High-fidelity annotated triploid genome of the quarantine root-knot nematode, Meloidogyne enterolobii

Marine Poullet, Hemanth Konigopal, Corinne Rancurel, Marine Sallaberry, Celine Lopez-Roques, Ana Paula Zotta Mota, Joanna Lledo, Sebastian Kiewnick, Etienne G J Danchin

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

High-fidelity annotated triploid genome of the quarantine root-knot nematode, *Meloidogyne enterolobii*

Marine Poullet[1]✉, Hemanth Konigopal[2], Corinne Rancurel[1], Marine Sallaberry[3], Celine Lopez-Roques[3], Ana Paula Zotta Mota[1], Joanna Lledo[3], Sebastian Kiewnick[2,4] & Etienne G. J. Danchin[1,4]✉

Root-knot nematodes (RKN) of the genus *Meloidogyne* are obligatory plant endoparasites that cause substantial economic losses to agricultural production and impact the global food supply. These plant parasitic nematodes belong to the most widespread and devastating genus worldwide, yet few measures of control are available. The most efficient way to control RKN is deployment of resistance genes in plants. However, current resistance genes that control other *Meloidogyne* species are mostly inefficient on *Meloidogyne enterolobii*. Consequently, *M. enterolobii* was listed as a European Union quarantine pest requiring regulation. To gain insight into the molecular characteristics underlying its parasitic success, exploring the genome of *M. enterolobii* is essential. Here, we report a high-quality genome assembly of *M. enterolobii* using the high-fidelity long-read sequencing technology developed by Pacific Biosciences, combined with a gap-aware sequence transformer, DeepConsensus. The resulting triploid genome assembly spans 285.4 Mb with 556 contigs, a GC% of 30 ± 0.042 and an N50 value of 2.11 Mb, constituting a useful platform for comparative, population and functional genomics.

Background & Summary

Root-knot nematodes (RKN) belong to the genus *Meloidogyne*, and are among the most destructive plant-parasitic nematodes[1]. Due to their extensive geographic distribution and ability to infest a wide range of host plants, they have a detrimental impact on the yield and quality of numerous economically valuable crops[2]. At present, the *Meloidogyne* genus comprises more than 100 described species. However, *M. arenaria, M. incognita, M. javanica* and *M. hapla* are considered the most widespread and damaging species[3]. In recent years, *M. enterolobii* has received increasing attention due to its unique ability to overcome several sources of resistance against the other RKN[2,4,5].

The species *M. enterolobii* was originally first described as *M. incognita* from a population obtained from the Pacara Earpod Tree (*Enterolobium contortisiliquum* [Vell.] Morong) in Hainan Island, China by Yang and Eisenback (1983)[6]. Later in 1988, Rammah and Hirschmann described a new species[7], *M. mayaguensis*, sampled from eggplant (*Solanum melongena* L.) roots from Puerto Rico. However, this species was later synonymized with *M. enterolobii*, based on identical esterase phenotype and mitochondrial DNA sequence[8,9].

M. enterolobii has an extremely high damage potential[10], surpassing many of the other RKN species studied so far[11,12]. The reports of severe damage in high-value crops have increased in the past years[13,14]. In 2009, the European Plant Protection Organization (EPPO) performed a risk analysis, which came to the conclusion that this species was recommended for regulation and placed on the EPPO A2 list in 2010. Following numerous interceptions over the years, it was concluded that *M. enterolobii* fulfilled the conditions provided in Article 3 and Section 1 of Annex I to Regulation (EU) 2016/2031 in respect of the Union territory and therefore should be listed in Part A of Annex II to Implementing Regulation (EU) 2019/2072 as Union quarantine pest[15]. However,

[1]Institut Sophia Agrobiotech, INRAE, Université Côte d'Azur, CNRS, 400 routes des Chappes, 06903, Sophia-Antipolis, France. [2]Julius Kühn-Institut, Institute for Plant Protection in Field Crops and Grassland, Messeweg 11-12, 38104, Braunschweig, Germany. [3]INRAE, GeT_PlaGe, Genotoul, 31326, Castanet – Tolosan, France. [4]These authors jointly supervised this work. ✉e-mail: poullet.m@hotmail.fr; etienne.danchin@inrae.fr

once damage is detected, *M. enterolobii* identification is challenging due to morphological resemblances it shares with other RKN species[11,14,16,17].

In that perspective, providing high-quality nuclear and mitochondrial genomes for this species can accelerate the development of reliable molecular markers and the understanding of the biology of *M. enterolobii*. A first *M. enterolobii* draft genome was published in 2017 as part of a comparative genomics analysis with other RKN[18]. The population named L30, originated from Burkina Faso and was sequenced using Illumina short reads. Consequently, the assembled genome was quite fragmented with $>$ 46,000 contigs and an N50 length $<$ 9.3 kb, precluding analyses of structural variants or conserved synteny with other *Meloidogyne* species. Nevertheless, this initial genome allowed confirming that this species was likely polyploid, similarly to other tropical parthenogenetic RKN[19]. Using k-mer statistics[20] on the Illumina reads, *M. enterolobii* L30 was further predicted to be triploid with relatively high divergence between its three subgenomes[21]. Using the same k-mer approaches, similar conclusions about subgenomes high divergence were drawn for the triploid genome of *M. incognita* and the tetraploid ones of *M. arenaria* and *M. javanica*[22]. In 2020, a *M. enterolobii* genome assembled from Pacific Biosciences (PacBio) RS long reads and polished with Illumina short reads was published[23]. The sequenced population named Mma-II, was isolated from infected tomatoes in a Swiss organic farm[24]. With $>$ 4,500 contigs and an N50 length of 143 kb, the genome assembly, predicted to be triploid, represented a substantial improvement compared to the only other assembly available at this time. However, recent progress in the quality and data volume of long-read sequencing technologies[25] promises more contiguous and higher-quality genomes even for complex polyploid species such as those present in the *Meloidogyne* genus[22,26]. Therefore, we used the PacBio HiFi, highly accurate long-read sequencing technology to produce a more complete, contiguous and reliable reference genome for this quarantine plant-parasitic nematode.

Using this technology and further improvement of the quality of the reads, we assembled the genome of the *M. enterolobii* population (E1834), originally isolated from the roots of eggplant collected in Puerto Rico, in 556 contigs with an N50 length surpassing 2 Mb. It should be noted here that as opposed to diploid genomes with low heterozygosity, for which the genome is represented as a haploid consensus, we aimed at representing the three divergent subgenomes in the assembly. The genome assembly size of 285.4 Mb was consistent with previous flow cytometry estimation of nucleus total DNA content on a population from Guadeloupe island (274.7 $\pm$ 18.52 Mb)[23]. This suggests the three subgenomes are represented in this new assembly.

Further quality checks of our genome assembly, along with comparisons to all previously available *M. enterolobii* genomes, including the isolate from Burkina Faso (L30)[18] and from Switzerland (Mma-II-24)[23], confirmed the correct species identification for population E1834. The accurate species identity was corroborated for the short-read genome from Burkina Faso (L30), but not for the PacBio genome from Switzerland (Mma-II-24). Indeed, our study revealed that the Mma-II Swiss population previously sequenced with PacBio RS underwent a contamination by *M. incognita*, which over several generations in a greenhouse, completely overtook the originally described *M. enterolobii* population Mma-II. As this population was not maintained as a single egg mass line, contamination by a highly virulent and equally pathogenic *M. incognita* population remained undetected. Mis-identification among *Meloidogyne* species is not uncommon as reported populations of *M. ethiopica* in Europe were later identified as *M. luci*[27]. Consequently, the genome assembly in that publication[23] mostly corresponded to *M. incognita*, implying no long-read-based contiguous genome for *M. enterolobii* was finally available so far.

This finding also motivated us to develop a methodology based on mitochondrial genomes reconstruction and relative coverage to detect contamination between closely related species which are not detectable with standard BlobTools[28] approaches based on nuclear genome contigs GC content and coverage. This methodology can be reused to confirm correct species identification in other sequencing projects.

Overall, we propose a high-quality contiguous triploid genome for *M. enterolobii* constituting a reliable resource for within- and between-species comparative genomics. The contiguity of the genome enables study of structural variations and conserved synteny, which will be essential towards comprehensive identification of genomic variations in relation with the host range of this quarantine nematode species in Europe.

Methods

Nematode collection and DNA extraction. The *M. enterolobii* population (E1834) was originally isolated from the roots of eggplant collected in Puerto Rico and has been maintained since 2005 in the *Meloidogyne* spp. reference collection at The Netherlands Institute for Vectors, Invasive plants and Plant health (NIVIP) Wageningen, Netherlands. In 2020, this population was kindly provided by NRC, for the research conducted in the framework of the project AEGONE (No. 431627824r) and has been maintained at the Julius Kühn Institut (JKI) in Braunschweig, Germany in a greenhouse on the tomato cultivar 'Phantasia', resistant to other tropical RKN. Nematodes used for DNA extraction were obtained from single egg mass (SEM) lines. To obtain these lines, 12 single females with egg masses were carefully picked from the infected roots of tomato and second stage juveniles (J2) were allowed to hatch in six well plates (SARSTEDT AG & Co. KG, Nümbrecht, DE) with 5 ml molecular grade water per well. After one week at room temperature (20 $\pm$ 1 °C) in the dark, 10 wells with the highest number of hatched J2s were selected for inoculation. In addition, two J2s from each egg mass were collected for DNA extraction and species verification by Real-time PCR[29] and SCAR species-specific markers[30,31]. For multiplication of the SEMLs, five-week-old tomato seedlings from the cultivar 'Phantasia' were transplanted into 1000 ml clay pots (Risa Pflanzgefässe GmbH, Germany) containing 750 ml quartz sand (0.3–1 mm) supplemented with slow-releasing fertilizer, Osmocote (1.5 g/L). Afterwards, tomatoes were inoculated with J2s obtained from the respective egg masses. Tomato plants were maintained in a greenhouse at 20 to 25°C with 16 h of light and 8 h of darkness. Plants were watered daily and fertilized once per week with Wuxal® super solution (8:8:6; N: P: K, Hauert MANNA, Nürnberg, DE). After 8 weeks, the galled roots were carefully washed free of sand and the eggs and juveniles (E&J) were extracted with 0.7% chlorine solution[32]. The resulting E&J suspension was counted to

Single Egg Mass (SEML)	Number of Eggs & Juveniles per root
SEML 1	43,200 ± 577.35
SEML 2	22,7800 ± 1285.82
SEML 3	56,600 ± 1604.16
SEML 4	454,800 ± 1442.22
SEML 5	230,400 ± 945.16
SEML 6	3000 ± 503.32
SEML 7	187,800 ± 2457.64
SEML 8	170,000 ± 416.33
SEML 9	109,000 ± 1222.02
SEML 10	189,800 ± 901.85

Table 1. Number of newly produced eggs and juveniles per root system of 10 single eggs mass lines of *M. enterolobii* population (E1834). Tabulated values are the mean count of Eggs and Juveniles with standard error for different SEML.

identify the line with the highest reproduction rate. The SEML number 4 was therefore selected (Table 1) for further experiments and production of DNA.

DNA extraction. The selected SEML 4 was multiplied on tomato plants to obtain J2 for DNA extraction. Galled tomato roots were carefully washed free of sand and placed in a mist chamber to collect freshly hatched J2 after 14 days. The J2 suspension was purified by the modified centrifuge-floatation method[33] with a 45% sugar solution to reduce contaminations such as root debris, bacteria, fungal spores, etc… Afterwards, approximately 50,000-70,000 J2 were transferred into 1.5 ml Eppendorf tube and washed 3 times with molecular grade water. After freezing in liquid nitrogen, DNA was extracted from the homogenized sample using the MasterPure Complete DNA & RNA Purification Kit (Lucigen) following the manufacturer's protocol. The DNA was suspended in 10 mM Tris-HCl buffer and the DNA concentration was determined with either a Qubit™ 4 fluorometer (Life Technologies, Singapore) or NanoDrop 2000™ spectrophotometer (Thermo Fisher Scientific, USA). The NanoDrop 2000™ was blanked using the respective elution buffer for the method. DNA concentration was measured using Qubit™ (1X dsDNA HS (High Sensitivity) Assay Kit, Invitrogen, #Q32853) and NanoDrop 2000™. Purity was measured using the 260/280 nm and 260/230 nm absorbance ratios of NanoDrop 2000™.

Genome sequencing and read processing. The long-fragment DNA library from the *M. enterolobii* population E1834 was constructed at the GeT-PlaGe core facility, INRAE Toulouse according to the manufacturer's instructions "Preparing whole genome and metagenome libraries using SMRTbell® prep kit 3.0". At each step, DNA was quantified using the Qubit dsDNA HS Assay Kit (Life Technologies). DNA purity was tested using the NanoDrop 2000™ and size distribution and degradation assessed using the Femto pulse Genomic DNA 165 kb Kit (Agilent). Purification steps were performed using AMPure PB beads (PacBio) and SMRTbell cleanup beads (PacBio). A DNA damage repair step was performed using the SMRTbell Damage Repair Kit SPV3 (PacBio). A total of 9.4 μg of DNA was purified and then sheared at 20 kb using the Megaruptor system (Diagenode). The SMRTbell® prep kit 3.0 (PacBio) was used on 8.3 μg of sample for library construction then nuclease treatment. Subsequently, blunt hairpin adapters were ligated to the library and a nuclease treatment was performed using the nuclease mix of "SMRTbell® prep kit 3.0". In order to produce an 11 kb library, a size selection step using a cutoff from 6 kb to 50 kb was performed on the BluePippin Size Selection system (Sage Science) with "0.75% DF Marker S1 6–10 kb vs3 Improved Recovery" protocol. Using Binding kit 2.2 and sequencing kit 2.0, the primer V5 annealed, and polymerase 2.2 bounded library was sequenced by diffusion loading with the adaptive-loading method onto 1 SMRTcell 8 M on sequel II instrument at 90 pM with a 2 h pre-extension and a 30 h movie.

The Sequel II sequencing system outputs 1 Tb of raw data into a subread file. This contains unaligned base calls from high-quality regions, the complete set of base quality values and kinetic measurements from the sequencing instrument. This subread file is used as input for the Circular Consensus Sequencing (CCS[34] v6.4.0) analysis to generate a draft consensus sequence. Very low-quality reads (≺ Q9) were filtered out by using the parameter–min-rq = 0.88. To further improve the quality of the PacBio Sequel II reads, we have used a gap-aware sequence transformer, DeepConsensus[35] (v1.1.0). As a final step, the previous subreads were aligned to the draft consensus sequence using ACTC[36] with default parameters (v0.2.0) and used as input to the DeepConsensus[35] transformer-encoder. The Phred-scale read accuracy score (Qconcordance) has been calculated according to Baid *et al.*[35] where Qconcordance = −10*log10 (1-identity) and identity = matches / (matches + mismatches + deletions + insertions).

Ploidy, heterozygosity, and genome size estimation. To infer the ploidy level of the *M. enterolobii* population E1834, a k-mer-based approach was employed to profile the genome. The k-mer frequencies in DeepConsensus[35] sequencing reads were analyzed using KMC[37] (v3.0.0, kmc -k21 -m100 -ci1 -cs10000). In accordance with the author's recommendations, canonical 21-mers were extracted using a hash and organized in a histogram file using the kmc_tools transform option. To determine the appropriate coverage thresholds required for the inference, the KMC histogram file is utilized as input for the cutoff option in Smudgeplot[20] (v0.2.4). Subsequently, we generated a smudge plot using the coverage of the identified k-mer pairs to determine ploidy.

Species	Length (bp)	GenBank accession	TaxIDs	Custom TaxIDs
M. graminicola[77]	19589	NC_056772	189291	4890 (Ascomycota)
M. arenaria[78]	17580	NC_026554	6304	6340 (Annelida)
M. enterolobii[47]	17053	NC_026555	390850	390850
M. javanica[79]	18291	NC_026556	6303	10190 (Rotifera)
M. incognita[76]	17662	NC_02409	6306	6656 (Arthropoda)
M. chitwoodi[76]	18201	KJ476150	59747	6447 (Mollusca)
M. oryzae[80]	17066	MK507908	325757	7711 (Chordata)

Table 2. Mitochondrial sequence data statistics for different nematodes and their correspondence with the modified BlobTools analyses. Seven *Meloidogyne* mitochondrial sequences have been analyzed for this study. For each species, a specific custom TaxID corresponding to a different phylum has been used.

To estimate genome size and heterozygosity prior to assembly, we used GenomeScope[20] (v2.0) on the histogram file generated from Jellyfish[38] (v2.3.0, jellyfish histo -h 1000000) and ploidy setting as well as the coverage thresholds produced by the Smudgeplot[20] cutoff tool. The total genome size is therefore obtained by multiplying the estimated haploid genome size by the estimated ploidy level.

Genome assembly, estimation of completeness and contamination. DeepConsensus[35] reads were trimmed of remaining adapters using HifiAdapterFilt[39] (v2.0.0) with default parameters. The trimmed reads were then used as input to the Peregrine-2021 assembler[40,41] (v0.4.11) while increasing the default number of best overlaps for each initial graph (parameter–bestn 8). We chose this parameter, optimized for highly heterozygous genomes, aiming at representing all the subgenomes in the assembly.

To further assess genome assembly completeness in a reference-free approach, we used Merqury's algorithm[42] (v1.3). This tool uses k-mer frequencies to evaluate a genome's base accuracy and completeness. This is achieved by counting and comparing the distribution of canonical 21-mers found in the assembled genome with those detected in the high-accuracy DeepConsensus[35] read set. Merqury's k-mer[42] analysis will therefore indicate how much the genome assembly has captured the information present in the HiFi reads.

The screening of the contig assembly for potential contaminants by non-nematode sequences was done with the BlobTools[28] standard pipeline (v3.2.6). DeepConsensus[35] polished long-reads were aligned to the contigs with Minimap2[43] (v2.24) and the map-hifi parameter. Each contig was then assigned to a taxonomic group based on the BLAST[44] (v2.13.0+) analysis results against the NCBI nucleotide (nt) database[45]. Particular attention was paid to contigs of non-nematode taxa or contigs with a GC percentage deviating from the average GC content (around 30%[18]) of the *M. enterolobii* population E1834 to detect possible contamination. A total of 39 contigs spanning ~2.5 Mb were discarded from the assembly. The resulting assembly of 556 contigs is used for downstream analyses.

Mitochondrion assembly and functional annotation. The circular mitochondrial genome sequence was reconstituted using the ALADIN[46] package (v1.1) and DeepConsensus[35] HiFi reads in input with default parameters. We employed as a reference seed sequence the complete mitochondrion of *M. enterolobii* previously downloaded from the GenBank database (GenBank accession: NC_026555.1[47]). The annotation was carried out using GeSeq.[48], encompassing both the tRNA, the rRNA and the protein-coding genes. We set the minimum threshold of 85% for the protein and non-coding DNA search identity, and we used seven *Meloidogyne* mitochondrial genomes as third-party references (Table 2). The tRNAs prediction was also performed using third-party predictors, such as tRNAscan-SE[49] (v2.0.7), ARAGORN[50] (v1.2.38), and ARWEN[51] (v1.2.3), with codon usage corresponding to Metazoan and Invertebrate Mitochondrial.

Gene prediction and genome structure determination. Gene models prediction was done with the fully automated pipeline EuGene-EP[52] (v1.6.5). EuGene has been configured to integrate similarities with known proteins of *Caenorhabditis elegans* (PRJNA13758) from WormBase Parasite[53] and "nematoda" section of UniProtKB/Swiss-Prot library[54], with the prior exclusion of proteins that were similar to those present in RepBase[55]. A dataset composed of both *M. incognita* and *M. enterolobii de novo* assembled transcriptomes[23] was aligned on the genome and used by EuGene as transcriptional evidence. Only the alignments of datasets on the genome spanning 30% of the transcript length with at least 97% identity were retained. The EuGene default configuration was edited to set the "preserve" parameter to 1 for all datasets, the "gmap_intron_filter" parameter to 1 and the minimum intron length to 35 bp. Finally, the Nematodes-specific Weight Array Method matrices were used to score the splice sites (available at this URL: http://eugene.toulouse.inrae.fr/WAM/).

Genome structure analysis was conducted using MCScanX[56], with default settings. First, the whole proteome of the *M. enterolobii* population E1834, predicted by EuGene, was self-blasted with an E-value cutoff of 1e-25, a maximum of 5 aligned sequences, and maximum 1 high-scoring pair (hsp). Subsequently, we used gene location information extracted from the GFF3 annotation file of EuGene, along with homology information based on the all-versus-all BLASTP analysis, to identify and categorize each duplicated protein-coding gene into one of five groups using the duplicate_gene_classifier program implemented in the MCScanX[56] package. These groups are: singleton, proximal, tandem, whole-genome or segmental duplications (WGD), and dispersed duplications. Singleton refers to cases where no duplicates are found in the assembly. Proximal duplicates refer to gene duplications that are on the same contig and separated by 1 to 10 genes. Tandem duplicates, on the other hand, are

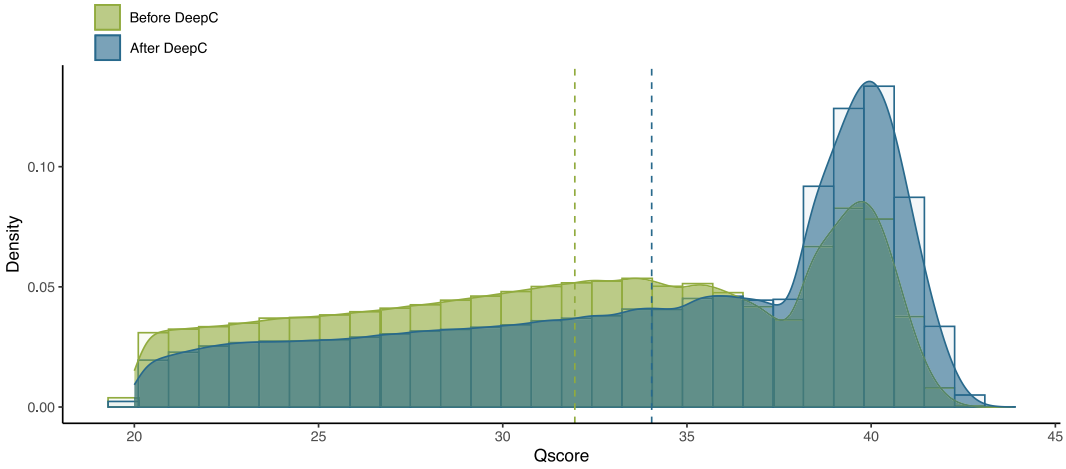

Fig. 1 Distribution of raw PacBio HiFi reads before and after DeepConsensus treatment. Comparison of the Concordance Qscore before and after DeepConsensus. The average phred-scale read accuracy score has increased by two points after treatment.

consecutive. Segmental/WGD are identified when they form collinear blocks with other pairs of duplicated genes. Finally, dispersed duplicates are those that cannot be assigned to any of the above-mentioned categories.

Species verification and validation. In the following step, we further screened the genomic reads for potential contamination, this time by other RKN sequences. BlobTools[28] allows the identification of potential contamination in genome assemblies, but, by default, only at distant taxonomic levels between different phyla (e.g., Chordata, Nematoda, Arthropoda, …). Therefore, although contamination can be detected and cleaned at this level, it remains undetectable at the intra-genus level (e.g. within *Meloidogyne*). To allow the detection of contamination by other closely related nematodes at the reads level, we adapted the BlobTools pipeline to work with mitochondrial genomes. The polished long-reads were aligned against complete mitochondrial sequences for seven *Meloidogyne* species downloaded from the NCBI database (Table 2), using the same procedure as above. Since BlobTools works by default at the phylum and not species rank, we used a script to create an additional hits file and assign a custom NCBI phylum TaxID to each species. The seven *Meloidogyne* samples have been then temporarily assigned to a different phylum for the BlobPlot visualization purpose only (Table 2).

Species-specific SCAR (sequence characterized amplified region) markers are routinely used to confirm species identity in plant-parasitic nematodes[57]. SCAR markers are locus-specific fragments of DNA that are amplified by PCR using specific 15–30 bp primers. In this study, we retrieved primer sequences of species-specific SCAR markers from the literature (Supplementary Table 1) for four *Meloidogyne* species with genome assemblies publicly available and belonging to the same clade (*M. arenaria*, *M. incognita*, *M. javanica* and *M. enterolobii*). We aligned all the primers to all the above-mentioned genomes with BLAST and when the primer pairs matched the same contig, we retrieved from the genome the 'virtual' PCR products. After verification of consistency with the lengths from the literature, the virtual PCR products were then aligned to the two previous and present versions of *M. enterolobii* genome assemblies with an E-value threshold of 1e-25.

Data Records

The PacBio HiFi sequence data as well as the nuclear and mitochondrial genome assemblies and gene predictions supporting the results of this paper have been deposited and are publicly available at the EMBL-EBI's European Nucleotide Archive (ENA) under accession number PRJEB69523 (https://www.ebi.ac.uk/ena/browser/view/PRJEB69523)[58–60]. All the scripts[61], and processed data, including genome assemblies[62], gene predictions[63], OrthoFinder analysis[64] and all the structural annotation[65] results have been deposited and are publicly available at the Recherche Data Gouv institutional collection[61–65] (https://entrepot.recherche.data.gouv.fr/dataverse/Ment-HiFI-E1834).

Technical Validation

Assessing read accuracy. After implementing the DeepConsensus[35] sequence transformer procedure, statistical analysis showed an increase in the number of high-quality reads obtained (Fig. 1, Table 3) with long fragment DNA reads of up to 26 kb in length. The average length of the reads is around 11 kb with a total number of 2.4 million reads, and a higher average Phred-scale read accuracy score (Qconcordance), which increased from 31.95 before to 34 after DeepConsensus. This transformer has elevated the PacBio HiFi read yield to a minimum Q30 by 10% and a minimum Q40 score (99.99% read accuracy) by 70%. Furthermore, we have retrieved 198,880 long reads that were initially dismissed prior to treatment in the filter, yielding more coverage of the *M. enterolobii* genome.

Profiling genome ploidy level, heterozygosity, and size. Prior to assembling a genome, it is crucial to evaluate its ploidy and size. Even though previous versions of *M. enterolobii* genomes suggested a triploid

	Before DeepConsensus	After DeepConsensus
Longest read	26 302 bp	26 302 bp
Mean length	11 194 bp	11 187 bp
Number of reads	2 250 199	2 449 079
Number of bases	25 442 730 700	27 277 356 063
Average Qscore	31.95	34.04

Table 3. Read statistics before and after the use of DeepConsensus sequence transformer. After DeepConsensus treatment, a higher number of reads with higher quality have been retrieved.

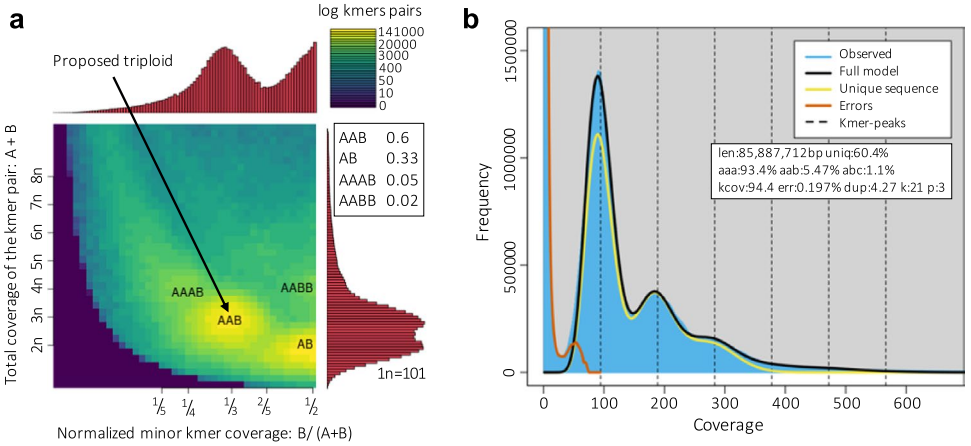

Fig. 2 Genome profiling of *M. enterolobii*. (**a**) Smudgeplot of *M. enterolobii* extracting 21-mers from DeepConsensus reads. The color intensity of each smudge reflects the approximate number of k-mers per bin. This *M. enterolobii* E1834 population is proposed as a triploid organism. (**b**) GenomeScope2 k-mer profile and estimated parameters for the triploid nematode *M. enterolobii*. Coverage (kcov), error rate (err.), haploid genome size estimation (len.), k-mer size (k) and ploidy level (p). The peak heights are proportional to the species' heterozygosity. *M. enterolobii* shows a high average heterozygosity between its three subgenomes.

structure with relatively high divergence between subgenomes[18,21,23], this was done on different populations, and we re-evaluated these features on the E1834 population. The distribution of k-mer frequencies within the DeepConsensus[35] sequencing reads allows estimating major genome features such as ploidy level, genome size, and heterozygosity rate. As GenomeScope2[20] can only precisely examine organisms when the ploidy is known, we utilized first, the results of Smudgeplot[20] (Fig. 2a) to provide GenomeScope2 with estimated ploidy level. Each smudge on the graph appears to be distinct, indicating sufficient sequencing coverage for further analysis. The most prevalent smudge corresponds to a predicted triploid AAB genome for the *M. enterolobii* population E1834 (Supplementary Table 2). This result is consistent with previous k-mer analysis performed on the short reads for the L30 population from Burkina Faso[18,21]. We remarked that the AB peak is as high as the AAB peak, suggesting an equally distributed divergence between the A1, A2 and B subgenomes. This is in contrast with the Smudge plots of *M. incognita*[22], another triploid RKN species, where the AAB peak is much higher than the AB peak, consistent with two A1, A2 subgenomes more closely related to one another than to a distant B subgenome. It should be noted here that Smudgeplot will never predict an ABC genome but always variations in the numbers of As and Bs, even in the case of triploidy with equally distant subgenomes.

Subsequently, we estimated the genome size using GenomeScope2 with a ploidy level of 3 (Fig. 2b). The genome size was determined by multiplying the estimated haploid genome length (85,887,712 bp) by the estimated ploidy level (p = 3), providing an estimated genome size of ca. 257.66 Mb.

Furthermore, the GenomeScope2 k-mer histogram of this polyploid population displays a distinct multimodal profile, with a substantial first peak located at roughly 95X, a smaller second peak at about 187X, and finally, an additional peak at 282X, typical for triploid genomes. Finally, GenomeScope2 estimated a high average heterozygosity between the subgenomes (6.6% estimated on average), consistent with a previous estimation of ca. 6.1% on the L30 population from Burkina Faso[18]. It should be noted that the term heterozygosity does not exactly apply here as we do not measure divergence between homologous chromosomes in a diploid genome but between the three subgenomes in a triploid species. Therefore, we will refer to average nucleotide divergence between subgenomes in the rest of the manuscript.

***De novo* genome assembly.** After filtering and elimination of the contaminated and mitochondrial contigs, the resulting triploid genome of the *M. enterolobii* population E1834 was assembled in 556 contigs with a total size of 285.4 Mb. The corresponding contig N50 length is equal to 2.11 Mb, with the longest being 8.30 Mb long. The genome assembly size of 285.4 Mb is consistent with previous flow cytometry estimation of nucleus total DNA content on a population from Guadeloupe island (274.7 ± 18.52 Mb)[23]. This suggests the three subgenomes

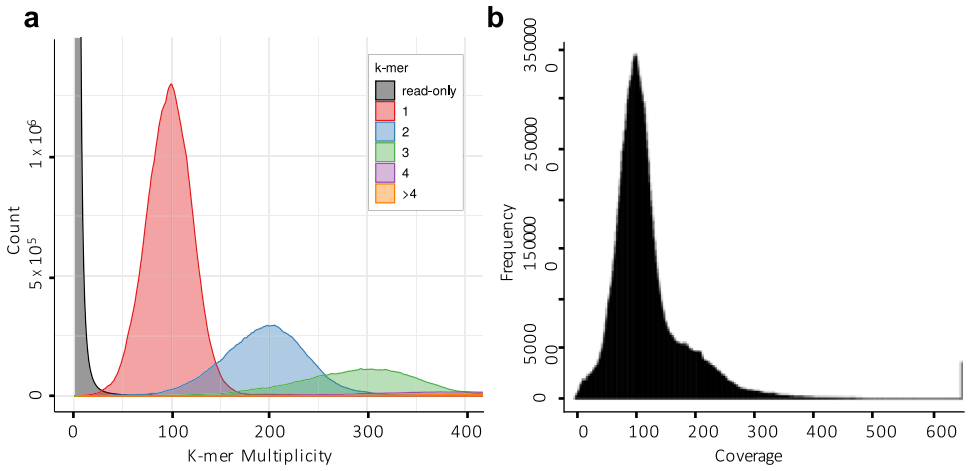

Fig. 3 Genome assembly spectra. (**A**) The Merqury spectrum plot using DeepConsensus reads tracks the multiplicity of each k-mer detected in the read set. The plot is color-coded according to the number of times a k-mer is found in an assembly. (**B**) Bedtools per-base reports coverage for the assembly.

are represented in this assembly, as opposed to diploid genomes with low heterozygosity which are usually represented as a collapsed haploid consensus.

However, the genome size estimated by analysis of k-mer distribution (257.66 Mb) is lower than the assembly size and in the lower range of the flow cytometry evaluation. A previous study showed that the accuracy of genome size estimation based on k-mer frequencies can be affected by repeats, high heterozygosity and sequencing errors[66]. This suggests that the high heterozygosity rate or repeat-richness in the *M. enterolobii* genome could have played a role in this underestimation.

To further assess genome assembly quality metrics and evaluate genome's base accuracy and completeness we used Merqury[42]. In the Merqury spectrum produced (Fig. 3a), the first and prominent 1-copy peak at a ~100X multiplicity corresponds to k-mers in the reads detected only one time in the assembly. This can be interpreted as heterozygous regions between the three subgenomes. The second peak at twice this multiplicity (~200X) corresponds to homozygous k-mers present in the reads and detected two times in the assembly. This most likely represents regions identical between two of the three subgenomes. Similarly, most of the k-mers detected 3 times in the assembly, probably represent regions identical between the three subgenomes. Conversely, the grey 'read-only' curve at low multiplicity represents rare k-mers which solely exist within the read set and are probably due to sequencing errors. This Merqury plot did not reveal missing content in the assembly as there was no subsequent grey peak neither at single coverage (~100X coverage) nor at double or triple coverage. Additionally, the three subgenomes seem to be divergent enough to have been mostly not collapsed during the assembly. This can be observed in the coverage plot provided by BEDTools[67] v2.29.0 (Fig. 3b), where the coverage depth for each base on each contig has been computed. We can clearly see a prominent peak located at roughly 101X corresponding to the haploid coverage found in the k-mers with Smudgeplot (Fig. 2a), along with a shoulder at roughly twice the haploid coverage. It seems likely that this shoulder represents a few identical regions between sub-genomes that have been collapsed during assembly. Furthermore, it should be noted that, although divergent enough, the three subgenomes have not been separated into three distinct FASTA files as this feature would require phasing and is not implemented in Peregrine[40]. Overall, the k-mer analysis with Merqury[42] indicates that all the information present in the HiFi reads has been captured in the genome assembly, thereby further suggesting a complete triploid genome assembly.

Validating species identity and purity. We confirmed the purity of the *M. enterolobii* population E1834 and the correct species identification by using, first, the standard BlobTools[28] pipeline. The pipeline generated BlobPlots, which are two-dimensional plots depicting contigs presented as circles, whose diameters are proportional to the sequence length and are colored based on their taxonomic affiliation, determined by the BLAST similarity search results against the NCBI nt database[45]. The relative positions of the circles are according to their GC content and coverage by the long reads. Following the removal of contaminant contigs, the resulting BlobPlot is shown in Fig. 4a. Any contigs lacking taxonomic annotation are labeled as 'no-hit'. For all the estimated 'non-Nematoda' contigs that fell within the estimated GC content range of *M. enterolobii*[18] (around 30%), the proposed assignments from BlobTools were disregarded and instead, a comprehensive manual verification was conducted. For each of the contigs falling into this category, we retained the highest-ranking result proposed by BLAST if the calculated percentage of identity was over 90%, the e-value did not exceed $1e^{-50}$, and the taxID belonged to the Nematoda phylum. Subsequently, eight contigs with a non-Nematoda taxonomic assignment by BLAST were retained, as no further evidence of contamination was identified upon application of the aforementioned threshold criteria. This resulted in a total of 556 contigs for the final assembly. The BlobTools pipeline is a valuable tool for detecting possible contaminations in a genome assembly, especially those originating from distant species of different phyla. However, if the contamination comes from a closely related species with a comparable GC content or has been sequenced at a similar coverage, the classical approach will not detect a

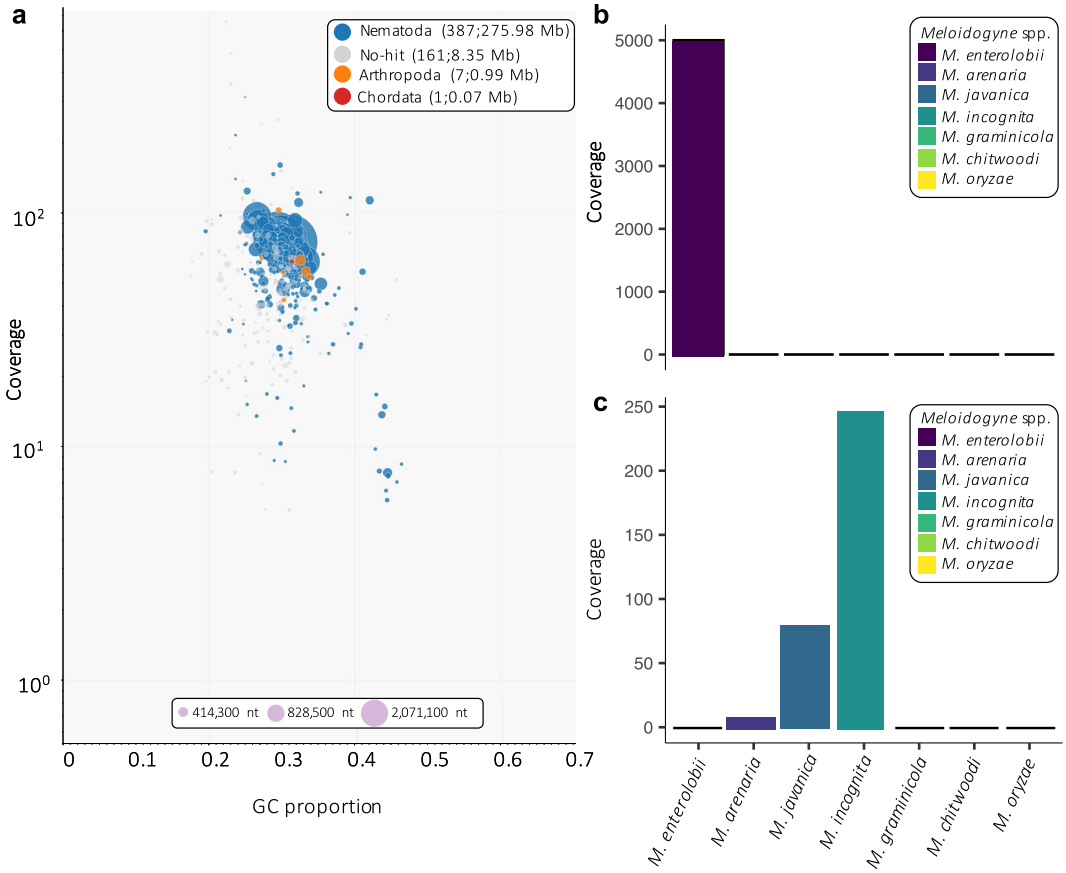

Fig. 4 BlobPlot of different *Meloidogyne* genome assemblies. (**a**) BlobPlot showing taxonomic affiliation at the phylum rank level for the E1834 population of *M. enterolobii*. After removing contamination and mitochondrion, 556 contigs were left. The average GC content for *M. enterolobii* is equal to 30 ± 0.042. (**b**) Coverage of different Meloidogyne mitochondrial genomes by the E1834 population PacBio HiFi long reads. No sign of contamination by other Meloidogyne species in the reads was identified. (**c**) Coverage of different Meloidogyne mitochondrial genomes by *M. enterolobii* PacBio RS reads from the Mma-II population[23]. This BlobPlot revealed a contamination by other *Meloidogyne* spp. and no coverage of the *M. enterolobii* mitochondrial genome.

contamination. For this reason, we made slight adjustments to the methodology (see the Species Verification and Validation Paragraph in Methods) to achieve a taxonomic classification based on different closely related species within the Nematoda phylum, instead of between phyla only (Fig. 4b,c). We focused our analysis on different species within the *Meloidogyne* genus because (i) they are difficult to differentiate based on the morphology, (ii) they live in the same environment, (iii) they have similar GC content. Therefore, a non-negligible possibility for undetected contamination exists.

Using this modified BlobTools[28] methodology, on the *M. enterolobii* population E1834 we have sequenced, we observed that the *M. enterolobii* reference mitochondrial genome from the NCBI was highly covered whereas all the other mitochondrial genomes from the other *Meloidogyne* species were not covered by our long reads. Hence, no evidence for contamination by other *Meloidogyne* species was found in the E1834 population (Fig. 4b).

For comparison, this method was applied to the previous long-read genome assembly of *M. enterolobii*[23], and surprisingly, it was found to be heavily contaminated by another *Meloidogyne* (Fig. 4c). Specifically, the *M. enterolobii* mitochondrial genome was not covered by the previous long reads while those of *M. incognita*, *M. javanica* and *M. arenaria* were all substantially covered. Approximately 60%, 30%, and 10% of the mitochondrial reads aligned with these mitochondrial genomes, respectively. Although this adjusted BlobTools approach suggested contamination of the previous Mma-II Swiss population by other RKN, this alone was not sufficient to discriminate between these three closely related species.

Consequently, we combined this approach with SCAR markers. All the pairs of primers for the SCAR marker of the four *Meloidogyne* species of interest were aligned to the previous and current assemblies of *M. enterolobii*. Both for the L30 population of Burkina Faso and the E1834 population from Puerto Rico sequenced here, the pair of primers for the *M. enterolobii* SCAR marker matched the genome assemblies with 100% identity in the correct orientation on one single contig. This allowed identification of a virtual amplified sequence of 537 bp, which is consistent with the ~520 bp estimated PCR product on the electrophoresis gel in Tigano *et al.*[30]. In contrast, neither the *M. enterolobii* SCAR primers nor the reconstructed corresponding PCR product matched the

BUSCO Categories	*M. enterolobii* (L30[81])	*M. enterolobii* (E1834) *
Complete	59.2% (151)	71.4% (182)
Single-copy	29.8% (76)	14.9% (38)
Duplicated	29.4% (75)	56.5% (144)
Fragmented	18.4% (47)	12.5% (32)
Missing	22.4% (57)	16.1% (41)

Table 4. BUSCO completeness at the genome level for *M. enterolobii* E1834 and L30 populations using lineage dataset eukaryota_odb10. *This work.

previous Mma-II genome assembly, confirming the genome was probably mostly not *M. enterolobii*. To further determine the possible source of contamination, we aligned the pairs of primers of the *M. incognita*, *M. javanica* and *M. arenaria* SCAR markers on the Mma-II genome. The *M. incognita* pair of primers matched perfectly on this previous Mma-II assembly in the correct orientation and allowed reconstructing a virtual PCR product of 1192 bp, consistent with the estimated size of the PCR product of ~1,200 bp for *M. incognita*[57]. Neither the pair of *M. incognita* primers nor the reconstructed PCR product matched the L30 or E1834 genome assemblies, and none of the *M. javanica* or *M. arenaria* pairs of primers matched any of the previously published or current *M. enterolobii* genomes.

Therefore, we can conclude that although no trace of contamination by closely related *Meloidogyne* species could be identified in the L30 or E1834 genome, there is clear evidence that the Mma-II population had been heavily contaminated by *M. incognita*.

The combination of SCAR marker analysis and a modification of BlobTools, specifically for 'mitochondrion', has resulted in a powerful tool for the examination and the verification of species purity.

Considering this detection of contamination in the Mma-II population, we further examined each sequencing library that had been produced in the corresponding BioProject (PRJEB36431). We found that apart from the Illumina mate-pair reads in which no contamination was detected, all the other datasets (genomic and transcriptomic) were contaminated at various degree by *M. incognita* (Supplementary File 1).

Consequently, we contacted the EBI's ENA, the NCBI's SRA and WormBase Parasite to ask them to remove all the contaminated data. In the rest of the manuscript, we also provide no further comparison with the contaminated Mma-II population as it is not representative of a *M. enterolobii* genome.

Genome completeness assessment. To evaluate the completeness of our genome assembly in terms of expected gene content among related species, we benchmarked nearly universal single-copy orthologs (BUSCO[68] v5.2.2) by using the eukaryote_odb10 lineage dataset in fast mode. Despite the presence of a nematode dataset in BUSCO, it only contains seven species and none of them belong to the same clade as the RKN. Therefore, we decided to use the more comprehensive Eukaryotic dataset, which encompasses 70 species. This procedure generates a report that indicates the number of nearly universal genes that are found within the assembly and classifies them into several groups: complete, fragmented, single-copy, or duplicated. The results show that 71.4% (182/255) of BUSCO genes are complete and 12.5% are fragmented. This is a substantial improvement compared to the genome assembly of the L30 *M. enterolobii* population from Burkina Faso that reached eukaryotic BUSCO completeness score of 59.2% (Table 4). This eukaryotic BUSCO completeness score of 71.4% is comparable to that obtained for *M. javanica* (69.5%) recently assembled from combination of PacBio HiFi, Nanopore and Hi-C data[69].

BUSCO is a valuable and robust tool for assessing completeness in a genome assembly in terms of a widely conserved gene set. Nevertheless, in the case of less studied species, the analysis may lack precision if the newly assembled genome comprises variations not included in the initial BUSCO gene set, such as true copy number or sequence variants[42]. We then used Merqury[42] once more to identify any copy-number errors and measure completeness and base accuracy via k-mers. Consequently, Merqury determined the proportion of reliable k-mers in the sequencing sample that were detected in the assembly, resulting in a completeness score of 99.60%. To establish Merqury's base accuracy score, a binomial model for k-mer survival was employed, resulting in a Qscore of 65.70. Higher Qscores indicate a more precise consensus. For instance, Q30 corresponds to an accuracy of 99.9%, Q40 to 99.99%, and so on. In contrast, the Burkina Faso isolate has a Qscore of 55.12 based on its own reads.

Gene prediction completeness and accuracy. Using the automated Eugene-EP[52] pipeline, a total of 49,870 genes were predicted, with 45,924 being protein-coding genes and 3,946 being non-protein-coding genes such as rRNA, tRNA, and splice leader genes. These genes cover 84 Mb (approximately 29.48%) of the genome assembly length, with the exons spanning 44.51 Mb (around 15.60%). On average, 5.26 exons were predicted per gene, and the gene length varies from a minimum of 150 bp to a maximum of 35,976 bp. The mean GC content is higher in both the protein-coding region (35.19%) or in the non-protein-coding gene regions (44.19%) compared to that of the whole genome (30.34%).

We used BUSCO[68] in proteome mode with the same eukaryota odb10 dataset to compare the completeness of the predicted proteome of *M. enterolobii* E1834 population and of the previously published L30 population (Table 5). Compared to the previously available L30 proteome, the overall completeness score was substantially improved, progressing from 68.2% to 83.5%. We also note that the proportion of duplicated complete BUSCO

BUSCO Categories	*M. enterolobii* (L30[81])	*M. enterolobii* (E1834)*
Complete	68.2% (174)	83.5% (213)
Single-copy	40.4% (103)	25.1% (64)
Duplicated	27.8% (71)	58.4% (149)
Fragmented	16.9% (43)	8.6% (22)
Missing	14.9% (38)	7.9% (20)

Table 5. BUSCO completeness at the proteome level for *M. enterolobii* E1834 and L30 populations using lineage dataset eukaryota_odb10. *This work.

Duplication depth	0	1	2	3	4	5+
Gene numbers	1992	4917	34720	2700	769	826
Percentage (%)	4.34	10.71	75.60	5.88	1.67	1.80

Table 6. Duplicate gene classifier program of MCScanX for a self-comparison of *M. enterolobii*. Genes with a duplication depth of 0 are not duplicated, while a depth of 1 indicates a maximum of one copy, a depth of 2 indicates two copies, and so forth.

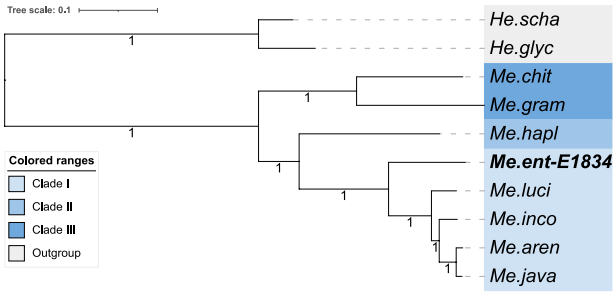

Fig. 5 Maximum likelihood phylogenetic tree produced by OrthoFinder from the concatenation of 2980 low-copy number and highly conserved orthogroups at the protein level. Number at branches represent support values of the groups.

genes increased from 27.8% in L30 to 58.4% in E1834, suggesting that a substantial portion of gene copies expected from a triploid genome were missing in the previous short-read based genome.

Finally, although we carefully checked for absence of contamination in the E1834 *M. enterolobii* genome, we wanted to verify that this absence of contamination was reflected at the predicted proteome level as well. Previous usage of the predicted proteome from the contaminated Mma-II genome in an OrthoFinder[70] comparative analysis with other nematode proteomes yielded an incorrect phylogenetic position as compared to the expected position for *M. enterolobii*[71]. Indeed, due to the high contamination level by *M. incognita*, the Mma-II population, was closely related to *M. incognita* and *M. floridensis*, instead of holding an outgroup position relative to the other tropical RKN as expected for *M. enterolobii*[72]. To check whether the new E1834 *M. enterolobii* predicted proteome solved this problem, we conducted an OrthoFinder analysis, including the predicted proteomes of seven other RKN as well as two cyst nematodes as outgroup species (Supplementary Table 3). The resulting phylogenetic tree built from multiple sequence alignment with MAFFT[73] and maximum likelihood phylogeny with FastTree[74] positioned with high support the *M. enterolobii* E1834 population as an outgroup to the rest of the tropical RKN (Clade I), exactly as expected for this RKN species (Fig. 5). The whole OrthoFinder analysis is available at (https://doi.org/10.57745/KGA7CI).

Confirmation of genome structure and ploidy level. The *M. enterolobii* population E1834 genome has been predicted to be triploid based on k-mer analyses and Smudgeplot[20]. Therefore, we further explored the genome structure and ploidy in the light of the annotation. The use of MCScanX[56] revealed that a majority of gene duplicates belong to whole genome duplication blocks, rather than dispersed independent duplications. Following the classification established by the duplicate_gene_classifier program implemented in the MCScanX package, 39,532 of the protein-coding genes (around 86.10%) are predicted to be duplicated at least once. As shown in Table 6, a majority of these coding genes (75.6%) display a duplication depth of two (meaning for these genes, two other copies exist), further reinforcing the idea that the genome is triploid. Furthermore, it was found that 69.76% of the protein-coding genes fall under the whole-genome duplication category of MCScanX, forming 516 syntenic blocks of collinear genes (see Fig. 6 for visualization of multiple syntenic blocks between different contigs). Besides, 12.61% of the genes are classified as dispersed duplicates, while 2.18% and 1.53% constitute proximal and tandem duplicates, respectively. These findings strongly suggest that the genome of *M. enterolobii* is triploid, confirming Smudgeplot results.

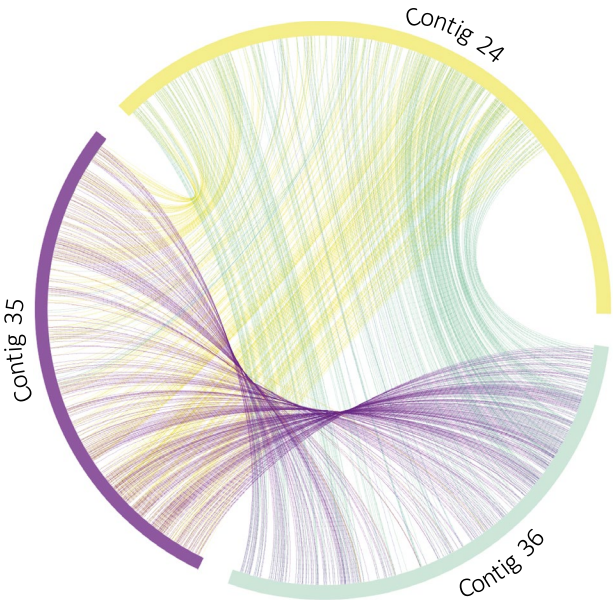

Fig. 6 *M. enterolobii* exhibits a triploid genome. The circle plot produced by MCScanX shows collinear gene pairs forming homologous duplicated regions between three contigs. All the collinear gene pairs are linked with different curved colored lines between and within each contig.

Considering the genome assembly size, which is in the range of the measured total nuclear DNA content, the k-mer analyses as well as the genome duplicated structure analysis, the *M. enterolobii* E1834 genome assembly seems to be representative of the three subgenomes in this triploid species. This situation is reminiscent of that observed in *M. incognita*, another triploid root-knot nematode[22]. In *M. incognita*, the AAB and AB smudge plots formed clearly distinct peaks with higher peak for the AAB peak, suggesting two relatively close AA subgenomes and a more distant B subgenome. An analysis of the median rates of synonymous substitutions between genes present in triplets of duplicated contigs, confirmed the k-mer results in *M. incognita*. Indeed, in triplets of contigs there were a two-peaks distribution in median Ks values, with one relatively low value (0.05) representing the relation between the two A subgenomes, and two relatively and equally high values (0.14) representing the relation between the B and each A subgenome.

Based on k-mer analyses, the situation in *M. enterolobii* E1834 seems to be different with equally high peaks for the AAB and AB smudge plots, suggesting a more symmetrically distributed divergence between the three subgenomes (Fig. 2). To further investigate the genome structure in the light of the MCScanX analysis, we performed the same analysis of median Ks values between triplets of contigs in *M. enterolobii*. As opposed to *M. incognita*, we did not observe a distribution of median Ks values with two clearly distinct peaks. Instead, the distribution formed a single higher peak with an overall median Ks value of 0.09. This distribution is consistent with the smudge plots k-mer analyses and suggests a triploid genome with three equally diverged (AAA or ABC) subgenomes. An illustration of this situation on two different triplets of big contigs is available as supplementary Figure 1.

Mitochondrial genome assembly and annotation. Using the Aladin package[46], the mitochondrial genome of the *M. enterolobii* population E1834 has been assembled and spanned a length of 19,193 bp with a GC content of 17.2% (Fig. 7). We have retrieved and annotated all the mitochondrially encoded genes involved in the mitochondrial respiratory chain (MRC), including the seven core subunits of the complex I, the cytochrome b of the complex III, the three cytochrome c, and the ATP synthase. Additionally, a comprehensive set of tRNA involved in amino acid synthesis was obtained, with some tRNAs present in multiple copies (e.g., trnA for Alanine, trnS for Serine, trnL for Leucine, and trnN for Asparagine). Furthermore, ribosomal RNAs (rrnS, rrn12, and rrn16) were also identified.

When blasted against the NCBI nt database, the reconstructed mitochondrial genome of the *M. enterolobii* population E1834 returned as first hit the complete mitochondrial reference genome of *M. enterolobii*[47], with 99.53% identity and an alignment length of 13,067 bp, as the primary hsp. The second-best hit corresponds to an incomplete mitochondrial genome from an *M. enterolobii* isolate discovered on sweet potatoes in the states of Carolina in the USA[75] (GenBank: MW246173.1).

In contrast, reconstruction of the mitochondrial genome using Aladin[46] on the Swiss population Mma-II yielded a ~23 kb genome which returned as first hit the *M. incognita* reference mitochondrial genome[76] with $>$ 99% identity covering $>$ 97% of the query while the *M. enterolobii* reference mitochondrial genome only emerged as the fifth hit with only 87% identity covering 78% of the length.

These results further confirm the E1834 population we have sequenced is indeed *M. enterolobii*.

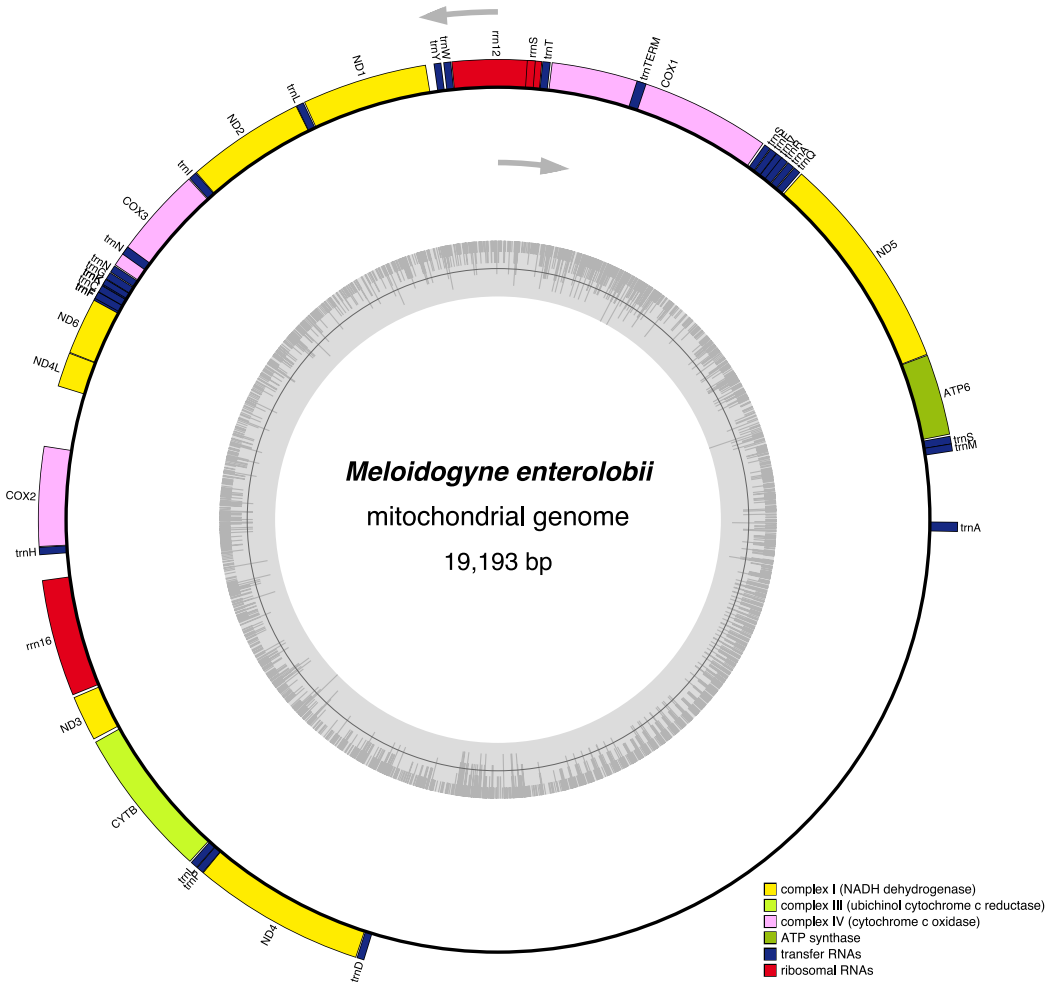

Fig. 7 Mitochondrial genome organization of *M. enterolobii*. The inner circle displays the GC content while grey arrows denote the transcription direction. The rRNAs and tRNAs are respectively colored in red and blue. The various complexes of the MRC are represented in yellow, light green, pink, and dark green.

Code availability

The codes used to run the different tools listed in the methods have been deposited and are publicly available at the Recherche Data Gouv institutional collection[61]: https://doi.org/10.57745/EGUUHK.

Received: 9 April 2024; Accepted: 7 January 2025;
Published online: 30 January 2025

References

1. Jones, J. T. *et al.* Top 10 plant-parasitic nematodes in molecular plant pathology. *Molecular Plant Pathology* **14**, 946–961 (2013).
2. Santos, D., Abrantes, I. & Maleita, C. The quarantine root-knot nematode Meloidogyne enterolobii – a potential threat to Portugal and Europe. *Plant Pathology* **68**, 1607–1615 (2019).
3. Moens, M., Perry, R. N. & Starr, J. L. Meloidogyne species - a diverse group of novel and important plant parasites. in (eds. Perry, R. N., Starr, J. L. & Moens, M.) 1–17 (CABI International, Wallingford, Oxon (CABI), 2009).
4. Kiewnick, S., Dessimoz, M. & Franck, L. Effects of the Mi-1 and the N root-knot nematode-resistance gene on infection and reproduction of Meloidogyne enterolobii on tomato and pepper cultivars. *J Nematol* **41**, 134–139 (2009).
5. Elling, A. A. Major Emerging Problems with Minor Meloidogyne Species. *Phytopathology*® **103**, 1092–1102 (2013).
6. Yang, B. & Eisenback, J. D. Meloidogyne enterolobii n. sp. (Meloidogynidae), a Root-knot Nematode Parasitizing Pacara Earpod Tree in China. *J Nematol* **15**, 381–391 (1983).
7. Rammah, A. & Hirschmann, H. Meloidogyne mayaguensis n. sp. (Meloidogynidae), a Root-knot Nematode from Puerto Rico. *J Nematol* **20**, 58–69 (1988).
8. Karssen, G., Liao, J., Kan, Z., van Heese, E. Y. & den Nijs, L. J. On the species status of the root-knot nematode Meloidogyne mayaguensis Rammah & Hirschmann, 1988. *Zookeys* 67–77, https://doi.org/10.3897/zookeys.181.2787 (2012).
9. Xu, J., Liu, P., Meng, Q. & Long, H. Characterisation of Meloidogyne species from China using Isozyme Phenotypes and Amplified Mitochondrial DNA Restriction Fragment Length Polymorphism. *European Journal of Plant Pathology* **110**, 309–315 (2004).
10. Sikandar, A., Jia, L., Wu, H. & Yang, S. Meloidogyne enterolobii risk to agriculture, its present status and future prospective for management. *Front. Plant Sci.* **13**, 1093657 (2023).
11. Castagnone-Sereno, P. Meloidogyne enterolobii (= M. mayaguensis): profile of an emerging, highly pathogenic, root-knot nematode species. *Nematology* **14**, 133–138 (2012).

12. Castillo, P. & Castagnone-Sereno, P. Meloidogyne enterolobii (Pacara earpod tree root-knot nematode). *CABI Compendium* **CABI Compendium**, 33238 (2020).
13. Schwarz, T., Li, C., Ye, W. & Davis, E. Distribution of Meloidogyne enterolobii in Eastern North Carolina and Comparison of Four Isolates. *Plant Health Progress* **21**, 91–96 (2020).
14. Philbrick, A. N., Adhikari, T. B., Louws, F. J. & Gorny, A. M. Meloidogyne enterolobii, a Major Threat to Tomato Production: Current Status and Future Prospects for Its Management. *Frontiers in Plant Science* **11** (2020).
15. THE EUROPEAN COMMISSION. COMMISSION IMPLEMENTING REGULATION (EU) 2021/2285 of 14 December 2021 amending Implementing Regulation (EU) 2019/2072 as regards the listing of pests, prohibitions and requirements for the introduction into, and movement within, the Union of plants, plant products and other objects, and repealing Decisions 98/109/EC and 2002/757/EC and Implementing Regulations (EU) 2020/885 and (EU) 2020/1292. Official Journal of the European Union L 458:173-283 https://eur-lex.europa.eu/eli/reg_impl/2021/2285/oj.
16. Blok, V. C. & Powers, T. O. Biochemical and molecular identification. *Root-knot nematodes* 98–118, https://doi.org/10.1079/9781845934927.0098 (2009).
17. Min, Y. Y., Toyota, K. & Sato, E. A novel nematode diagnostic method using the direct quantification of major plant-parasitic nematodes in soil by real-time PCR. *Nematology* **14**, 265–276 (2012).
18. Szitenberg, A. *et al*. Comparative Genomics of Apomictic Root-Knot Nematodes: Hybridization, Ploidy, and Dynamic Genome Change. *Genome Biol Evol* **9**, 2844–2861 (2017).
19. Blanc-Mathieu, R. *et al*. Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes. *PLOS Genetics* **13**, e1006777 (2017).
20. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**, 1432 (2020).
21. Jaron, K. S. *et al*. Genomic Features of Parthenogenetic Animals. *Journal of Heredity* **112**, 19–33 (2021).
22. Mota, A. P. Z. *et al*. Unzipped genome assemblies of polyploid root-knot nematodes reveal unusual and clade-specific telomeric repeats. *Nat Commun* **15**, 773 (2024).
23. Koutsovoulos, G. D. *et al*. Genome assembly and annotation of Meloidogyne enterolobii, an emerging parthenogenetic root-knot nematode. *Sci Data* **7**, 324 (2020).
24. Kiewnick, S., Karssen, G., Brito, J. A., Oggenfuss, M. & Frey, J.-E. First Report of Root-Knot Nematode Meloidogyne enterolobii on Tomato and Cucumber in Switzerland. *Plant Disease* **92**, 1370–1370 (2008).
25. Marx, V. Method of the year: long-read sequencing. *Nat Methods* **20**, 6–11 (2023).
26. Dai, D. *et al*. Unzipped chromosome-level genomes reveal allopolyploid nematode origin pattern as unreduced gamete hybridization. *Nat Commun* **14**, 7156 (2023).
27. Gerič Stare, B., Strajnar, P., Susič, N., Urek, G. & Širca, S. Reported populations of Meloidogyne ethiopica in Europe identified as Meloidogyne luci. *Plant Disease* **101**, 1627–1632 (2017).
28. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. Preprint at https://doi.org/10.12688/f1000research.12232.1 (2017).
29. Kiewnick, S., Frey, J. E. & Braun-Kiewnick, A. Development and Validation of LNA-Based Quantitative Real-Time PCR Assays for Detection and Identification of the Root-Knot Nematode Meloidogyne enterolobii in Complex DNA Backgrounds. *Phytopathology®* **105**, 1245–1249 (2015).
30. Tigano, M. *et al*. Genetic diversity of the root-knot nematode Meloidogyne enterolobii and development of a SCAR marker for this guava-damaging species. *Plant Pathology* **59**, 1054–1061 (2010).
31. Ye, W., Zeng, Y. & Kerns, J. Molecular Characterisation and Diagnosis of Root-Knot Nematodes (Meloidogyne spp.) from Turfgrasses in North Carolina, USA. *PLOS ONE* **10**, e0143556 (2015).
32. Gómez-González, G. *et al*. Meloidogyne enterolobii egg extraction in NaOCl versus infectivity of inoculum on cucumber. *J Nematol* **53**, e2021–57 (2021).
33. Jenkins, W. R. A rapid centrifugal-flotation technique for separating nematodes from soil. *Plant Dis. Rep.* **48**, 692 (1964).
34. Töpfter, A. PacificBiosciences/pbbioconda: PacBio Secondary Analysis Tools on Bioconda. *Github* https://identifiers.org/github:PacificBiosciences/pbbioconda (2022)
35. Baid, G. *et al*. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat Biotechnol* **41**, 232–238 (2023).
36. Töpfter, A. PacificBiosciences/ACTC: Align subreads to CCS reads. *Github* https://identifiers.org/github:PacificBiosciences/actc (2022).
37. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
38. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
39. Sim, S. B., Corpuz, R. L., Simmonds, T. J. & Geib, S. M. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics* **23**, 157 (2022).
40. Chin, J. Peregrine-2021: A faster and minimum genome assembler. (2023).
41. Chin, C.-S. & Khalak, A. Human Genome Assembly in 100 Minutes. 705616 Preprint at https://doi.org/10.1101/705616 (2019).
42. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245 (2020).
43. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
44. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* **32**, W20–W25 (2004).
45. Sayers, E. W. *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **50**, D20–D26 (2021).
46. Koutsovoulos, G. GDKO/Aladin: Aladin (mitochondrial circular DNA reconstitution). *Github* https://identifiers.org/github:GDKO/aladin (2021).
47. Humphreys-Pereira, D. A. & Elling, A. A. *Genbank* https://identifiers.org/nucleotide:NC_026555.1 (2015).
48. Tillich, M. *et al*. GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Research* **45**, W6–W11 (2017).
49. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol Biol* **1962**, 1–14 (2019).
50. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research* **32**, 11–16 (2004).
51. Laslett, D. & Canbäck, B. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. *Bioinformatics* **24**, 172–175 (2008).
52. Sallet, E., Gouzy, J. & Schiex, T. EuGene: An Automated Integrative Gene Finder for Eukaryotes and Prokaryotes. in *Gene Prediction: Methods and Protocols* (ed. Kollmar, M.) 97–120, https://doi.org/10.1007/978-1-4939-9173-0_6 (Springer, New York, NY, 2019).
53. Howe, K. L., Bolt, B. J., Shafie, M., Kersey, P. & Berriman, M. WormBase ParaSite − a comprehensive resource for helminth genomics. *Mol Biochem Parasitol* **215**, 2–10 (2017).
54. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **46**, 2699 (2018).
55. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).

56. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**, e49 (2012).
57. Zijlstra, C., Donkers-Venne, D. T. H. M. & Fargette, M. Identification of Meloidogyne incognita, M. javanica and M. arenaria using sequence characterised amplified region (SCAR) based PCR assays. *Nematology* **2**, 847–853 (2000).
58. *European Nucleotide Archive* https://identifiers.org/ena.embl:PRJEB69523 (2024).
59. *European Nucleotide Archive*. https://identifiers.org/insdc.sra:ERP154457 (2024)
60. Poullet, M. *Genbank*. https://identifiers.org/insdc.gca:GCA_963681835.1 (2024)
61. Poullet, M. Meloidogyne enterolobii scripts. *Recherche Data Gouv* https://doi.org/10.57745/EGUUHK.
62. Poullet, M. & Danchin, E. G. J. Meloidogyne enterolobii assemblies. *Recherche Data Gouv* https://doi.org/10.57745/5MXZSJ (2023).
63. Poullet, M. & Danchin, E. G. J. & Rancurel C. Meloidogyne enterolobii E1834 gene prediction. Recherche Data Gouv https://doi.org/10.57745/Y0O2LP (2023).
64. Danchin, E. Meloidogyne genus OrthoFinder analysis. *Recherche Data Gouv* https://doi.org/10.57745/KGA7CI (2024).
65. Poullet, M. & Danchin, E. G. J. Meloidogyne enterolobii E1834 structural annotation. Recherche Data Gouv https://doi.org/10.57745/VEKHQS (2023).
66. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. Preprint at https://doi.org/10.48550/arXiv.1308.2012 (2020).
67. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
68. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **38**, 4647–4654 (2021).
69. Winter, M. R. *et al*. Phased chromosome-scale genome assembly of an asexual, allopolyploid root-knot nematode reveals complex subgenomic structure. *PLOS ONE* **19**, e0302506 (2024).
70. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238 (2019).
71. Grynberg, P. *et al*. Comparative Genomics Reveals Novel Target Genes towards Specific Control of Plant-Parasitic Nematodes. *Genes* **11**, 1347 (2020).
72. Álvarez-Ortega, S., Brito, J. A. & Subbotin, S. A. Multigene phylogeny of root-knot nematodes and molecular characterization of Meloidogyne nataliei Golden, Rose & Bird, 1981 (Nematoda: Tylenchida). *Sci Rep* **9**, 11788 (2019).
73. Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490–2492 (2018).
74. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).
75. Rutter, W. B. Phillip, W. A., Mueller, J. D. & Paula, A. *Genbank* https://identifiers.org/nucleotide:MW246173.1 (2020)
76. Humphreys-Pereira, D. A. & Elling, A. A. Mitochondrial genomes of Meloidogyne chitwoodi and M. incognita (Nematoda: Tylenchina): Comparative analysis, gene order and phylogenetic relationships with other nematodes. *Molecular and Biochemical Parasitology* **194**, 20–32 (2014).
77. Sun, L., Zhuo, K., Lin, B., Wang, H., & Liao, J. *Genbank* https://identifiers.org/nucleotide:NC_056772.1 (2014).
78. Humphreys-Pereira, D. A. & Elling, A. A. *Genbank* https://identifiers.org/nucleotide:NC_026554.1 (2015).
79. Humphreys-Pereira, D. A. & Elling, A. A. *Genbank* https://identifiers.org/nucleotide:NC_026556.1 (2015).
80. Besnard, G. *et al*. On the close relatedness of two rice-parasitic root-knot nematode species and the recent expansion of Meloidogyne graminicola in Southeast Asia. *Genes* **10**(2), 175, https://doi.org/10.3390/genes10020175 (2019).
81. Lunt, D. H. Genetic tests of ancient asexuality in Root Knot Nematodes reveal recent hybrid origins. *BMC Evol Biol* **8**, 194 (2008).

Acknowledgements

We are grateful to the colleagues from the Netherlands Institute for Vectors, Invasive Plants and Plant Health (NIVIP) for providing the *M. enterolobii* population E1834 for this study. We thank the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) for providing computing resources. We are grateful to the bioinformatics and genomics platform, BIG, Sophia Antipolis (ISC plantBIOs, https://doi.org/10.15454/qyey-ar89) for computing and storage resources. We thank Claire Caravel for help in providing the primer sequences for SCAR identification of *Meloidogyne* species. We would like to thank Kamil Jaron for his help in interpreting the smudge plots of k-mers relative to genome structure. Our *M. enterolobii* genome research was financially supported by a Franco-German bilateral grant ANR-DFG "AEGONE", reference ANR-19-CE35-0017 and reference No 431627824.

Author contributions

E.G.J.D. and S.K. conceived the research idea and acquired the funding. E.G.J.D. supervised all the bioinformatics analyzes, performed OrthoFinder analysis as well as SCAR marker virtual PCRs, analyzed Smudgeplot and Ks results, contributed to manuscript writing, reviewing and revision. S.K. supervised all the nematode rearing and DNA extraction experiments, contributed to manuscript writing and reviewing. M.P. performed reads processing, genome assembly, contamination and purity check, completeness assessment, ploidy and genome size and structure estimation and wrote the manuscript. H.K. generated single egg mass lines, performed maintenance of the nematode collection, DNA extraction experiments, and contributed to manuscript writing. C.R. performed gene prediction and wrote the corresponding method section. M.S., C.L.R. and J.L. performed library and PacBio HiFi sequencing and contributed to manuscript writing. A.P.Z.M. performed the analysis of Ks distribution between triplets of duplicated regions.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04434-w.

Correspondence and requests for materials should be addressed to M.P. or E.G.J.D.