Single worm long read sequencing reveals genome diversity in free-living nematodes

Yi-Chien Lee[1,2,3], Hsin-Han Lee[1], Huei-Mien Ke[4], Yu-Ching Liu[1], Min-Chen Wang[5], Yung-Che Tseng[5], Taisei Kikuchi[6] and Isheng Jason Tsai[1,2,*]

[1]Biodiversity Research Center, Academia Sinica, Taipei 115, Taiwan
[2]Biodiversity Program, Taiwan International Graduate Program, Academia Sinica and National Taiwan Normal University, Taipei, Taiwan
[3]Department of Life Science, National Taiwan Normal University, 116 Wenshan, Taipei, Taiwan
[4]Department of Microbiology, Soochow University, Taipei, Taiwan
[5]Marine Research Station (MRS), Institute of Cellular and Organismic Biology, Academia Sinica, I-Lan County, Taiwan
[6]Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba, 277-8562, Japan

* Email: ijtsai@gate.sinica.edu.tw

Abstract

Obtaining sufficient genetic material from a limited biological source is currently the primary operational bottleneck in studies investigating biodiversity and genome evolution. In this study, we employed multiple displacement amplification (MDA) and Smartseq2 to amplify nanograms of genomic DNA and mRNA, respectively from individual *Caenorhabditis elegans*. Although reduced genome coverage was observed in repetitive regions, we produced assemblies covering 98% of the reference genome using long-read sequences generated with Oxford Nanopore Technologies (ONT). Annotation with the sequenced transcriptome coupled with the available assembly revealed that gene predictions were more accurate, complete and contained far fewer false positives than *de novo* transcriptome assembly approaches. We sampled and sequenced the genomes and transcriptomes of 13 nematodes from Dorylaimia, Enoplia, and early-branching species in Chromadoria. These free-living species had larger genome sizes, ranging from 147-792 Mb, compared to those of the parasitic lifestyle. Nine mitogenomes were fully assembled and displaying a complete lack of synteny to other species. Phylogenomic analyses based on the new annotations revealed strong support for Enoplia as sister to the rest of Nematoda. Our result demonstrates the robustness of MDA in combination with ONT, paving the way for the study of genome diversity in the phylum Nematoda and beyond.

Introduction

A genome reference is a prerequisite for a complete understanding of the biology and evolution of a species. Advances in long-read sequencing, together with increasing affordable costs (1), have paved the way for the ambition to study and generate genomes for the entire group of species, including every bird, vertebrate, insect, or eukaryote on Earth (2-5). However, the major challenge remains to obtain high-quality DNA and RNA from the majority of organisms across the tree of life (6). Such requirements are challenging to meet in microscopic organisms that cannot be cultured, leading to sampling bias and loss of their biological information on genetic and evolutionary studies (7-9). Recent advances in whole genome amplification (WGA) have enabled single cells or limited samples to generate sufficient DNA for sequencing (10,11), and have been applied to eukaryotic microorganisms, including fungi (9,12), marine phytoplankton (13) and parasitic nematodes (14,15) for genomic and population genetic studies (9,16).

This study investigated the feasibility of WGA combined with long-read sequencing for nematodes, which are the most abundant metazoans on Earth. More than one million nematode species are estimated to exist, but only approximately 30,000 species have been described so far (17,18). The Nematoda phylum has been classified on the basis of 18S ribosomal RNA (18S rRNA) into three lineages and five major clades: Dorylaimia (clade I), Enoplia (clade II), and Chromadoria (clade III-V). Chromadoria further includes Spirurina (clade III), Tylenchina (clade IV) and Rhabditina (clade V) as well as early derived lineages including Araeolaimida, Chromadorida, Desmodorida, Monhysterida, and Plectida (19-21). The roundworm *Caenorhabditis elegans* was the first animal to have its genome sequenced, with a size of 100.2 Mb. Since then, over 200 nematode genomes and mitogenomes of have been published (22,23). Of these, ~72% are mainly terrestrial parasites belonging in Dorylaimia, Spirurina, and Tylenchina because of their importance in plant crop and animal health. The remaining species are terrestrial free-living nematodes in Rhabditina (17,24,25). In contrast, only one genome of a marine nematode *Plectus sambesii*, is available (26-28), despite the fact that marine nematodes comprise half of all recorded nematodes and play a crucial role in benthic communities as decomposers, predators, food sources, and bioindicators (29). Only a few marine nematode species belonging to Monhysterida and Rhabditida (30) can be cultured. Thus, obtaining enough genomic DNA for

sequencing is a challenging for most of these species, making them potential candidates for the WGA techniques.

The Enoplia clade and the early derived Chromadoria lineage (19), found primarily in marine habitats, currently lack genomic data and have several important implications. Of particular interest is the phylogenetic relationship of basal nematodes, which remains unresolved due to insufficient sampling and limited resolution of 18s rRNA (20). Resolving the phylogenetic relationship of nematodes can help to understand the genomic basis of nematode diversity and the processes of evolution from free-living to parasitic life style (21,31). Increased sampling of marine nematodes and phylogenomic analyses based on mitogenomes or *de novo* transcriptomes (19,22,27,32,33) have shown better resolution of Enoplia sister to the rest of the Nematoda, but not entirely resolved (19).

Here, we have developed an assembly and annotation workflow capable of generating long genome sequences and accurate gene predictions from single nematodes. We quantified the coverage biases in the amplified sequences, produced assemblies, and assessed the accuracy of the annotations using this workflow on *C. elegans* compared to those generated *de novo* without a genome available. The finalized workflow was applied to 13 free-living marine nematodes isolated from the Taiwanese coasts. Despite obtaining lower genome coverage in these nematodes, phylogenomic analyses were performed using a total of 210,777 newly annotated genes to resolve the positions of basal clades in the Nematoda phylum. Comparisons of the genomes and complete mitogenomes of these nematodes revealed remarkable variation in genome features not observed in well-studied clades and shed light on the early evolution of nematodes.

Materials and Methods

Single worm DNA extraction

C. elegans strain N2 was grown at 22°C on NGM plates with *Escherichia coli* strain OP50 and *Aphelenchoides besseyi* APVT strain was grown at 22°C on PDA plates with *Alternaria citri*. Worms were either washed with M9 buffer from NGM plate and pelleted in 15 ml centrifuge tube or washed with the same M9 buffer for three times and starved in M9 buffer included 1% Antibiotic-Antimycotic (Thermo Fisher Scientific, Massachusetts, 15240062) in 15 ml centrifuge tube for 24 hours. A total of 13 nematode species (*Epsilonema* sp., *Enoplolaimus lenunculus*, *Linhomoeus* sp., *Microlaimidae* sp., *Mesodorylaimus* sp. , *Paralinhomoeus* sp., *Ptycholaimellus* sp., *Trileptium ribeirensis*, *Sabatieria punctata*, *Rhynchonemsa* sp., *Theristus* sp., *Trissonchulus latispiculum*, *Trissonchulus* sp.) were collected in Taiwan between November 2020 to May 2022 (**Supplementary Table S1**). The sampling sites encompassed various seashores throughout Taiwan, as well as the seabed at a depth of 15-18 meters surrounding Guishan Island. Individual nematode was isolated and washed with 10% bleach for 10 seconds and transferred into with 200µl PCR tube containing lysis buffer (8µl direct PCR lysis reagent (Viagen, #102-T), 1 µl 5mg/ml Proteinase K, 1 µl 200mM DTT) and incubated in 65 ℃ for 20 minutes and 95 ℃ for 5 minutes. The DNA concertation were quantified with Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Massachusetts, 2339927) following manufacturer's instructions.

Whole genome amplification

Various sources of extracted genomic DNA (gDNA) were used for amplification (**Supplementary Figure S1**): (i) Whole worm: Single worm or ten adult worms were cut into pieces with 22-gauge needle in 200µl PCR tube. In one amplification instance, single whole worm was prepared with 5% DMSO was added in polymerase mix. Genomic DNA extraction and amplification was performed with Qiagen REPLI-g Kit (150023,150043,150343, Qiagen, German). (ii) Purified DNA: single worm DNA extract with lysis buffer (8µl direct PCR lysis reagent (Viagen, #102-T), 1 µl 5mg/ml Proteinase K, 1 µl 200mM DTT) and incubate in 65 ℃ for 20 minutes and 95 ℃ for 5 minutes. The samples were further purified using Ampure XP (1:1) (A63882, Beckman Coulter, US) and diluted in 10 µl elution buffer. The multiple displacement amplification

step was performed with REPLI-g Kit (150023, Qiagen) following manufacturer's instructions.

Genomic DNA library preparation, sequencing and assembly

For Oxford Nanopore sequencing, two digestion times were initially tested on *C. elegans* using 1.5 µg amplified genomic DNA from single or ten adult worms. Template were first digested with T7 endonuclease I (M0302L, NEB, USA) as suggested in the protocol available on the ONT community website, which was initially designed based on sequencing whole-genome-amplified genomic DNA in *E. coli* (SQK-LSK109; ver. WAL_9070_v109_revN_14Aug2019). Namely 15 minutes (as recommended in the original ONT protocol) and 30 minutes were tested. For the other nematodes, 3 µg amplified templates were digested with T7 endonuclease I for 30 minutes and subjected to library preparation according to the manufacturer's instructions. Oxford Nanopore libraries were prepared according to SQK-LSK109 and SQK-LSK110 protocols, and sequenced on a GridION instrument. Basecalling of Nanopore raw signals was performed using Guppy (ver. 6.1.2) into a total 224.5 Gb of raw sequences at least 1 kb or longer. A summary of the sequencing data is shown in **Supplementary Table S2**. For short read sequencing, the amplified genomic DNA was sent to Biotools Co., Ltd (New Taipei City, Taiwan) for library preparation and sequencing of Illumina 150bp paired-end reads on a Novaseq 6000 sequencer.

The Flye (ver. 2.9.1) assembler (34) was used to assemble the raw ONT reads, which were then polished by four iterations of Racon (35) (ver. 1.4.11), followed by Medaka (-m r941_min_sup_g507 or r103_sup_g507, ver. 1.2.0; https://github.com/nanoporetech/medaka). The consensus sequences were further corrected with Illumina reads using NextPolish (36)(ver. 1.4.0) and haplotigs were removed using Purge Dups (37) (v1.2.5). Contigs with non-nematode origins were excluded. Genome completeness was assessed using nematode dataset of BUSCO (38) (ver. 5.1.2). Raw Illumina reads were assembled using the Spades assembler (ver. v3.14.1; spades_sc)(39). The mitochondrial genome was assembled separately using by aligning Oxford Nanopore reads to mitochondrial protein-coding gene of 52 nematode species listed in **Supplementary Table S3** using DIAMOND (40), following the approach described in (41). The circlised assembly were further annotated and manually curated using two versions of Mitos(v1.0.5) and Mitos2(2.1.0) (42).

204

205 **Single worm RNA transcriptome sequencing and assembly**

206 The Smart-seq 2 protocol (43) was used to extract and amplify RNA from single adult
207 worms. The resulting cDNA was sent to Biotools Co., Ltd (New Taipei City, Taiwan)
208 for library preparation using the NEBNext® DNA Library Prep Kit (NEB,
209 USA,20015828,20015829), and sequenced for 150bp paired-ends on an Illumina
210 Hiseq 2500 sequencer. Individual sample statistics are provided in **Supplementary**
211 **Table S2**. Sequencing reads were quality and adaptor trimmed using Trimmomatic
212 (ver. 0.39) (44), and reads mapped to nematode ribosomal RNA genes in the NCBI
213 database were removed. On average, 30% of the reads in each sample were identified
214 as ribosomal RNA sequences, and after removal, 21-86% of the reads remained. *De*
215 *novo* transcriptome assemblies were generated using the Spades assembler (ver.
216 v3.14.1; option: -k=55,77)(39). The best protein encoding predictions from *de novo*
217 assembled transcripts were produced using Transdecoder (v 5.5.0)(45) integrating
218 homology information from the UniProt database (Release version 2021_03).

219

220 **Genome annotation**

221 Repetitive elements were identified using RepeatModeler (ver 2.1) (46),
222 TransposonPSI (ver 1.0.0; https://github.com/NBISweden/TransposonPSI) and
223 USEARCH (ver 11.0)(47) based on the protocol by Berriman *et al*. (48). Repetitive
224 DNA sequences were identified and masked using Repeatmasker (ver 4.1.2)(49).
225 Proportions of repeat content along the non-overlapped 100kb window were
226 calculated using BEDTools (ver 2.26.0)(50).

227

228 The proteomes of 11 representative nematodes were obtained from WormBase
229 WBP16 (51) and are listed in **Supplementary Table 4**. Single worm transcriptome
230 reads were mapped to the corresponding genome assemblies using STAR (ver.
231 2.7.7a) (52) and assembled using Trinity(ver 2.13.2; guided approach) (53), Stringtie
232 (ver 2.1.7) (54) and Cufflinks (ver 2.2.1)(55). Transcripts generated by Trinity were
233 mapped to the corresponding genome assemblies using Minimap2 (ver 2.1, options: -
234 ax splice) (56), and splice junctions were quantified using Portcullis(ver 1.2.3)(56). The
235 gene predictor Augustus (ver 3.4.0) and gmhmm(57) were trained using
236 BRAKER2(ver. 2.1.6) (52,58) and SNAP (59) with proteomes and RNAseq mappings
237 as evidence hints to generate an initial set of annotations. The assembled transcripts

selected by MIKADO(ver 2.3.3) (60), proteome homology, and BRAKER2 annotations were combined as evidence hints for input into the MAKER2 annotation pipeline(61) to produce a final annotation for each species. Comparison of all annotation results to the reference protein-coding gene were conducted using Gffcompare (ver. v0.11.2) (62).

Decontamination

To identify contigs of non-nematode origin, we used a combination of three methods, given the lack of uncontaminated nematode sequences in the database. First, we employed Kraken2 (ver. 2.1.2) (62) to determine the kingdom and phylum of scaffolds based on k-mers. We rebuilt the Kraken2 database to include six kingdoms and phyla: Archaea, Bacteria, Nematoda, Eukaryota (Annelida, Arthropoda, Cnidaria ,Chordata, Porifera, Placozoa, Platyhelminthes), Outgroup (Human), Viruses, and Undefined. A list of species in the reconstructed database is provided in **Supplementary Table S5**. Second, we annotated genes using Braker2 and aligned them against the NCBI nr database using BLAST to assign phylum categories, such as Nematoda, Bacteria, Eukaryota, Eukaryotea-undef, Candidatus, Fungi, Planta, Viruses, Algae, Archaea, and Unclassified. Third, we aligned RNAseq reads for each species to the corresponding genome assemblies using STAR and calculated the RNAseq mapping rate of each scaffold using BedTools. We excluded scaffolds assigned to bacteria by Kraken2 and those with genes that contained 90% or more bacterial proteins. For scaffolds that could not be identified by Kraken2 or the nr database, we removed those with RNAseq mapping rates below 1,000 reads.

Phylogenomics of nematodes

Protein data sets from 13 representative nematodes were download from WormBase WBP16 (51) (**Supplementary Table S6**). We also downloaded the assembled transcripts of one Araeolaimida, one Plectia, six Enoplia, and an outgroup Nematodmorpha from Smythe *et al* (19) (**Supplementary Table S6**). Orthogroups (OGs) were identified using OrthoFinder (63,64). Sequences in each OGs were aligned using MAFFT(v7.515)(65). A gene tree was inferred from each OG alignment using FastTree(ver 2.1.11)(66), and a species tree was inferred from all OGs gene trees using ASTRAL-Pro (67). Two sets of data were used to construct the nematode species phylogeny: (i) proteomes of nematode species downloaded from wormbase,

13 free-living nematodes sequenced in this study, and outgroup *Priapulus cauatus,* comprising a total of 27 species with 58,531 OGs (**Supplementary Table S6**), and (ii) the 27 species from (i), *de novo* transcriptomes from Smythe *et al.*, *Drosophila melanogaster, Priapulus cauatus,* and *Hypsibius dujardini*, comprising a total of 39 species with 77,411 OGs.

Results

Whole genome amplification facilitates sufficient DNA for long read sequencing from single nematodes

To study the genome diversity of free-living nematodes, we isolated nematodes from a variety of marine environments in Taiwan and extracted the genomic DNA (gDNA) from individual adults across 11 taxa, including the Enoplia clade, for which genome sequences are currently unavailable (**Supplementary Table S1**). Highlighting the challenge of obtaining sufficient gDNA for long-read sequencing across the Nematoda phylum, we found that yields ranged from 1.5 to 3.8 ng for individual adults and were not associated with worm size (tau = 0.2, P = 0.33, Kendall's tau test, **Supplementary Table S7**). To mitigate this problem, we used multiple displacement amplification (MDA) method to amplify whole genomes from individual nematodes, yielding 6.2-51.5 µg, corresponding to approximately 4,700X and 22,000X amplification using REPLI-g mini kit, REPLI-g midi or sc kit, respectively (**Supplementary Table S2**). To assess potential amplification bias, we first sequenced the genome of the model nematode *Caenorhabditis elegans* N2. A total of 35 µg genomic DNA was obtained after MDA from an initial 1.56 ng, and sequencing using an ONT 9.4.1 flow cell yielded 7.36 Gb with a N50 of 7.74 kb, corresponding to 73.6X depth of coverage (68) (**Supplementary Table S2**).

Whole genome amplification disparity in repetitive regions

MDA suffers from several challenges, the most important of which is highly uneven amplification (69,70). This issue can lead to incomplete genome assembly and reduced coverage in certain genome regions, and affect analyses such as copy number variants (70,71). We aligned the amplified and published unamplified ONT

reads (72) against the *C. elegans* genome, and the former clearly displayed an uneven depth of coverage with 19.2% of the non-overlapping 100kb window showing less than half of the genome-wide median (**Figure 1A**). Sequencing of the amplified DNA using Illumina short reads also exhibited similar patterns, suggesting that the MDA rather than the sequencing platforms were causing the bias (**Figure S2A**). This unevenness was clearly associated with the arm-center-arm domain structures exhibited by the nematode autosomes. In the six *C. elegans* chromosomes, the arm region contains a significantly higher number of repeat sequences ($P<0.001$, Wilcoxon rank sum test) and a lower read coverage ($P<0.001$, Wilcoxon rank sum test; **Supplementary Figure S2B,2C**) compared to the center region. However, in the X chromosome, the read coverage is more balanced due to the lower percentage of repeats in the arm region (Median of repeats 16-21% vs. 15-39% in autosomes; **Supplementary Figure S2D**).

We calculated the proportions of 17 genomic features in 100kb non-overlapping windows and found that the presence of rolling-circle transposable elements (RC/Helitron) contributed least to sequence coverage ($R^2=0.36$,$P<$ 22e-16, Pearson test), followed by unclassified repeats, DNA transposable elements and Satellite (**Supplementary Figure S3**), suggesting that the reduced coverage displayed at autosome arms were mainly due to enriched repeats (68). Within repeat class, small repeats were the primary repeat class affecting coverage (**Figure 1B; Supplementary Figure S4**). Of the 6,332 rolling-circle transposable elements regions totaling 1,409,947 bp, on average a reduced 36.7% of coverage compared to the genome median (68X vs. 25X; **Supplementary Figure S4**). The coverage of amplified data on the gene, retrotransposon, and other small repeats were significantly lower than unamplified data (**Figure 1B**). Finally, we determined 0.49 Mb that were not sequenced at all ranging from 2 to 40,137 bp. As expectedly, 17% were repeats and 80 (0.3%) of genes were affected. Ten genes located at near the arm of III, IV, V, and X chromosome were not completely sequenced (**Figure 1C**). At the exon level, 0.09% were have also been affected, including 96 not sequenced CDS.

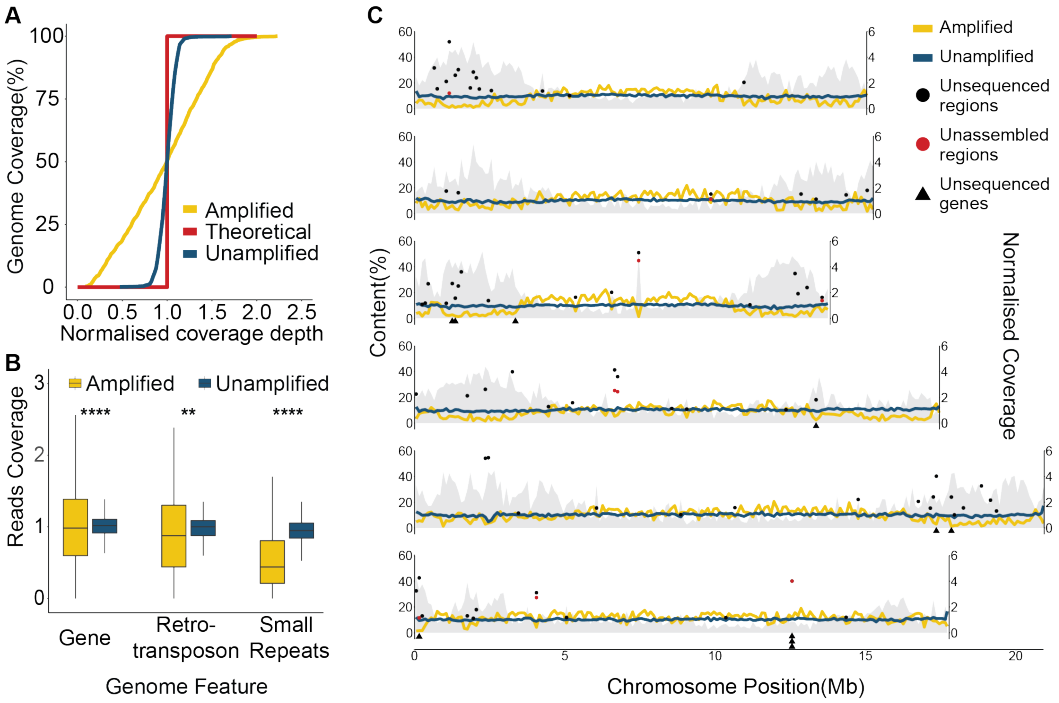

Figure 1. Sequencing coverage of *C. elegans* genomic DNA A. Cumulative genome coverage versus genome wide median. The red line indicates the theoretical coverage of unbiased coverage. More derivation away from this line suggest less uniformity across the genome. **B.** Normalized read coverage on genes and repeats. Repeats were categorised into two groups: retrotransposons and small repeats. Retrotransposons include long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeats (LTRs). DNA transposons, RC Helitron, rRNA, snRNA, tRNA, satellite, simple repeat, and unknown repeats were labeled as small repeats. **: $P < 0.01$, ****: $P < 0.0001$. **C.** Lines represent normalized read coverage of amplified and unamplified data. The black and red dots represent the top 10% regions (11.1-44.8kbp and 10.1-54.5 kbp) that were not sequenced and not assembled. The triangles represent the position of genes that were not sequenced at all. The shaded area indicates the proportions of repeats along the chromosomes.

Due to differences in repeat content between species, we sought to evaluate the impact of MDA by analysing the amplified reads from the genome of the plant parasitic nematode *Aphelenchoides besseyi*, which has a smaller genome (44.7 Mb) and lower repeat content (5.38%) compared to other nematodes (73). A similar but less pronounced pattern of uneven coverage was observed in this nematode compared to

C. elegans. (**Supplementary Figure S5A**). In the *A. besseyi* chromosomes, the chromosome arm also contains significantly higher percentage of repeats sequence and lower reads coverage ($P < 0.001$, Wilcox test; **Supplementary Figure S5B, 5C**). While we observed reduced coverage in repeat regions, we were able to capture the entire genome with only 0.06% (29,230 bp) of the genome not sequenced, with missing regions ranging from 1 to 2,540 bp and affecting only six repeats and four genes. Taken together, these observations suggest that the MDA approach is a robust approach for capturing the entire nematode genome, although additional sequencing may be required to rescue repetitive regions in some species.

Longer T7 endonuclease digestion time increase ONT sequencing performance

To enhance the yield and minimize the bias of sequencing amplified samples, we attempted to optimise different stages of the workflow (**Supplementary Figure S1**). One of the challenges was that the branching templates generated by MDA were unsuitable for Oxford Nanopore sequencing as they can block sequencing pores and reduce sequencing yield (74). T7 endonuclease I was used in the existing protocol to generate linear templates by cutting the junction of the branching template. We found that the digestion time of T7 endonuclease I affected the dropout rate of sequencing pores and sequencing output (**Supplementary Figure 6A**) (75). A longer digestion time (30 versus 15 mins) improved the sequencing performance by reducing the dropping rate of sequencing pores and increasing the sequencing yield (**Supplementary Figure 6B, 6C**). In addition, we tried three approaches to address the influence of repeats that reduce amplification efficiency (**Supplementary Figure S1**), but no improvement was observed (**Supplementary Figure S2A**), indicating that uneven coverage is more related to the polymerase efficiency in amplifying repeats.

Complete genome assemblies from amplified sequences

To evaluate the feasibility of generating assemblies from amplified sequences, we generated genome assemblies based on different data types and sources (**Supplementary Table S8**). On average 112.5X Illumina and 32.6X-88.6X ONT reads were used and the initial assemblies produced under the default options yielded were 72.5-115.9Mb and the recently updated size of 100.2Mb (76), suggesting that the biased genome coverage in the amplified reads remains a challenge in the assembly process. The final assemblies were produced using the meta option of the Flye

assembler, with haplotigs removed, screened for contamination and polished using Illumina reads (**Table 1;** Methods), resulting in more similar genome sizes (100.9 to 98.3 Mb, **Table 1**). Compared to the assembly from unamplified long sequences (72), the N50 of the genome-amplified assembly is 20% shorter than the unamplified data, presumably due to the shorter ONT sequence length achieved by MDA (**Table 1**).

We assessed the completeness of the *de novo* assemblies of single and ten pooled nematode(s) by first aligning the contigs back to the reference. The unamplified data covered 99.7% of the reference genome, compared to 98% for the single-worm amplified assembly (**Table 1**). 2.5 - 4.4 Mb of the reference genome was not covered by the amplified genomes. We benchmarked the completeness of the assemblies using universal single-copy orthologs (BUSCO,(38), which were similar regardless of amplification (**Table 1**). The unassembled regions coincided with not sequenced or highly repetitive regions which were mostly located on the chromosome arms (**Figure 1C**). Taken together, the results show that the capability of sequencing the genome of the nematode with only a single worm using the WGA method is equivalent to using multiple worms. Interestingly, we observed a decrease in reference coverage (98.0 vs. 95.8%) and BUSCO completeness (96.8 vs. 95.1%) as the number of worms increased from single to ten worms prior to MDA. In addition, more contaminated sequences were present (17.1 Mb vs. 0.6 Mb in single worm).

Table 1. Statistics of *C. elegans* genome assemblies

	Reference	Amplified	Unamplified
Reads N50(Kb)	-	7.7	21.1
Depth	-	73.5	88.6
Size (Mb)	100.2	100.9	103.8
Seq number	7	484	79
Longest (Mb)	20.9	2.7	15.1
Minimum(bp)	13,794	1,041	815
N50(Kb)	17,494	614	6,740
L50	3	45	5
N90(Kb)	13,784	126	2,691
L90	6	197	15
BUSCO completeness (%)	98.7	96.8	92.8
Assembly covered (%)	-	98.0	99.7

High quality annotations from single nematode genome and transcriptome

To quantify the difference between annotation based on a single worm genome and transcriptome, we profiled the transcriptome from a single *C.elegans* adult based on the Smartseq2 protocol and generated ~10Gb of Illumina reads (43); **Supplementary Table S2**) We generated annotations either by *de novo* assembly of these reads or mapped these reads to genome assembly and used these reads as evidences in the MAKER pipeline (61). To evaluate the accuracy of the annotations from different approaches and datasets, all gene predictions were aligned back to the *C. elegans* reference and compared to the most recent annotation of 20,184 protein encoding genes from WormBase (ver WBP14, (51) ; **Figure 2 and Supplementary Table S9**). Comparing two versions of the *C.elegans* reference (WBP14 vs.WS100), more loci (0.6% vs 9.1%) were annotated in the current reference, demonstrating the improventment in gene annotation over time. Gene structure predicted using single worm RNA-seq mapped to the reference genome by Stringtie (54) had lower sensitivity and precision compared to the reference proteome, and the MAKER pipeline produced a more accurate prediction when using these mappings as hints. More importantly, the same pipeline produced predictions with only a slightly reduced accuracy of 1.2% on average when the genome-amplified assembly was used instead (**Figure 2A**). As expected, these annotations were more sensitive and accurate than ones that was originally published (WS100) and *de novo* transcriptomes for all metrics (**Supplementary Table. S9**). The *de novo* transcriptome had more missing exons (36.6% vs. 7.0%-23.9%), missing loci (35.7% vs 2.5-23.9 %) and fewer matching transcripts (7,227 vs 10,163 - 13,294)(**Figure 2B, 2C**). Of particular note, the proportions of 50% and 95% assembled genes, which are imparative for phylogenomic analyses, were 76.8% and 48.3% lower in the *de novo* transcriptome compared to the annotation produced with the available genome-amplified assembly (**Supplementary Table. S9**). These results indicate that the genome-amplified assemblies can be annotated with reasonable accuracy using the existing pipeline.

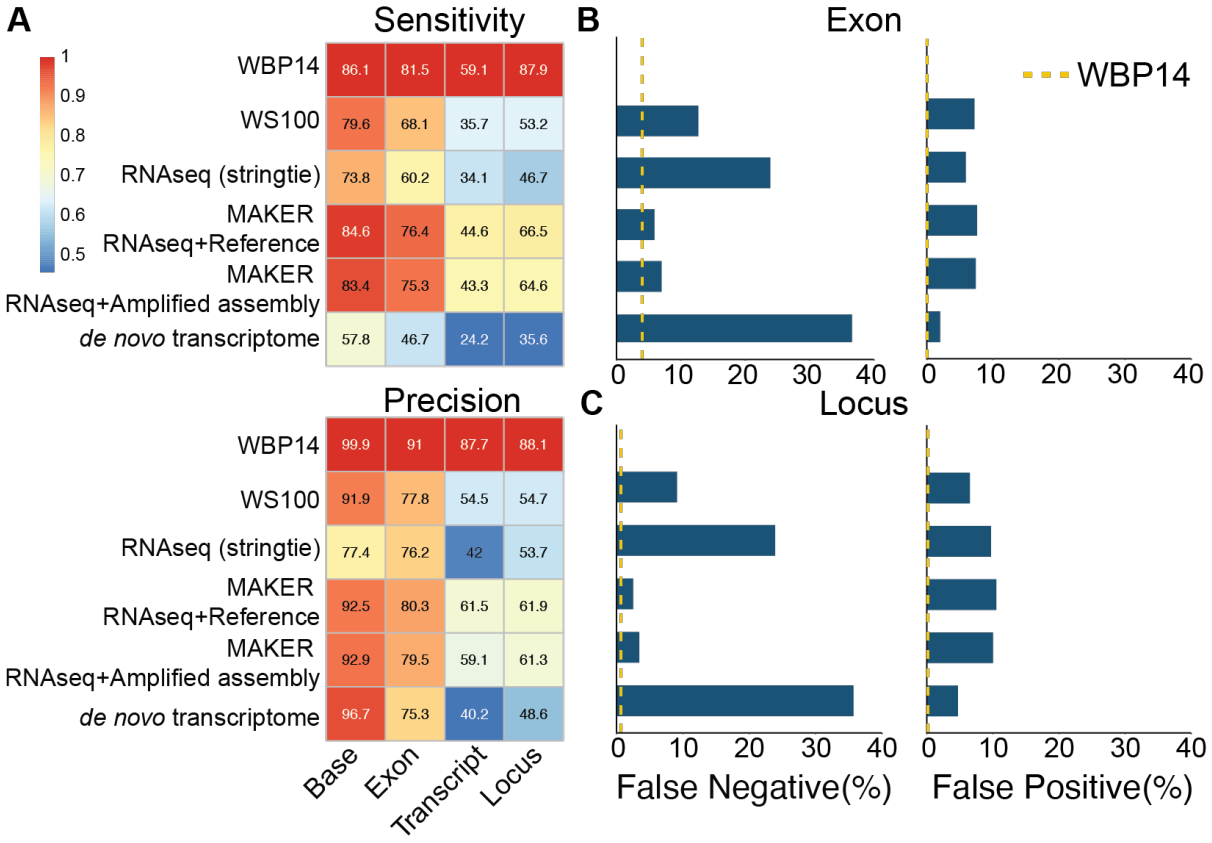

Figure 2. Comparison of annotations using different approaches and datasets. **A.** Sensitivity and precision on base, exon, transcript and locus level. WBP14 indicates the baseline performance when the reference proteome was aligned back to the reference genome using Minimap2. Colour of the heatmap is the value of each sample divided by the values in the WBP14 comparison. **B,C.** Percentage of false negatives and false positives in exon and locus. False negatives is the precentage of reference genes missing from the predictions, whereas false positives are new genes in the predictions that are not present in the reference proteome. Yellow dash line represent the value of the WBP14 comparison as baseline.

Genome characteristics of free-living nemtodes

We applied our optimised sequencing and annotation protocol to 13 free-living nematodes from three clades collected from the north coast of Taiwan (**Table 2**). On average, 12 Gb of ONT genomic reads and 5 Gb of transcriptome reads were sequenced from two adults per species (**Supplementary Table S2**). Strikingly, the assemblies ranged from 147.9 to 792.4 Mb indicating that the genome size of free-living nematodes is larger than that of most published parasite genomes. In particular, nematodes belonging to Enoplia tend to be larger than other clades (**Figure 3A**), with

the 792.4Mb assembly of *T. latispiculum* being the largest currently recorded in nematodes. Some of the assemblies may underestimate the true genome size, as the sequence coverage was 11-80X, warranting additional sequencing. Using the MAKER pipeline, 22,422 - 59,888 protein-coding genes were annotated in 13 free-living nematode genomes and were 34.5 to 92.6% complete based on BUSCO analysis (**Supplementary Table S10**). The intron distribution of the Dorylaimia and Chromadoria lineages sequenced in this study had a similar pattern, peaking around 50 bp (**Table 2, Supplementary Figure S7**) and were similar to previously published nematodes in these clades (77). Interestingly, there are fewer but longer introns in the four Enoplia species, suggesting a different intron distribution in the last common ancestor of this clade compared to the rest of nematodes (**Figure 3B, Supplementary Figure S8**). Orthology inference using Orthofinder (20, 21) placed these gene models and those of 13 other nematode genomes into 54,890 orthologous groups. Within these orthologous groups, 45.9 % (25,218 orthologous groups) were shared between two or more species, consistent with previous observations of extensive clade-specific families in nematode lineages (78). In addition, 16-69% of the genes in 13 free-living nematode species in were species-specific.

Of of the 13 nematode species sequenced in this study, nine were able to assemble a circular mitochondrial genome (mitogenome) with a read coverage depth of 39-2,656X. Interestingly, only five species be completely predicted all 12 proteins typically found in nematode mitogenomes. These mitogenomes are highly rearranged compared to other clades (Rhabditina, Spirurina, and Tylenchina) (**Supplementary Figure S9**). The gene order in *Mesodorylaimus* sp. showed lack of synteny compared to other species in Dorylaimia, consistent with previous observations of a high rearrangement rate in the Dorylaimia species (79).

The free-living nematodes contained significantly more repeats than nine representative parasitic nematodes across three clades ($P<0.001$, Wilcoxon ranked sum test; **Figure 3C** and **Supplementary Figure S10A**). This was particularly evident in Enoplia and Chromadoria, which contained a significantly higher proportion of repeats (24.1% to 69.5%) in their genomes compared to parasites (0.8% to 31.4%). In contrast to most published genomes in the Dorylaimia and Rhabditina clades, which were enriched in DNA transposons (78), (**Supplementary Figure S10B**), LTR or LINE

497 repeats were more abundant in six free-living nematode species, especially in the two
498 *Trissonchulus* species (7.4-6.8% vs. 1.2-5.4% other species), both LTR and LINEs are
499 enriched (**Supplementary Figure S10C-E**). A total of 16,631 unknown repeat families
500 were identified in the 13 nematodes, which were clustered into 16,557 sequence
501 groups based on 90% identity suggesting that they were species specific. These
502 unclassfied repeats were evenly distributed across the genome with the exception of
503 three species in Enoplia (*Trissonchulus* sp. *T. latispiculum and T. ribeirensis*), which
504 each had a dominant family comprising 2.2-5.6% and 0.7-1.5% of the unclassified
505 repeats and genomes, respectively.
506

507 **Table 2. Genome statistics of 13 free-living nematode genomes assembly.**
508

Species	clade	Size (Mb)	Seq (num.)	Longest (Kb)	N50 (Kb)	Gene (num.)	Intron length median(bp)
Mesodorylaimus sp.	I	148	1,727	2,531.2	409.2	16,194	64
Trissonchulus sp.	II	447	21,105	252.3	36.3	14,926	538
T. ribeirensis	II	746	30,783	291.3	42	12,305	1,796
E. lenunculus	II	300.9	4,131	891	164.7	10,199	564
T. latispiculum	II	792.4	32,069	326.8	40.5	20,486	437
Theristus sp.	C	170.7	5,900	427.2	58.1	15,115	248
S. punctata	C	390.8	8,097	959.4	90.9	23,376	98
Ptycholaimellus sp.	C	238.3	12,730	319.1	33.3	15,899	58
Linhomoeus sp.	C	240.5	14,458	316.5	28.9	20,474	88
Paralinhomoeus sp.	C	177.3	3,869	553.7	80	18,659	76
Microlaimidae sp.	C	575.8	35,431	240.1	24.1	14,273	981
Epsilonema sp.	C	171.9	7,781	250.3	36.9	17,522	245
Rhynchonema sp.	C	222	9,378	328.1	43.5	11,349	211

509

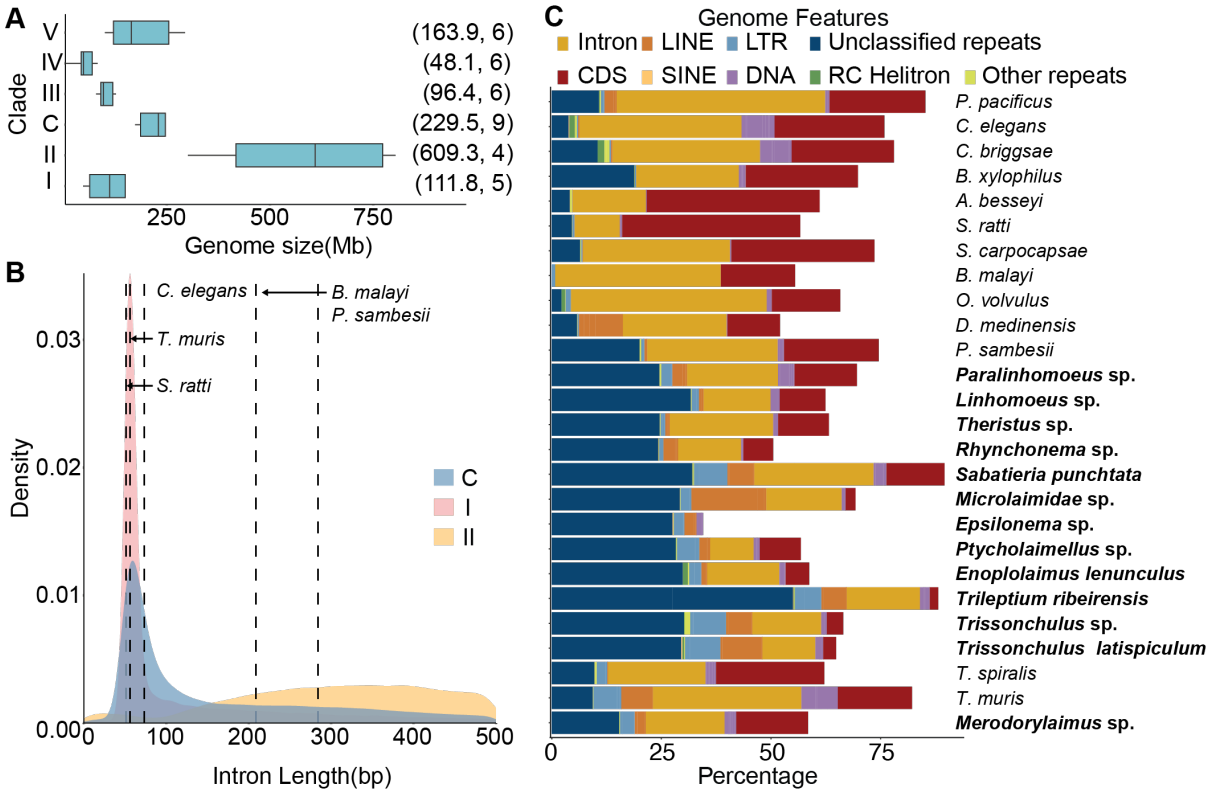

Figure 3. Nematode genome size, intron distribution and genome structure. A. Genome size variation between different nematode clades. The brackets denote the median genome size and number of nematodes used for analysis, respectively. **B.** Intron distribution of nematodes. Dashed lines are the intron median of the representative species in clades I, III, IV, V and Chromadoria(C). **C.** The proportion of different genome features in the nematode genomes. Bold letters represent the nematode genome assembled in this study. The 'Other repeats' feature include the sum of tRNA, snRNA, rRNA, simple repeats and satellite.

Enoplia is sister to the rest of the nematode classes.

To resolve uncertainties in the basal branch order of the Nematoda phylogeny, in particular the relative placement of Enoplia and Dorylaimia, a species tree was inferred based on a coalescent-based analysis (67,80) of 58,531 paralogous gene trees from 26 representative nematode species and the cactus worm *Priapulus cauatus* as an outgroup. The species phylogeny separated nematodes into six groups, including five clades and groups of the early derived Chromadoria lineage (20), and placed Enoplia as a sister group to Dorylaimia and Chromadoria lineage, both with strong boostrap support (**Supplementary Figure S11**). The topology remained similar when we included an additional nine *de novo* transcriptomes from Enoplia and Chromadoria

and a further three outgroups (**Figure 4**) (19). The combined phylogeny shows that nematodes can be separated into the previously designated five clades (20) and support for the placement of Enoplia remained robust (Astro-pro, bs=100). In the early derived Chromadoria lineages, most of the lineages were grouped by order, with the exception of the Monhysterida lineage which was paraphyletic and separated by Areaolmida.

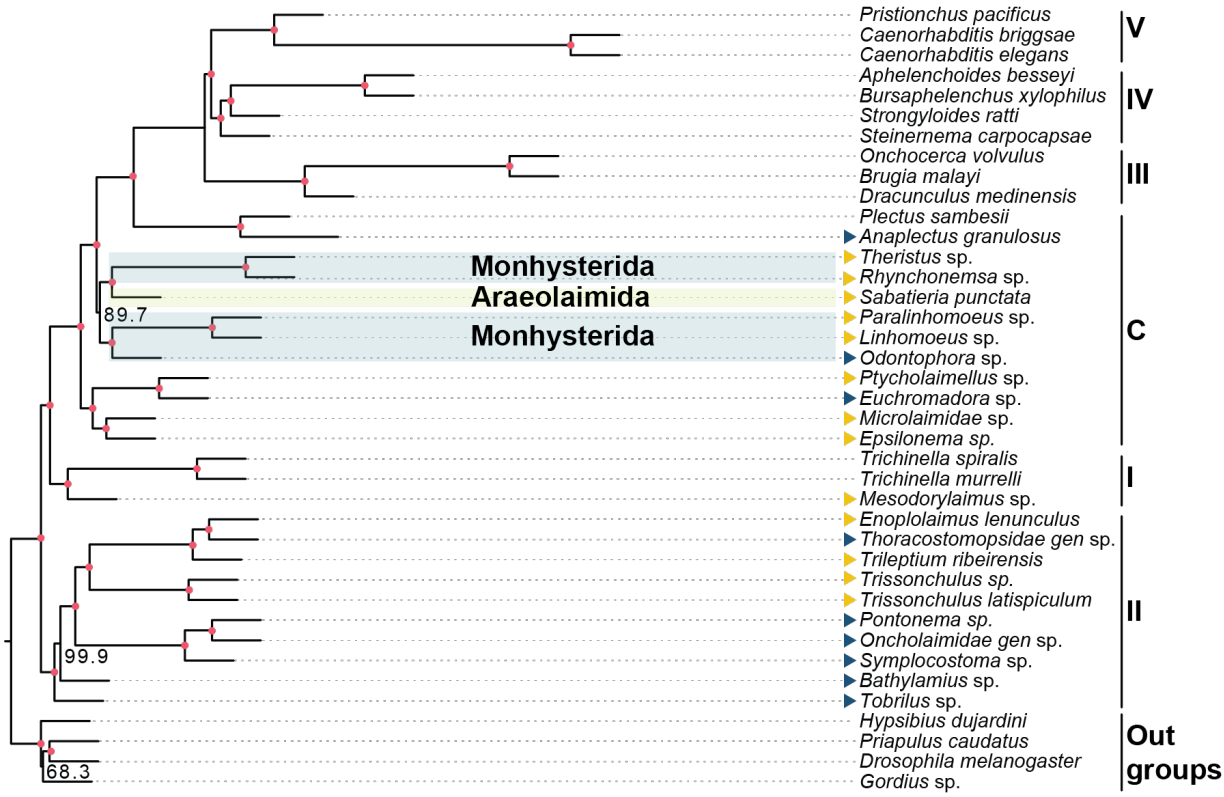

Figure 4. Phylogenetic analysis combining nematode transcriptome and genome. The bold font represents the Nematoda order in the shade area. The Roman numerals on right represent the five clades of Nematoda and C represent the basal Chromadoria lineage. The colours of triangles denote data source. The red dots in branches denote a bootstrap support value of 100.

Discussion

In this study, we demonstrated the feasibility of generating genome assemblies from single adult nematodes using multiple displacement amplification. By testing the protocols on *C. elegans,* we were able to fully quantify the extent of bias and address it with existing analysis pipelines. With a genome size of 148-792.4 Mb in 13 nematodes, sequencing on a single MinION flowcell can be expected to provide approximately 37.8X depth of coverage. We demonstrate that a genome assembly and accurate gene annotations can be achieved with this workflow and further sequenced the genomes of 13 free-living nematodes. Of these genomes, four are the first reported in the Enoplia clade revealing their unusually large genome sizes and structures (**Figure 3**). Through phylogenomics, we established Enoplia as sister to the phylum Nematoda, suggesting a marine origin in the last common ancestor of nematodes (19). We overcame the stage of obtaining axenic cultures (8) whilst assembly and annotation can be achieved within two weeks of nematode isolation. Assuming that 1ug is required for long-read sequencing, combining MDA with ONT sequencing thus provides a cost- and labour-effective solution (1) to generate complete assemblies in organisms with as little as 50 picograms of starting material.

The advantage of using a single individual for whole genome sequencing is also seen in the sequencing of organisms such as obligate symbionts and helminth eggs, where it is possible to overcome obstacles such as inability to culture and inaccessibility in the live host (9,81). The use of a single nematode had several benefits over pooling multiple worms, for instance closely related nematode species have imperceptible morphological differences that increased the risk of mixing different species (82,83). For example, a host can be infected with multiple *Anisakis* species with no morphological differences (84). In addition, natural populations are likely to have high levels of heterozygosity, which also affects the quality of assembly and annotation, as observed in this study.

The MDA method used in this study is known to result in uneven read coverage (85). This unevenness is thought to be caused by the formation of secondary structures that reduce the efficiency of the phi29 polymerase used in the amplification process, particularly in repetitive sequences that are prone to forming such structures (81, 82). Despite these challenges, only a small portion of the *C. elegans* genome remained

unsequenced, including ten genes representing only 0.4% of the genome. This approach allowed us to effectively assemble the genome, with only 2.5% missing due to the combined challenges posed by repetitive sequences and reduced coverage. One existing limitation was that the N50 of the amplified data was capped to ~8kb, resulting in a more fragmented assembly compared to assembly from unamplified sample. Nevertheless, the BUSCO completeness values suggest that the amplified assembly is complete and capable of generating high-quality annotation compared to reference genome (96.8% vs 98%) (86). In addition, we show that gene prediction is superior when a complete genome is available compared to *de novo* transcriptome assemblies especially in the number of 95% assembled loci (82.9 vs 40.0%), and is ideal for subsequent phylogenomics and comparative genomics analyses.

To conclude, we demonstrate the feasibility of incorporatng whole genome amplification into investigation the study of microbial biodiversity from sampling to comparative genomic analysis. By thoroughly characterising and accounting for the inherent nature of this approach, complete assemblies and accurate gene predictions can be generated. The availability of the new free-living genomes has allowed us to address outstanding questions and offer new biological insights. As long-read sequencing advances in accuracy and affordability, we envision that a complete assembly will be available for any species that were once considered inaccessible.

Data Availability

All sequences generated from this study were deposited on NCBI under BioProject PRJNA953805 and accession numbers of the free-living nematodes can be found in **Supplementary Table S3**.

Funding

This work was supported by Academia Sinica grant (AS-CDA-107-L01) and National Science and Technology Council (111-2628-B-001-021) to IJT. YiCL is supported by the doctorate fellowship of the Taiwan International Graduate Program, Academia Sinica of Taiwan.

Authors' Contributions

IJT conceived and the study; YiCL, YuCL, MCW and YCT carried out the sampling; YiCL isolated the nematodes; YiCL and HMK conducted the experiments and ONT sequencing; YiCL performed the assemblies and annotations with help from HHL and YuCL; YiCL carried out analyses. YiCL and IJT wrote the manuscript with input from TK and others. All authors read and approved the final manuscript.

Acknowledgements

We thank Wei-An Liu for testing out the initial MDA protocols in yeast. We thank NGS Genomics core lab of Academia Sinica of Taiwan for sequencing the initial single worm RNAseq data. We would like to thank the National Center for High-performance Computing (NCHC) of the National Applied Research Laboratories (NARLabs) of Taiwan for providing computational resources and storage resources.

References

1. Faulk, C. (2023) De novo sequencing, diploid assembly, and annotation of the black carpenter ant, Camponotus pennsylvanicus, and its symbionts by one person for $1000, using nanopore sequencing. *Nucleic Acids Res*, **51**, 17-28.
2. He, K., Minias, P. and Dunn, P.O. (2021) Long-Read Genome Assemblies Reveal Extraordinary Variation in the Number and Structure of MHC Loci in Birds. *Genome Biol Evol*, **13**.
3. Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J. *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, **592**, 737-746.
4. Hotaling, S., Sproul, J.S., Heckenhauer, J., Powell, A., Larracuente, A.M., Pauls, S.U., Kelley, J.L. and Frandsen, P.B. (2021) Long Reads Are Revolutionizing 20 Years of Insect Genome Sequencing. *Genome Biol Evol*, **13**.
5. Runnel, K., Abarenkov, K., Copot, O., Mikryukov, V., Koljalg, U., Saar, I. and Tedersoo, L. (2022) DNA barcoding of fungal specimens using PacBio long-read high-throughput sequencing. *Mol Ecol Resour*, **22**, 2871-2879.
6. Lewin, H.A., Richards, S., Lieberman Aiden, E., Allende, M.L., Archibald, J.M., Balint, M., Barker, K.B., Baumgartner, B., Belov, K., Bertorelle, G. *et al.* (2022) The Earth BioGenome Project 2020: Starting the clock. *Proc Natl Acad Sci U S A*, **119**.
7. Ekblom, R. and Galindo, J. (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity (Edinb)*, **107**, 1-15.
8. Hongoh, Y. and Toyoda, A. (2011) Whole-genome sequencing of unculturable bacterium using whole-genome amplification. *Methods Mol Biol*, **733**, 25-33.
9. Montoliu-Nerin, M., Sanchez-Garcia, M., Bergin, C., Grabherr, M., Ellis, B., Kutschera, V.E., Kierczak, M., Johannesson, H. and Rosling, A. (2020) Building de novo reference

genome assemblies of complex eukaryotic microorganisms from single nuclei. *Sci Rep*, **10**, 1303.

10. Deleye, L., Tilleman, L., Vander Plaetsen, A.S., Cornelis, S., Deforce, D. and Van Nieuwerburgh, F. (2017) Performance of four modern whole genome amplification methods for copy number variant detection in single cells. *Sci Rep*, **7**, 3422.
11. Santoro, A.E., Kellom, M. and Laperriere, S.M. (2019) Contributions of single-cell genomics to our understanding of planktonic marine archaea. *Philos Trans R Soc Lond B Biol Sci*, **374**, 20190096.
12. Sahraei, S.E., Sanchez-Garcia, M., Montoliu-Nerin, M., Manyara, D., Bergin, C., Rosendahl, S. and Rosling, A. (2022) Whole genome analyses based on single, field collected spores of the arbuscular mycorrhizal fungus Funneliformis geosporum. *Mycorrhiza*, **32**, 361-371.
13. Lepere, C., Demura, M., Kawachi, M., Romac, S., Probert, I. and Vaulot, D. (2011) Whole-genome amplification (WGA) of marine photosynthetic eukaryote populations. *FEMS Microbiol Ecol*, **76**, 513-523.
14. Nyaku, S.T., Sripathi, V.R., Lawrence, K. and Sharma, G. (2021) Characterizing Repeats in Two Whole-Genome Amplification Methods in the Reniform Nematode Genome. *Int J Genomics*, **2021**, 5532885.
15. Eccles, D., Chandler, J., Camberis, M., Henrissat, B., Koren, S., Le Gros, G. and Ewbank, J.J. (2018) De novo assembly of the complex genome of Nippostrongylus brasiliensis using MinION long reads. *BMC Biol*, **16**, 6.
16. Dillman, A.R., Mortazavi, A. and Sternberg, P.W. (2012) Incorporating genomics into the toolkit of nematology. *J Nematol*, **44**, 191-205.
17. Christoph Dieterich, R.J.S. (2009) How to become a parasite - lessons from the genomes of nematodes. *Trends in Genetics*, **25**, 203-209.
18. Blaxter, M. (2011) Nematodes: the worm and its relatives. *PLoS Biol*, **9**, e1001050.
19. Smythe, A.B., Holovachov, O. and Kocot, K.M. (2019) Improved phylogenomic sampling of free-living nematodes enhances resolution of higher-level nematode phylogeny. *BMC Evol Biol*, **19**, 121.
20. Blaxter, M.L., De Ley, P., Garey, J.R., Liu, L.X., Scheldeman, P., Vierstraete, A., Vanfleteren, J.R., Mackey, L.Y., Dorris, M., Frisse, L.M. *et al.* (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature*, **392**, 71-75.
21. De Ley, P. (2006) A quick tour of nematode diversity and the backbone of nematode phylogeny. *WormBook*, 1-8.
22. Kern, E.M.A., Kim, T. and Park, J.K. (2020) The Mitochondrial Genome in Nematode Phylogenetics. *Front Ecol Evol*, **8**.
23. Consortium, C.e.S. (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, **282**, 2012-2018.
24. Kikuchi, T., Eves-van den Akker, S. and Jones, J.T. (2017) Genome Evolution of Plant-Parasitic Nematodes. *Annu Rev Phytopathol*, **55**, 333-354.
25. Mahfouz M.M. Abd-Elgawad, T.H.A. (2015) *Impact of phytonematodes on agriculture economy.* CAB International.
26. Semprucci, F. (2013) Marine nematodes from the shallow subtidal coast of the Adriatic Sea: species list and distribution. *International Journal of Biodiversity*.
27. Bik, H.M., Lambshead, P.J., Thomas, W.K. and Lunt, D.H. (2010) Moving towards a complete molecular framework of the Nematoda: a focus on the Enoplida and early-branching clades. *BMC Evol Biol*, **10**, 353.

28. Beltran, T., Barroso, C., Birkle, T.Y., Stevens, L., Schwartz, H.T., Sternberg, P.W., Fradin, H., Gunsalus, K., Piano, F., Sharma, G. *et al.* (2019) Comparative Epigenomics Reveals that RNA Polymerase II Pausing and Chromatin Domain Organization Control Nematode piRNA Biogenesis. *Dev Cell*, **48**, 793-810 e796.
29. Gingold, R., Moens, T. and Rocha-Olivares, A. (2013) Assessing the Response of Nematode Communities to Climate Change-Driven Warming: A Microcosm Experiment. *PLoS One*, **8**, e66653.
30. Moens, T. and Vincx, M. (2000) Temperature, salinity and food thresholds in two brackish-water bacterivorous nematode species: assessing niches from food absorption and respiration experiments. *Journal of Experimental Marine Biology and Ecology*, **243**, 137-154.
31. Viney, M. (2017) How Can We Understand the Genomic Basis of Nematode Parasitism? *Trends Parasitol*, **33**, 444-452.
32. Meldal, B.H., Debenham, N.J., De Ley, P., De Ley, I.T., Vanfleteren, J.R., Vierstraete, A.R., Bert, W., Borgonie, G., Moens, T., Tyler, P.A. *et al.* (2007) An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa. *Mol Phylogenet Evol*, **42**, 622-636.
33. Ahmed, M., Roberts, N.G., Adediran, F., Smythe, A.B., Kocot, K.M. and Holovachov, O. (2022) Phylogenomic Analysis of the Phylum Nematoda: Conflicts and Congruences With Morphology, 18S rRNA, and Mitogenomes. *Front Ecol Evol*, **9**.
34. Kolmogorov, M., Yuan, J., Lin, Y. and Pevzner, P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*, **37**, 540-546.
35. Vaser, R., Sović, I., Nagarajan, N. and Šikić, M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*, **27**, 737-746.
36. Hu, J., Fan, J., Sun, Z. and Liu, S. (2020) NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics*, **36**, 2253-2255.
37. Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y. and Durbin, R. (2020) Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, **36**, 2896-2898.
38. Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210-3212.
39. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*, **19**, 455-477.
40. Buchfink, B., Reuter, K. and Drost, H.-G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, **18**, 366-368.
41. De Vivo, M., Lee, H.-H., Huang, Y.-S., Dreyer, N., Fong, C.-L., de Mattos, F.M.G., Jain, D., Wen, Y.-H.V., Mwihaki, J.K., Wang, T.-Y. *et al.* (2022) Utilisation of Oxford Nanopore sequencing to generate six complete gastropod mitochondrial genomes as part of a biodiversity curriculum. *Scientific Reports*, **12**, 9973.
42. Bernt, M., Donath, A., Juhling, F., Externbrink, F., Florentz, C., Fritzsch, G., Putz, J., Middendorf, M. and Stadler, P.F. (2013) MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol Phylogenet Evol*, **69**, 313-319.

43. Serra, L., Chang, D.Z., Macchietto, M., Williams, K., Murad, R., Lu, D., Dillman, A.R. and Mortazavi, A. (2018) Adapting the Smart-seq2 Protocol for Robust Single Worm RNA-seq. *Bio Protoc*, **8**.
44. Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114-2120.
45. Haas, B.P.A. (2020).
46. Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C. and Smit, A.F. (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*, **117**, 9451-9457.
47. Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460-2461.
48. Matthew Berriman, A.C., Isheng Jason Tsai (2018) Creation of a comprehensive repeat library for a newly sequenced parasitic worm genome. *PROTOCOL (Version 1) available at Protocol Exchange*
49. Smit, A., Hubley, R & Green, P. (2013-2015) RepeatMasker Open-4.0.
50. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
51. Howe, K.L., Bolt, B.J., Shafie, M., Kersey, P. and Berriman, M. (2017) WormBase ParaSite - a comprehensive resource for helminth genomics. *Mol Biochem Parasitol*, **215**, 2-10.
52. Dobin, A. and Gingeras, T.R. (2015) Mapping RNA-seq Reads with STAR. *Curr Protoc Bioinformatics*, **51**, 11 14 11-11 14 19.
53. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, **29**, 644-652.
54. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. and Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*, **33**, 290-295.
55. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, **7**, 562-578.
56. Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859-1875.
57. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O. and Borodovsky, M. (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res*, **18**, 1979-1990.
58. Tempel, S. (2012) Using and understanding RepeatMasker. *Methods Mol Biol*, **859**, 29-51.
59. Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.
60. Venturini, L., Caim, S., Kaithakottil, G.G., Mapleson, D.L. and Swarbreck, D. (2018) Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience*, **7**.
61. Holt, C. and Yandell, M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, **12**, 491.
62. Pertea, G. and Pertea, M. (2020) GFF Utilities: GffRead and GffCompare. *F1000Res*, **9**.

63. Emms, D.M. and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*, **16**, 157.
64. Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*, **20**, 238.
65. Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, **30**, 772-780.
66. Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
67. Zhang, C., Scornavacca, C., Molloy, E.K. and Mirarab, S. (2020) ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. *Mol Biol Evol*, **37**, 3292-3307.
68. Carlton, P.M., Davis, R.E. and Ahmed, S. (2022) Nematode chromosomes. *Genetics*, **221**.
69. Blainey, P.C. and Quake, S.R. (2011) Digital MDA for enumeration of total nucleic acid contamination. *Nucleic Acids Res*, **39**, e19.
70. Sidore, A.M., Lan, F., Lim, S.W. and Abate, A.R. (2016) Enhanced sequencing coverage with digital droplet multiple displacement amplification. *Nucleic Acids Res*, **44**, e66.
71. Kim, S.C., Premasekharan, G., Clark, I.C., Gemeda, H.B., Paris, P.L. and Abate, A.R. (2017) Measurement of copy number variation in single cancer cells using rapid-emulsification digital droplet MDA. *Microsyst Nanoeng*, **3**.
72. Tyson, J.R., O'Neil, N.J., Jain, M., Olsen, H.E., Hieter, P. and Snutch, T.P. (2018) MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. *Genome Res*, **28**, 266-274.
73. Lai, C.K., Lee, Y.C., Ke, H.M., Lu, M.R., Liu, W.A., Lee, H.H., Liu, Y.C., Yoshiga, T., Kikuchi, T., Chen, P.J. *et al.* (2023) The Aphelenchoides genomes reveal substantial horizontal gene transfers in the last common ancestor of free-living and major plant-parasitic nematodes. *Mol Ecol Resour*.
74. Calus, S.T., Ijaz, U.Z. and Pinto, A.J. (2018) NanoAmpli-Seq: a workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform. *Gigascience*, **7**.
75. Muller, B., Jones, C. and West, S.C. (1990) T7 endonuclease I resolves Holliday junctions formed in vitro by RecA protein. *Nucleic Acids Res*, **18**, 5633-5636.
76. Yoshimura, J., Ichikawa, K., Shoura, M.J., Artiles, K.L., Gabdank, I., Wahba, L., Smith, C.L., Edgley, M.L., Rougvie, A.E., Fire, A.Z. *et al.* (2019) Recompleting the Caenorhabditis elegans genome. *Genome Res*, **29**, 1009-1022.
77. Ma, M.Y., Xia, J., Shu, K.X. and Niu, D.K. (2022) Intron losses and gains in the nematodes. *Biol Direct*, **17**, 13.
78. International Helminth Genomes, C. (2019) Comparative genomics of the major parasitic worms. *Nat Genet*, **51**, 163-174.
79. Hyman, B.C., Lewis, S.C., Tang, S. and Wu, Z. (2011) Rampant gene rearrangement and haplotype hypervariation among nematode mitochondrial genomes. *Genetica*, **139**, 611-615.
80. Smith, M.L., Vanderpool, D. and Hahn, M.W. (2022) Using all Gene Families Vastly Expands Data Available for Phylogenomic Inference. *Molecular Biology and Evolution*, **39**.
81. Doyle, S.R., Sankaranarayanan, G., Allan, F., Berger, D., Jimenez Castro, P.D., Collins, J.B., Crellen, T., Duque-Correa, M.A., Ellis, P., Jaleta, T.G. *et al.* (2019) Evaluation of

DNA Extraction Methods on Individual Helminth Egg and Larval Stages for Whole-Genome Sequencing. *Front Genet*, **10**, 826.

82. Bogale, M., Baniya, A. and DiGennaro, P. (2020) Nematode Identification Techniques and Recent Advances. *Plants (Basel)*, **9**.

83. Subbotin, S.A., Oliveira, C.J., Alvarez-Ortega, S., Desaeger, J.A., Crow, W., Overstreet, C., Leahy, R., Vau, S. and Inserra, R.N. (2021) The taxonomic status of Aphelenchoides besseyi Christie, 1942 (Nematoda: Aphelenchoididae) populations from the southeastern USA, and description of Aphelenchoides pseudobesseyi sp. n. *Nematology*, **23**, 381-413.

84. Van Hien, H., Thi Dung, B., Ngo, H.D. and Doanh, P.N. (2021) First morphological and molecular identification of third-stage larvae of Anisakis typica (Nematoda: Anisakidae) from marine fishes in Vietnamese water. *J Nematol*, **53**.

85. Tsai, I.J., Hunt, M., Holroyd, N., Huckvale, T., Berriman, M. and Kikuchi, T. (2014) Summarizing specific profiles in Illumina sequencing from whole-genome amplified DNA. *DNA Res*, **21**, 243-254.

86. Denton, J.F., Lugo-Martinez, J., Tucker, A.E., Schrider, D.R., Warren, W.C. and Hahn, M.W. (2014) Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput Biol*, **10**, e1003998.