SCIENTIFIC DATA

Check for updates

OPEN

DATA DESCRIPTOR

Genome assembly and annotation of *Meloidogyne enterolobii*, an emerging parthenogenetic root-knot nematode

Georgios D. Koutsovoulos[1], Marine Poullet[1], Abdelnaser Elashry[2], Djampa K. L. Kozlowski[1], Erika Sallet[3], Martine Da Rocha[1], Laetitia Perfus-Barbeoch[1], Cristina Martin-Jimenez[1], Juerg Ernst Frey[4], Christian H. Ahrens[4,5], Sebastian Kiewnick[6,7] ✉ & Etienne G. J. Danchin[1,7] ✉

Root-knot nematodes (genus *Meloidogyne*) are plant parasites causing huge economic loss in the agricultural industry and affecting severely numerous developing countries. Control methods against these plant pests are sparse, the preferred one being the deployment of plant cultivars bearing resistance genes against *Meloidogyne* species. However, *M. enterolobii* is not controlled by the resistance genes deployed in the crop plants cultivated in Europe. The recent identification of this species in Europe is thus a major concern. Here, we sequenced the genome of *M. enterolobii* using short and long-read technologies. The genome assembly spans 240 Mbp with contig N50 size of 143 kbp, enabling high-quality annotations of 59,773 coding genes, 4,068 non-coding genes, and 10,944 transposable elements (spanning 8.7% of the genome). We validated the genome size by flow cytometry and the structure, quality and completeness by bioinformatics metrics. This ensemble of resources will fuel future projects aiming at pinpointing the genome singularities, the origin, diversity, and adaptive potential of this emerging plant pest.

Background & Summary

The root-knot nematode *Meloidogyne enterolobii* (syn. *M. mayaguens*[1]) is a polyphagous species, attacking a wide range of life-sustaining and ornamental plants[2]. *M. enterolobii* is considered one of the most pathogenic and virulent root-knot nematode (RKN) species, as it is able to develop and reproduce on host plants carrying resistance to the major tropical RKN species[3–5]. *M. enterolobii* is already widespread, damaging cotton and soybean in the United States[6], causing detrimental effects for watermelon production areas in Mexico[7], and potato in Africa[8].

In 2010, *M. enterolobii* was added to the European Plant Protection Organisation (EPPO) A2 list and is now recommended for regulation as a quarantine species[9]. Based on the widespread distribution and potential to establish in the Mediterranean and subtropical regions, several countries now have designated *M. enterolobii* as a quarantine pest[10,11]. Currently, only a few potential sources of resistance such as the *Ma* gene from the Myrobalan plum (*Prunus cerasifera*)[12] or in guava (*Psidium* spp.) and pepper accessions were reported[13,14]. However, these genetic resources were only tested against a single regional *M. enterolobii* population and not the full range of isolates from various sources and host plants.

M. enterolobii is described as reproducing clonally via mitotic parthenogenesis, similarly to the most damaging RKN to worldwide agriculture (*M. incognita*, *M. javanica*, and *M. arenaria*). The absence of meiosis and sexual reproduction in these species was based on cytological observations in the 80's, where no pairing of homologous chromosomes had been observed in hundreds of isolates[15]. Recent population genomics analysis confirmed the

[1]Université Côte d'Azur, INRAE, CNRS, Institut Sophia Agrobiotech, Sophia Antipolis, France. [2]Strube Research GmbH & Co. KG, Hauptstraße 1, 38387, Söllingen, Germany. [3]Laboratoire des Interactions Plantes-Microorganismes, INRAE, CNRS, 31326, Castanet-Tolosan, France. [4]Agroscope, Molecular Diagnostics, Genomics & Bioinformatics, Wädenswil, Switzerland. [5]Swiss Institute of Bioinformatics, Bioinformatics, Wädenswil, Switzerland. [6]Julius Kühn Institute (JKI) - Federal Research Centre for Cultivated Plants Federal Research Centre for cultivated Plants, Messeweg 11/12, 38104, Braunschweig, Germany. [7]These authors jointly supervised this work: Sebastian Kiewnick, Etienne G. J. Danchin. ✉e-mail: sebastian.kiewnick@julius-kuehn.de; etienne.danchin@inrae.fr

Dataset	Mean read length	Insert Size	Number of reads	Sum (bp)
Illumina Paired-end	101	250	382,534,946	38,636,029,546
Illumina Mate-pair	51	3,000	430,206,972	21,940,555,572
PacBio	7,008	NA	1,854,399	12,996,274,301

Table 1. Genomic Sequence Data Statistics.

absence of sexual meiotic recombination in *M. incognita*[16]. The genomes of the three above-mentioned tropical RKN revealed some interesting characteristics. All three are polyploid with highly diverged genome copies most likely resulting from hybridization events[17]. This peculiar genome structure was shown to provide these species with diverged homoeologous gene copies that presented different gene expression patterns. These diverged gene copies might be involved in the extreme polyphagy of these species despite their asexual reproduction. More recently, it was shown that convergent gene copy losses were associated with the breaking down of a resistance gene in tomato by *M. incognita*[18], suggesting gene copy number variations are involved in adaptive processes. *M. enterolobii* belongs to the same Clade I[19] as the most damaging RKN species and has the same reproductive mode. However, this species is not controlled by the resistance genes deployed in plants for protection against the other RKN. Therefore, it is of great interest to explore the genome of *M. enterolobii* and determine similarities and differences between the species of the same RKN clade.

A first draft genome of *M. enterolobii* from Burkina Faso, based solely on Illumina short reads, was previously available[20]. However, the genome assembly was quite fragmentary.

Here, we used both PacBio long reads and Illumina short read sequences to study the genome of the first *M. enterolobii* strain officially reported in Europe. Compared to the previous draft genome of *M. enterolobii*, the present genome assembly is more contiguous, reducing the number of fragments by more than 10 times (from >46,000 scaffolds to <4,500 contigs) and improving the N50 length by >15x (from 9.3 kb to 143 kb).

We produced RNA-seq transcriptome data that we used as a source of evidence to predict protein-coding and non-coding genes. We also annotated transposable elements (TE) as well as other repeats, and used the annotated genome to determine the ploidy level, the degree of divergence between genome copies and TE abundance. So far, this new genome is one of the most contiguous and most complete annotated one, publicly available, for a root-knot nematode species.

Providing a robust reference genome for this species constitutes an important resource for further analyses with important evolutionary and agro-economic implications. This resource will pave the way for comparative as well as population genomics towards pinpointing the genome singularities, the origin, diversity, and adaptive potential of this emerging plant pest.

Methods

Nematode collection and DNA/RNA extraction. For genome and transcriptome sequencing, the Swiss *M. enterolobii* population was originally isolated from infected tomato root-stock obtained from an organic farm[21]. The population was maintained in a greenhouse at 25 °C and 16 h supplemental light on the cultivar Oskar F1 carrying the Mi-1 gene, conferring resistance to other Meloidogyne species (e.g. *M. incognita*, *M. javanica* and *M. arenaria*). We confirmed the purity of the population and correct species identification by species-specific PCR and DNA barcoding[11,22] both on the bulk population and on randomly picked individual juveniles obtained from separate egg masses.

Heavily infected (i.e. galled) tomato roots were collected and carefully washed free from soil to prepare the nematode material. Through incubation in a mist chamber, freshly-hatched second-stage juveniles (J2) were collected every 2 to 3 days over two weeks and used for DNA (PacBio long reads and Illumina short-read sequencing) or RNA (Illumina short read) extraction. For eggs, three galled tomato roots were extracted using the NaOCl method[23]. In order to obtain pure egg suspensions, density-gradient centrifugation was used to separate eggs from organic debris, according to Schaad *et al.*[24]. All J2 and egg suspensions were checked using a microscope and impurities were removed before DNA or RNA extraction.

After collecting eggs and J2s, samples were frozen in liquid nitrogen. Total DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen, The Netherlands), while total RNA was extracted using the Nucleospin RNA plant kit with the protocol modified for maximum yield (Macherey-Nagel, Oensingen, Switzerland). The RNA concentration was determined with a Qubit 3.0 (Thermofisher).

Genome and transcriptome sequencing. The Swiss *M. enterolobii* population long-fragment DNA libraries were sequenced with the PacBio RS technology at the FGCZ (ETH, Switzerland) sequencing center and were complemented by additional PacBio and Illumina short read data at GATC Biotech company. DNA extracted from nematode egg suspensions was used to generate the long-read genomic libraries. The libraries were subsequently size-selected with the BluePippin approach and then sequenced using the P6C4 chemistry. This procedure allowed us to obtain long DNA reads of up to 35 kb in length, an average length around 7 kb, and a total of ca. 13 Gigabases of long read sequence data (Table 1). Genomic Illumina short read data were obtained from J2 pre-parasitic juveniles. Transcriptomic data from eggs and J2, separately, were also produced in order to be used as a source of evidence for gene prediction. A total of one µg RNA for each stage (egg and J2) was used for sequencing with the Illumina MiSeq reagent kit V3 with 2 x 300 bp paired-end reads (Table 2).

Dataset	Mean read length	Insert Size (range)	Number of reads	Sum (bp)
Eggs Illumina Paired-end	196	300 (56–336)	54,424,802	10,660,505,616
J2 Illumina Paired-end	196	300 (57–347)	56,468,708	11,083,687,102

Table 2. Transcriptomic Sequence Data Statistics.

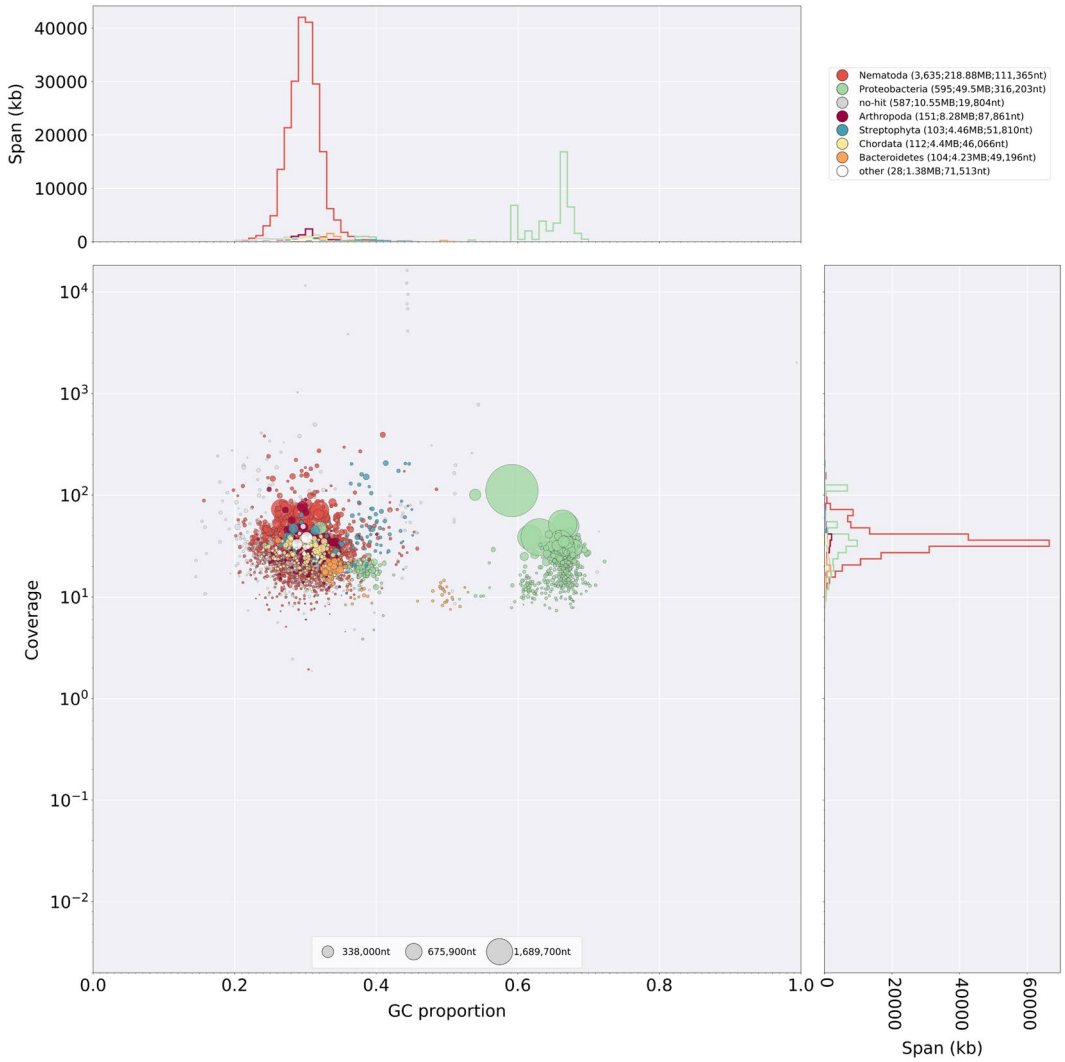

Fig. 1 BlobPlot of the genome assembly before removing contamination. Each circle is a contig proportionally scaled by contig length and coloured by taxonomic annotation based on BLAST similarity search results. Contigs are positioned based on the GC content (X-axis) and the coverage of PacBio reads (Y-axis). There are some contigs of Proteobacteria origin at high GC and variable coverage indicating possible contamination. These contigs were removed from the assembly.

Genome assembly. Pacbio reads were corrected and trimmed using sprai with default parameters (http://zombie.cb.k.u-tokyo.ac.jp/sprai/). The trimmed reads were then used as input to Canu assembler[25] for a first preliminary assembly. This assembly was then used to check for contamination with the blobtools pipeline[26,27]. Briefly, Illumina genomic reads from both this study and the previously produced *M. enterolobii* draft genome[20] were mapped to the genome with bwa[28], and each contig was given a taxonomy affiliation based on BLAST results against the NCBI nt database (Fig. 1). Contigs that had coverage only in one Illumina library, GC percentage outside of the range of the estimated *M. enterolobii* GC content, and affiliation to different taxa, were considered possible contaminants. The GC estimate was calculated to be around ~30% by taking into account the GC% of nematode contigs from the blobplot analyses, the GC% of the assembly done by Szitenberg *et al.*[20], and the GC% estimate of *Meloidogyne* species as shown in Table 3. Careful investigation of each such contig was then employed to limit the false positive rate. After, only the Pacbio reads that mapped to the clean contigs were retained, and a new assembly was created with Canu. Similarly, as described above, this assembly was checked for contamination,

Species/Feature	Assembly size (Mb)	Nuclear DNA content (Mb)	% complete CEGMA (C) (copies) %Partial (P)	% complete BUSCO (C) % fragment (F) % missing (M)	N50 (kb)	# contigs/ scaffolds	GC %
*M. enterolobii (Swiss)**	240	275 ± 19	C: 94.76 (3.30) P: 96.77	C:87.5% F:3.6%; M:8.9%	143	4,437	30
M. enterolobii (L30)[20]	162.4	NA	C: 81.45 (2.66) P: 89.92	C:79.9% F:8.9%; M:11.2%	9.3	46,090	30.2
M. incognita (Morelos, 2017)[17]	183.5	189 ± 15	C: 94.76 (2.93) P :96.77	C:88.5% F:3.0%; M:8.5%	38.6	12,091	29.8
M. incognita (W1)[20]	122	NA	C: 82.66 (2.34) P :89.52	C:80.2% F:7.9%; M:11.9%	16.5	33,735	30.6
M. incognita (Morelos, 2008)[59]	86	189 ± 15	C: 74.6 (1.77) P :78.63	C:71.3% F:7.3%; M:21.4%	82.8	2,817	31.4
M. javanica (Avignon)[17]	235.8	297 ± 27	C: 92.74 (3.68) P :95.56	C:90.1% F:2.3%; M:7.6%	10.4	31,341	30
M. javanica (VW4)[20]	142.6	NA	C :89.52 (2.71) P: 95.16	C:87.5% F:4.3%; M:8.2%	14.1	34,394	30.2
M. arenaria (Guadeloupe)[17]	258.1	304 ± 9	C :94.76 (3.66) P :95.56	C:87.1% F:4.3%; M:8.6%	16.5	26,196	30
M. arenaria (A2-O)[60]	284.05	NA	C: 94.76 (3.57) P :96.77	C:87.1% F:2.6%; M:10.3%	204.6	2,224	30
M. arenaria (HarA)[20]	163.8	NA	C :91.53 (2.74) P :95.97	C: 78.2 F :12.2; M:9.6	10.5	46,509	30.3
M. hapla (VW9)[61]	53.6	121 ± 3	C :93.55 (1.19) P: 95.56	C:87.4% F:4.3%; M:8.3%	83.6	1,523	27.4
M. floridensis (JB5)[62]	99.9	NA	C: 56.45 (1.95) P: 76.61	C:54.1%F:28.7%; M:17.2%	3.5	81,111	29.7
M. floridensis (SJF1)[20]	74.9	NA	C: 77.42 (1.71) P: 83.87	C:76.5% F:7.6%; M:15.9%	13.3	9,134	30.2
M. luci (SI-Smartno)[63]	209.2	NA	C: 95.56 (2.92) P :96.77	C:87.8% F:4.0%; M:8.2%	1,712	327	30.2
M. graminicola (IARI)[64]	38.18	NA	C: 84.27 (1.34) P: 90.73	C:73.6% F:15.2%; M:11.2%	20.4	4,304	23.05

Table 3. Root-knot nematode genome features. *This work.

and after these two steps of contamination removal a total of 676 contigs spanning ~73 Mbp were eliminated. The clean assembly was then corrected with pilon[29] using the transcriptomic reads generated in this study (–fix snps parameter), which resulted in the final frozen assembly used for downstream analyses.

Transcriptome assembly. Adapters, low-quality regions (Phred score <30), and regions with two or more consecutive ambiguous nucleotides, were cropped using prinseq[30], resulting in 56,468,708 and 54,424,802 cleaned paired-end reads for J2 and egg libraries, respectively. The clean reads were assembled using CLC Genomics Workbench 9.0 (https://www.qiagenbioinformatics.com/) with automatic bubble size and word size estimation, in simple contig mode, and minimum contig length of 200. The transcriptome assemblies consisted of 110,068 and 110,263 contigs for J2 (J2 stage transcriptome assembly of *Meloidogyne enterolobii* (INRA/JKI). available in Figshare[31]) and Egg (Egg transcriptome assembly of *Meloidogyne enterolobii* (INRA/JKI). available in Figshare[31]) libraries, respectively.

Gene prediction and annotation. Detection of gene models was done with the fully automated pipeline EuGene-EP version 1[32]. EuGene has been configured to integrate similarities with known proteins from Wormpep 221[33], *Meloidogyne incognita* predicted proteome[17] and UniProtKB/Swiss-Prot database[34], with the prior exclusion of proteins that were similar to those present in RepBase[35].

Three datasets of *Meloidogyne enterolobii* transcribed sequences were aligned on the genome and used by EuGene as transcription evidence: (i) the *de novo* assemblies of the egg and J2 transcriptomes (ii) the egg stage RNAseq clean reads and (iii) the J2 stage RNAseq clean reads. The alignments of dataset (i) on the genome, spanning 50% of the transcript length with at least 95% identity were retained. The alignments of datasets (ii) and (iii) spanning 90% of the read length with 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, the minimum intron length to 35 bp, and to allow the non canonical donor splice site "GC". Finally, the Nematode specific Weight Array Method matrices were used to score the splice sites (http://eugene.toulouse.inra.fr/Downloads/WAM_nematodes_20171017.tar.gz).

Functional annotation of predicted proteins. We analysed all the predicted proteins in order to identify their putative functions and the putative secretory subset.

Prediction of conserved protein domains and gene ontology assignment. We used PfamScan[36] on the whole set of *M. enterolobii* proteins against the Pfam-A v32 library of HMM profiles[37] to identify conserved

protein domains. We used the Pfam annotation as well as the set of predicted proteins as an input to BLAST2GO pro v.1.12.11[38] to assign standard and 'slim' gene ontology terms according to the presence of Pfam domains, to BLASTp[39] homology searches against the NCBI's nr database and to Interproscan[40] annotation performed internally in BLAST2GO.

Secretome prediction. Signal peptides for secretion were detected in the set of *M. enterolobii* predicted proteins using Signalp4.1[41] and cleaved. From the cleaved proteins, transmembrane regions were predicted using TMHMM2.0c[42]. We considered as possibly secreted all the proteins that featured a predicted signal peptide and no transmembrane region.

Prediction and annotation of transposable elements. Prediction and annotation of transposable elements (TE) were performed with the REPET meta-pipeline, which combines TEdenovo and TEannot pipelines[43].

Data pre-processing. To correctly perform the all against all comparisons of the genome contigs step (see below), REPET automatically trims stretches of Ns of length 11 or more. However, this can result in a higher fragmentation of the genome with a risk of identification of spurious short repetitive matches. The *M. enterolobii* genome comprises few unresolved 'N' bases (1,401). We first created a modified version of the genome by splitting it at N stretches of length 11 or more and then trimming all N, using dbchunk.py from the REPET tools collection. To circumvent the potential problem with shorter DNA fragments, we then only kept chunks of length above 8,659 bp, defined as the L99 chunk length threshold.

TE de-novo prediction. The previously pre-processed genome was used to build a TE consensus library using the TEdenovo pipeline (REPET 2.5 configuration files used for TE prediction and annotation in the *M. enterolobii* genome (JKI/INRA). available in Figshare[31]). The genome was aligned to itself using Blaster[44] and High Scoring Segment Pairs (HSPs) were detected. The repetitive HSPs were clustered by Recon[45] and Grouper[44]. A consensus sequence was created for each cluster based on the corresponding multiple alignment with MAP[46]. Consensus sequences were analysed by PASTEClassifier[47] in order to find homology with known TE and to detect TE-related features such as specific HMM-profiles and structural characteristics. For identification of TE features, we used the curated libraries provided by the REPET development team (URGI): Repbase 20.05 (aa and nt), a concatenation of Pfam 27.0 and GyDB 2.0 hmm-profiles databanks and rRNA Eukaryota databank. The detected features were used by PASTEClassifier to classify consensus sequences in the different TE orders defined in Wicker's classification[48]. Simple sequence repeats (SSR) and under-represented unclassified consensuses were filtered out.

Semi-automated TE consensus filtering. The TE consensus library was filtered as follows in order to minimise false positives and redundancy. A draft annotation of the whole genome based on the consensus library was performed using TEannot (steps 1, 2, 3, 7; (REPET 2.5 configuration files used for TE prediction and annotation in the *M. enterolobii* genome (JKI/INRA) available in Figshare[31]). Only consensus sequences with at least one Full-Length Copy (FLC) annotated on the genome were retained to constitute a new filtered TE consensus library (*M. enterolobii* cleaned TE consensus sequence library available in Figshare[31]). Consensuses with at least one FLC annotated on the genome were identified with PostAnalyzeTELib.py and the corresponding sequences were extracted using GetSequencesFromAnnotations.py (REPET tools).

TE whole genome annotation. The TEannot pipeline was used to annotate the whole *M. enterolobii* genome from the previously created 'filtered TE consensus library' (steps 1, 2, 3, 4, 5, 7, 8; (REPET 2.5 configuration files used for TE prediction and annotation in the *M. enterolobii* genome (JKI/INRA). available in Figshare[31]). TE consensus sequences from the 'filtered TE consensus library' were aligned on the genome using Blaster, Censor[49], and RepeatMasker. The results of the three methods were concatenated and MATCHER[44] was used to remove overlapping HSPs and make connections with the "join" procedure. SSRs were detected by TRF[50], Mreps[51], and RepeatMasker, and then merged. Eventually, after some redundancy removal and application of the "long join" procedure (distant fragments belonging to the same copies are joined), raw annotations (Unfiltered transposable elements annotation of *M. enterolobii* genome (INRA/JKI). available in Figshare[31]) were exported for post-processing.

Annotations filtering and post-processing. We used in-house Python scripts (Python script used to filter TE annotations in the *M. enterolobii* genome (INRA/JKI). available in Figshare[31]) to extract canonical TE annotations from the whole repeatome annotation, but also increase the consistency between TE consensus sequences and their annotated copies.

Only TE annotation with >85% identity to the consensus sequence, a length >250 nucleotides, and classified as retrotransposon (Wicker's class I) or DNA-transposon (Wicker's class II) were retained. TE-annotations covering less than 1/3 of the consensus sequence length were discarded. Each genomic sequence corresponding to TE annotations were individually blasted against the TE consensus library, and only the ones with their consensus as the best hit were retained. Eventually, overlapping annotations on the same strand were removed. This ensemble of filters yielded the final TE annotation (Filtered final transposable elements annotation of *M. enterolobii* genome (INRA/JKI). available in Figshare[31]) used to compute statistics on the percentage of the genome occupied by TE and relative repartition in TE orders.

These scripts also allowed computing basic statistics and to describe TE coverage on the genome regarding Wicker's classification.

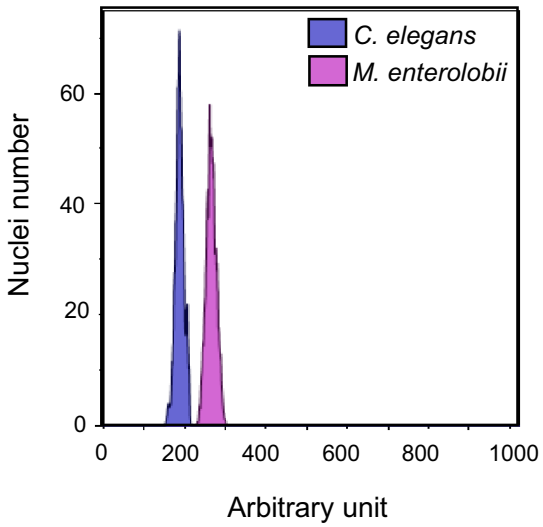

Fig. 2 Relative DNA staining in nuclei of *M. enterolobii*. Cytogram example obtained after gating on G0/G1 nuclei (arbitrary units) from *M. enterolobii* when processed mixed together with *C. elegans*, the internal standard (diploid genome size is 200 Mb).

Data Records

All the Illumina and PacBio sequence data supporting the results of this article as well as the genome assembly and gene models have been deposited and are publicly available at the EMBL-EBI's European Nucleotide Archive[52] and at the NCBI[53,54]. All the processed data, including genome and transcriptome assemblies and all the annotation results have been deposited and are publicly available as a Figshare collection[31]. The genome assembly as well as the predicted gene models, a blast server and a genome browser are publicly available at https://meloidogyne.inrae.fr/.

Technical validation

De novo genome assembly. After elimination of the few contaminant contigs, we assembled the *M. enterolobii* Swiss strain in 4,437 contigs for a total genome size of 240 Mb with a contig N50 length of 143 kb (*Meloidogyne enterolobii* genome assembly (V1, INRA/JKI). available in Figshare[31]). Overall, the genome assembly we produced constitutes a significant improvement in terms of contiguity and completeness compared to the previous *M. enterolobii* draft genome from a Burkina Faso isolate[20]. The genome size grew from 162.4 to 240 Mb. The number of contigs / supercontigs was divided by >10 (46,090 to 4,437) while the N50 length was multiplied by >15 (9.3 to 143 kb). This N50 length is in the top 3 highest for a *Meloidogyne* genome and the best for a publicly available annotated one (Table 3).

Validation of the genome size. We used flow cytometry to perform accurate measurement of cells DNA contents in *M. enterolobii* compared to internal standards with known genome sizes: *Caenorhabditis elegans* strain Bristol N2 (approximately 200 Mb at diploid state) and *Drosophila melanogaster* strain Cantonese S. (approximately 350 Mb at diploid state) as previously described[17]. Briefly, nuclei were extracted from two hundred thousand J2 infective juvenile larvae as described in[55] and stained with 75 µg/mL propidium iodide and 50 µg/mL DNAse-free RNAse. Flow cytometry analyses were carried out using an LSRII / Fortessa (BD Biosciences) flow cytometer operated with FACSDiva v6.1.3 (BD Biosciences) software. The DNA contents of the *M. enterolobii* samples were calculated by averaging the values obtained from three biological replicates. For estimation of total nuclear DNA content, we used the *M. enterolobii* strain Godet (from Guadeloupe France; N°75 available in Institut Sophia Agrobiotech collection).

The calculated nuclear DNA content *via* flow cytometry experiments was 274.69 ± 18.52 Mb (Fig. 2). With 240 Mb, the de novo genome assembly size is close to the estimated total DNA content in *M. enterolobii* cells and also represents a substantial improvement compared to the previous genome assembly (162.4 Mb, Table 3)[20].

This suggests that most of the *M. enterolobii* genome has been captured in this assembly. The difference between the genome assembly size and the total estimated DNA content ranges from 16 to up to 53 Mb and could be due either to duplicated and repetitive genome regions that have not been correctly separated during assembly or to differences in genome sizes between the Swiss *M. enterolob*ii strain we have sequenced and the 'Guadeloupe' strain used for DNA content measurement via flow cytometry.

Genome completeness assessment. To assess the completeness of the genome assembly, we ran CEGMA 2.5[56] and BUSCO v3.0[57] with the Eukaryotic dataset (odb9) in fast mode and *C. elegans* as a model for Augustus predictions[58]. Both pipelines search in genome assemblies for genes universally or largely conserved in Eukaryotes and produce reports on the number of genes found in complete, partial, single copy or duplicated versions in the genome under consideration. Although a nematode dataset exists in BUSCO, it only includes 8

BUSCO genes (303)	*M. enterolobii* (Swiss)*	*M. enterolobii* (L30)[20]
Complete	94.7% (287)	78.2% (237)
Single-copy complete	9.2% (28)	45.9% (139)
Duplicated complete	85.5% (259)	32.3% (98)
Fragmentary	3.0% (9)	11.2% (34)
Missing	2.3% (7)	10.6% (32)

Table 4. Estimation of the protein set completeness with BUSCO. *This work.

genomes from just three of the 12 described nematode clades (2, 8 and 9)[59]. Because the root-knot nematodes belong to Clade 12, which is not represented at all, we decided to use the Eukaryotic dataset which is more comprehensive (65 species, including the 8 nematodes).

Overall, 94.76 and 87.5% of the CEGMA and eukaryotic BUSCO genes, respectively, were found in complete length.

For comparison purposes, we also ran CEGMA and BUSCO with the same parameters on all the root-knot nematode genome assemblies that are publicly available (Table 3) The CEGMA and BUSCO completeness scores are both among the highest obtained for a *Meloidogyne* genome to date. For comparison, the model nematode *C. elegans* reaches CEGMA and eukaryotic BUSCO completeness scores of 96.37 and 94.7, respectively. This is higher than all the *Meloidogyne* genomes but it should be noted that the whole protein set of *C. elegans* is part of the training set for both CEGMA and BUSCO.

Gene prediction. Using the Eugene-EP pipeline, 63,841 genes were predicted, of which 59,773 were protein-coding with the exons spanning 61.9 Mb (~26%) of the genome assembly length. The GC percentage was higher in the coding portion (34.3%) than in the whole genome (30.0%). Spliceosomal introns were detected in 88% of protein-coding genes, with an average of 6.2 exons per gene. Similarly to the other tropical RKN genomes[17], less than 1% of splice sites have a non-canonical GG donor dinucleotide (canonical is GT).

Eugene-EP also predicted 4,068 non protein-coding genes (e.g. ribosomal, tRNA, splice leader genes). None of them had predicted intron but similarly to protein-coding genes the average GC content was higher (34.1%) than for the rest of the genome.

The overall statistics of the EUGENE gene predictions (Statistics of the gene predictions of *M. enterolobii* (genome V1, INRA/JKI).) and GFF3 annotations (GFF3 of the EUGENE gene prediction of *M. enterolobii* (genome V1, INRA/JKI).) as well as the fasta files of the predicted protein-coding genes (Predicted protein-coding genes of *M. enterolobii* (genome V1, INRA/JKI).), coding sequences (Predicted coding sequences (CDS) in *M. enterolobii* (genome V1, INRA/JKI).), proteins (Predicted proteins of *M. enterolobii* (genome V1, INRA/JKI).), messenger RNAs (Predicted messenger RNAs of *M. enterolobii* (genome V1, INRA/JKI).) and non-coding RNAs (Predicted non coding RNAs (ncRNA) of *M. enterolobii* (genome V1, INRA/JKI).) are all available in Figshare[31].

Validation of the predicted proteins. We ran a BUSCOV3 analysis in protein mode with the eukaryotic (odb9) dataset of 303 highly conserved genes to assess the completeness and quality of the set of predicted proteins in *M. enterolobii* and compared to the previous version of the *M. enterolobii* predicted proteins from the Burkina Faso isolate. Overall, 94.7% (287/303) of the eukaryotic BUSCO genes were found in complete length in the *M. enterolobii* proteome. This is a substantial improvement compared to the previously available *M. enterolobii* proteome, in which only 78.2% of complete eukaryotic BUSCO genes were found (Table 4). This improvement was due to both a reduction of the number of fragmentary BUSCO genes (from 11.2% to 3.0%) and of the number of missing BUSCO genes (from 10.6% to 2.3%).

Predicted secreted proteins. RKN and other plant-parasitic nematodes secrete *in planta*, proteins called effectors, where they support parasitism by manipulating plant structure and functions[60]. We assessed whether the set of predicted proteins in *M. enterolobii* could support effector discovery. To do so, we studied whether some gene ontology terms enriched in predicted secreted proteins corresponded to obvious known effector functions. We identified 4,973 *M. enterolobii* proteins (~8.3% of the total) as possibly secreted by having a predicted signal peptide for secretion and no transmembrane domain (List of predicted secreted proteins of *M. enterolobii* (genome V1, INRA/JKI). available in Figshare[31]).

Using BLAST2GO pro functional annotation (Functional annotation of predicted proteins in *M. enterolobii* (genome V1, INRA/JKI). available in Figshare[31]), we investigated whether some functions were enriched in the set of predicted secreted proteins. We used a gene set enrichment analysis (GSEA) as described in[61] to identify significantly overrepresented gene ontology (GO) terms in the set of predicted secreted proteins in comparison to the rest of the proteins. We considered as significantly overrepresented, the GO terms that returned a false discovery rate (FDR) value < 0.05 in one-tailed Fisher's exact test. We identified 49 overrepresented GO 'slim' terms in the predicted secreted proteins (Over-represented GO terms in predicted secreted proteins *M. enterolobii* (genome V1, INRA/JKI). available in Figshare[31]).

In total, 20, 24 and 5 enriched GO terms were identified in the 'biological process' (BP), 'molecular function' (MF) and 'cellular component' (CC) ontologies, respectively. In the BP category, enriched terms encompassed several enzymatic / catabolic processes in relation with plant cell wall macromolecules (e.g. 'cell wall macromolecule catabolic process', 'cellulose catabolic process') or other macromolecules (e.g. 'chitin metabolic process', 'peptidoglycan catabolic process'). These terms echo enriched terms in the MF category such as 'pectate lyase activity',

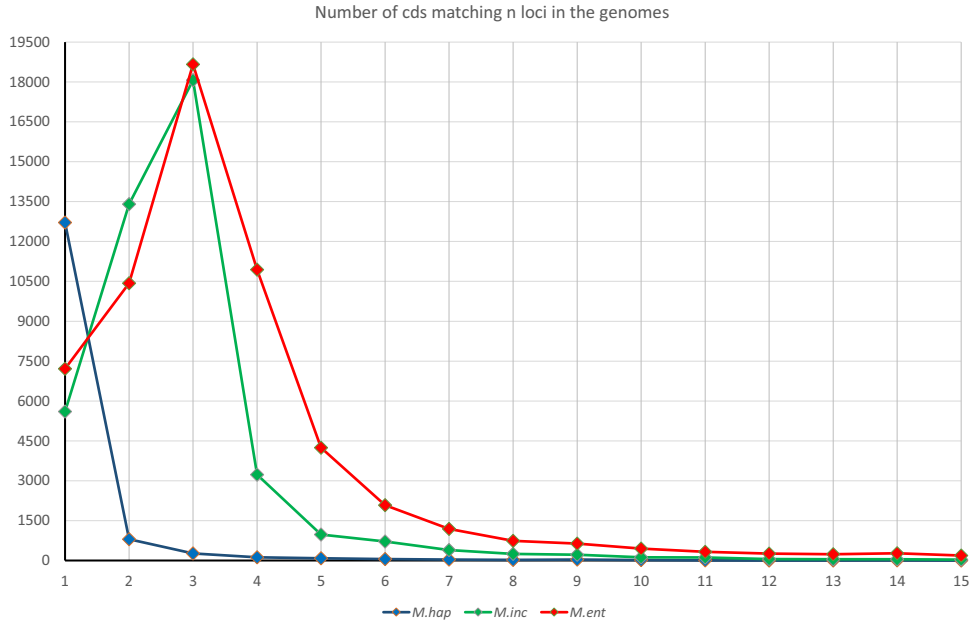

Fig. 3 Number of CDS mapping n loci in *Meloidogyne* genomes. Number of CDS mapping with >95% identity on >66% on their length at n loci on the *M. enterolobii* (red), *M. incognita* (green) and *M. hapla* (blue) genomes.

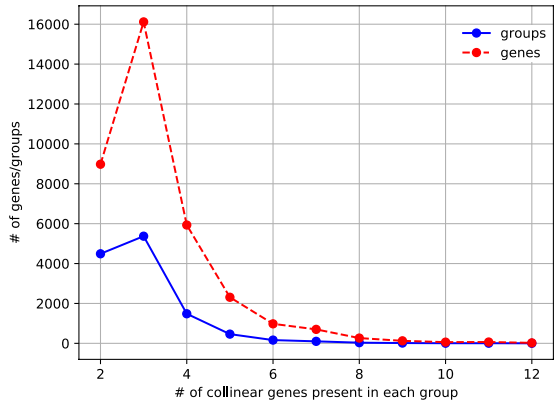

Fig. 4 Distribution of genes based on the duplication depth in the collinear blocks. The peak of the distribution is observed at a depth of three, suggesting a triploid genome.

'polysaccharide binding', 'cellulase activity', all related to degradation / modification of the plant cell wall. Plant cell wall-degrading enzymes are one category of known effectors secreted by plant-parasitic nematodes with the clearest function[62,63]. The above-mentioned enriched GO terms in the BP and MF ontologies, validate the set of predicted secreted proteins an interesting resource towards discovery and description of *M. enterolobii* effectors.

As expected for predicted secreted proteins, in the CC category, the two most significantly enriched GO terms were 'extracellular space' and 'endoplasmic reticulum lumen'.

Genome structure and ploidy level. The genomes of *M. incognita*, *M. javanica* and *M. arenaria*, other mitotic parthenogenetic RKN from Clade I, have been described as allopolyploids[17]. Because *M. enterolobii* belongs to the same clade and is also described as a mitotic parthenogenetic species, it is important to estimate the ploidy level of the genome assembly.

In the genomes of *M. incognita*, *M. javanica* and *M. arenaria*, respectively described as triploid and degenerate tetraploids[17], the CEGMA genes were found in 2.93, 3.68 and 3.66 copies, on average, respectively (Table 3). In contrast, the genomes of *M. hapla* and *C. elegans* have an average number of CEGMA gene copies of 1.19 and 1.09, respectively, which is consistent with their homozygous diploid nature and the corresponding haploid genome assemblies. Thus, the average number of CEGMA gene copies in RKN genomes seem to recapitulate the estimated ploidy levels. In *M. enterolobii*, the average number of copies per CEGMA gene was 3.3, suggesting a polyploid (at least triploid) genome structure as well. To further investigate and validate the ploidy level, we performed additional analyses using gene predictions as markers.

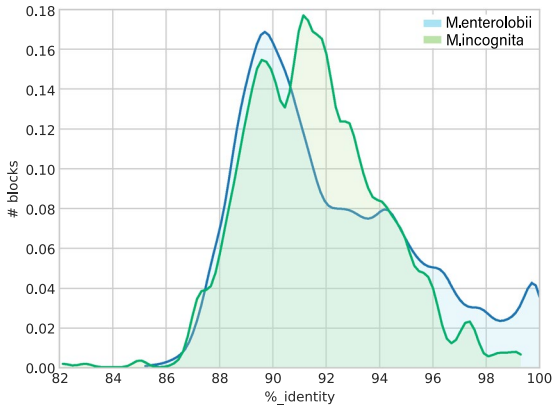

Fig. 5 Nucleotide identity computed from NUCmer alignments between homologous duplicated blocks. *M. enterolobii* (blue) shows one major peak at 89.9% and two smaller ones at 94.2% and 99.9% nucleotide identity, whereas *M. incognita* (green) shows almost two overlapping peaks at 89.5% and 91.4% of identity.

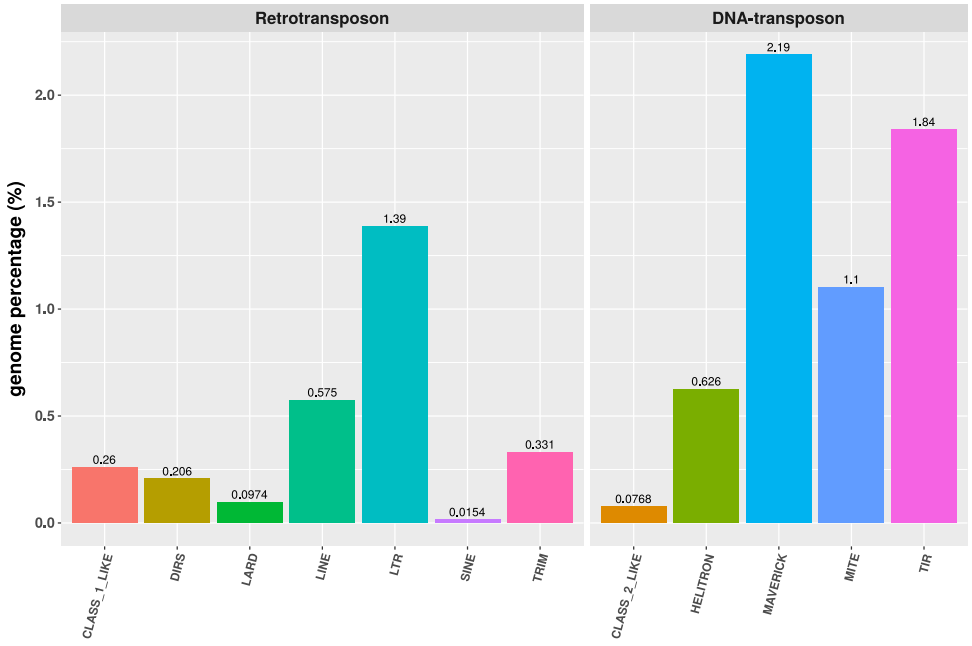

Fig. 6 Transposable Elements annotation in the genome of *M. enterolobii*. Canonical-TE annotations distribution as a genome percentage. Annotations are grouped according to Wicker's classification[48].

Duplication level estimation via mapping of coding sequences (CDS). We used BLAT[64] to align the ensemble of predicted coding sequences (CDS) to the respective reference genomes of *M. enterolobii*, *M. incognita* Morelos[17] and the meiotic facultative sexual species *M. hapla* VW9[65]. We only retained genome matches that covered at least 2/3 of the CDS length at a minimum 95% identity, and counted the number of matched loci per CDS query. In the diploid meiotic species *M. hapla*, 90% of the CDS mapped to a unique locus in the reference genome and can be considered single-copy genes. In contrast, only 12 and 13% of *M. enterolobii* and *M. incognita* CDS map a unique locus on their respective genome assemblies. Hence, the vast majority of protein-coding genes are in multiple copies in *M. enterolobii* and *M. incognita* (Fig. 3). Furthermore, both *M. enterolobii* and *M. incognita* show a peak of CDS alignments at 3 different loci. This is consistent with the average CEGMA gene copy number of 2.93 and 3.3 for *M. incognita* and *M. enterolobii*, respectively, and reinforces the possibility of triploidy.

Genome structure estimation by classification of gene duplicates. Although CEGMA gene copy numbers and alignment of CDS to the genome suggest most of the genes are triplicated, this does not necessarily imply a triploid genome. To further validate ploidy, it is important to show that gene duplicates are forming whole duplicated blocks and not only dispersed independent duplications.

We used MCScanX[66] to detect and classify duplications and study the genome structure (MCScanX duplication analysis results on the *M. enterolobii* genome (INRA/JKI). available in Figshare[31]). In a first step, all the

	order	nb. of features	sum of features length (bp)	median features length (bp)	median identity with consensus (%)
Retro-transposons	CLASS_1_LIKE	86	624,868	7,197	98.65
	LTR	647	3,329,701	4,012	97.55
	DIRS	58	494,483	8,323.5	97.9
	LINE	343	1,379,884	3,798	97.1
	SINE	36	36,906	1,076	99.5
	LARD	108	233,797	1,795.5	93.83
	TRIM	469	794,644	690	97.9
DNA-transposons	CLASS_2_LIKE	46	184,474	1,193.5	99
	TIR	3,974	4,415,242	918	96.95
	HELITRON	179	150,764	8,328	98.9
	MAVERICK	546	5,253,812	8,954	97.3
	MITE	4,452	264,544	590	96.4
	Total	**10,944**	**17,163,119**		

Table 5. Summary of *M. enterolobii* Canonical-TE annotations statistics.

predicted protein sequences were self-blasted to determine homologous relationships between them, with an e-value threshold of $1E^{-10}$. Using homology information from the all-against-all blast output and gene location information from the GFF3 annotation file, MCScanX detects duplicated protein-coding genes and classifies them in the following categories, (i) singleton when no duplicates are found in the assembly, (ii) proximal when duplicates are on the same contig and separated by 1 to 10 genes, (iii) tandem when duplicates are consecutive, (iv) whole genome duplication (WGD) or segmental when duplicates form collinear blocks with other pairs of duplicated genes, and (v) dispersed when the duplicates cannot be assigned to any of the above-mentioned categories. For detection of WGD / segmental duplications we required at least 3 collinear gene pairs.

MCScanX analyses showed that 94.8% of the protein-coding genes are duplicated in *M. enterolobii*. Overall, the majority of the genes (61.8%) are part of the whole genome duplication category and form 2,892 collinear blocks, that span 164 Mb (ca. 68% of the genome). This reinforces the idea that the genome is polyploid. Then, 27.6% of the genes have been classified as dispersed duplicates, 3.1% and 2.2% form proximal or tandem duplicates, respectively. By comparison, in *M. incognita*, only 28.5% of the genes formed duplicated collinear blocks and 59.4% were dispersed copies. This highlights the gain of resolution allowed by improved genome assemblies. Indeed, the proportion of genes forming whole duplicated blocks seems to be correlated to the genomes N50 lengths. We investigated the duplication depth of genes in the collinear blocks and observed a peak at a depth of three (Fig. 4), which is consistent with the peak at three loci matched by the CDS on the genome and the average copy number of CEGMA genes. This ensemble of results strongly suggests that the genome of *M. enterolobii* is triploid.

Divergence level between duplicated genome blocks. In the mitotic tropical RKN *M. incognita*, *M. javanica* and *M. arenaria*, the duplicated genomic regions have a high pairwise average nucleotide divergence of ~8%[17]. To check whether a similar high nucleotide divergence existed between the *M. enterolobii* genome copies, we used nucmer[58] and aligned the 2,892 duplicated blocks identified by MCScanX (WGD/segmental category). The average pairwise nucleotide identity between duplicated blocks was 92%, thus validating a similar 8% nucleotide divergence. The distribution of pairwise identities between the *M. enterolobii* duplicated genomic blocks shows a major peak at 89.9% identity and a second minor peak at 94.2% (Fig. 5). This distribution is different from that of *M. incognita*, where almost two overlapping peaks at 89.5% and 91.4% identity are observed. This suggests despite similar average divergence, different evolutionary trajectories in *M. incognita* and *M. enterolobii* and independent polyploidization events.

Transposable and other repetitive elements. Repetitive elements span 47.62% of the *M. enterolobii* genome assembly size, but canonical Transposable Elements (TE) annotations occupy only 8.70% (Fig. 6, Table 5). Using Wicker's classification[48], Class-I Retrotransposons span 2.87% of the genome assembly while class-II DNA-transposons occupy a twice higher proportion (5.83%). Using the same methodology, a much higher proportion of DNA transposons compared to retrotransposons was also observed in the genomes of *M. incognita* and *C. elegans*[67]. Because TE annotation is highly dependent on genome assembly quality and on the methodology used, we will not make direct comparison of the percentage of the genome occupied by TE with other nematode genomes. Nonetheless, re-annotating the *C. elegans* genome with the same methodology as a control yielded a very similar proportion of the genome occupied by TE than reported in previous studies[67].

As a further validation of the accuracy of TE annotation, we observed that all the previously reported TE orders in RKN genomes[17,68,69] (i.e. class-I Retrotransposons:DIRSs, LINEs, LTRs, SINEs; and class-II DNA-transposons:HELITRONs, TIRs, MAVERICKs, and MITEs) were detected in *M. enterolobii* as well. Furthermore, TRIM and LARD non-autonomous retrotransposons, previously reported in RKN genomes as 'unclassified'[17], were also now correctly identified and classified in the genome of *M. enterolobii*.

Received: 7 February 2020; Accepted: 27 August 2020;
Published online: 05 October 2020

References

1. Karssen, G., Liao, J., Kan, Z., van Heese, E. & den Nijs, L. On the species status of the root-knot nematode Meloidogyne mayaguensis Rammah & Hirschmann, 1988. *ZooKeys* **181**, 67–77 (2012).
2. Anonymous. Meloidogyne enterolobii. *EPPO Report. Serv*. 159–163 (2014).
3. Brito, J. A., Stanley, J. D., Mendes, M. L., Cetintas, R. & Dickson, D. W. Host status of selected cultivated plants to Meloidogyne mayaguensis in Florida. *Nematropica* **37**, 65–72 (2007).
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. Hallmann, J. & Kiewnick, S. Virulence of Meloidogyne incognita populations and Meloidogyne enterolobii on resistant cucurbitaceous and solanaceous plant genotypes. *J. Plant Dis. Prot.* **125**, 415–424 (2018).
6. Ye, W. M., Koenning, S. R., Zhuo, K. & Liao, J. L. First Report of Meloidogyne enterolobii on Cotton and Soybean in North Carolina, United States. *Plant Dis.* **97**, 1262–1262 (2013).
7. Ramírez-Suárez, A., Rosas-Hernández, L., Alcasio-Rangel, S. & Powers, T. O. First Report of the Root-Knot Nematode Meloidogyne enterolobii Parasitizing Watermelon from Veracruz, Mexico. *Plant Dis.* **98**, 428–428 (2013).
8. Onkendi, E. M., Kariuki, G. M., Marais, M. & Moleleki, L. N. The threat of root-knot nematodes (Meloidogyne spp.) in Africa: a review. *Plant Pathol.* **63**, 727–737 (2014).
9. Castagnone-Sereno, P. Meloidogyne enterolobii (=M. mayaguensis): profile of an emerging, highly pathogenic, root-knot nematode species. *Nematology* **14**, 133–138 (2012).
10. Elling, A. A. Major emerging problems with minor meloidogyne species. *Phytopathology* **103**, 1092–1102 (2013).
11. Kiewnick, S. *et al*. Comparison of two short DNA barcoding loci (COI and COII) and two longer ribosomal DNA genes (SSU & LSU rRNA) for specimen identification among quarantine root-knot nematodes (Meloidogyne spp.) and their close relatives. *Eur. J. Plant Pathol.* **140**, 97–110 (2014).
12. Claverie, M. *et al*. The Ma gene for complete-spectrum resistance to Meloidogyne species in Prunus is a TNL with a huge repeated C-terminal post-LRR region. *Plant Physiol.* **156**, 779–792 (2011).
13. Freitas, V. M. *et al*. Resistant accessions of wild Psidium spp. to Meloidogyne enterolobii and histological characterization of resistance. *Plant Pathol.* **63**, 738–746 (2014).
14. Gonçalves, L. S. A. *et al*. Resistance to root-knot nematode (Meloidogyne enterolobii) in Capsicum spp. accessions. *Rev. Bras. Ciênc. Agrár*. **9**, (2014).
15. Triantaphyllou, A. C. Cytogenetics, cytotaxonomy and phylogeny of root-knot nematodes. In *An advanced treatise on Meloidogyne: Biology and control* (eds. Sasser, J. N. & Carter, C. C.) vol. 1, 113–26 (North Carolina State University Graphics, 1985).
16. Koutsovoulos, G. D. *et al*. Population genomics supports clonal reproduction and multiple independent gains and losses of parasitic abilities in the most devastating nematode pest. *Evol. Appl.* **13**, 442–457 (2020).
17. Blanc-Mathieu, R. *et al*. Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes. *PLOS Genet.* **13**, e1006777 (2017).
18. Castagnone-Sereno, P. *et al*. Gene copy number variations as signatures of adaptive evolution in the parthenogenetic, plant-parasitic nematode Meloidogyne incognita. *Mol. Ecol.* **28**, 2559–2572 (2019).
19. De Ley, I. T. *et al*. Phylogenetic Analyses of Meloidogyne Small Subunit rDNA. *J Nematol* **34**, 319–27 (2002).
20. Szitenberg, A. *et al*. Comparative Genomics of Apomictic Root-Knot Nematodes: Hybridization, Ploidy, and Dynamic Genome Change. *Genome Biol. Evol.* **9**, 2844–2861 (2017).
21. 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 Dis.* **92**, 1370–1370 (2008).
22. 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 Pathol.* **59**, 1054–1061 (2010).
23. Hussey, R. S. & Janssen, G. J. W. Root-knot nematodes: Meloidogyne species. In *Plant resistance to plant parasitic nematodes*. 43–70 (J. L. Starr, R. Cook and J. Bridge, 2002).
24. Schaad, N. W. & Walker, J. T. The Use of Density-Gradient Centrifugation for the Purification of Eggs of Meloidogyne spp. *J. Nematol.* **7**, 203–204 (1975).
25. Koren, S. *et al*. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
26. Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M. & Blaxter, M. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front. Genet*. **4**, (2013).
27. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**, 1287 (2017).
28. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
29. Walker, B. J. *et al*. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE* **9**, e112963 (2014).
30. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinforma. Oxf. Engl.* **27**, 863–864 (2011).
31. Danchin, E. *et al*. The polyploid genome of the mitotic parthenogenetic root-knot nematode Meloidogyne enterolobii. *figshare* https://doi.org/10.6084/m9.figshare.c.5007182.v1 (2020).
32. Sallet, E., Gouzy, J. & Schiex, T. EuGene: An Automated Integrative Gene Finder for Eukaryotes and Prokaryotes. *Methods Mol. Biol. Clifton NJ* **1962**, 97–120 (2019).
33. Lee, R. Y. N. *et al*. WormBase 2017: molting into a new stage. *Nucleic Acids Res.* **46**, D869–D874 (2018).
34. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699–2699 (2018).
35. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
36. Li, W. *et al*. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.* **43**, W580–W584 (2015).
37. Finn, R. D. *et al*. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).
38. Götz, S. *et al*. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).
39. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–402 (1997).
40. Quevillon, E. *et al*. InterProScan: protein domains identifier. *Nucleic Acids Res* **33**, W116–20 (2005).
41. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011).
42. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. *J. Mol. Biol.* **305**, 567–580 (2001).
43. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLoS ONE* **6**, e16526 (2011).
44. Quesneville, H., Nouaud, D. & Anxolabéhère, D. Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes. *J. Mol. Evol.* **57**(Suppl 1), S50–59 (2003).
45. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).

46. Huang, X. On global sequence alignment. *Comput. Appl. Biosci. CABIOS* **10**, 227–235 (1994).
47. Hoede, C. *et al*. PASTEC: An Automatic Transposable Element Classification Tool. *PLoS ONE* **9**, e91929 (2014).
48. Wicker, T. *et al*. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
49. Jurka, J., Klonowski, P., Dagman, V. & Pelton, P. Censor - A program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* **20**, 119–121 (1996).
50. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–80 (1999).
51. Kolpakov, R., Bana, G. & Kucherov, G. mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res* **31**, 3672–8 (2003).
52. *European Nucleotide Archive* https://identifiers.org/insdc.sra:ERP119623 (2020).
53. *NCBI Assembly* https://identifiers.org/insdc.gca:GCA_903994135.1 (2020).
54. *NCBI Assembly* https://identifiers.org/insdc.gca:GCA_903797545.1 (2020).
55. Perfus-Barbeoch, L. *et al*. Elucidating the molecular bases of epigenetic inheritance in non-model invertebrates: the case of the root-knot nematode Meloidogyne incognita. *Front. Physiol.* **5**, (2014).
56. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
57. Waterhouse, R. M. *et al*. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
58. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
59. van Megen, H. *et al*. A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. *Nematology* **11**, 927–950 (2009).
60. Vieira, P. & Gleason, C. Plant-parasitic nematode effectors — insights into their diversity and new tools for their identification. *Curr. Opin. Plant Biol.* **50**, 37–43 (2019).
61. Subramanian, A. *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
62. Danchin, E. G. *et al*. Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. *Proc Natl Acad Sci U A* **107**, 17651–6 (2010).
63. Haegeman, A., Jones, J. T. & Danchin, E. G. Horizontal gene transfer in nematodes: a catalyst for plant parasitism? *Mol Plant Microbe Interact* **24**, 879–87 (2011).
64. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res* **12**, 656–64 (2002).
65. Opperman, C. H. *et al*. Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism. *Proc Natl Acad Sci U A* **105**, 14802–7 (2008).
66. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**, e49 (2012).
67. Kozlowski, D. K. *et al*. Transposable Elements are an evolutionary force shaping genomic plasticity in the parthenogenetic root-knot nematode *Meloidogyne incognita*. bioRxiv, 2020.04.30.069948, ver. 4 peer reviewed and recommended by PCI Evolutionary Biology. https://doi.org/10.1101/2020.04.30.069948 (2020).
68. Abad, P. *et al*. Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. *Nat. Biotechnol.* **26**, 909–915 (2008).
69. Szitenberg, A. *et al*. Genetic Drift, Not Life History or RNAi, Determine Long-Term Evolution of Transposable Elements. *Genome Biol. Evol.* **8**, 2964–2978 (2016).

Acknowledgements

M. Oggenfuss and M. Holterman, both Agroscope, are acknowledged for their support in DNA extraction and quantification. We would like to thank Jérôme Gouzy for help and advice in the gene prediction step. Julie Cazareth from "Imaging and Cytometry platform" at IPMC for tech support. GDK has received the support of the EU in the framework of the Marie-Curie FP7 COFUND People Programme, through the award of an AgreenSkills+ fellowship (under grant number 609398). CMJ received funding from the ASEXEVOL project, grant number 392 ANR-13-JSV7-0006 which also funded bioinformatics equipment used for this paper. This work has been supported by the French government, through the UCA-JEDI "Investments in the Future" project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01. Our research on the *M. enterolobii* genome is financially supported by a France-Germany bilateral funding ANR-DFG "AEGONE" with reference number ANR-19-CE35-0017-01.

Author contributions

E.G.J.D., S.K., .G.D.K., D.K.K., M.P. and L.P.B. wrote the manuscript. L.P.B. and C.M.J. performed measurements of the nuclear DNA content and analysed the results. G.D.K. cleaned and assembled the genome and participated in the analysis of the genome structure. E.G.J.D. participated in the analysis of the genome structure and performed the comparative assessment of genome completeness. M.P. performed the genome structure and divergence analysis. E.S. performed the gene prediction and analysed the results. D.K.K. performed the transposable elements annotation and analysis. J.E.F., S.K. produced material and conducted Illumina and PacBio short read sequencing at GATC Biotech (Konstanz, DE). C.A. initiated together with J.E.F. and S.K. the project on using long reads of the PacBio RSI platform at the Functional Genomics Center in Zurich (CH). S.K. and A.E.A. produced the material for transcriptome sequencing and A.E.A. analysed the results. A.E.A. and M.D.R. assembled the transcriptome and analysed the results. S.K. and J.E.F. identified, discovered and originally reared the *M. enterolobii* Swiss strain used for genome and transcriptome sequencing.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.K. or E.G.J.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files associated with this article.

© The Author(s) 2020