Sequence analysis

PPNID: a reference database and molecular identification pipeline for plant-parasitic nematodes

Xue Qing[1,*,†], Meng Wang[1,†], Gerrit Karssen[2], Patricia Bucki[1], Wim Bert[3] and Sigal Braun-Miyara[1,*]

[1] Department of Entomology, Nematology and Chemistry units; Agricultural Research Organization (ARO), the Volcani Center, Rishon LeZion, P.O. Box 15159, Israel, [2]National Plant Protection Organization, Wageningen Nematode Collection, P.O. Box 9102, 6700 HC, Wageningen, The Netherlands, [3] Nematology Research Unit, Department of Biology, Ghent University, K.L. Ledeganckstraat 35, 9000 Ghent, Belgium.

*To whom correspondence should be addressed.

† These authors contribute equally

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The phylum Nematoda comprises the most cosmopolitan and abundant metazoans on Earth and plant-parasitic nematodes represent one of the most significant nematode groups, causing severe losses in agriculture. Practically, the demands for accurate nematode identification are high for ecological, agricultural, taxonomic and phylogenetic researches. Despite their importance, the morphological diagnosis is often a difficult task due to phenotypic plasticity and the absence of clear diagnostic characters while molecular identification is very difficult due to the problematic database and complex genetic background.

Results: The present study attempts to make up for currently available databases by creating a manually-curated database including all up-to-date authentic barcoding sequences. To facilitate the laborious process associated with the interpretation and identification of a given query sequence, we developed an automatic software pipeline for rapid species identification. The incorporated alignment function facilitates the examination of mutation distribution and therefore also reveals nucleotide autapomorphies, which are important in species delimitation. The implementation of genetic distance, plot and maximum likelihood phylogeny analysis provides more powerful optimality criteria than similarity searching and facilitates species delimitation using evolutionary or phylogeny species concepts. The pipeline streamlines several functions to facilitate more precise data analyses, and the subsequent interpretation is easy and straightforward.

Availability: The pipeline was written in vb.net, developed on Microsoft Visual Studio 2017 and designed to work in any Windows environment. The PPNID is distributed under the GNU General Public License (GPL). The executable file along with tutorials is available at https://github.com/xueqing4083/PPNID.

Contact: xueqing4083@gmail.com; sigalhor@volcani.agri.gov.il

1 Introduction

The phylum Nematoda comprises the most cosmopolitan and abundant metazoans on Earth, with an estimated number of species extending to 10^6 (Boucher & Lambshead, 1995; Coomans, 2002). Plant-parasitic nematodes (PPN) represent one of the most significant nematode groups, over 4100 (Decraemer & Hunt, 2006) of the 27,000 described species (Quist *et al.*, 2015) are known to be plant-parasites. PPN parasitize

practically every higher plant species, causing agricultural losses estimated at 80 billion US dollars per year (Nicol *et al.*, 2011).

Despite their importance, the morphological diagnosis of plant-parasitic nematodes is often a difficult task. This is due to their high phenotypic plasticity (Coomans, 2002; Nadler, 2002) and the absence of clear diagnostic characters (Wijova *et al.*, 2005; Derycke *et al.*, 2008), especially in the most-frequently encountered juveniles (Anderson, 2000). Molecular barcoding is a diagnostic technique that is not reliant on morphology and therefore holds great promise as a tool to simplify and standardize nematode identification (Hebert *et al.*, 2003; Savolainen *et al.*, 2005). Indeed, the use of molecular barcoding has already been widely reported for phylogenies and species identifications in the field of nematology (Floyd *et al.*, 2002; Powers, 2004; De Ley *et al.*, 2005; Holterman *et al.*, 2006; Subbotin *et al.*, 2011).

However, molecular barcoding is underpinned by the assumptions that (i) the reference database represents a satisfactory taxonomic sampling of sequences; (ii) the sequences in the reference database have been correctly identified and annotated; and (iii) the link to species name is standardized, universally adopted, and not easily misunderstood (Blaxter *et al.*, 2005). Unfortunately, the commonly-used PPN sequence repository (INSD: International Nucleotide Sequence Databases, e.g. GenBank, EMBL, DDBJ) not only fails to fully respond to any of the above assumptions, but also includes numerous errors of various types: (1) *Erroneous sequencing*: certain supposed nematode sequences are in reality fungal sequences or reveal fungus-nematode chimera (e.g. KF568416, KF568433). For example, the entries for *Heterodera latipons* (FJ151164) and *Criconema* sp. (MG994946) appear to have been contaminated by fungus *Malassezia* sp. and *Vanrija* sp.; (2) *Species mislabeling*, e.g. the bacterivorous nematode incorrectly labeled as *Pratylenchus goodeyi* resulted in a cascade of erroneous interpretations, as shown by the reports of "plant-parasitic" nematodes on important crops (Janssen *et al.*, 2017); (3) *Legacy naming*: the identity of some sequences has changed with the improvement of identification techniques, but the old name still remains, e.g. several *Helicotylenchus* species (Subbotin *et al.*, 2005; Subbotin *et al.*, 2011); (4) *Data mishandling*: raw data were not properly assembled or trimmed, e.g. *Atetylenchus minor* (KP730045) contains *ca.* 200 bp unrelated fragment resulting in an erroneous long branch in phylogeny (Yaghoubi *et al.*, 2015); (5) *Incomplete referencing*: some species have been synonymized or the genus name has a new replacement but their sequence identity has not been changed accordingly, e.g. *Heterodera cynodontis* (DQ328698, EU284037, EU284024, EU284023 AF274386) has been synonymized with *H. cardiolata* (Subbotin et al., 2010) causing different interpretations in phylogenies (Sekimoto *et al.*, 2017). The genus *Rhizonema* has been replaced by *Rhizonemella* (Andrássy, 2007), but both names are used in the database; (6) *Incomplete entries*: certain sequences are labeled as having an unidentified species/genus, while they have been formally described to species level, e.g. *Tylenchorhynchus* sp. 1 and 2 (KJ461559–KJ461562) were actually identified as *T. agri* and *T. thermophilus* in Handoo *et al.* (2014); (7) *Arbitrary entries*: OTU short reads from metabarcoding studies were substantially submitted. The sequences identity were arbitrarily given based on highest BLAST match and subsequently submitted (Qing *et al.*, 2018). While Bridge *et al.* (2003) estimated that up to 20% of fungal sequences are actually misidentified in the INSD, the exact number of problematic PPN sequences remains unknown. It is clear that the INSD contains substantial sequence faults, which has the dual effect of both reducing its accuracy and increasing the complexity of downstream barcoding analyses.

Although molecular techniques are informative and powerful tools, inappropriate barcoding gene selection and data interpretation, next to the known flaws of existing databases, can result in serious limitations. The best-known example to illustrate this point is the case of "tropical root-knot nematodes", or the *M. incognita* group (MIG). For these species, the conventional barcoding genes (e.g. *ITS*, *18S*, *28S* rRNA) either lack the required resolution or contain intragenomic variations (Hugall *et al.*, 1999), and a set of less commonly-used mitochondrial genes are needed (Pagan *et al.*, 2015; Janssen *et al.*, 2016). Furthermore, compiling and interpreting molecular data involves a significant amount of manual intervention, and solely using DNA-similarity searches like BLAST can give misleading results. For example, there is no certain intra- or interspecific threshold available in PPN. Several species can share the same similarity score while the same species can be quite divergent among the population, and it is impossible to compare single nucleotide polymorphism (SNP) in MIG identification.

Finally, given the fact that expertise in nematode taxonomy and the correct interpretation of barcodes is declining, an appropriate molecular identification, although theoretically relatively easy, can become a cumbersome endeavor. A simple and efficient identification method is clearly vitally important and much-needed.

2 Methods

2.1 Manually-curated reference database

The present study attempts to make up for any INSD deficiencies by creating a manually-curated database including all up-to-date authentic PPN sequences. All PPN sequences from GenBank were extracted and reference sequences were pre-selected based on weighted criteria:

(1) Most priority is given to sequences from its original species description, or topotype. If type sequences are absent.

(2) Sequences with morphological information from a taxonomic article.

(3) Sequences from a non-taxonomic but peer-viewed journal.

(4) Exceptionally, if above are unsatisfied, sequences from other resource are used (but with more strict downstream quality control)

From the pre-selected candidates, entries were screened for contamination and inappropriate assembly by using BLAST searches or alignments with other related species. Subsequently, possible misidentifications were examined by reconstructing phylogenies similar to Janssen *et al.* (2017). Potential synonyms and name updates were checked by searching corresponding taxonomic references. In cases where different synonym proposals were in conflict with each other, those entries supported by molecular and detailed morphology (e.g. SEM) data were given priority. Finally, a name was added to the unlabeled species based on published literature or after contact with corresponding authors.

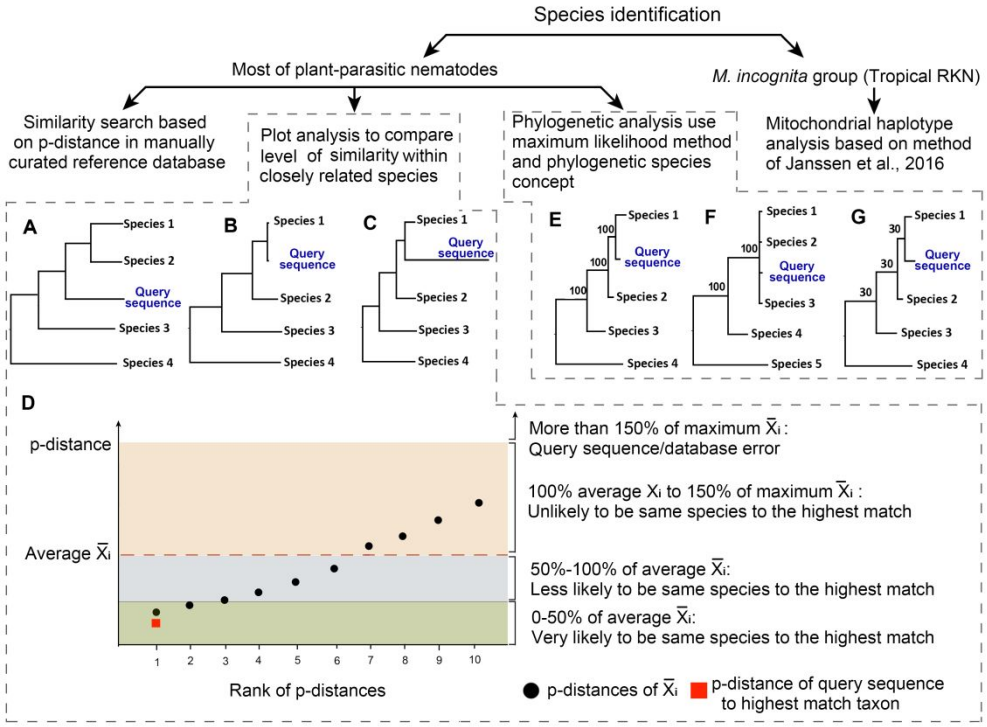

Fig. 1. The proposed species identification methods use in *PPNID* program. Tree branch length represents relative genetic distance. (**A-C**) Graphic explanations of the method used in plot analysis to distinguish inter- and intraspecific variations. (**E-G**) Phylogeny species concept used in phylogenetic analyses. (**A**) The query sequence is at a similar level of interspecific variation to other related species, suggesting this variation is likely to be interspecific and query sequence may be new species or known species without molecular data. (**B**) The query sequence to its similar species (Species 1) has significant lower genetic variations compare with other related species, suggesting this variation is likely to intra-specific and query sequence may be the highest p-distance hit. (**C**) The query sequence is exceptionally divergent, suggesting the input sequence may contain error or the inappropriate database was selected. (**D**) The p-distance plot that gives graphic suggestions for species identification. Suggestions are given corresponding to the placements of the red square. (**E**) The query sequence is likely to be the highest p-distance hit (Species 1), as they form a well-supported monophyletic clade. (**F**) The query sequence is not placed in a monophyletic clade, more than one species are closely related to input sequence. In this case, the barcoding gene is less informative and the user query sequence may not be the same as top p-distance hit. (**G**) The query sequence may not be the top p-distance hit (Species 1) as monophyletic clade is not well-supported.

2.2 Molecular identification pipeline

To facilitate the laborious process associated with the interpretation and identification of a given query sequence, we have developed an automatic software pipeline for rapid PPN identification (Fig. 1). For the majority of species, rRNA genes (*ITS*, *18S*, *28S* rRNA) are used as barcoding markers and the identity of species was evaluated at three steps:

(1) The query sequence is aligned with the corresponding database in MUSCLE (Edgar, 2004). The species with the highest p-distance similarity (HPS) are selected as a candidate. Comparable to the similarity-based identification in BLAST searches, HPS species can be considered as query species in many cases.

(2) However, this is not always true, like genetic variations to HSP are sometimes interspecifically (query species without molecular data) and query species may contain errors (*e.g.* low sequence quality; reverse and complementary; inappropriate database selected). Given there is no certain threshold for inter- or intraspecific variations in PPN, a proper interpretation for similarity comparison in step 1 is difficult. Here the newly proposed method distinguish inter- or intraspecific variations or error input by compare variation among other related species. This method that can be mathematically described as:

$$\bar{x}_j = \frac{1}{5}\sum_{k=1}^{m}{}' d_{j,k}$$

where the average p-distances x were calculated with j = {species 1, species 2, species 3, …, species 10}, and species 1 to 10 are ten top p-distance hit of the query sequence. d is the pairwise p-distance between j and k, and m = {species 1, species 2, species 3, species 4, species 5}, where species 1 to 5 are five most similar sequences to each of j. The ten returned x values are plotted in increasing order (Fig. 1D). Based on empirical evidence we further split p-distances into four groups, and the query sequence placed in each of group is defined as "likely to be same species", "less likely to be same species", "unlikely to be same species" to "error input" (Fig. 1D). Notice this approach provides induction rather than a definitive species delimitation method.

(3) Aside from similarity comparison, the proposed pipeline analyzed phylogenetic placement using the maximum likelihood approaches implied in PhyML (Guindon & Gascuel, 2003). In

the light of phylogeny species concept, the query sequence is identified as its HPS when they form a well-supported (*e.g.* bootstrap ≥ 80) monophyletic clade in maximum likelihood tree (Fig. 1E). Other cases like an unresolved clade contain several similar species (Fig. 1F) or a monophyletic but less supported clade (Fig. 1G) indicate query sequence may not be the same species to its HPS.

For MIG root-knot nematodes, a higher sensitivity is needed and therefore relatively fast-evolving mitochondrial genes were incorporated (Janssen *et al.*, 2016). The program automatically aligns a query sequence with the *Nad5* haplotypes database and returns informative SNP. In case *Nad5* is insufficient, the program will automatically guide users to supply a *Cytb* gene.

2.3 Test of identification performance

The performance of the proposed identification pipeline was tested by using newly generated sequences as well as the sequences available from GenBank. For newly generated sequences, soil and root samples were collected from several farms or grassland in Israel and nematodes were extracted by a Baermann funnel. Genomic DNA was extracted from a single fresh nematode by transferring each a PCR tube with 10 µl of 0.05 N NaOH and 1 µl of 4.5% (w/v) Tween 20. PCR amplification and sequencing is same to the protocol detailed in Qing *et al.* (2019). Contigs were assembled using Geneious R6.1.8. Prior to the tested, species were identified with traditional morphological methods. Extracted fresh nematodes were fixed in 4% formalin solution at 65℃, and gradually transferred to glycerin and mounted in a glass slide. Morphological and morphometric analyses were made with a Nikon Eclipse Ni light microscope equipped with differential interference contrast. For MIG root-knot nematodes, *Nad5* and *Cytb* sequences were acquired to follow the protocol in Janssen *et al.* (2017). For GenBank extracted sequences, depends on data availability 1–7 sequences that belong to the PPN-databases-included species were selected with the aforementioned criteria, but excluding those with 100% similar to PPN databases.

3 Results

The architecture and user interface of *PPNID* program is presented in Fig. 2. In general, program consists of three parts: standard identification pipeline for general plant-parasitic nematodes, pipeline for the MIG nematode, and reference database for molecular barcoding. For database, we manually screened PPN barcoding sequences in GenBank and 2407 sequences were extracted as a reference. These sequences were assigned to 78 databases according to their taxonomic groups and genes (*ITS*, *28S*, *18S*, *COI*, *Nad5*, and *Cytb*). The database will continuously be updated by the first author twice a year.

A total of 21 newly generated and 100 GenBank extracted sequences (belong to 17 and 56 species respectively) covering all involved barcoding genes were tested in the proposed methods. The details for identifications are listed in Supplement 1. Briefly, 95% (20 out of 21) of newly generated and 89% (89 out of 100) of GenBank extracted sequences were successfully identified and interpreted in all steps (p-distance similarity, plot analysis, and phylogeny). Those misidentifications were due to either exceptionally elevated intraspecific divergences (six sequences) or inadequate interspecific species variations (six sequences). In former scenario plot analysis inappropriately rejects query sequence to be same as HPS species, while in later case query sequence is the same as several species and can be any of them. In compare to our pipeline, BLAST searches return similar results in p-distance similarity step but fail to identify MIG species, and the further plot and phylogeny steps are completely missing in BLAST.

4 Discussion

The proposed pipeline streamlines several functions to facilitate more precise data analyses, and the subsequent interpretation is easy and straightforward. The incorporated alignment function facilitates the examination of mutation distribution and therefore also reveals nucleotide autapomorphies, which are important in species delimitation (Adams, 1998). The implementation of genetic distance, plot analysis, and maximum likelihood phylogeny provides more powerful optimality criteria than similarity searching (Nilsson *et al.*, 2004; Hanekamp *et al.*, 2007) and facilitates species delimitation using evolutionary or/and phylogeny species concepts (Adams, 1998, 2000). The SNP searching function allows automatic MIG haplotype identification at a single click, which represents a completely new approach. Finally, the pipeline also includes practical information on topics such as gene selection, commonly-used primers, and PCR conditions, to be found in the user guidelines. However, notice that the identification of PPN is difficult, and relies on experience and expertise. Furthermore, a consensus among taxonomists is often lacking, meaning that a standardized identification pipeline cannot completely replace conclusions based on a comprehensive taxonomical investigation.

The pipeline is especially useful for non-specialists working on PPN identification, as it is a fast and relatively easy tool to use. At the same time, it improves the interpretation of the data precision by increasing its resolution and by applying a more standardized methodology.

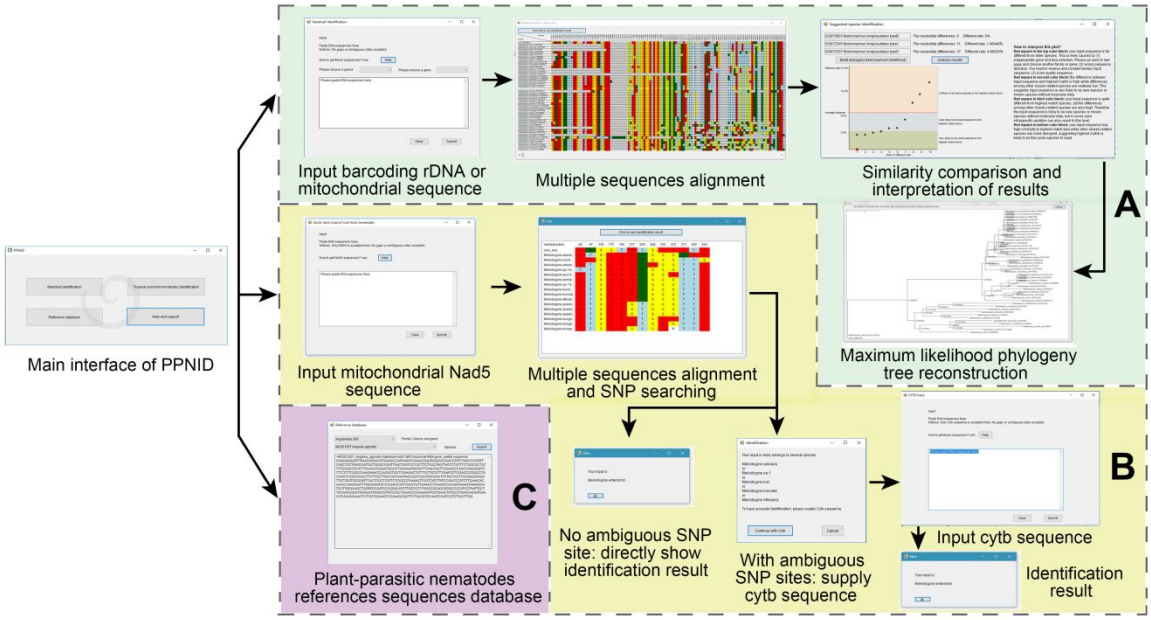

Fig. 2. The architecture and user interface of *PPNID* program. (**A**) Identification pipeline for general plant-parasitic nematodes. (**B**) Identification pipeline for the MIG root-knot nematodes. (**C**) Plant-parasitic nematodes reference database for molecular barcoding.

Acknowledgement

We thank Mr. Zaichen Ding for his technical supports in program development.

Funding

This work was supported by the Chief Scientist of the Ministry of Agriculture and Rural Development, Israel, grant no. 20-07-0012.

Conflict of Interest: none declared.

References

Adams,B.J. (1998) Species concepts and the evolutionary paradigm in modem nematology. *J. Nematol.,* **30**, 1–21.

Adams,B.J. (2001) The species delimitation uncertainty principle. *J. Nematol.,* **33**, 153–160.

Anderson, R.C. (2000) *Nematode parasites of vertebrates: their development and transmission*. Wallingford, UK, CABI Publishing. 650pp.

Andrássy,I. (2007) *Free-living nematodes of Hungary II (Nematoda errantia)*. Budapest, Hungary, Hungarian Natural History Museum and Systematic Zoology Research Group of the Hungarian Academy of Sciences, 496 pp.

Blaxter,M. *et al.* (2005) Defining operational taxonomic units using DNA barcode data. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.,* **360**, 1935–1943.

Boucher,G. and Lambshead,P.J.D. (1995) Ecological biodiversity of marine nematodes in samples from temperate, tropical, and deep-sea regions. *Conserv. Biol.*, **9**, 1594–1604.

Bridge,P.D. *et al.* (2003) On the unreliability of published DNA sequences. *New Phytol.*, **160**, 43–48.

Coomans,A. (2002) Present status and future of nematode systematics. *Nematology*, **4**, 573–582.

De Ley,P. *et al.* (2005) An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **360**, 1945–1958.

Decraemer,W. and Hunt, D.J. (2006) Structure and classification. In Perry, R. N. & Moens, M. (eds.). *Plant Nematology.* Wallingford, UK, CABI Publishing, p3-32.

Derycke,S. *et al.* (2008) Disentangling taxonomy within the *Rhabditis* (Pellioditis) *marina* (Nematoda, Rhabditidae) species complex using molecular and morhological tools. *Zool. J. Linn. Soc.,* **152**, 1–15.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Floyd,R.A.E. *et al.* (2002) Molecular barcodes for soil nematode identification. *Mol. Ecol.,* **11**, 839–50.

Guindon,S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696-704.

Handoo,Z.A. *et al.* (2014) Integrative taxonomy of the stunt nematodes of the genera *Bitylenchus* and *Tylenchorhynchus* (Nematoda, Telotylenchidae) with description of two new species and a molecular phylogeny. *Zool. J. Linn. Soc.,* **172**, 231–264.

Hanekamp,K. *et al.* (2007) PhyloGena—a user-friendly system for automated phylogenetic annotation of unknown sequences. *Bioinformatics*, **23**, 793-801.

Hebert,P.D.N. *et al.* (2003) Biological identifications through DNA barcodes. *Philos. Trans. R. Soc. Lond., B, Biol. Sci*., **270**, 313–321.

Holterman,M. *et al.* (2006) Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades. *Mol. Biol. Evol.,* **23**, 1792–1800.

Hugall,A. *et al.* (1999) Reticulate evolution and the origins of ribosomal internal transcribed spacer diversity in apomictic *Meloidogyne*. *Mol. Biol. Evol.,* **16**, 157–164.

Janssen,T. *et al.* (2016) Mitochondrial coding genome analysis of tropical root-knot nematodes (*Meloidogyne*) supports haplotype based diagnostics and reveals evidence of recent reticulate evolution. *Sci. Rep.*, **6**, 22591.

Janssen,T. *et al.* (2017) The pitfalls of molecular species identification: a case study within the genus *Pratylenchus* (Nematoda: Pratylenchidae). *Nematology*, **19**, 1179–1199.

Nadler,S.A. (2002) Species delimitation and nematode biodiversity: phylogenies rule. *Nematology*, **4**, 615–625.

Nicol,J.M. *et al.* (2011) Current nematode threats to world agriculture. In Jones, J. T., Gheysen, G. & Fenoll, C. (eds.). *Genomics and Molecular Genetics of Plant-Nematode Interactions*. Heidelberg, Springer, p21–43.

Nilsson,R.H. *et al.* (2004) galaxieEST: addressing EST identity through automated phylogenetic analysis. *BMC Bioinformatics*, **5**, 87.

Pagan,C. *et al.* (2015) Mitochondrial haplotype-based identification of ethanol-preserved root-knot nematodes from Africa. *Phytopathology*, **105**, 350–357.

Powers,T. (2004) Nematode molecular diagnostics: from bands to barcodes. *Annu. Rev. Phytopathol.*, **42**, 367–383.

Qing,X. *et al.* (2018) A new species of *Malenchus* (Nematoda: Tylenchomorpha) with an updated phylogeny of the Tylenchidae. *Nematology*, **20**, 815–836.

Qing,X. *et al.* (2019) Phylogeography and molecular species delimitation of *Pratylenchus capsici* n. sp., a new root-lesion nematode in Israel on pepper

(*Capsicum annuum*). *Phytopathology,* **109,** 847–858.

Quist,C. *et al.* (2015) Evolution of plant parasitism in the phylum Nematoda. *Annu. Rev. Phytopathol*., **53**, 289–310.

Savolainen,V. *et al.* (2005) Towards writing the encyclopaedia of life: an introduction to DNA barcoding. *Philos. Trans. R. Soc. Lond., B, Biol. Sci*., **360**, 1805–1811.

Sekimoto,S. *et al.* (2017) Morphological and molecular characterisation of *Heterodera koreana* (Vovlas, Lamberti & Choo, 1992) Mundo-Ocampo, Troccoli, Subbotin, Del Cid, Baldwin & Inserra, 2008 (Nematoda: Heteroderidae) from bamboo in Japan. *Nematology*, **19**, 333–350.

Subbotin,S.A. *et al.* (2005) Phylogeny of Criconematina Siddiqi, 1980 (Nematoda: Tylenchida) based on morphology and D2-D3 expansion segments of the 28S-rRNA gene sequences with application of a secondary structure model. *Nematology,* **7**, 927–944.

Subbotin,S.A., Mundo-Ocampo, M. & Baldwin, J.G. (2010). Systematics of cyst nematodes (Nematoda: Heteroderinae). In Hunt, D.J. & Perry, R.N. (eds.). *Nematology Monographs and Perspectives 8A*. Leiden, The Netherlands, Brill, 351 pp.

Subbotin,S.A. *et al.* (2011) Diversity and phylogenetic relationships within the spiral nematodes of *Helicotylenchus* Steiner, 1945 (Tylenchida: Hoplolaimidae) as inferred from analysis of the D2-D3 expansion segments of 28S rRNA gene sequences. *Nematology,* **13**, 333–345.

Wijova,M. *et al.* (2005) Phylogenetic position of *Dracunculus medinensis* and some related nematodes inferred from 18S rRNA. *Parasitol. Res.,* **96**, 133–135.

Yaghoubi,A. *et al.* (2015) Description of *Atetylenchus minor* n. sp.(Tylenchina: Tylenchidae) and data on two other species of the family. *Nematology*, **17**, 981–994.