Hi-C reads are then paired and sorted based on mapping quality to produce a quality filtered, paired-end BAM file. Picard Tools is then used to flag PCR duplicates which are then discarded using SAMtools (Camacho et al. 2009). The quality filtered BAM file was then used as input for scaffolding. Scaffolding was performed in YaHS: yet another Hi-C scaffolding tool (https://github.com/c-zhou/yahs) (Zhou et

al. 2022). The output of YaHS was then converted into .hic and .assembly files using Juicer tools 1.9.9_jcuda.0 (Durand et al. 2016). The Hi-C contact map generation was done using Juicebox Assembly Tools 1.9.1 (Robinson et al. 2018).

Quality Control Checks

Several quality control checks were conducted after each analytical step (i.e., unpurged assembly, assembly after purging, decontaminated assembly, and scaffolded genome). Quast 5.0.2 (Gurevich et al. 2013) was used to determine statistical parameters of the primary genome assembly. Using Merqury 1.3 (Rhie et al. 2020), a meryl database was generated, and quality statistics such as base pair quality and k-mer completeness retrieved. The different assemblies were benchmarked against the 954 universal single-copy orthologs of the metazoa_odb10 dataset using BUSCO+ 5.5.0 (Simão et al. 2015, Manni et al. 2021). In preparation for BlobToolKit, Blast+ 2.13.0 (Camacho and Madden 2013) was used to map each contig of the assembly against a local copy of the NCBI nucleotide (nt) database downloaded as part of the pipeline. Additionally, HiFi reads were mapped against the primary assembly with Minimap2 2.17 (Li 2018), and further prepared for BlobTools with SAMtools 1.10 (Camacho et al. 2009). The BUSCO scores, BLAST results, and read coverage were uploaded and further analyzed within BlobToolKit 3.1.4. After assembly, mitochondrial genome sequences were retrieved using Blast+ and the amino-acid sequences of all protein-coding genes of the mitochondrial genome NC_000931. These mitochondrial sequences were queried against the nt database with Blast+ 2.13.0 to confirm species identification and possible sources of contamination.

Genome Annotation

For structural annotation, RepeatModeler 2.0.1 (Flynn et al. 2020) was used to model repeat content followed by soft repeat masking utilizing RepeatMasker 4.1.2 with default settings (Smith et al. 2015). As no transcriptome was available for this species, annotation was performed using the Braker 3 (Hoff et al. 2019) pipeline based on protein sequences from closely related species. 37 publicly available nemertean transcriptomes belonging to 25 species (Table S1) were downloaded from NCBI, assembled with Trinity (Grabher et al. 2011), and translated with TransDecoder (https://github.com/TransDecoder/). The translated transcriptomes, combined with the OrthoDB v10 metazoa dataset, were used to generate protein prediction hints with ProtHint 2.6.0 (Hoff et al. 2019). The ProtHint mapping pipeline was used by Braker 3 to produce protein hints to train the model. The soft-masked and decontaminated primary genome assembly and protein databases were used as input to Braker 3. The protein set derived from our annotation was isoform filtered to only include one protein per locus using "AGAT, Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format" (Dainat, 2020). The completeness of gene annotations was evaluated using BUSCO+ 5.5.0 (Simão et al. 2015) with the metazoa odb_10 database, along with the designated nemertean transcriptomes listed in Table S1.

Data Availability

All unprocessed sequence data, as well as the Refseq genome assembly can be found at the NCBI Sequence Read Archive under Bioproject PRJNA1077883. The genome sequence is released openly for reuse. All custom scripts are available at GitHub https://github.com/torstenstruck/InvertOmics.

Ethical approval

The Nagoya protocol does not apply to this work. Both sample collection and molecular work were done in Norway.

Acknowledgements

This work was funded by the Research Council of Norway project "InvertOmics – Phylogeny and evolution of lophotrochozoan invertebrates based on genomic data" (Project number: 300587 to THS). KMK, MKY, and NGR were funded by NSF DEB-1846174. We thank Matz Berggren (University of Gothenburg) for assisting and allowing AVG to use the camera setup used for taking the picture of a specimen of *E. gracile* included in Figure 1A, and Miguel Ángel Naranjo Ortiz (University of Oslo) and Emanuela Di Martino (University of Catania) for discussion. We thank Ann Mc Cartney, Giulio Formenti, and Alice Mounton whose efforts were instrumental in establishing and coordinating the ERGA Pilot Project.

Author Contributions

AVG and THS collected and preserved the specimens of *Emplectonema gracile*, identified by AVG. AVG carried out the molecular parts of this work which were not conducted at the Norwegian Sequencing Center, except for barcoding the voucher specimen N53, done by ATC. AVG ran the different genome profiling, genome assembly, and quality control checks developed by THS. MY-C performed Hi-C scaffolding of the genome plotted by AVG. NR and AVG performed the genome structural annotation. AVG, KMK, MM, and THS conceived the study. AVG wrote the first draft of the manuscript, and all authors contributed to the submitted version.

References

Ament-Velásquez SL,etal. Population genomics of sexual and asexual lineages in fissiparous ribbon worms (Lineus, Nemertea): hybridization, polyploidy and the Meselson effect. Mol Ecol. 2016:25(14):3356–3369. https://doi.org/10.1111/mec.13717.

Andrade SC, et al. A transcriptomic approach to ribbon worm systematics (Nemertea): resolving the Pilidiophora problem. Mol Biol Evol. 2014:31(12): 3206–3215. https://doi.org/10.1093/molbev/msu253.

Astrin JJ, Stüben PE. Phylogeny in cryptic weevils: molecules, morphology and new genera of western Palaearctic Cryptorhynchinae (Coleoptera: Curculionidae). Invertebr Syst. 2008:22(5):503–522. https://doi.org/10.1071/IS07057.

Bleidorn C. Recent progress in reconstructing lophotrochozoan (spiralian) phylogeny. Org Divers Evol. 2019:19(4):557–566. https://doi.org/10.1007/s13127-019-00412-4.

Brown, M, González De la Rosa PM, Mark B. A Telomere Identification Toolkit. Zenodo. 2023. https://doi.org/10.5281/zenodo.10091385.

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 Genom Bioinform. 2021:3(1):lqaa108. https://doi.org/10.1093/nargab/lqaa108.

Camacho C, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009:10:1–9. https://doi.org/10.1186/1471-2105-10-421.

Camacho C, Madden T. BLAST+ release notes. BLAST® Help. 2013 [Internet].

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit–interactive quality assessment of genome assemblies. G3 (Bethesda). 2020:10(4):1361–1374. https://doi.org/10.1534/g3.119.400908.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021:18(2):170–175. https://doi.org/10.1038/s41592-020-01056-5.

Coe WR. *On the development of the pilidium of certain nemerteans* (Vol. 10). 1899 Connecticut Academy of Arts and Sciences, USA.

Drábková M, et al. Different phylogenomic methods support monophyly of enigmatic 'Mesozoa' (Dicyemida+Orthonectida, Lophotrochozoa). Proc R Soc Lond B Biol Sci. 2022:289(1978):20220683. https://doi.org/10.1098/rspb.2022.0683.

Dudchenko O, et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. BioRxiv. 2018: 254797. https://doi.org/10.1101/254797.

Durand NC, et al. Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016:3(1):95–98. http://dx.doi.org/10.1016/j.cels.2016.07.002.

Egger B, et al. A transcriptomic-phylogenomic analysis of the evolutionary relationships of flatworms. Curr Biol. 2015:25(10):1347–1353. http://dx.doi.org/10.1016/j.cub.2015.03.034.

Flynn JM, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA. 2020:117(17): 9451–9457. https://doi.org/10.1073/pnas.1921046117.

Gibson R. *Nemerteans* 2nd ed. 1994: vol. 24. Field Studies Council, USA.

Giribet G, Edgecombe GD. *The Invertebrate Tree of Life*. 2020. Princeton University Press, USA.

Grabherr MG, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2011:29(7):644. http://dx.doi.org/10.1038/nbt.1883

Guan D, et al. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020:36(9):2896–2898. https://doi.org/10.1093/bioinformatics/btaa025.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013:29(8):1072–1075. https://doi.org/10.1093/bioinformatics/btt086.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016:32(5):767–769. https://doi.org/10.1093/bioinformatics/btv661.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. *Gene Prediction: Methods and Protocols*. Humana, New York, NY. 2019, p. 65–95. https://doi.org/10.1007/978-1-4939-9173-0_5.

Hoff KJ, Stanke M. Predicting genes in single genomes with AUGUSTUS. Curr Protoc Bioinformatics. 2019:65(1):e57. https://doi.org/10.1002/cpbi.57.

Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. Hum Immunol. 2021:82(11):801–811. https://doi.org/10.1016/j.humimm.2021.02.012.

Jiang Z, Rokhsar DS, Harland RM. Old can be new again: HAPPY whole genome sequencing, mapping and assembly. Int J Biol Sci. 2009:5(4):298. https://doi.org/10.7150/ijbs.5.298.

Kocot KM, et al. Phylogenomics of Lophotrochozoa with consideration of systematic error. Syst Biol. 2017:66(2):256–282. https://doi.org/10.1093/sysbio/syw079.

Kvist S, Laumer CE, Junoy J, Giribet G. New insights into the phylogeny, systematics and DNA barcoding of Nemertea. Invertebr Syst. 2014:28(3):287–308. https://doi.org/10.1071/IS13061.

Kwiatkowski D, Blaxter M, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines

collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium. The genome sequence of the bootlace worm, *Lineus longissimus* (Gunnerus, 1770). Wellcome Open Res. 2021:6. https://doi.org/10.12688/wellcomeopenres.17193.1.

Laumer CE, et al. Spiralian phylogeny informs the evolution of microscopic lineages. Current Biol. 2015:25(15):2000–2006. http://dx.doi.org/10.1016/j.cub.2015.06.068.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018:34(18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009:326(5950):289–293. https://doi.org/10.1126/science.118136.

Luo YJ, et al. Nemertean and phoronid genomes reveal lophotrochozoan evolution and the origin of bilaterian heads. Nat Ecol Evol. 2018:2(1):141–151. https://doi.org/10.1038/s41559-017-0389-y.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. MBE. 2021:38(10):4647–4654. https://doi.org/10.1093/molbev/msab199.

Marcais G, Kingsford C. Jellyfish: A fast k-mer counter. Version 1.1.4. Tutorialis e Manuais. 2012:1:1–8.

Marlétaz F, Peijnenburg KT, Goto T, Satoh N, Rokhsar DS. A new spiralian phylogeny places the enigmatic arrow worms among gnathiferans. Current Biol. 2019:2012:29(2):312–318. https://doi.org/10.1016/j.cub.2018.11.042.

Maslakova SA, Hiebert TC. From trochophore to pilidium and back again-a larva's journey. Int J Dev Biol. 2015:58(6–8):585–591. https://doi.org/10.1387/ijdb.140090sm.

Paule J, von Döhren J, Sagorny C, Nilsson MA. Genome Size Dynamics in Marine Ribbon Worms (Nemertea, Spiralia). Genes. 2021:12(9):1347. https://doi.org/10.3390/genes12091347.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020:11(1):1432. https://doi.org/10.1038/s41467-020-14998-3.

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020:21(1):1–27. https://doi.org/10.1186/s13059-020-02134-9.

Schmidt GA. Issledovania po embryologii nemertin. II. Pilidii *Cerebratulus pantherinus* i *marginatus*. Russkii Zool Zh. 1930:10:113–127.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015:31(19):3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2015:2013–2015.

Smit AFA, Hubley S, Green P. RepeatMasker. 2021:Version 4.1. 2.

Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008:24(5):637–644. https://doi.org/10.1093/bioinformatics/btn013.

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. 2006:7(1):1–11. https://doi.org/10.1186/1471-2105-7-62.

Stricker SA, Cloney RA. The ultrastructure of venom-producing cells in *Paranemertes peregrina* (Nemertea, Hoplonemertea). J Morphol. 1983:177(1):89–107. https://doi.org/10.1002/jmor.1051770108.

Struck TH, Fisse F. Phylogenetic position of Nemertea derived from phylogenomic data. MBE. 2008:25(4):728–736. https://doi.org/10.1093/molbev/msn019.

Struck TH, et al. Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of Spiralia. MBE. 2014:31(7):1833–1849. https://doi.org/10.1093/molbev/msu143.

Uliano-Silva M, et al. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. BMC Bioinformatics. 2023:24(1):288. https://doi.org/10.1186/s12859-023-05385-y.

Vlasenko AE, Kuznetsov VG, Magarlamov TY. Investigation of Peptide Toxin Diversity in Ribbon Worms (Nemertea) Using a Transcriptomic Approach. Toxins. 2022:14(8):542. https://doi.org/10.3390/toxins14080542.

von Reumont BM, et al. Proteo-transcriptomic analysis identifies potential novel toxins secreted by the predatory, prey-piercing ribbon worm *Amphiporus lactifloreus.* Mar Drugs. 2020:18(8):407. https://doi.org/10.3390/md18080407.

Vurture GW, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017:33(14):2202–2204. https://doi.org/10.1093/bioinformatics/btx153.

Zattara EE, Fernández-Álvarez FA, Hiebert TC, Bely AE, Norenburg JL. A phylum-wide survey reveals multiple independent gains of head regeneration in Nemertea. Proc R Soc Lond B Biol Sci. 2019:286(1898):20182524. https://doi.org/10.1098/rspb.2018.2524.

Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 2023:39(1):btac808. https://doi.org/10.1093/bioinformatics/btac808.

Figures and Tables

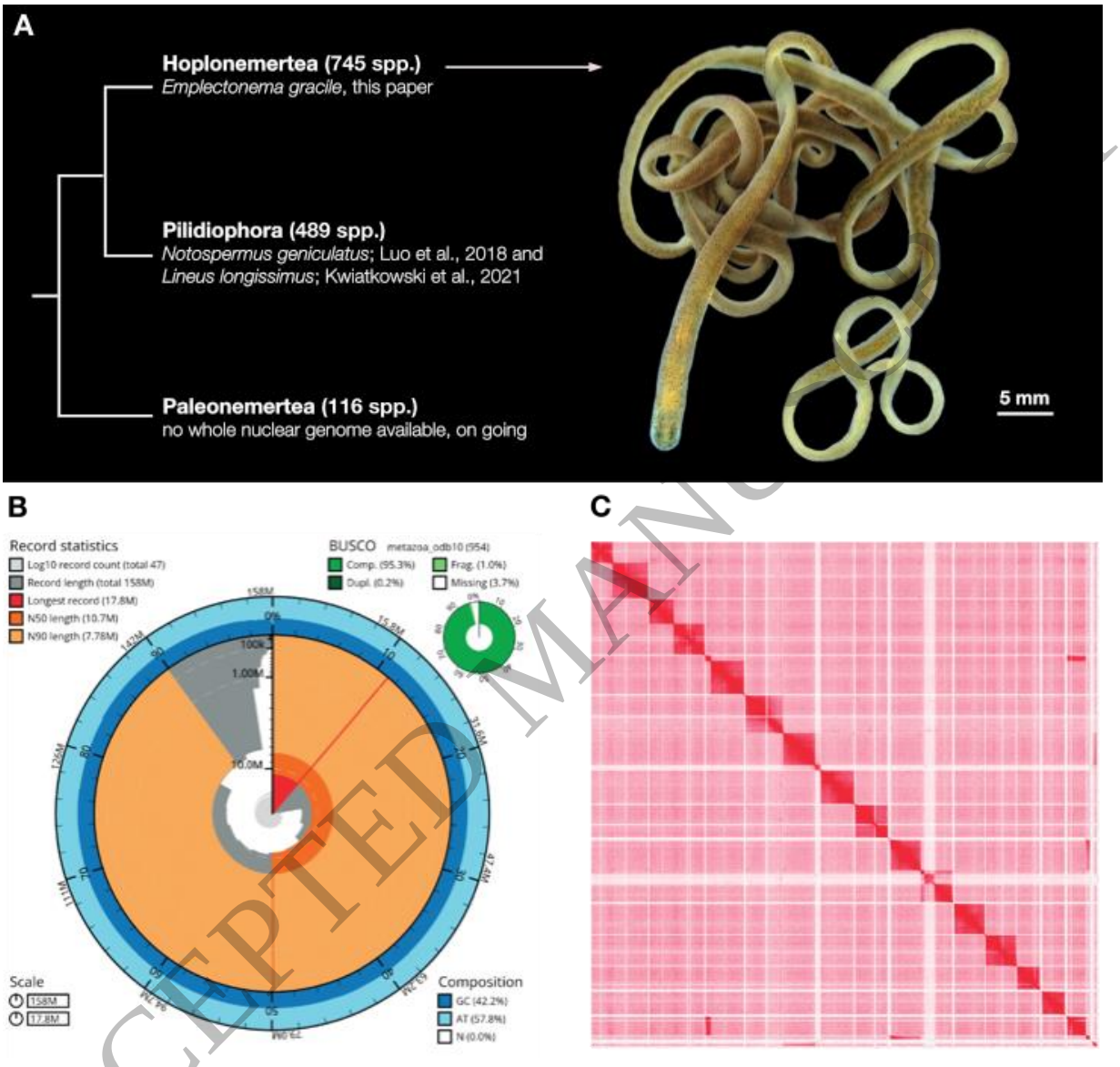

Fig. 1. – The ribbon worm *Emplectonema gracile* and its genome. **(A)** Phylogeny of subgroups within the Nemertea phylum, with the name of the species for which a whole genome is available (left), and adult specimen of *Emplectonema gracile* (right), photo taken by AVG. Species count based on Catalog of Life (https://www.catalogueoflife.org/data/taxon/BMH45). (**B)** "Snailplot" produced with BlobToolKit, illustrating N50 metrics and BUSCO gene completeness. **(C)** Hi-C contact map representing the final genome assembly of *E. gracile*, visualized with Juicebox 1.9.1.

ASSEMBLY METRICS	Unpurged primary genome assembly	Primary genome assembly after purging haplotypic duplications	Decontaminated primary genome assembly	HiC scaffolded genome	Benchmark
Span (Mb)	161.8	158.5	157.9	157.9	–
Number of contigs	135	49	22	–	–
Contig N50 length (Mb)	10.0	10.0	10.0	–	≥ 10
Longest contig (Mb)	13.1	13.1	13.1		
Contig L50 length	8	8	8	–	–
Consensus quality (QV)	61.6 (primary only)/ 62.8 (primary & alternative)	65.0 (primary only)	67.0 (primary only)	67.0 (primary only)	≥ 50
k-mer completeness	79.1% (primary only)/ 99.8% (primary & alternative)	79.1% (primary only)	79.1% (primary only)	79.1% (primary only)	≥ 95%
BUSCO scores (n:954)	C:95.6%[S:95.3%,D:0.3%], F:0.7%,M:3.7%	C:95.6%[S:95.3%,D:0.3%], F:0.7%,M:3.7%	C:95.5%[S:95.2%,D:0.3%], F:0.8%,M:3.7%	C:95.3%[S:95.1%,D:0.2%], F:1.0%,M:3.7%	C ≥ 95%
BUSCO protein	–	–	–	C:89.6%[S:84.6%,D:5%], F:3.1%,M:7.3%	–
Percentage of assembly mapped to chromosomes	–	–	–	99.1%	≥ 95%
Number of scaffolds	–	–	–	47	–
Scaffold N50 length (Mb)	–	–	–	10.7	≥
Longest scaffold (Mb)	–	–	–	17.8	–
Scaffold L50 length	–	–	–	6	–
Organelles Mitochondrial genome assembled	complete single contig	–	–	–	–

Table 1. – Project accession data and Assembly information for *E. gracile*. *BUSCO scored based on the metazoan_odb10 BUSCO set using v5.1.2. C=complete [S=single copy, D=duplicated], F=fragmented, M=missing, n=number of orthologues in comparison.