frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 07 February 2024
DOI 10.3389/fgene.2024.1308527

OPEN ACCESS

EDITED BY
Richard D. Emes,
Nottingham Trent University, United Kingdom

REVIEWED BY
Thomas Hackl,
University of Groningen, Netherlands
Iraad F. Bronner,
Wellcome Sanger Institute (WT),
United Kingdom

*CORRESPONDENCE
Nadège Guiglielmoni,
nguiglie@uni-koeln.de

RECEIVED 06 October 2023
ACCEPTED 04 January 2024
PUBLISHED 07 February 2024

CITATION
Guiglielmoni N, Villegas LI, Kirangwa J and Schiffer PH (2024), Revisiting genomes of non-model species with long reads yields new insights into their biology and evolution. *Front. Genet.* 15:1308527. doi: 10.3389/fgene.2024.1308527

Revisiting genomes of non-model species with long reads yields new insights into their biology and evolution

Nadège Guiglielmoni*, Laura I. Villegas, Joseph Kirangwa and Philipp H. Schiffer

Institut für Zoologie, Universität zu Köln, Cologne, Germany

High-quality genomes obtained using long-read data allow not only for a better understanding of heterozygosity levels, repeat content, and more accurate gene annotation and prediction when compared to those obtained with short-read technologies, but also allow to understand haplotype divergence. Advances in long-read sequencing technologies in the last years have made it possible to produce such high-quality assemblies for non-model organisms. This allows us to revisit genomes, which have been problematic to scaffold to chromosome-scale with previous generations of data and assembly software. Nematoda, one of the most diverse and speciose animal phyla within metazoans, remains poorly studied, and many previously assembled genomes are fragmented. Using long reads obtained with Nanopore R10.4.1 and PacBio HiFi, we generated highly contiguous assemblies of a diploid nematode of the Mermithidae family, for which no closely related genomes are available to date, as well as a collapsed assembly and a phased assembly for a triploid nematode from the Panagrolaimidae family. Both genomes had been analysed before, but the fragmented assemblies had scaffold sizes comparable to the length of long reads prior to assembly. Our new assemblies illustrate how long-read technologies allow for a much better representation of species genomes. We are now able to conduct more accurate downstream assays based on more complete gene and transposable element predictions.

KEYWORDS

Nematoda, genomics, long reads, genome assembly, genome annotation

1 Introduction

Over the past decade, the field of genome assembly has experienced major improvements fueled by the development of high throughput sequencing techniques and major increases in the length and accuracy of reads. Short-read sequencing prompted the release of many draft assemblies for a large variety of species. The limited length of these reads could only yield highly fragmented assemblies, which were sufficient for initial analyses of gene content, but could not account for the structure of genomes and often fell short on resolving repetitive regions (Rice and Green, 2019). Recent advances in genome assembly have been driven by the availability of long reads offered by Pacific Biosciences (PacBio) and Oxford Nanopore. While these reads initially had a high error rate, improvements of these technologies have drastically increased their accuracy to over 99%, with the release of PacBio HiFi reads (based on circular consensus sequencing) (Wenger et al. 2019) and Nanopore Q20+ (Sereika et al. 2021) reads

obtained using R10.4.1 flow cells. These developments have brought draft assemblies to Megabase-level N50s (Guiglielmoni et al. 2022), illustrating their high contiguity, and have opened new possibilities for genome analyses. Assemblies obtained with long-read data not only have a higher gene completeness, but they can also provide a more comprehensive overview of repetitive regions and potentially allow for a better understanding of their structure, activity and dynamics (Shahid and Slotkin 2020). In addition, high-accuracy long reads can be used to discriminate alleles and generate phased assemblies, including all haplotypes (Cheng et al. 2021; Rautiainen et al. 2023), while low-accuracy long reads were only sufficient for collapsed assemblies (in which homologous chromosomes are represented by a single sequence) as errors could not be distinguished from alternative haplotypes.

Some hundred genome assemblies have been released thus far for the phylum Nematoda, yet only a few are high-quality assemblies and they offer a poor representation of the diversity of the taxon for which over 30,000 species have been described (Hodda 2022). In particular, efforts have focused on *Caenorhabditis* and parasitic species, leaving incomplete resources for understudied clades (Kumar et al. 2012b; Kumar et al. 2012a). In this paper, we focus on the genomes of two species at two extremities of the nematode phylogeny: the basal *Romanomermis culicivorax* (clade I) and the derived *Panagrolaimus* sp. PS1159 (clade IV).

The Enoplean nematode *Romanomermis culicivorax* is a member of the mermithidae family which includes over 100 described species (Presswell et al. 2015). It is an obligate parasite of various species of mosquito larvae (Giblin and Platzer 1985). Along other mermithid nematodes, it is presently employed for the biological control of malaria (Petersen et al. 1978; Abagli et al. 2019). Enoplean research often revolves around *Trichinella spiralis*, given its significance as a mammalian parasite (Mitreva et al. 2011). Among mermithidae, only two genomes are currently publicly available (Bhattarai et al. 2022; Schiffer et al. 2013). In contrast with the published assembly of the sexual *R. culicivorax*, the long-read genome assembly of the parthenogenetic *Mermis nigrescens* is more contiguous and contains approximately twice the repeat content and heterozygosity. The need for additional high-quality genomes is evident, not only to address resource gaps in the Enoplean class, but also to enable investigations into sexual evolution, genome structural variations, and host-parasite interactions within the mermithidae family.

Panagrolaimus sp. PS1159 is a free-living nematode belonging to the Panagrolaimidae family. Members of this family have various reproductive modes including hermaphroditism, outcrossing between males and females and asexual reproduction through parthenogenesis (Lewis et al. 2009); *Panagrolaimus* sp. PS1159 is parthenogenetic. This strain has been isolated in North Carolina, United States by Paul Sterneberg, and is thought to be a triploid allopolyploid (3n = 12) (Schiffer et al. 2019). Previous studies have found it shares a common origin of parthenogenesis with most *Panagrolaimus* asexual strains, from a hybridization event estimated to have occurred 1.3–8.5 Million years ago (Schiffer et al. 2019; Shatilovich et al. 2023). To date, over 140 strains of the genus have been documented (NCBI Taxonomy Browser), however only nine, largely fragmented, draft genome assemblies are available on GenBank (accessed on 06.10.2023). This widely distributed group includes strains isolated from extreme environments such as Antarctica, the volcanic island of Surtsey and the Russian permafrost. Representatives of the genus from these locations have been found to be freezing-tolerant undergoing cryptobiosis (Shatilovich et al. 2023; McGill et al. 2015), and *Panagrolaimus* sp. PS1159 has also shown anhydrobiotic potential as a fast desiccation strategist (Shannon et al. 2005).

Short-read genome assemblies are available for both species, yet their high fragmentation impedes downstream analyses. Their scaffold N50s are limited to 17.6 kb for *R. culicivorax* and 9.9 kb for *P.* sp. PS1159. By contrast, these values would be the expected length for unassembled long reads nowadays. Although these draft assemblies provided a first insight into the genomics of these species, more contiguous assemblies can now be obtained using long reads. To reassemble these species, we chose to generate both PacBio HiFi and Nanopore sequencing data and to leverage distinct advantages of these technologies. For PacBio HiFi, we used an ultra-low input protocol with DNA extracted from only a few individuals and whole genome amplification. For Nanopore sequencing, we extracted DNA from large pools of individuals and selected the largest fragments. Using these heterogeneous long-read datasets, we produced new highly contiguous assemblies with increased completeness.

2 Materials and methods

2.1 Pacific Biosciences HiFi sequencing

Up to 10 individuals were collected and washed in water, then flash-frozen using liquid nitrogen in a salt-based extraction buffer (Tris-HCl 100 mM, ethylenediaminetetraacetic acid 50 mM, NaCl 0.5 M and sodium dodecylsulfate 1%). Samples were incubated overnight at 50°C after addition of 5 μL of proteinase K (Zymo Research D3001-2). DNA was precipitated using NaCl 5 M, yeast tRNA and isopropanol, and incubated at room temperature for 30 min, then pelleted at 18,000 g for 20 min (4°C). The DNA was washed twice with 80% ethanol and spinned at 18,000 g for 10 min (4°C). The DNA pellet was eluted in elution buffer (D3004-4-10 Zymo Research) and incubated at 50°C for 10 min. RNA was removed by incubating with RNAse (Qiagen, 19101) for 1 h at (37°C). DNA concentrations were quantified using a Qubit 4 fluoremeter with 1X dsDNA kit. HiFi libraries were prepared with the Express 2.0 Template kit (Pacific Biosciences, Menlo Park, CA, United States) and sequenced on a Sequel II/Sequel IIe instrument with 30 h movie time. HiFi reads were generated using SMRT Link (v10, Pacific Biosciences, Menlo Park, CA, United States) with default parameters. Sequencing results are presented in Supplementary Table S1.

2.2 Nanopore sequencing

Romanomermis culicivorax worms were picked from moss material supplied by Prof. Dr Edward Platzer at University of California Riverside. *Panagrolaimus* sp. PS1159 worms (isolate from North Carolina, United States) were harvested from agar plates with water and pelleted at 5,000 g for 5 min. The *P.*

sp. PS1159 pellet was re-suspended in a 1 M sucrose solution used for bacterial decontamination (sucrose flotation). The sample was centrifuged at 1,000 g for 3 min, the upper 1 mL of the supernatant containing the live clean worms was transferred to a new tube and diluted with nuclease-free water. The worms where pelleted again at 5,000 g for 5 min for further processing. Due to the large input, different DNA extraction protocols were tested as the salting-out protocol used for ultra-low input DNA extraction led to poor purity with many worms. Extractions with the Monarch DNA extraction kit also resulted in suboptimal OD260/230 values. Samples were incubated in cetyltrimethylammonium bromide (CTAB) buffer (polyvinylpyrrolidone 2%, Tris-HCl 100 mM, ethylenediaminetetraacetic acid 25 mM, NaCl 2 M, CTAB 2%) supplemented with 25 μL of proteinase K (Zymo Research D3001-2) for 1 h (*P.* sp. PS1159) or 2 h (*R. culicivorax*), until the individuals were dissolved. After further incubation for 10 min with 1.0 M potassium acetate, extracts were purified with phenol-chloroform-isoamyl alcohol 25:24:1, chloroform-isoamyl alcohol 24:1, centrifugation at 16,000 g for 10 min (room temperature) and AMPure XP beads (Agencourt). DNA was then incubated with RNAse cocktail enzyme mix (Thermo Fischer, AM2286) for 1 h at 37 °C. Prior trials of the same protocol without the potassium acetate step led to low OD260/230 values. DNA was fragmented in a 2 mL low-bind round bottom Eppendorf tube using a sterile 3 mm borosilicate bead (Z143928-1EA Merck) by vortexing for 1 min at maximum speed as described in Koetsier and Cantor (2021). Short fragments were removed using the Short Reads Eliminator (SRE) (Circulomics, Pacific Biosciences). The DNA samples were incubated with SRE buffer for 1 h (50°C), then the long fragments of DNA were pelleted at 10,000 g for 30 min (room temperature) and re-suspended in elution buffer. DNA concentrations were quantified using a Qubit 4 fluoremeter with 1X dsDNA kit.

Nanopore libraries were prepared using the Ligation Sequencing Kit LSK114 (Oxford Nanopore Technologies). The *R. culicivorax* library was loaded a first time on one R10.4 MinION flowcell. The library was recovered from the flowcell and reloaded after nuclease flush. The *P.* sp. PS1159 library was loaded 4 times (with nuclease flushes and fresh library loads) on one R10.4 MinION flowcell. Fast5 files were converted to Pod5 using pod5 v0.2.2. Basecalling was performed using Dorado v0.3.1 (Oxford Nanopore Technologies 2022) in duplex mode with model dna_r10.4.1_e8.2_400bps_supv4.2.0 and the reads were converted to fastq using SAMtools v1.6 (Danecek et al. 2021) with the module `samtools fastq`. This resulted in 5.7 Gb of Nanopore reads for *R. culicivorax* (N50: 15.9 kb) and 10.7 Gb for *P.* sp. PS1159 (N50: 33.4 kb) (Supplementary Table S2). Adapters were trimmed using chopper v0.5.0 (De Coster and Rademakers 2023) with minimum quality `-q` set to default (for *R. culicivorax* and *P.* sp. PS1159) or 20 (for *P.* sp. PS1159).

2.3 RNA sequencing

RNA was extracted from *R. culicivorax* adults using a modified version of the protocol established by Chomczynski and Sacchi (1987). Tissue pellets of approximately 10 mg were transferred into 1 mL Trimix and lysed using a homogeniser (Ultra-Turrax, IKA Werke GmbH) for 10 min on ice. After addition of 200 μL chloroform and incubation at room temperature for 5 min, the sample was centrifuged for 10 min at 15,000 g. The aqueous phase was collected and supplemented with 0.025 volumes of 1 M acidic acid and 0.5 volumes of pre-cooled 100% EtOH (−20°C). RNA was precipitated overnight at −20°C and then centrifuged at 15,000 g for 20 min. After removing the supernatant, the RNA pellet was dried for 10 min and resuspended in 125 μL of GU-mix and added 3.125 μL 1M acidic acid, vortexed the sample and added 70 μL 100% EtOH. RNA was precipitated overnight at −20°C and then centrifuged at 15,000 g for 20 min and washed twice with 500 μL EtOH (70%). The RNA pellet was resuspended in 20 μL DEPC-H2O and incubated at 65°C for 5 min. The quality of the total RNA was assessed using degenerative agarose-gel electrophoresis and a Nanodrop 1000 photometer (Agilent Inc.). RNA libraries were prepared using a TrueSeq RNA Sample Prep kit v2 (Illumina Inc.) and sequenced on Illumina HiSeq and MiSeq platforms (Illumina Inc.) at the Cologne Center for Genomics (CCG, Cologne, Germany). For *Panagrolaimus* sp. PS1159, publicly available Illumina RNA sequencing reads were used (SRR5253560) (Schiffer et al. 2019).

2.4 Long-read preliminary analyses

Quality and length of PacBio HiFi and Nanopore reads were plotted using Nanoplot v1.41.3 (De Coster and Rademakers 2023). Ploidy was estimated using Smudgeplot v0.2.2 (Ranallo-Benavidez et al. 2020) with the PacBio HiFi reads.

2.5 *Romanomermis culicivorax* long-read assembly

PacBio HiFi reads were assembled using Flye v2.9 (Kolmogorov et al. 2019) with parameter `–pacbio-hifi`, hifiasm v0.19 (Cheng et al. 2021) with parameter `-l 3`, NextDenovo v2.5 (NextOmics 2019) with parameters `genome_size = 300m read_type = hifi`, and wtdbg2 v2.5 (Ruan and Li 2020) with parameter `-x ccs`. For Nanopore reads, Canu v2.2 (Koren et al. 2017) was run with parameters `-nanopore genomeSize = 300m`, Flye v2.9 (Kolmogorov et al. 2019) with parameter `–nano-hq`, NextDenovo v2.5 (NextOmics 2019) with parameters `genome_size = 300m read_type = raw`, and wtdbg2 v2.5 (Ruan and Li 2020) with parameter `-x ont`. To combine PacBio HiFi and Nanopore reads, Nanopore reads longer than 15 kb were selected using seqtk v1.3 (Li 2012) with the module `seqtk seq` and the parameter `-L 15000`. hifiasm v0.19 was run using the PacBio HiFi reads and Nanopore reads > 15 kb with parameter `-l 3`. Assembly using Verkko v1.4 with default parameters failed.

2.6 *Panagrolaimus* sp. PS1159 long-read assembly

PacBio HiFi reads were assembled using Flye v2.9 (Kolmogorov et al. 2019) with parameter `–pacbio-hifi` and with the option `–keep-haplotypes`, hifiasm v0.19 (Cheng et al. 2021) was run

with parameters–n-hap 3 and -l set to 0 and 3, NextDenovo v2.5 (NextOmics 2019) with parameters genome_size = 300m read_type = hifi, and wtdbg2 v2.5 (Ruan and Li 2020) with parameter -x ccs. Nanopore reads with a quality higher than Q20 were selected using chopper. Different parameters were tested to adapt to the high accuracy and assemblies with highest contiguity and completeness were selected. Canu v2.2 (Koren et al. 2017) was run with parameters -nanopore -corrected genomeSize = 300m, Flye v2.9 (Kolmogorov et al. 2019) was run with parameters–nano-corr and with the option –keep-haplotypes, NextDenovo v2.5 (NextOmics 2019) was run with parameters genome_size = 300m read_type = hifi and wtdbg2 v2.5 (Ruan and Li 2020) was run with parameter -x ont. To combine PacBio HiFi and Nanopore reads, Nanopore reads longer than 30 kb were selected using seqtk v1.3 (Li 2012) with the module seqtk seq and the parameter -L 30000. hifiasm v0.19 (Cheng et al. 2021) was run using both datasets with parameters–n-hap 3 and -l set to 0 and 3. Verkko v1.4 (Rautiainen et al. 2023) was run with default parameters.

2.7 Assembly evaluation and post-processing

Assembly statistics were calculated using assembly-stats v1.0.1 (Sanger-Pathogens 2014). Ortholog completeness was computed using the Benchmarking Universal Single-Copy Orthologs (BUSCO) (Manni et al. 2021) tool v5.4.7 with parameter -m genome against the Metazoa odb10 and Nematoda odb10 lineages. PacBio HiFi reads were mapped against HiFi assemblies using minimap2 v2.24 (Li 2018) with parameters -ax map-hifi and Nanopore reads were mapped against the Nanopore and hybrid assemblies with parameters -ax map-ont. Mapped reads were sorted using SAMtools v1.6 with the module samtools sort. Contigs were aligned against the nt database using BLAST v2.13.0 (Altschul et al. 1990). The outputs were provided as input to BlobToolKit v4.1.5 (Challis et al. 2020), and contaminants identified as Proteobacteria, Actinobacteria, Actinomycetota and Bacteroidetes were subsequently removed; bacterial DNA is expected as these nematodes feed on bacteria. Reads were mapped again using minimap2 v2.24 and the output was provided to purge_dups v1.2.5 (Guan et al. 2020) to remove uncollapsed haplotypes. PacBio HiFi reads were used to purge HiFi-based assemblies, Nanopore reads for Nanopore-based assemblies, Nanopore reads for hybrid assemblies of *Panagrolaimus* sp. PS1159, and PacBio HiFi reads for hybrid assemblies of *Romanomermis culicivorax* (due to the low coverage of Nanopore reads).

2.8 Final scaffolding

Romanomermis culicivorax was assembled following two pipelines: 1) the decontaminated NextDenovo PacBio HiFi contigs were purged once using purge_dups; 2) the decontamined hifiasm PacBio HiFi + Nanopore contigs were purged twice; the assembly 1) was then scaffolded using RagTag v2.1.0 (Alonge et al. 2022) and the assembly 2) as reference. *Panagrolaimus* sp. PS1159 was also assembled using two pipelines: 1) the decontaminated hifiasm -l 3 PacBio HiFi + Nanopore contigs were purged twice using purge_dups; 2) the decontaminated Flye –keep-haplotypes Nanopore contigs were purged twice; the assembly 1) was then scaffolded using RagTag v2.1.0 and the assembly 2) as reference. The decontaminated hifiasm -l 0 PacBio HiFi + Nanopore contigs were retained as a phased assembly.

2.9 Repeat and gene annotation

Repeats were annotated using the Extensive *De novo* TE Annotator (EDTA) pipeline v2.0.1 (Ou et al. 2019) with parameters–sensitive 1 –anno 1. This pipeline filters and combines predictions from LTRharvest (Gremme et al. 2013), LTR_FINDER (Xu and Wang 2007) LTR_retriever (Ou and Jiang 2018), HelitronScanner (Xiong et al. 2014), Generic Repeat Finder (Shi and Liang 2019), TIR-learner (Su et al. 2019) and produces a final transposable element library using RepeatModeler (Flynn et al. 2020). The output hardmasked assembly was converted into a softmasked assembly. RNA-seq reads were trimmed using Trim Galore v0.6.10 and mapped to the assemblies using hisat2 v2.2.1 (Kim et al. 2019). After sorting using SAMTools v1.6 (Danecek et al. 2021), the mapped reads were provided as input to BRAKER v3.0.3 (Gabriel et al. 2023) with parameters –gff3 –UTR off.

2.10 Downstream analyses

BUSCO v5.4.7 (Manni et al. 2021) was run on the annotated protein-coding genes using the option -m proteins against the Metazoa odb10 and Nematoda odb10 lineages. *k*-mer completeness of the assemblies was assessed based on the PacBio HiFi dataset using Merqury v1.3 (Rhie et al. 2020).

3 Results

3.1 Initial long-read analyses

PacBio HiFi sequencing resulted in 37.5 Gb of reads (N50: 12.6 kb) for *Romanomermis culicivorax* and 29.2 Gb (N50: 15.8 kb) for *Panagrolaimus* sp. PS1159 (Supplementary Table S1). Nanopore sequencing yielded 5.7 Gb (N50: 15.9 kb) for *R. culicivorax* and 10.7 Gb (N50: 33.4 kb) for *P.* sp. PS1159 (Supplementary Table S2). While PacBio HiFi reads have a higher quality, Nanopore reads reach longer lengths, including some reads of 100+ kb (Figure 1A). Ploidy analyses using Smudgeplot predicts *R. culicivorax* as a diploid genome, while *P.* sp. PS1159 is expected to be triploid (Figure 1B). Nanopore reads with Q20+ quality were selected for initial assembly of *P.* sp. PS1159, but no quality threshold was applied for *R. culicivorax* Nanopore reads due to their limited amount. All PacBio HiFi reads were used for initial assemblies.

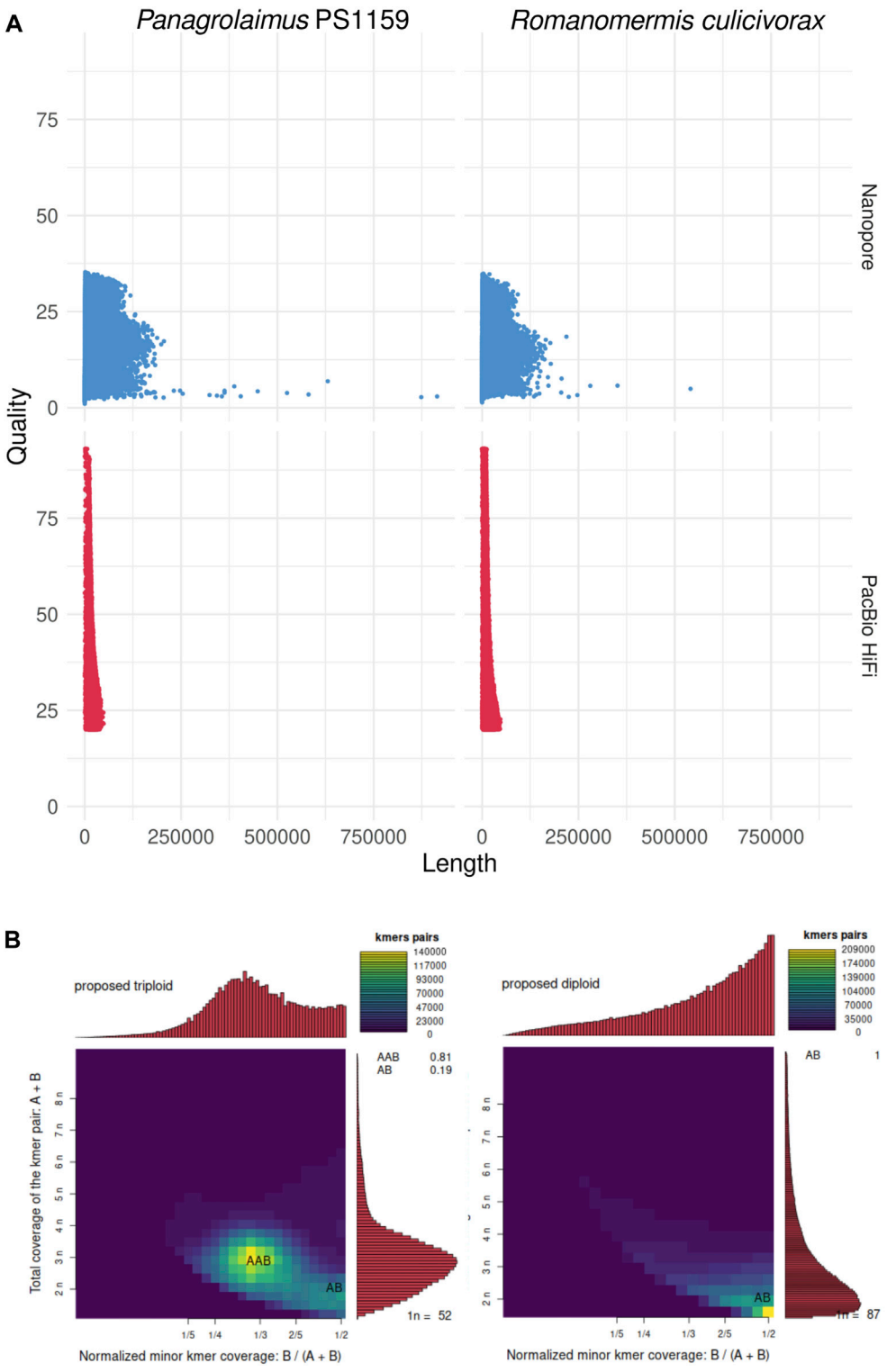

FIGURE 1
Initial analyses of the long reads. **(A)** Quality and length of Nanopore and PacBio HiFi reads. **(B)** *k*-mer analysis of the ploidy of the genomes using PacBio HiFi reads.

3.2 High-quality long-read assemblies

Depending on the program used, assemblies of PacBio HiFi, and Nanopore reads yielded contigs with variable contiguity, and cumulative size (Figure 2). For *Romanomermis culicivorax*, some PacBio HiFi assemblies had a size moderately above the Illumina assembly size of 322.8 Mb (467.2 for NextDenovo, 404.8 Mb for wtdbg2), but hifiasm and Flye produced oversized assemblies

FIGURE 2
Draft assembly statistics of PacBio HiFi reads (red), Nanopore reads (blue) and the two combined (purple) with assembly size (in Mb), N50 (in kb) and BUSCO completeness against the Metazoa and Nematoda lineages.

(969.3 Mb and 1.1 Gb respectively). These large genome sizes could not be explained by bacterial contamination in the data coming from their environment, as there was almost none in HiFi assemblies (Supplementary Figure S1). Nanopore assemblies were smaller: Flye and wtdbg2 assembly sizes were above the Illumina assembly size (499.3 Mb and 398.2 Mb), and Canu and NextDenovo assemblies were much shorter (114.9 Mb and 101.7 Mb). This is likely due to the low coverage of the Nanopore dataset, which was aggravated by a high amount of contamination from Proteobacteria and Bacteroidetes (Supplementary Figure S2), and led to a suboptimal sequencing coverage for these assemblers. Therefore, it is expected for Flye and wtdbg2 to yield the most qualitative assemblies as they have been shown to be more robust with low-coverage datasets (Guiglielmoni et al. 2021). The hybrid assembly obtained using hifiasm is oversized (1.1 Gb), similar to the PacBio-HiFi-only hifiasm assembly. N50s ranged from 108 kb (wtdbg2, Nanopore) to 550 kb (hifiasm, hybrid); although these values do not reach the Megabase level, they are still one order of magnitude larger than for the Illumina assembly (17.6 kb).

For *Panagrolaimus* sp. PS1159, assemblies ranged from 128.4 Mb (wtdbg2, PacBio HiFi) to 473.9 Mb (Flye, Nanopore). Shorter assemblies correlated with a low number of duplicated BUSCO orthologs, suggesting that they would be collapsed assemblies, in which homologous chromosomes are represented by one sequence. Larger assemblies have a high number of duplicated BUSCO orthologs, indicating that haplotypes are separated. These values would match the expectation of a phased assembly with a size three times larger than a collapsed assembly, for a triploid genome. These draft assemblies were overall more contiguous than for *R. culicivorax*, with a minimum of 240 kb (wtdbg2, Nanopore) and a maximum of 1.1 Mb (Flye, Nanopore). In addition, Nanopore assemblies had fewer bacterial contaminants than PacBio HiFi assemblies (Supplementary Figures S3, S4), likely owed to the supplementary sucrose decontamination step during library preparation. These read sets overall suffered much less from bacterial contamination than the Illumina data used in Schiffer et al. (2019).

After decontamination, haplotig purging and scaffolding, high-quality assemblies were obtained for both species. Although long reads were not sufficient to reach chromosome level, the final assemblies had an N50 over 1 Mb (1.1 Mb for *R. culicivorax* and 3.1 Mb for *P.* sp. PS1159) and their contiguity is drastically improved compared to Illumina assemblies (Table 1). Furthermore,

TABLE 1 Assembly statistics of previous (v1) and new (v2) versions of *Romanomermis culicivorax* and *Panagrolaimus* sp. PS1159.

	Romanomermis culicivorax *v1*	*Romanomermis culicivorax* v2	*Panagrolaimus* PS1159 v1	*Panagrolaimus* PS1159 v2
Assembly size	322.8 Mb	359.1 Mb	85.0 Mb	101.2 Mb
Number of scaffolds	62,537	595	17,628	67
N50	17.6 kb	1.0 Mb	9.9 kb	3.1 Mb
L50	4,624	114	2,232	11
N90	2.2 kb	315.0 kb	2.0 kb	1.1 Mb
L90	26,088	344	9,419	32
Number of gaps	303,605	716	49,960	140
Number of Ns	55.1 Mb	70.6 kb	1.7 Mb	12.7 kb
BUSCO score (Metazoa)	66.7%	68.4%	60.2%	65.9%
Single-copy orthologs	66.5%	65.2%	57.0%	57.2%
Duplicated orthologs	0.2%	3.2%	3.2%	8.7%
Fragmented orthologs	14.7%	9.2%	10.4%	7.4%
BUSCO score (Nematoda)	35.2%	39.4%	59.7%	66.6%
Single-copy orthologs	34.0%	37.3%	57.1%	58.7%
Duplicated orthologs	1.2%	2.1%	2.6%	7.9%
Fragmented orthologs	4.4%	4.1%	4.8%	4.4%

their BUSCO scores against the Metazoa and Nematoda lineages were also improved. Interestingly, the nematode BUSCO score of *R. culicivorax* remained low (35.2%), despite a higher metazoan BUSCO score. This suggests that the genome could be lacking many orthologs that would be expected in nematodes. The assembly of *R. culicivorax* has a QV score of 54.97; the *k*-mer spectrum shows a mostly collapsed assembly with yet some remaining artefactual duplications (Supplementary Figure S5). The assembly of *P.* sp. PS1159 has a QV score of 47.73 and the *k*-mer spectrum also supports a mostly collapsed assembly with limited artefactual duplications (Supplementary Figure S6).

3.3 Repeat and gene annotation

Repetitions were better resolved in the new long-read assemblies, than in the originally published ones (Figure 3). 68.2% of repeats were identified in the assembly of *R. culicivorax*, bringing it closer to the repetitive content of *Mermis negrescens*. The assembly of *P.* sp. PS1159 only has 16.0%, which is also higher than the 7.2% of repeats in the Illumina assembly. Many transposable elements (TE) were recovered in these improved assemblies that were undetected in Illumina assemblies. Notably, more long terminal repeats (LTR) were identified in *R. culicivorax*, the number of target inverted repeats was greatly increased, and 6.5 Mb of polintons were uncovered while they were almost absent in the Illumina assembly (Supplementary Table S3). The load of transposable elements is much lower in *P.* sp. PS1159 but still has a wider variety of LTRs, TIRs, helitrons and other elements than the Illumina assembly. Gene prediction resulted in 16,689 annotated genes for *R. culicivorax*, with overall BUSCO scores of 77.6% (Metazoa) and 56.2% (Nematoda), and 27,203 annotated genes for *P.* sp. PS1159 with overall BUSCO scores of 77.3% (Metazoa) and 78.6% (Nematoda). As expected, these annotations are more complete than the ones published with the previous Illumina assemblies.

3.4 Orthology analyses of the phased assembly of *Panagrolaimus* sp. PS1159

To re-analyse the *Panagrolaimus* sp. PS1159 in regard to being a triploid genome, we selected the hybrid hifiasm assembly as a phased candidate. After decontamination, this assembly has a size of 264.7 Mb, 876 contigs and an N50 of 559 kb. The assembly's BUSCO scores have high numbers of duplicated orthologs: 1.8% single-copy orthologs and 65.3% duplicated orthologs against Metazoa; 2.1% single-copy orthologs and 66.3% duplicated orthologs against Nematoda. The *k*-mer spectrum shows that the assembly has *k*-mers represented once, twice, or in three copies in the three different peaks at 50X, 100X and 150X (Supplementary Figure S7), which is expected for a phased triploid genome assembly. In addition, the QV score reaches 48.18. Annotation resulted in 70,448 predicted genes, with BUSCO scores of 78.1% (77.4% duplicated) against Metazoa and 79.7% (78.7% duplicated) against Nematoda. We analyzed the number of ortholog copies from the annotated genes in the collapsed and phased

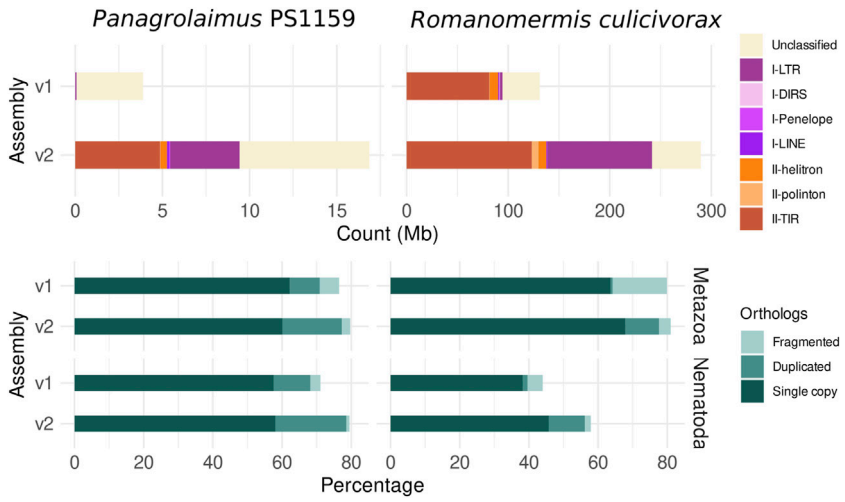

FIGURE 3
Comparison of assemblies based on TE (top) count and BUSCO ortholog (bottom) statistics for the protein annotations (against the Metazoa and Nematoda lineages) shows higher repeat and gene completeness of the new assemblies.

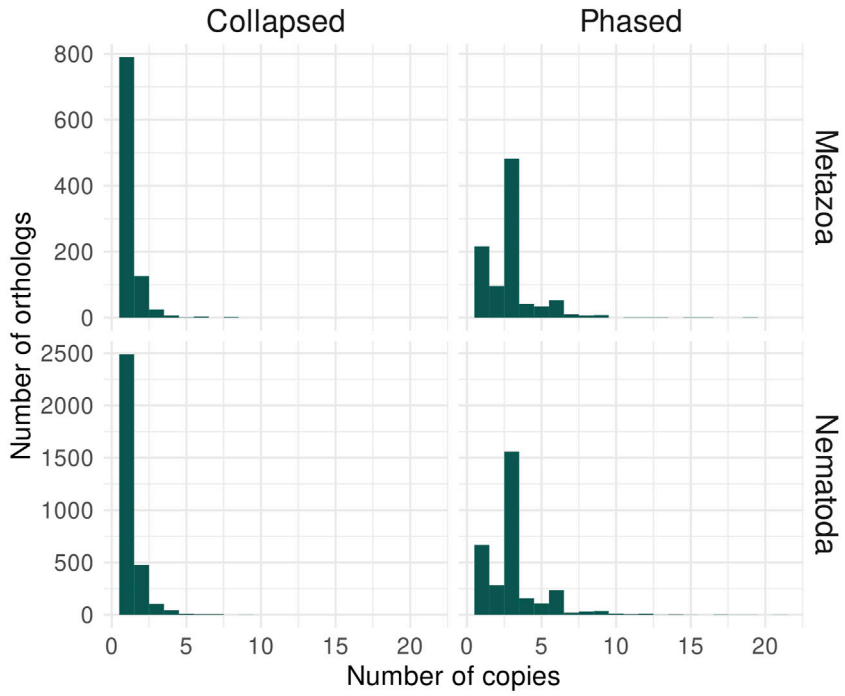

FIGURE 4
Ortholog analysis of *Panagrolaimus* sp. PS1159 collapsed and phased assemblies supports triploidy. The histograms represent the number of orthologs with their copy number from the lineages Metazoa and Nematoda identified in the protein-coding genes annotated for *P.* sp. PS1159. In the collapsed assembly, the majority of orthologs are present in a single copy, while most orthologs are in three copies in the phased assembly.

assemblies (Figure 4), considering that orthologs used by BUSCO are expected as single copy. For the collapsed assembly, most orthologs are in only one copy. In the phased assembly, the majority of orthologs are in three copies, as there would be one copy for each haplotype. This brings further support to the triploidy of *P.* sp. PS1159.

4 Discussion

Our new long-read assemblies for *Romanomermis culicivorax* and *Panagrolaimus* sp. PS1159 provide a drastic improvement to the previously published short-read-based assemblies, with higher contiguity, improved repeat resolution, and more accurate gene

annotation. Furthermore, we generated a draft phased assembly of *Panagrolaimus* sp. PS1159, which opens new possibilites for haplotype-specicific analyses. With the addition of long-range sequencing data, such as chromosome conformation capture, we can expect to scaffold these high-quality assemblies into chromosome candidates and further investigate genome structures.

The first challenge consisted in generating PacBio HiFi and Nanopore sequencing data for these two non-model species. The resulting reads clearly highlight the strengths of these technologies: while Nanopore reads provide an advantage on length, PacBio HiFi reads have the highest accuracy. It should be noted however that the overall accuracy of Nanopore reads has increased compared to data from R9.4.1 flowcells (Guiglielmoni et al. 2021) and were sufficient to produce assemblies with a high BUSCO completeness. Ultra-low input PacBio HiFi sequencing resulted in large datasets (over 29 Gb) despite the use of only a few individuals, and also led to high-quality draft assemblies. This amplification-based approach can be favored when the DNA availability for a species is limited, and for instance for nematodes which cannot be cultured. It should be considered however that amplification protocols can lead to a bias in the sequencing data. To better understand the impact of amplification bias on assemblies, additional PacBio HiFi reads without amplification from a large pool of individuals could be generated in future experiments.

Most initial assemblies improved on the published Illumina assemblies of the two species. The oversized PacBio HiFi assemblies of *R. culicivorax* could be attributed to the use of several individuals combined with the high accuracy of PacBio HiFi reads, leading to the separation of multiple haplotypes in heterozygous regions. Based on the quality of the assemblies obtained from the ultra-low input PacBio HiFi reads, we can expect that further improvements would enable the generation of data from a single individual, which would prevent issues introduced by alternative haplotypes and could additionally be used to generate a phased assembly. Nanopore assemblies did not have similar large sizes which may be owed to the lower accuracy of Nanopore reads which did not discriminate alternative haplotypes. For *P.* sp. PS1159, the Nanopore dataset was large enough to select for the more accurate Q20+ reads; therefore, haplotypes could be separated in both PacBio HiFi and Nanopore assemblies. In fact, almost all draft assemblies had the three haplotypes mostly separated with sizes close to 300 Mb (which would be the expected phased assembly size) and most BUSCO orthologs in multiple copies. Regarding contiguity, *R. culicivorax* assemblies were generally less contiguous than *P.* sp. PS1159 assemblies, which might be attributed to the higher repetitive content of this genome.

The most striking improvement in these assemblies lies in the resolution of repetitive regions. For both species, the percentage of repetitions in the genomes increased and revealed a wider variety of transposable elements. The comparison highlights that these transposable elements were in fact almost absent in the assembly of *P.* sp. PS1159 and very partially recovered in the assembly of *R. culicivorax*. Considering that TEs represent 289 Mb of the 359-Mb genome, we can estimate that a large aspect of this genome was completely overlooked in the past. A recent study has shown that genome assemblies from basal nematodes contain more repeats (ranging from 23.4% up to 50.6% repeats) than nematodes belonging to other clades (ranging from 0.8 %p to 31%) (Lee et al. 2023). The results here presented are consistent with the previous findings as *R. culicivorax*, a basal nematode, showed a high repeat and TE content and the derived *P.* sp. PS1159 has a low repeat and TE content. These variations and the better resolution of repetitions in long-read assembly should prompt further investigation into TE contents through nematode evolution.

The numbers of annotated genes for version 1 and 2 of *P.* sp. PS1159 are similar (26,760 genes v. 27,203), yet this number shrank for *R. culicivorax*: while the first assembly had 48,376 annotated genes, the long-read assembly has 16,689. This did not lead to a decrease in ortholog completeness as the BUSCO scores of the new assemblies and annotations both reached higher values. Interestingly, the score of *R. culicivorax* against the Metazoa lineage is slightly higher than *P.* sp. PS1159, but its score against the Nematoda lineage is low with a value of only 39.4%. As a matter of fact, the Nematoda dataset is composed of seven nematode species, out of which only one is a basal nematode (*Trichinella spiralis*). The lack of representation of early branching nematodes could explain the lower BUSCO completeness on basal nematodes genomes when compared to representatives of higher clades like *P.* sp. PS1159, and illustrates the bias of current genomics resources. Early branching nematode genomes are scarce: even at the subclass level, genome assemblies are available on GenBank for only 15 Dorylaimia species and four Enoplia species (accessed on 06.10.2023). Therefore this study brings crucial resources to guide future sequencing projects for understudied nematodes and to fill the gaps among available assemblies.

The use of high-accuracy long reads permitted the generation of a first draft phased assembly of *P.* sp. PS1159. This assembly, combined with *k*-mer predictions based on PacBio HiFi reads and the analysis of Schiffer et al. (2019), confirms that this species has a triploid genome. Considering the potential hybridization which could have introduced this third copy, a haplotype-resolved assembly is especially warranted to identify the original and newly acquired alleles. These analyses demonstrate the feasibility of long-read collapsed and phased assemblies for challenging genomes of understudied nematode species, including in the context of high repetitiveness and polyploidy. We gained new insights into these genomes regarding their gene and repeat content, which pave the way for more in-depth comparative genomics.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ebi.ac.uk/ena, PRJEB66727.

Author contributions

NG: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Visualization, Writing–original draft, Writing–review and editing. LV: Data curation, Formal Analysis, Investigation, Methodology, Visualization, Writing–original draft, Writing–review and editing. JK: Data curation, Formal Analysis, Investigation, Visualization, Writing–original draft,

Writing–review and editing. PS: Conceptualization, Funding acquisition, Supervision, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This project was supported through a DFG Emmy Noether Program (ENP) Projekt (434028868) and the DFG funded project B08 in the CRC1211 (268236062) to PHS. NG's position was first funded through a Deutsche Forschungsgemeinschaft (DFG) grant (458953049) to PHS and subsequently through the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101110569.

Acknowledgments

We thank Christopher Kraus for his contribution to RNA sequencing and the Cologne Center for Genomics and the Genomics and Transcriptomics Laboratory for generation of sequencing data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1308527/full#supplementary-material

References

Abagli, A. Z., Alavo, T. B., Perez-Pacheco, R., and Platzer, E. G. (2019). Efficacy of the mermithid nematode, *Romanomermis iyengari*, for the biocontrol of *Anopheles gambiae*, the major malaria vector in sub-saharan africa. *Parasites & Vectors* 12, 253–8. doi:10.1186/s13071-019-3508-6

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., et al. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology* 23, 258–19. doi:10.1186/s13059-022-02823-7

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215, 403–410. doi:10.1016/S0022-2836(05)80360-2

Bhattarai, U. R., Poulin, R., Gemmell, N. J., and Dowle, E. (2022). Genome assembly and annotation of the mermithid nematode *Mermis nigrescens*. bioRxiv. doi:10.1101/2022.11.05.515230

Challis, R., Richards, E., Rajan, J., Cochrane, G., and Blaxter, M. (2020). Blobtoolkit–interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics* 10, 1361–1374. doi:10.1534/g3.119.400908

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature Methods* 18, 170–175. doi:10.1038/s41592-020-01056-5

Chomczynski, P., and Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical Biochemistry* 162, 156–159. doi:10.1006/abio.1987.9999

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10, Giab008. doi:10.1093/gigascience/giab008

De Coster, W., and Rademakers, R. (2023). NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* 39, btad311. doi:10.1093/bioinformatics/btad311

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* 117, 9451–9457. doi:10.1073/pnas.1921046117

Gabriel, L., Bruna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., et al. (2023). BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. bioRxiv. doi:10.1101/2023.06.10.544449

Giblin, R. M., and Platzer, E. G. (1985). *Romanomermis culicivorax* parasitism and the development, growth, and feeding rates of two mosquito species. *Journal of Invertebrate Pathology* 46, 11–19. doi:10.1016/0022-2011(85)90124-7

Gremme, G., Steinbiss, S., and Kurtz, S. (2013). GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10, 645–656. doi:10.1109/TCBB.2013.68

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. doi:10.1093/bioinformatics/btaa025

Guiglielmoni, N., Houtain, A., Derzelle, A., Van Doninck, K., and Flot, J.-F. (2021). Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. *BMC Bioinformatics* 22, 303–23. doi:10.1186/s12859-021-04118-3

Guiglielmoni, N., Rivera-Vicéns, R., Koszul, R., and Flot, J.-F. (2022). A deep dive into genome assemblies of non-vertebrate animals. *Peer Community Journal* 2, e29. doi:10.24072/pcjournal.128

Hodda, M. (2022). Phylum nematoda: trends in species descriptions, the documentation of diversity, systematics, and the species concept. *Zootaxa* 5114, 290–317. doi:10.11646/zootaxa.5114.1.2

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* 37, 907–915. doi:10.1038/s41587-019-0201-4

Koetsier, P. A. G., and Cantor, E. J. (2021). A simple approach for effective shearing and reliable concentration measurement of ultra-high-molecular-weight DNA. *BioTechniques* 71, 439–444. doi:10.2144/btn-2021-0051

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* 37, 540–546. doi:10.1038/s41587-019-0072-8

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Research* 25, 722–736. doi:10.1101/gr.215087.116

Kumar, S., Koutsovoulos, G., Kaur, G., and Blaxter, M. (2012a). Toward 959 nematode genomes. *Worm* 1, 42–50. doi:10.4161/worm.19046

Kumar, S., Schiffer, P. H., and Blaxter, M. (2012b). 959 nematode genomes: a semantic wiki for coordinating sequencing projects. *Nucleic Acids Research* 40, D1295–D1300. doi:10.1093/nar/gkr826

Lee, Y.-C., Ke, H.-M., Liu, Y.-C., Lee, H.-H., Wang, M.-C., Tseng, Y.-C., et al. (2023). Single-worm long-read sequencing reveals genome diversity in free-living nematodes. *Nucleic Acids Research* 51, 8035–8047. doi:10.1093/nar/gkad647

Lewis, S. C., Dyal, L. A., Hilburn, C. F., Weitz, S., Liau, W.-S., LaMunyon, C. W., et al. (2009). Molecular evolution in *Panagrolaimus* nematodes: origins of parthenogenesis, hermaphroditism and the Antarctic species *P. davidi*. *BMC Evolutionary Biology* 9, 15. doi:10.1186/1471-2148-9-15

Li, H. (2012). *Seqtk*. Available at: https://github.com/lh3/seqtk.

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

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* 38, 4647–4654. doi:10.1093/molbev/msab199

McGill, L. M., Shannon, A. J., Pisani, D., Félix, M.-A., Ramløv, H., Dix, I., et al. (2015). Anhydrobiosis and freezing-tolerance: Adaptations that facilitate the establishment of panagrolaimus nematodes in polar habitats. *PLOS ONE* 10, e0116084. doi:10.1371/journal.pone.0116084

Mitreva, M., Jasmer, D. P., Zarlenga, D. S., Wang, Z., Abubucker, S., Martin, J., et al. (2011). The draft genome of the parasitic nematode *Trichinella spiralis*. *Nature Genetics* 43, 228–235. doi:10.1038/ng.769

NextOmics (2019). *NextDenovo*. Available at: https://github.com/Nextomics/NextDenovo.

Ou, S., and Jiang, N. (2018). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology* 176, 1410–1422. doi:10.1104/pp.17.01310

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R., Hellinga, A. J., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology* 20, 275–18. doi:10.1186/s13059-019-1905-y

Oxford Nanopore Technologies (2022). *Dorado*. Available at: https://github.com/nanoporetech/dorado.

Sanger-Pathogens (2014). *Pathogen Informatics, Wellcome Sanger Institute*. Assembly-Stats. Available at: https://github.com/sanger-pathogens.

Petersen, J., Chapman, H., Willis, O., and Fukuda, T. (1978). Release of *Romanomermis culicivorax* for the control of *Anopheles albimanus* in El Salvador II. Application of the nematode. *The American Journal of Tropical Medicine and Hygiene* 27, 1268–1273. doi:10.4269/ajtmh.1978.27.1268

Presswell, B., Evans, S., Poulin, R., and Jorge, F. (2015). Morphological and molecular characterization of Mermis nigrescens Dujardin, (Nematoda: Mermithidae) parasitizing the introduced European earwig (Dermaptera: Forficulidae) in New Zealand. *Journal of Helminthology* 89, 267–276. doi:10.1017/S0022149X14000017

Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* 11, 1432. doi:10.1038/s41467-020-14998-3

Rautiainen, M., Nurk, S., Walenz, B. P., Logsdon, G. A., Porubsky, D., Rhie, A., et al. (2023). Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nature Biotechnology* 41, 1474–1482. doi:10.1038/s41587-023-01662-6

Rhie, A., Walenz, B. P., Koren, S., and Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* 21, 245–27. doi:10.1186/s13059-020-02134-9

Rice, E. S., and Green, R. E. (2019). New approaches for genome assembly and scaffolding. *Annual Review of Animal Biosciences* 7, 17–40. doi:10.1146/annurev-animal-020518-115344

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods* 17, 155–158. doi:10.1038/s41592-019-0669-3

Schiffer, P. H., Danchin, E. G., Burnell, A. M., Creevey, C. J., Wong, S., Dix, I., et al. (2019). Signatures of the Evolution of Parthenogenesis and Cryptobiosis in the Genomes of Panagrolaimid Nematodes. *iScience* 21, 587–602. doi:10.1016/j.isci.2019.10.039

Schiffer, P. H., Kroiher, M., Kraus, C., Koutsovoulos, G. D., Kumar, S., R Camps, J. I., et al. (2013). The genome of *Romanomermis culicivorax*: revealing fundamental changes in the core developmental genetic toolkit in Nematoda. *BMC Genomics* 14, 923–16. doi:10.1186/1471-2164-14-923

Sereika, M., Kirkegaard, R. H., Karst, S. M., Michaelsen, T. Y., Sørensen, E. A., Wollenberg, R. D., et al. (2021). Oxford Nanopore R10.4 long-read sequencing enables near-perfect bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. bioRxiv. doi:10.1101/2021.10.27.466057

Shahid, S., and Slotkin, R. K. (2020). The current revolution in transposable element biology enabled by long reads. *Current Opinion in Plant Biology* 54, 49–56. doi:10.1016/j.pbi.2019.12.012

Shannon, A. J., Browne, J. A., Boyd, J., Fitzpatrick, D. A., and Burnell, A. M. (2005). The anhydrobiotic potential and molecular phylogenetics of species and strains of *Panagrolaimus* (Nematoda, Panagrolaimidae). *Journal of Experimental Biology* 208, 2433–2445. doi:10.1242/jeb.01629

Shatilovich, A., Gade, V. R., Pippel, M., Hoffmeyer, T. T., Tchesunov, A. V., Stevens, L., et al. (2023). A novel nematode species from the siberian permafrost shares adaptive mechanisms for cryptobiotic survival with *C. elegans* dauer larva. *PLOS Genetics* 19, e1010798. doi:10.1371/journal.pgen.1010798

Shi, J., and Liang, C. (2019). Generic Repeat Finder: a high-sensitivity tool for genome-wide *de novo* repeat detection. *Plant Physiology* 180, 1803–1815. doi:10.1104/pp.19.00386

Su, W., Gu, X., and Peterson, T. (2019). TIR-Learner, a new ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. *Molecular Plant* 12, 447–460. doi:10.1016/j.molp.2019.02.008

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P., Hall, R. J., Concepcion, G. T., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* 37, 1155–1162. doi:10.1038/s41587-019-0217-9

Xiong, W., He, L., Lai, J., Dooner, H. K., and Du, C. (2014). HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proceedings of the National Academy of Sciences* 111, 10263–10268. doi:10.1073/pnas.1410068111

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* 35, W265–W268. doi:10.1093/nar/gkm286