Ghent University

Faculty of Sciences

Department of Biology – Marine Biology

Academic Year 2015-2016

DNA barcoding and metabarcoding marine nematodes: from voucher to sequence and back

Febe Noppe

Master's dissertation submitted to obtain the degree of

Master of Science in Biology

Major Biodiversity and Evolutionary Biology

Promoter: Dr. Sofie Derycke

Co-promoter: Dr. Katja Guilini

Table of contents

1. Introduction……..4

2. Objectives……6

3. Material and methods……...7
3.1. Building a high quality database……...7
3.2 Pinpointing cryptic diversity……..9
3.3. Answering the meiofauna paradox……..10
3.4. Improving identification of meiofaunal communities.....10

4. Results……11
4.1. Database……...11
4.2. Pinpointing cryptic diversity (including phylogenetic analysis).....16
4.3. Answering the meiofauna paradox……..27
4.4. Improving identification of meiofaunal communities.....29

5. Discussion……37
5.1. Distance and threshold values……..37
5.2. Pinpointing cryptic diversity (including phylogenetic analysis).....37
5.3. Answering the meifauna paradox……....40
5.4. Improving identification of meiofaunal communities……40

6. Conclusion……...41

7. Summary……..43
7.1 English summary……...43
7.2 Dutch summary……46

8. Acknowledgments……...48

9. References…….....49

10. Appendix……...52
10.1. Taxonomic list……52
10.2. Excluded specimens…….....54
10.3. Species having both gene sequences available……55
10.4. Specimen couples with an interspecific distance value of zero for 18S……...55
10.5. Artificial community composition…….....58
10.6. QIIME commands……59
10.7. Complete maximum likelihood trees……60
10.8. Python scripts…….....72

1. Introduction

Nematodes account for 80-90% of all metazoans on Earth, yet only a fraction of the estimated more than 1 million species are formally known and described (Creer, et al., 2010). Free-living nematodes are dominant in both density (10^5-10^7 individuals per m^2) and diversity (> 10 species per cm^2) in marine sediments (Heip, et al., 1985). They fulfil important ecological roles, are a high quality food source for higher trophic levels (Leduc, 2009) and influence the composition of lower trophic groups (De Mesel, et al., 2004; Hamels, et al., 2001). Their eggs are deposited in situ, they do not have planktonic or pelagic larvae, and their small body size likely limits active dispersal over large distances, even though they can actively move in the sediment or swim in the water column (Jensen, 1981; Schratzberger, et al., 2004). However, many species seem to be cosmopolitan, with their distribution spanning the globe. This is a contra-intuitive curiosity, considering the small size and presumably low dispersing capacity of nematodes. This has been called the "meiofauna paradox" (Giere, 2008).

Identification is essential for ecological studies. One cannot unravel the complete complexity of a community without knowledge about the species it comprises, just as a chemist cannot completely understand a chemical reaction without knowing which chemical elements are involved. However, these biological units are not as easily definable as their chemical counterparts, and the definition of a "species" is still being discussed, often called "the species problem" (Byron J Adams, 2001; Van Regenmortel, 2010). This is a problem for a wide range of taxa, including nematodes, but the latter bumps into several other identification issues. These microscopic metazoans possess little diagnostic characters, which makes morphological identification very time consuming. This is often avoided by identifying only to genus, family or even just feeding guild. However, a considerable amount of information is lost this way (De Mesel, et al., 2004). Isolating randomly selected individuals from a sample is traditionally done by hand using a needle and a binocular microscope. When loaded onto a microscope slide, they can be identified with a microscope at 100x10x magnification, before extracting DNA. Because of their abundance, picking out every individual in a sample is far too time-expensive, so this method often misses less abundant species and represents only a fraction of the total amount of species. The presence of cryptic species (different species that are classified as a single species because they are morphologically indistinguishable), enlarges this problem even further (Bickford, et al., 2007; Lawton, et al., 1998). It might cause nematode diversity to be seriously underestimated (Apolonio Silva De Oliveira, et al., 2012).

Molecular tools provide a promising solution to the troublesome morphological identification, both in terms of speed and reliability, and have been rapidly advancing (Bhadury, et al., 2006; Floyd, et al., 2002; Porazinska, et al., 2009). It can identify all life stages, it is cost-efficient and it can be used to identify (environmental) samples containing multiple species (called 'metabarcoding') without the time-consuming handling of the individual species (Powers, 2004). DNA barcoding has become a popular identification tool (Cowart, et al., 2015; Lallias, et al., 2015). One or more genes, often the small subunit ribosomal DNA (18S) or the mitochondrial cytochrome oxidase c subunit 1 (COI), are sequenced and compared against sequences from identified specimens (reference sequences). If a close enough match is found (e.g. at least 95% identical for COI), the unknown specimen can be considered the same species as the one linked to the matched reference sequence. The idea of comparing gene sequences to each other to unravel phylogeny is not new. For example, it was through analysis of 16S rRNA that Carl R. Woese defined the Archaea as a separate kingdom (now "domain"), next to the Bacteria (in which it was previously included) and the Eukarya in 1977 (Woese & Fox, 1977). The concept of specimen identification by comparing its gene sequences against reference sequences is relatively recent. It got the name 'barcoding' only after Hebert *et al.* (2003) compared the concept with the use of Universal Product Codes. The latter is the twelve-digit code and its accompanying strip of black bars and white spaces, used to identify items with a quick scan at the store checkout.

However, DNA barcoding still has several weak points. Whether or not the specimens are considered the same species, depends on the similarity percentage that is used as threshold (and be considered a 'match'). This is often 95% for COI, and goes up to 99,5% for 18S. The choice of this similarity percentage in turn depends on the presence of a 'barcoding gap', a high degree of separation between intraspecific and interspecific distance distributions (variability in base pair composition of sequences of respectively the same or different species) (Hebert, et al., 2003). Errors and bias in sequence composition can also be introduced during amplification and sequencing. To be able to use barcoding identification at all, a reference database is needed that includes the DNA sequences linked to morphologically identified species. Sequences in public databases such as Genbank are often not identified to species level and do not have any guarantee for correct identification. To be able to verify the identification later on, a voucher must be included, containing images or videos of the specimen (Derycke, De Ley, et al., 2010). Such vouchers are however very often lacking. Moreover, the public databases are often biased towards the most popular genes. They provide for example a wide array of 18S sequences, but only a limited amount of COI sequences. As of April 29 2016, Genbank provoides 4773 COI and 19 478 18S sequences for nematodes, but most of these are parasitic, and only a small part are free-living marine nematodes.

18S rDNA has been traditionally used because of the availability of universal nematode primers, thanks to the conserved flanking regions of the sequence, and its phylogenetic resolution at genus and higher taxon levels (Floyd, et al., 2002; Meldal, et al., 2007). Its major downside is that it lacks resolution at the species level (De Ley, et al., 2005). To compensate for this shortcoming, it can be used in combination with the mitochondrial COI gene. Except for some taxa (e.g. Anthozoa, Porifera…), the COI gene has been proven efficient in identifying Metazoan species (Bowles, et al., 1992; Hebert, et al., 2003). However, nuclear copies of the COI gene ("numts"), often inactive and rapidly mutating, may cause an overestimation of the taxonomic diversity (Song, et al., 2008). Amplification success for free-living nematodes is known to give problems (De Ley, et al., 2005), but primers developed specifically for this phylum perform better and allow a range of nematode species to be amplified (Derycke, Vanaverbeke, et al., 2010).

The primers for the COI gene give different amplification results in two different regions of the gene: the M1-M6 (often called Folmer region) and I3-M11. The first is amplified with the so-called 'universal' invertebrate primers, but the latter has been proven to outperform the first, while still being reliable in identification and being able to reveal cryptic diversity (Derycke, Vanaverbeke, et al., 2010). However, identification might get confused because of the previously mentioned "numts", endosymbionts or contamination (Derycke, Vanaverbeke, et al., 2010). Heteroplasmy (presence of more than one type of mitochondrial genome within a cell or individual) and incomplete lineage sorting (different allele trees not matching the overall species pattern) can also obscure phylogenetic relationships. Public databases provide little COI sequences for nematodes, and mainly focus on 18S.

In this project, we will build a marine nematode reference database containing species all across the phylum that are identified, vouchered and sequenced for 18S and/or COI, from six different regions around the globe. We will calculate the intra-interspecific distance gap (the 'barcoding gap') for both genes based on p-distance. This p-distance has recently been found equally or even more suitable than the corrected Kimura-2-parameter (K2P) model (Collins, et al., 2012; Srivathsan & Meier, 2012), and calculate a threshold distance value for species identification. This allows us to identify presumably cosmopolitan "species" and track down cryptic species. Finally, we will test the applicability of both our database and the calculated threshold values for identification of nematode communities. This will be done using a metagenetic approach, with an artificial community with known species that will be compared against our database.

2. Objectives

1. Building a high quality database

Sanger sequences of COI and 18S from several studies will be combined. This includes both major clades and taxonomic levels of free-living marine nematodes, as well as closely related and cryptic species. Starting from 11 FAS and FASTA files, custom Python scripts will be written to set all sequence labels to a standardised version "code_location_genus_species". They will then be written to two total FASTA files, one for 18S and one for COI. Using further custom scripts, these two total FASTA files will be converted to one well-ordered table. This table will be used as a base for further goals, by calculating the intra-interspecific variability threshold for both genes. We expect to find a similar, yet more refined value as used in literature. For COI, this is approximately 0.05, but depends on the metrics used and the number of taxa sampled (Derycke, Vanaverbeke, et al., 2010). The threshold used for 18S varies from 0,04 (Lallias, et al., 2015) to 0,005 (Floyd, et al., 2002). All data of the gathered species is obtained from previous studies and submitted to public databases.

2. Pinpointing cryptic diversity

Having far and closely related species from a variety of regions will allow us to more clearly define the species boundaries used for barcoding. This can also be useful in tracking down cryptic species. By using both COI and 18S, we will search for high intraspecific distances and flag these as potential cryptic species. Afterwards, we will check the vouchers of these specimens to confirm or reject their identification as the same species, and conclude them to be cryptic or not. At the same time, this also allows us to track down any misidentifications.

3. Answering the meiofauna paradox

The small size and presumably low dispersing capacity of nematodes should intuitively result in differentiation through isolation in species with a wide to even global range. Species lists from different regions will be cross-referenced to find species that occur in multiple locations. P-distance will be calculated between specimens of the same species from different regions. We expect to find that former cosmopolitan "species" will often consist of multiple cryptic species (Boeckner, et al., 2009).

4. Improving identification of meiofaunal communities

The efficiency of the 18S short fragment (G18S4-22R primer set) and the COI I3-M11 region (JB2-JB5GED and JB3-JB5 primer sets) for identifying marine nematodes will be tested in a metagenetic setting. This will be done using an artificial marine nematode community of Ion Torrent sequences of known species and our calculated threshold values. Two replicates of the artificial community will be used, that are amplified using the three primer sets, in two replicate PCR runs for each, resulting in 12 PCR products. We expect 18S to be better amplified, but COI to be able to identify to a lower taxonomic level.

We will use an artificial marine nematode community of Ion Torrent sequences of known species, the efficiency of the 18S short fragment (G18S4-22R primer set) and the COI I3-M11 region (JB2-JB5GED and JB3-JB5 primer sets) for identifying marine nematodes will be tested in a metagenetic setting, using our calculated threshold values. Two replicates of the artificial community will be used, that are amplified using the three primer sets, in two replicate PCR runs for each, resulting in 12 PCR products. We expect 18S to be better amplified, but COI to be able to identify to a lower taxonomic level.

3. Materials and methods

3.1. Building a high quality database

Origin of the sequence data

All sequences used in this study were obtained from six previous studies, each with a different sampling sites: Cuba, Panarea (part of the Aeolian Islands, a volcanic island chain north of Sicily, Italy), Papua New Guinea, Paulina (polder by the Scheldt river, in the southwest of the Netherlands), Tunesia and Vietnam. An overview of the original publication of each location is given in Table 1. No new sampling was done in this study. However, to get familiar with the work leading to the sequence data, the process from specimen isolation, identification and vouchering, to DNA extraction and amplification was performed on 15 Brazilian specimens belonging to genera not yet present in our database.

Specimens were identified by professional taxonomists, vouchered as outlined in Derycke et al. (2010) and sequenced for 18S and/or COI with Sanger sequencing. The primer sets G18S4-4R (18S long fragment, 925bp) and JB3-JB5 (COI I3-M11 region, 426bp) were used for all locations, with the exception of Tunesia, where the G18S4-22R (18S short) primer pair was used for 18S. An overview of the primers used is given in Table 1, and their sequences Table 2. Both genes were sequenced for all locations, with the exception of Papua New Guinea, where no COI sequences were obtained. The author stated that "there was an attempt to amplify the COI fragment, but this was not successful". For the Panarea dataset, the G18S4-22R primer pair was used in addition to the previously mentioned one, amplifying the 18S short fragment of 400bp.

Location	**18S**	**COI**	**Original study**
Cuba	G18S4-4R (18S long)	JB3-JB5 (I3-M11 region)	Armenteros *et al.* 2014
Panarea	G18S4-4R (18S long) , G18S4-22R (18S short)	JB3-JB5 (I3-M11 region)	Unpublished master's thesis, Kanfra X. 2015
Papua New Guinea	G18S4-4R (18S long)	-	Unpublished master's thesis, D'Hont A. 2014
Paulina	G18S4-4R (18S long)	JB3-JB5 (I3-M11 region)	Unpublished master's thesis, Eche C. O. 2012
Tunesia	G18S4-22R (18S short)	JB3-JB5 (I3-M11 region)	(Unpublished, data gathered by Boufahja F.)
Vietnam	G18S4-4R (18S long)	JB3-JB5 (I3-M11 region)	Unpublished master's thesis, Nguyen Thi X. P. 2014

Table 1. The locations of origin of the sequence data, the corresponding primer sets that were used to amplify specimen DNA of that location and the original study providing all details.

Primer	**Sequence (5'-3')**	**Source**
G18S4 (F)	GCT TGT CTC AAA GAT TAA GCC	Blaxter *et al.* 1998
22R (R)	GCC TGC TGC CTT CCT TGG A	Blaxter *et al.* 1998
4R (R)	GTA TCT GAT CGC CKT CGA WC	Creer *et al.* 2010
JB3 (F)	TTT TTT GGG CAT CCT GAG GTT TAT	Bowles *et al.* 1992
JB5 (R)	AGC ACC TAA ACT TAA AAC ATA ATG AAA ATG	Derycke *et al.* 2005
JB2 (F)	ATG TTT TGA TTT TAC CWG CWT TYG GTG T	Derycke *et al.* 2007
JB5GED (R)	AGC ACC TAA ACT TAA AAC ATA RTG RAA RTG	Derycke *et al.* 2007

Table 2. Each primer used to obtain the data for this study (F= forward, R=reverse), its sequence and the publication of the primer (source).

For further details on materials and methods used, we refer to the corresponding publications or theses, listed in Table 1.

Creating the database

Sequences were provided as a FASTA file per location per gene, resulting in a total of 11 files. Before using this data, all sequence labels needed to be set to a standardized form: "code_location_genus_species". For this purpose, a custom Python script was written that corrected several common inconsistencies and errors in the sequence labels, and then wrote all the sequences with standardized label to a new FASTA file, one for 18S and one for COI (Appendix 10.8.1). Some last remaining errors and typos in the taxonomic names were corrected by hand. We then listed all genera represented in our data in a Microsoft Excel file, and for each added the higher taxonomic ranks (family, order and class), based on the World Database of Free-Living Marine Nematodes (NeMys: http://nemys.ugent.be/).

Next, another custom script was written that read the sequences from the total FASTA file one by one (Appendix 10.8.2). It recognised each unique voucher code and added the following information for each specimen:
- the genus and species name from the label
- the location from the label
- 2 checkboxes to indicate if there was a 18S and/or COI sequence available
- 2 columns for the 18S and/or COI sequence
- a column for each higher taxonomic rank, for which the correct information was searched in the previously created genus list

The script also replaced the PCR numbers of Panarea and Tunesia with the correct specimen code (listed in separate files) and added an "A" or an "F" behind the Papua New Guinea and Tunesia codes respectively, to avoid double specimen codes.

Sequence alignment

A Muscle alignment was made for both genes in Molecular Evolutionary Genetics Analysis version 6 (MEGA 6) (Tamura, et al., 2013) with default settings. In the 18S alignment, we saw that all sequences from the Vietnam dataset and four from the Panarea dataset were reversed and corrected this. Three different versions of the 18S alignment were then created: one full alignment (having 1133 bp), one run through a strict Gblocks (Castresana, 2000) filter (default parameters with allowed gap positions set to 'with half'; leaving only 163 bp) and one with a mild Gblocks run (Conserved Position: 231; Flank Position: 231; Contiguous Nonconserved Positions: 8; Min. Length Of A Block: 5; Allowed gap positions: with half; leaving 675 bp). The latter turned out to be the most suited one, with enough unreliable position deleted, without deleting too much of the sequence. For COI, we translated the sequences using invertebrate mitochondrial code before aligning and translated back to nucleotides after alignment. This ensured a correct gap/insertion placement, as the COI gene is a coding one. The first and last three base pairs were removed, to trim off the uninformative primer regions, and the last ones were sometimes separated from the sequence by a gap, as the one or two leftover base pairs did not form a complete, translatable codon.

The barcoding gap

The final alignments were then used to make an exploring neighbor-joining (NJ) tree in MEGA6, using default parameters, to check for abnormal clustering and identification errors. P-distances (Collins, et al., 2012; Srivathsan & Meier, 2012) were also calculated. For this, sequences from unidentified specimens or sequences that were too short to give an overlap with at least one other sequence were removed. There were also some specimens only identified to genus level from which we could not be sure if they were the same species as another one with the same name, also only identified to the genus level (for example two specimens named "Daptonema_sp" from different locations). The vouchers of these specimens were

consulted to resolve these cases. Two specimens that were identified as different species based on the vouchers have been renamed: 41P_Viet_Dichromadora_sp1 was renamed to "sp3" and 103P_Viet_Oxystomina_sp to "sp1". Eight 18S sequences showing strange clustering on the exploring neighbor-joining tree were checked against Genbank (BLAST, Basic Local Alignment Search Tool) and turned out to be fish DNA contamination. These were excluded from further analysis. A list of specimens left out when the problem could not be resolved is given in Appendix10.2. The resulting distance matrices were then used in the program ExCaliBAR, which sorts the pairwise distance comparisons into a file containing intra- and interspecific distances. To be able to use the alignment in ExCaliBAR, locations were deleted from the sequence labels before calculating the distance, because the program would otherwise interpret the locations as genus and the genus names as species. Histograms contrasting the intra- and interspecific distances were created in Excel.

The R package *Adhoc* was then used to calculate the distance threshold, relying on an estimated probability of relative identification error (Sonet, et al., 2013). We used 0.05 as maximum significant relative error value, and set the "ambiguous" option to "correct". A short script was written to convert the sequence labels to a form with genus and species name first, that could be read by the program. We did not find a significant overall value for 18S. Because a variable substitution rate for different groups could be the cause (Holterman, et al., 2006), we looked at the 2 best represented families from each of the 3 best represented orders (containing more than 100 specimens) separately (a complete list of the number of specimens for each family in these orders is given in Appendix 10.1): Chromadoridae, Cyatholaimidae, Oncholaimidae, Oxystominidae, Sphaerolaimidae and Xyalidae. The sequences from all specimens belonging to each of these six families were extracted from the alignment using FaBox 1.41 (Villesen, 2007). We then calculated the distances again, sorted them with ExCaliBAR, made the histograms and ran *Adhoc* on them.

3.2. Pinpointing cryptic diversity

Intraspecific COI values higher than 0.05 (often used threshold value for COI; (Derycke, Vanaverbeke, et al., 2010), were marked as potentially pointing to cryptic species. This threshold was chosen based on previous results from marine nematodes (Derycke, Vanaverbeke, et al., 2010). We applied three "double evidence" rules to decide if the considered specimens were different cryptic species or not, in decreasing evidence value: 1) the intraspecific distance value was high for both COI and 18S (sequence divergence of one gene through chance can be ruled out), 2) the compared specimens having a high distance value came from different regions (divergence through isolation), 3) the high distance values showed consistent patterns that divided the specimens in clear groups (the same small values within and large values between groups). The suspected different cryptic species were indicated as such in the alignments by adding a letter to the species name (for example "Genus_spB" is indicated as a separate cryptic species from "Genus_sp").

The potential cryptic species were listed and the vouchers of these specimens were checked to confirm or reject their identification as the same species, and conclude them to be cryptic or not. After flagging cryptic species, p-distances were recalculated, sorted with ExCaliBAR, and used to create the histograms of intra- and interspecific distances. A new NJ with default parameters was build using MEGA6, to check if the flagged cryptic species clustered according to the group they were assigned to. A maximum-likelihood (ML) tree was also built with MEGA6, using the T92+G as best-fit substitution model (ML model search in MEGA6), 100 bootstrap replications and further default parameters, to study the clustering and compare with the NJ tree. The vouchers of specimens that did not cluster as expected were checked to validate their identification. A BLAST against Genbank sequences was done in addition to the voucher check, or if there was no voucher available.

3.3. Answering the meiofauna paradox

A custom Python script was written that listed all species for each location for each gene. These lists were then provided to the Venn diagram web tool available on the bioinformatics site of the UGent (http://bioinformatics.psb.ugent.be/webtools/Venn/). This web tool gives an overview of the number of species per location and the species shared between locations, the latter visualized in a Venn diagram. The same was done for the species lists of both genes per location, to generate an overview of the species that had both sequences (see Appendix 10.3). The species that were identified as shared species between locations were then checked using the vouchers. Extra care was taken with species that were only identified to genus level, because two researchers from different locations can both call a specimen "Dichromadora sp1", without them necessarily being the same species. Species that were contamination, wrongly identified or had no voucher available could not be used for inter-location comparison. Some other species that could not be identified with certainty as the same species based on the vouchers, were also excluded. A full list of the left out specimens is given in Appendix 10.2. The p-distance of species occurring in multiple locations was calculated to identify cryptic diversity in different locations.

3.4. Improving identification of meiofaunal communities

An artificial nematode community (mock community) was used to test the efficiency of 18S and COI barcoding as identification method in comparison to each other. This was done using the previously calculated threshold values and the ones often used in literature. A varying but known volume of PCR product of 50 different species (including four different strains for two species, *Litoditis marina* and *Halomonhystera disjuncta*, comprising a total of 56 specimens) was collected in an Eppendorf tube. Each of the specimens used was identified by a professional taxonomist and vouchered. This represented far and closely related species present in a meiofaunal community in varying abundances, using volume of DNA extract varying from minimally one to maximally six µl per replicate. Two replicates were made of this artificial community ("A" and "B") and three primer sets were used for each: G18S4-22R (18S short fragment), JB3-JB5 (I3-M11 region COI) and the degenerated JB2-JB5GED (I3-M11 region COI), amplifying a fragment of 400, 364 and 393bp respectively. Lastly, 2 PCR replicates were made for each combination ("1" and "2"), resulting in 12 PCR runs, each marked with a unique barcode in front of the primer to allow distinction. The PCR products were sequenced using Ion Torrent. The resulting Standard flowgram format (SFF) file was converted to a FASTQ format and a quality control was conducted. The resulting sequences were then provided to this study as three FASTA files, one for each primer set.

The obtained sequences were then run through the QIIME bioinformatics pipeline(Caporaso, et al., 2010), performing a open-reference OTU (Operational Taxonomic Unit) picking. During this process, reads are clustered against a reference sequence collection. Any reads which do not hit the reference sequence collection are subsequently clustered de novo. The reference collection is either our own database of 18S and COI sequences, or the Silva database (http://www.arb-silva.de), containing 29669 eukaryote including 1268 nematodes sequences, for 18S. This clustering was done using our calculated threshold values (objective 1) and a number of threshold values that have been used in previous studies. The clustering for 18S will be done on 96%, 97%, 99% and 99,5% sequence similarity (Armenteros, et al., 2014; Floyd, et al., 2002; Lallias, et al., 2015; Ratnasingham & Hebert, 2007). For COI, 93%, 95% and 96,66% (Armenteros, et al., 2014; Derycke, Vanaverbeke, et al., 2010; Meier, et al., 2006) will be used in addition to our threshold value. Taxonomy was assigned to the OTU's using the reference sequence collection and a file containing the taxonomy. Two different algorithms were used: UCLUST (Edgar, 2010), dividing the sequences into clusters, and BLAST (Altschul, et al., 1990), comparing by aligning. Sequences that did not return a significant hit (<90% sequence similarity) were labelled as "Unassigned" for UCLUST and "No blast hit" for BLAST. The resulting biological observation matrix (BIOM) file gives us a table of each OTU, its taxonomy and the number of sequences assigned to it for each community replicate. This file was converted to a text file and opened in Microsoft Excel for

examination. The assigned taxonomy was then summarized using the script "summarize_taxonomy_through_plots.py", providing us a list of all identified taxa and the proportion of sequences assigned to it for each of the mock community replicates. Each of these taxa were cross-referenced against the composition of the artificial community, to check how well it was identified. In a last step, these results were visualized in graphs. An overview of the QIIME commands used per script are given in Appendix 10.6.

4. Results

4.1. Database

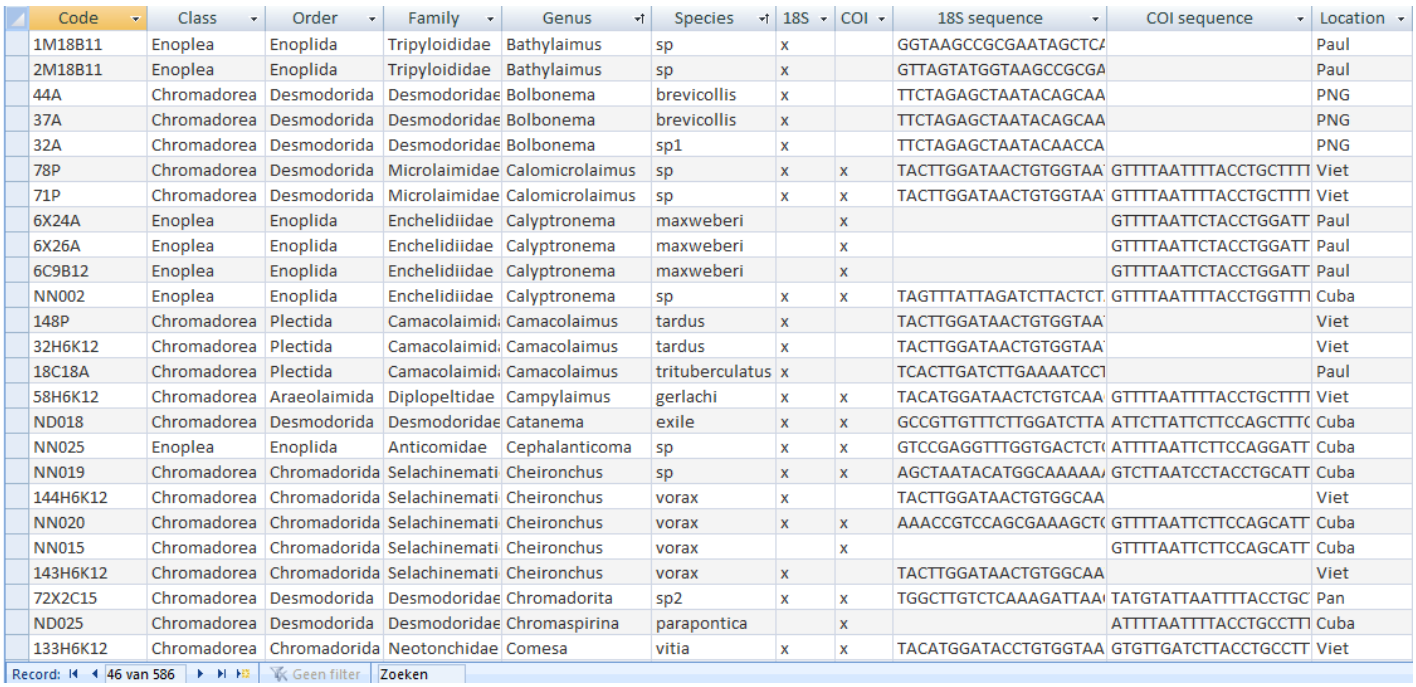
The database contains 586 specimens and 756 sequences (461 for 18S and 295 for COI), including representatives of 115 genera from 37 families. The number of specimens and sequences for each location is given in Table 3.

Location	Number of 18S sequences	Number of COI sequences	Total number of specimens
Cuba	27	34	37
Panarea	54	35	55
Paulina	101	66	147
Papua New Guinea	30	-	30
Tunesia	24	53	69
Vietnam	225	107	248
Total	461	295	586

Table 3. An overview of the six locations that were sampled, showing the number of 18S and COI sequences and the number of specimens available in our database per location. 170 specimens had both gene sequences. There were no COI sequences available for Papua New Guinea.

Three specimens in our database were unidentified. A few specimen duplicates were removed.

The table resulting from the Python scripts was saved in text format so it could be opened in several programs. By default, it was ordered by specimen code, but opened in another program like Microsoft Access or Excel, sorting can be done in any preferable way (see Fig. 1).



Code	Class	Order	Family	Genus	Species	18S	COI	18S sequence	COI sequence	Location
1M18B11	Enoplea	Enoplida	Tripyloididae	Bathylaimus	sp	x		GGTAAGCCGCGAATAGCTCA		Paul
2M18B11	Enoplea	Enoplida	Tripyloididae	Bathylaimus	sp	x		GTTAGTATGGTAAGCCGCGA		Paul
44A	Chromadorea	Desmodorida	Desmodoridae	Bolbonema	brevicollis	x		TTCTAGAGCTAATACAGCAA		PNG
37A	Chromadorea	Desmodorida	Desmodoridae	Bolbonema	brevicollis	x		TTCTAGAGCTAATACAGCAA		PNG
32A	Chromadorea	Desmodorida	Desmodoridae	Bolbonema	sp1	x		TTCTAGAGCTAATACAACCA		PNG
78P	Chromadorea	Desmodorida	Microlaimidae	Calomicrolaimus	sp	x	x	TACTTGGATAACTGTGGTAA	GTTTTAATTTTACCTGCTTTT	Viet
71P	Chromadorea	Desmodorida	Microlaimidae	Calomicrolaimus	sp	x	x	TACTTGGATAACTGTGGTAA	GTTTTAATTTTACCTGCTTTT	Viet
6X24A	Enoplea	Enoplida	Enchelidiidae	Calyptronema	maxweberi		x		GTTTTAATTCTACCTGGATT	Paul
6X26A	Enoplea	Enoplida	Enchelidiidae	Calyptronema	maxweberi		x		GTTTTAATTCTACCTGGATT	Paul
6C9B12	Enoplea	Enoplida	Enchelidiidae	Calyptronema	maxweberi		x		GTTTTAATTCTACCTGGATT	Paul
NN002	Enoplea	Enoplida	Enchelidiidae	Calyptronema	sp	x	x	TAGTTTATTAGATCTTACTCT	GTTTTAATTTTACCTGGTTTT	Cuba
148P	Chromadorea	Plectida	Camacolaimid	Camacolaimus	tardus	x		TACTTGGATAACTGTGGTAA		Viet
32H6K12	Chromadorea	Plectida	Camacolaimid	Camacolaimus	tardus	x		TACTTGGATAACTGTGGTAA		Viet
18C18A	Chromadorea	Plectida	Camacolaimid	Camacolaimus	trituberculatus	x		TCACTTGATCTTGAAAATCCT		Paul
58H6K12	Chromadorea	Araeolaimida	Diplopeltidae	Campylaimus	gerlachi	x	x	TACATGGATAACTCTGTCAA	GTTTTAATTTTACCTGCTTTT	Viet
ND018	Chromadorea	Desmodorida	Desmodoridae	Catanema	exile	x	x	GCCGTTGTTTCTTGGATCTTA	ATTCTTATTCTTCCAGCTTTC	Cuba
NN025	Enoplea	Enoplida	Anticomidae	Cephalanticoma	sp	x	x	GTCCGAGGTTTGGTGACTCT	ATTTTAATTCTTCCAGGATT	Cuba
NN019	Chromadorea	Chromadorida	Selachinemati	Cheironchus	sp	x	x	AGCTAATACATGGCAAAAA	GTCTTAATCCTACCTGCATT	Cuba
144H6K12	Chromadorea	Chromadorida	Selachinemati	Cheironchus	vorax	x		TACTTGGATAACTGTGGCAA		Viet
NN020	Chromadorea	Chromadorida	Selachinemati	Cheironchus	vorax	x	x	AAACCGTCCAGCGAAAGCT	GTTTTAATTCTTCCAGCATT	Cuba
NN015	Chromadorea	Chromadorida	Selachinemati	Cheironchus	vorax		x		GTTTTAATTCTTCCAGCATT	Cuba
143H6K12	Chromadorea	Chromadorida	Selachinemati	Cheironchus	vorax	x		TACTTGGATAACTGTGGCAA		Viet
72X2C15	Chromadorea	Desmodorida	Desmodoridae	Chromadorita	sp2	x	x	TGGCTTGTCTCAAAGATTAA	TATGTATTAATTTTACCTGC	Pan
ND025	Chromadorea	Desmodorida	Desmodoridae	Chromaspirina	parapontica		x		ATTTTAATTTTACCTGCCTTT	Cuba
133H6K12	Chromadorea	Chromadorida	Neotonchidae	Comesa	vitia	x	x	TACATGGATACCTGTGGTAA	GTGTTGATCTTACCTGCCTT	Viet

Record: 46 van 586 | Geen filter | Zoeken

Figure 1. A screenshot of the database layout in Microsoft Access, with the specimens sorted by scientific name.

The barcoding gap

2,24% and 9,28% of the intraspecific distance values were higher than 0,05 for respectively 18S and COI. 31,49% of the intraspecific distance values for 18S were higher than 0,01. More than half of the intraspecific values, respectively 57,98% and 55,68% for 18S and COI, were equal to zero. Intraspecific distances ranged from 0-0,1939 and from 0-0,2883 for 18S and COI respectively. Interspecific distances ranged from 0-0,3820 and from 0,0025-0,5455 for 18S and COI respectively. Interspecific minimum values equal to zero were found both within (e.g. *Terschellingia longicaudata* vs. *T. sp.nov*) and between (*Tubolaimoides* sp. vs. *Paracanthonchus* sp2) genera for 18S. The interspecific distance values for COI smaller than 0,05 were within genus in all cases and only for *Metachromadora* and *Sphaerotheristus*. A complete list of specimen comparisons that yielded an interspecific distance value of zero is given in Appendix 10.4. A histogram for both genes, comparing the overall intra- and interspecific distances for each is given in Fig. 2 .

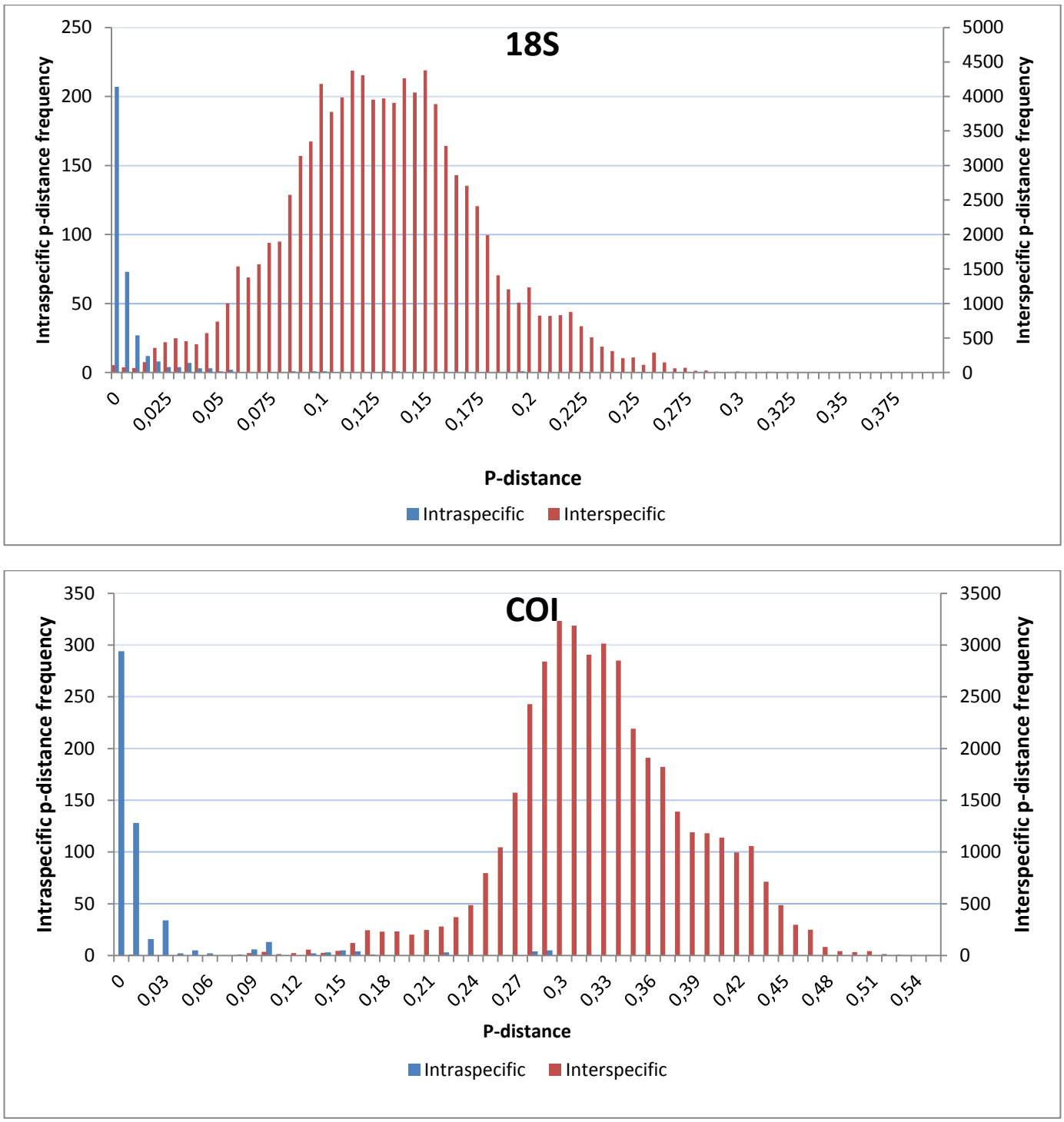


Figure 2. P-distance histograms for 18S (above) and COI (below), showing the frequency of each distance value if all specimens are pair-wise specimen compared. The primary (intraspecific, blue) and secondary Y-axis (interspecific, red) have a different scale.

The intra- and interspecific distances show a relatively large overlap for 18S, suggesting the threshold value for species discrimination is most likely somewhere between 0,015 - 0,05. The overlap in COI is much smaller. 95,80% of the intraspecific values for 18S were lower than 0,04 and 94,98% of interspecific values were higher than 0,205. For COI, 94,89% of the intraspecific values were lower than 0,1 and 95,66% of the interspecific values were higher than 0,21.

Using linear regression, *Adhoc* found a significant threshold value of 0,0562 for COI, with a relative error (RE; the number of incorrect identifications divided by the total number of identifications) of less than 0,05 (see Fig. 3). For 18S, we could not find a significant value. Here the lowest distance value (0,00597)

other than zero still had a RE of 0,247, the lowest to be found. This means that a quarter of all identifications done using this distance value would be wrong.

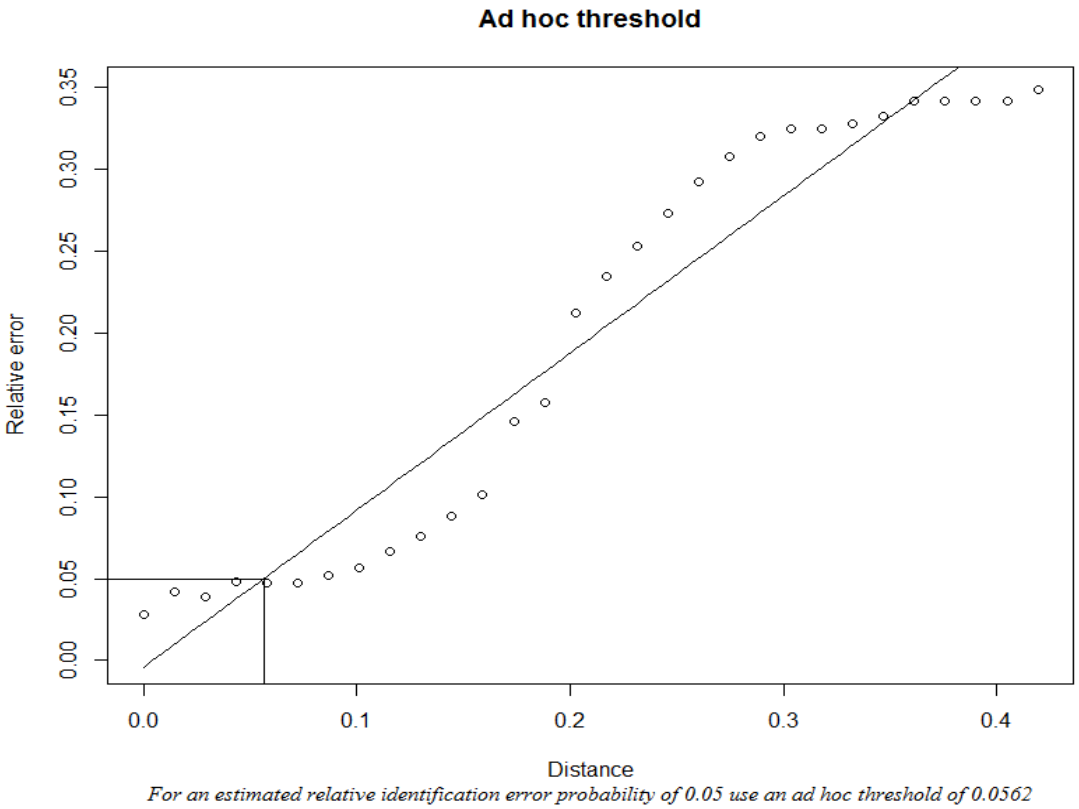


Figure 3. The calculated relative error for different distance values by *Adhoc* for COI. Using linear regression, the program found a threshold value of 0,0562 to be the most suited to use as species boundary for identification.

As we did not find a significant threshold value for 18S, we looked at six major families separately. Histograms of the calculated p-distances are given in Fig. 5 and the distance value(s) with the lowest RE listed by *Adhoc* are given in Table 4 for all six families. Only for Oncholaimidae and Oxystominidae we found a significant threshold value of respectively 0,0038 and 0,0016 (see Fig. 4). The histograms of the six families did not all approach the form of the total histogram for 18S. Xyalidae seems to approach the ideal expected form the best, with a clear gap and little overlap between two nicely distributed curves (Fig. 5f). Chromadoridae and Oncholaimidae (Fig. 5a and c) histograms also resulted in elegant graphs, answering our expectations. The histogram of Cyatholaimidae (Fig. 5b), with a long tail of high interspecific values, hangs somewhere in the middle between the previous two and Oxystominidae (Fig. 5d) which is more sloppy, with the two curves smeared in a large overlap. Lastly, the histogram of Sphaerolaimidae (Fig. 5e) is too chopped up to give us much useful information.

The threshold values for Oncholaimidae and Oxystominidae are both very low, but the first value is still more than the double of the latter. For the other families (except for Cyatholaimidae; see further), all minimum RE's of the lowest distance values given by the program are larger than the considered significant 0,05. For Cyatholaimidae, the minimum RE of 0.0357 counts for a very large interval of distance values (Table 4), that probably resulted in a clustering of distance values to one side in the regression graph, preventing the program to use the graph to find a suitable threshold value.

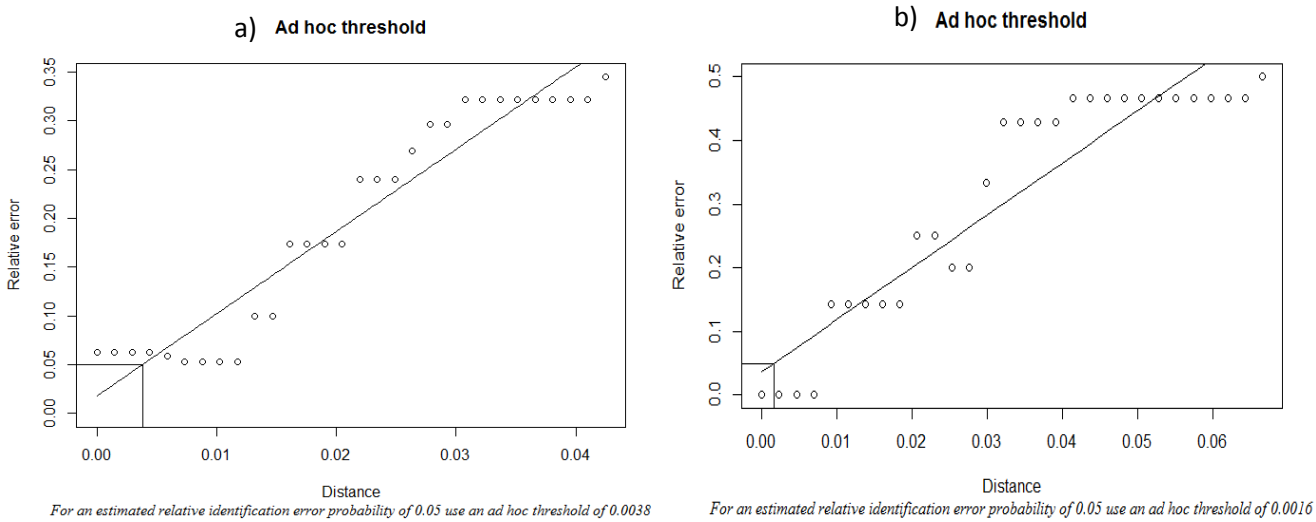


Figure 4. The calculated relative error for different distance values within (a) Oncholaimidae and (b) Oxystominidae by *Adhoc*. Using linear regression, the program found a threshold value of respectively 0,0038 and 0,0016 to be the most suited to use as species boundary for identification.

Family	**Distance value**	**Relative error**
Chromadoridae	0.002490	0.238095
Cyatholaimidae	[0.006121- 0.012242]	0.035714
Oncholaimidae	[0.007320 - 0.011711]	0.052632
Oxystominidae	[0.002296 - 0.006889]	0
Sphaerolaimidae	[0.005266 - 0.014746]	0.071429
Xyalidae	[0.009278 - 0.013918]	0.121212

Table 4. The 6 major families we used in Adhoc for 18S. For each family, the distance value other than zero with the smallest relative error is given. If there were multiple distance values for the smallest relative error, an interval containing these is given.

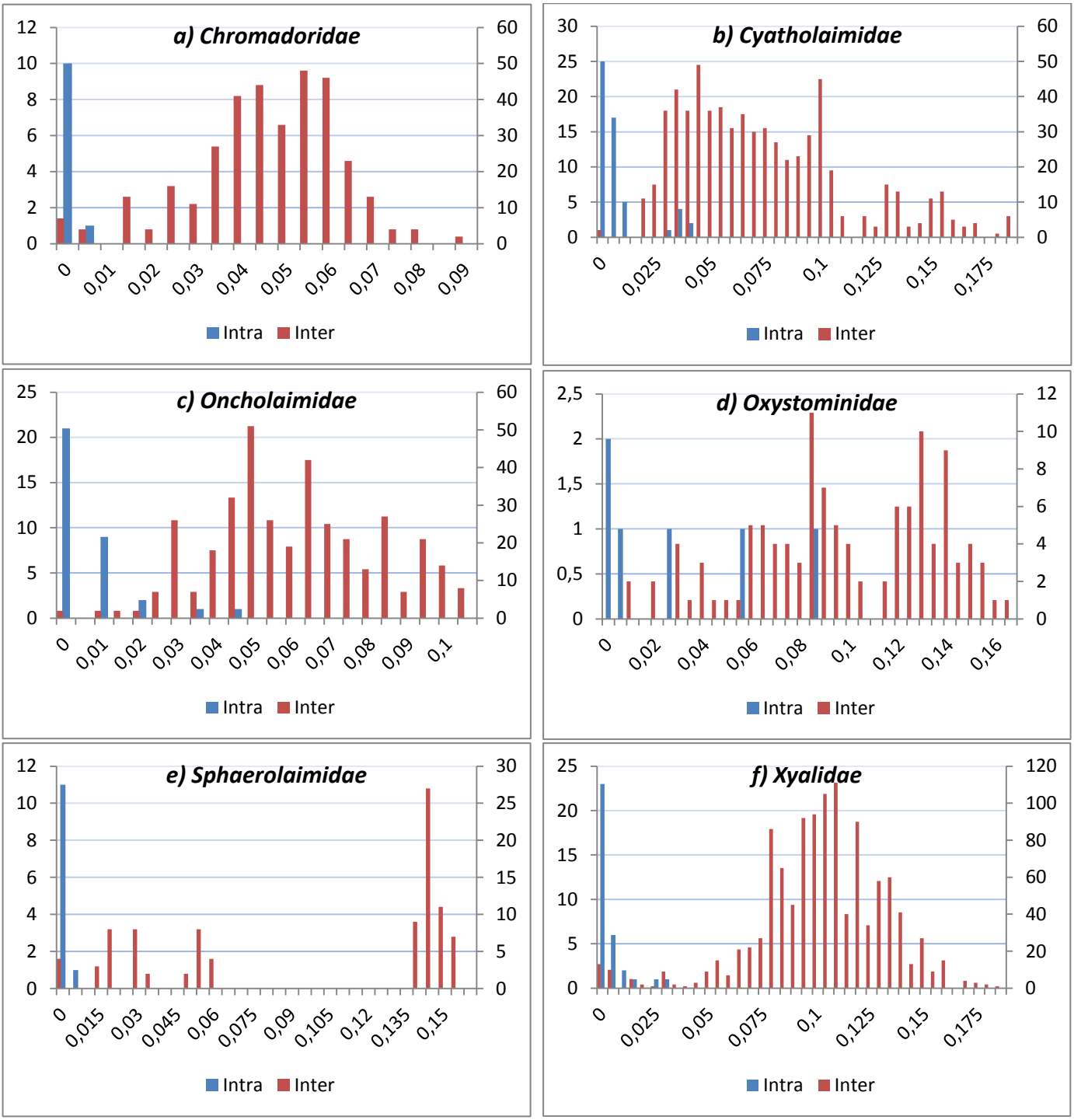


Figure 5. The p-distance histograms for the six families for 18S, showing the frequency of each distance value if all specimens are pair-wise specimen compared. The primary (intraspecific, blue) and secondary Y-axis (interspecific, red) have a different scale.

4.2. Pinpointing cryptic diversity

Phylogenetic analysis

An overview of the most important clusters in the maximum likelihood (ML) trees for both genes are given on the next pages (Fig 6 - 15). The bootstrap values (BV) of the corresponding neighbor-joining (NJ) tree are provided between brackets. The complete trees can be consulted in Appendix 10.7.

For 18S, 16 clusters were strongly supported (BV ≥ 95) and five were well supported (BV ≥ 90). For COI, 59 clusters were strongly supported (BV ≥ 95), three were well supported (BV ≥ 90). Not only were more strongly supported clusters found for COI as opposed to 18S, but specimens of the same genus,

family or order more often clustered together without any specimens not belonging to the group (outsiders). This is especially the case for Araeolaimida. In the COI ML tree, almost all COI sequences from species from this order grouped together in two clusters, one for the family Comesomatidae and one for the Axonolaimidae, without any outsiders. The only two specimens that did not cluster in one of these, were 49H6K12 (*Pseudolella* sp., Fig. 9a)) and 58H6K12 (*Campylaimus gerlachi*, Fig. 9c)), who clustered within the Monhysterida. No vouchers were available for these specimens, so we could not validate their identification. A BLASTx (comparing translated sequences) against Genbank suggested the 49H6K12 specimen to be *Daptonema* or *Trichotheristus*, which is perfectly in line with the specimens it clusters with. For the 58H6K12 specimen, the BLASTx yielded no results with a sequence similarity higher than 70%. This specimen clustered in a different way for both genes, so we cannot give an alternative identification with any confidence. A similar case was specimen NN004 from Cuba (see Fig. 6e and Fig. 7d). This specimen was named *Valvaelaimus* sp. for 18S and *Sphaerolaimus* sp. for COI, and clustered in the Enoplida, to which neither genera belong. Naturally, at least one of these identifications was wrong. No voucher was available for this specimen, but a BLASTx suggested *Bathylaimus* or *Tripyloides* to be the correct genus, matching the specimens it clusters with.

Few large clusters contained only specimens from one family like for the Comesomatidae and the Axonolaimidae, but some remarkably large clusters were formed that contained only specimens from one particular order. The best example for 18S is the Enoplida (Fig. 6a), but some moderately large clusters are also found in the Monhysterida. The COI ML tree yielded more of these large single-order clusters, as found in the Monhysterida (Fig. 7a), Chromadorida (Fig. 10a) and Desmodorida (Fig. 11a). For some orders, the COI tree was nearly the only one yielding well supported clusters. This is the case for Chromadorida (Fig. 10) and Desmodorida (Fig. 11), which did each only give one small 18S cluster, and for Araeolaimida, for which none were found.

For Rhabditida (Fig. 14) and Plectida (Fig. 15), there was only one strongly supported cluster found. For the former, it was for 18S and for the latter for COI. For the Rhabditida, this is because there were only 18S sequences available for these specimens, and for the Plectida because the 18S cluster did not have a high enough bootstrap value.

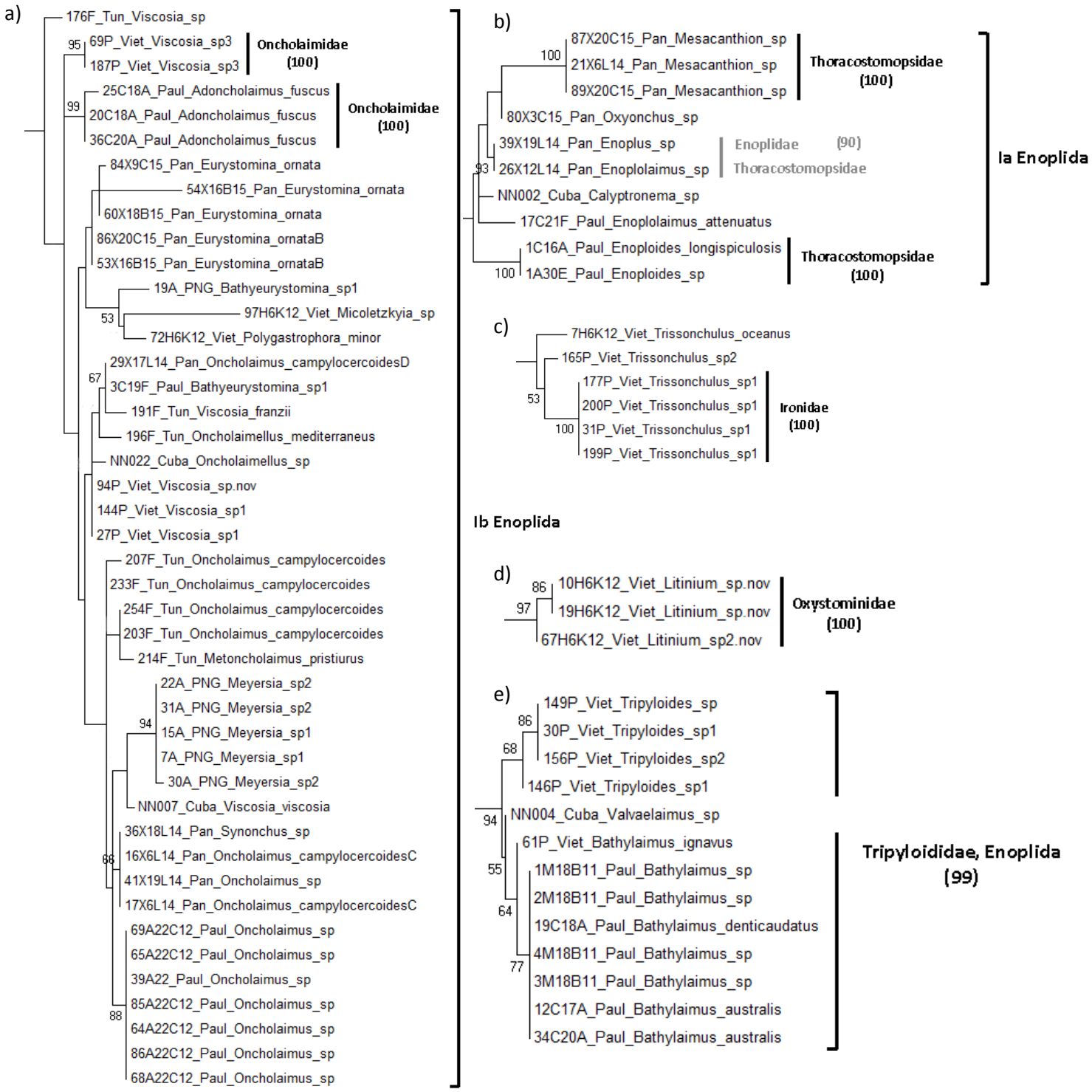


Figure 6. The most important clusters of the 18S ML tree for specimens within the order Enoplida. Specimen NN004 (*Valvaelaimus* sp. From Cuba)* in cluster e) is the only specimen not belonging to this order. Bootstrap values (BV) of the corresponding NJ tree are given between brackets. Clusters that are supported by a BV of at least 95 for both trees (ML and NJ) are indicated in by a black line. Clusters that are supported by a BV of at least 90 are indicated by a grey line. Clusters with specimens belonging to the same order are indicated by a square bracket. BV's smaller than 50 are not shown.

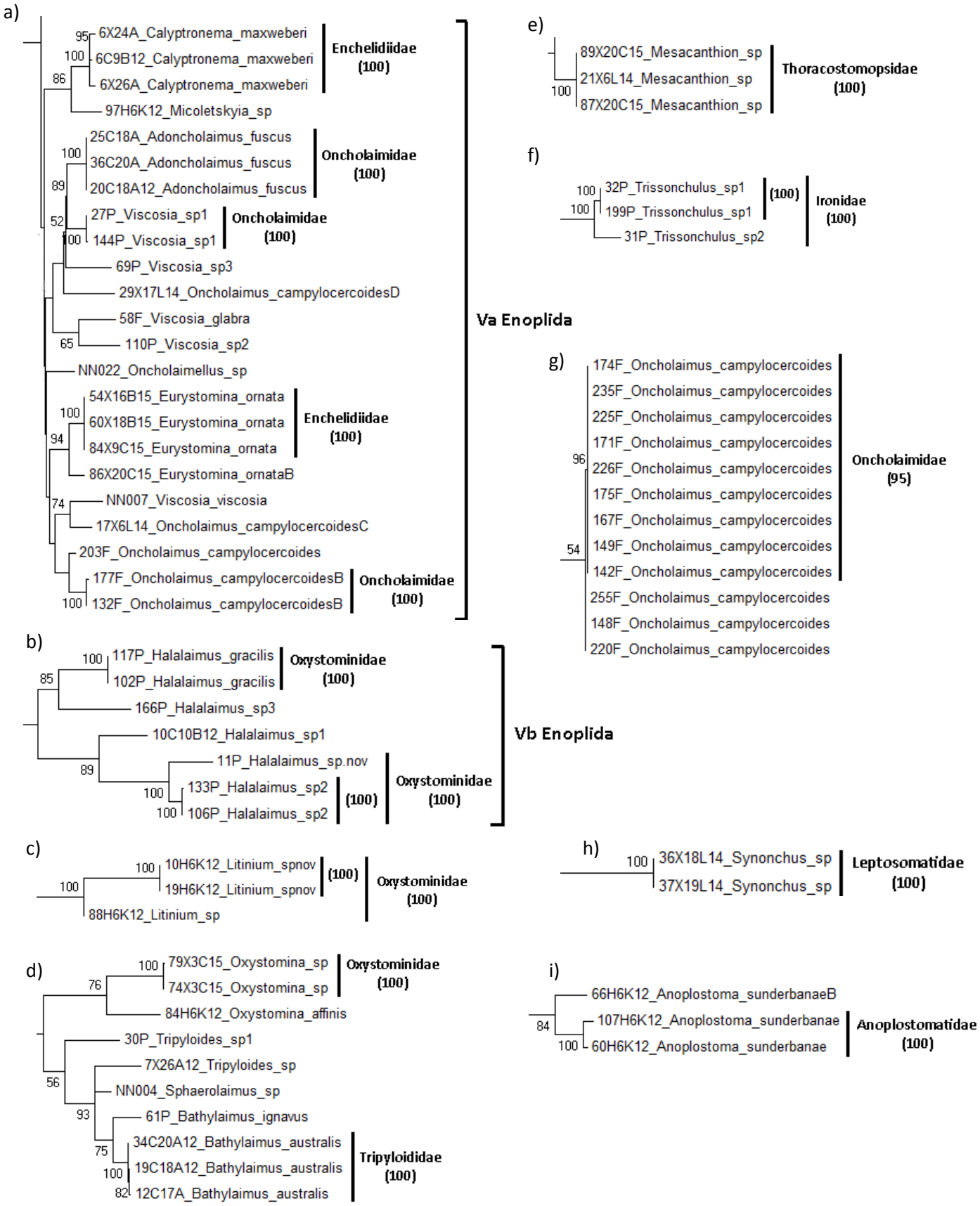


Figure 7. The most important clusters of the COI ML tree for specimens within the order Enoplida. Specimen NN004 (*Sphaerolaimus* sp. From Cuba)* in cluster d) is the only specimen shown not belonging to this order. Bootstrap values (BV) of the corresponding NJ tree are given between brackets. Clusters that are supported by a BV of at least 95 for both trees (ML and NJ) are indicated in by a black line. Clusters with specimens belonging to the same order are indicated by a square bracket. BV's smaller than 50 are not shown.

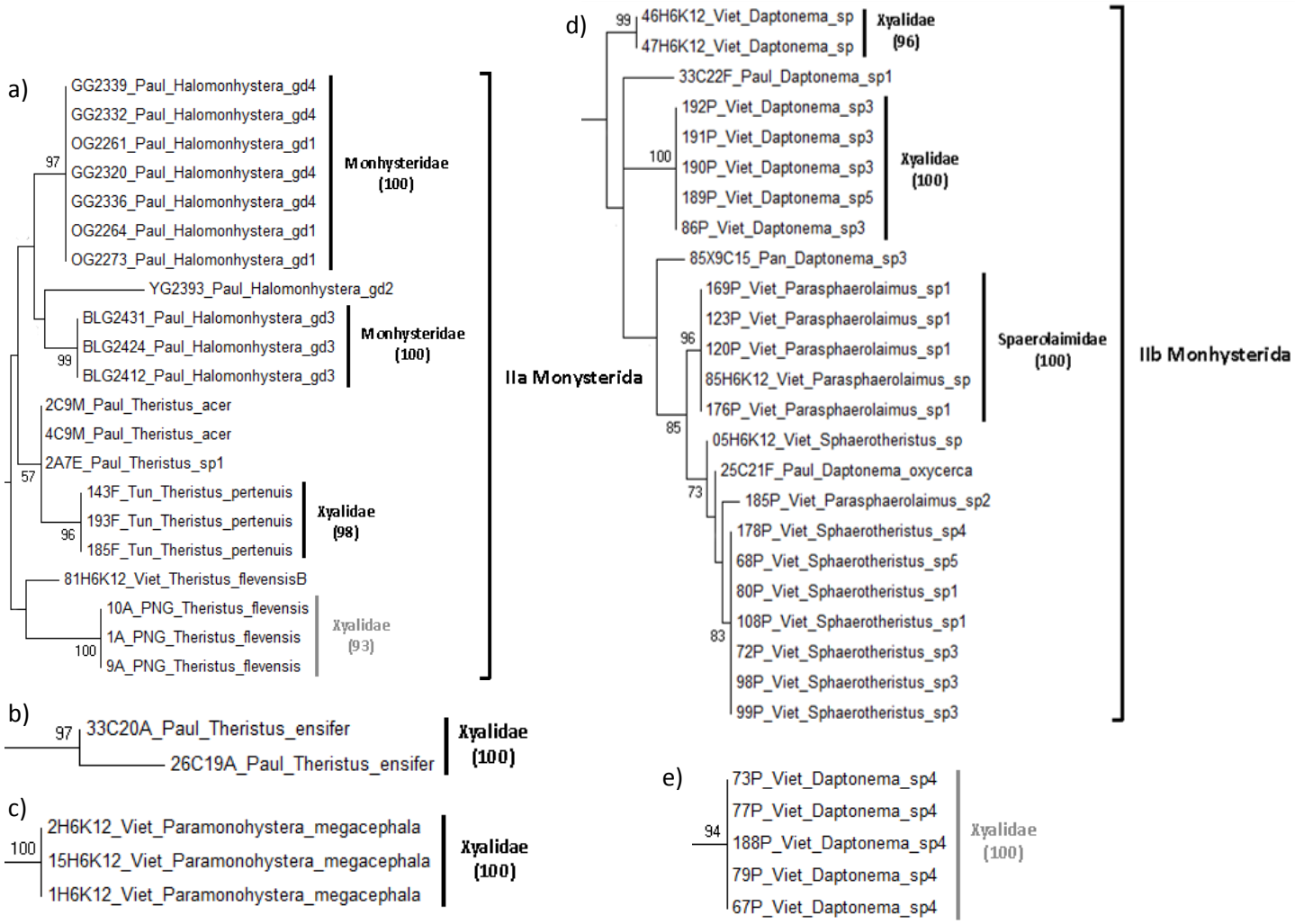


Figure 8. The most important clusters of the 18S ML tree for specimens within the order Monhysterida. Bootstrap values (BV) of the corresponding NJ tree are given between brackets. Clusters that are supported by a BV of at least 95 for both trees (ML and NJ) are indicated in by a black line. Clusters that are supported by a BV of at least 90 are indicated by a grey line. Clusters with specimens belonging to the same order are indicated by a square bracket. BV's smaller than 50 are not shown.

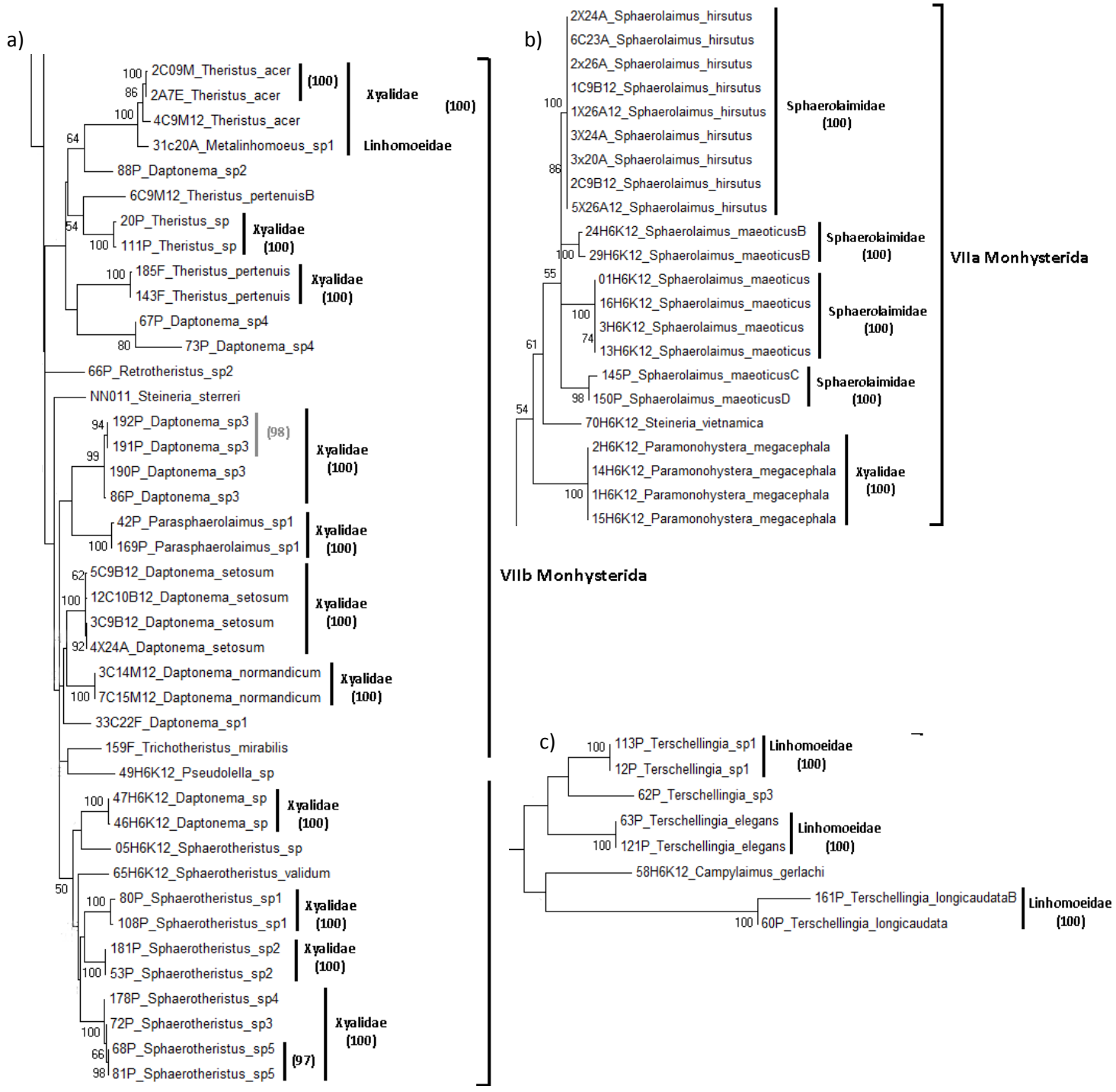


Figure 9. The most important clusters of the COI ML tree for specimens within the order Monhysterida. Specimens 49H6K12 (*Pseudolella* sp. from Vietnam, cluster a)) and 58H6K12 (*Campylaimus gerlachi* from Vietnam, cluster c)) are the only specimens shown here that do not belong to this order, but to the Araeolaimida. Bootstrap values (BV) of the corresponding NJ tree are given between brackets. Clusters that are supported by a BV of at least 95 for both trees (ML and NJ) are indicated in by a black line. Clusters with specimens belonging to the same order are indicated by a square bracket. BV's smaller than 50 are not shown.

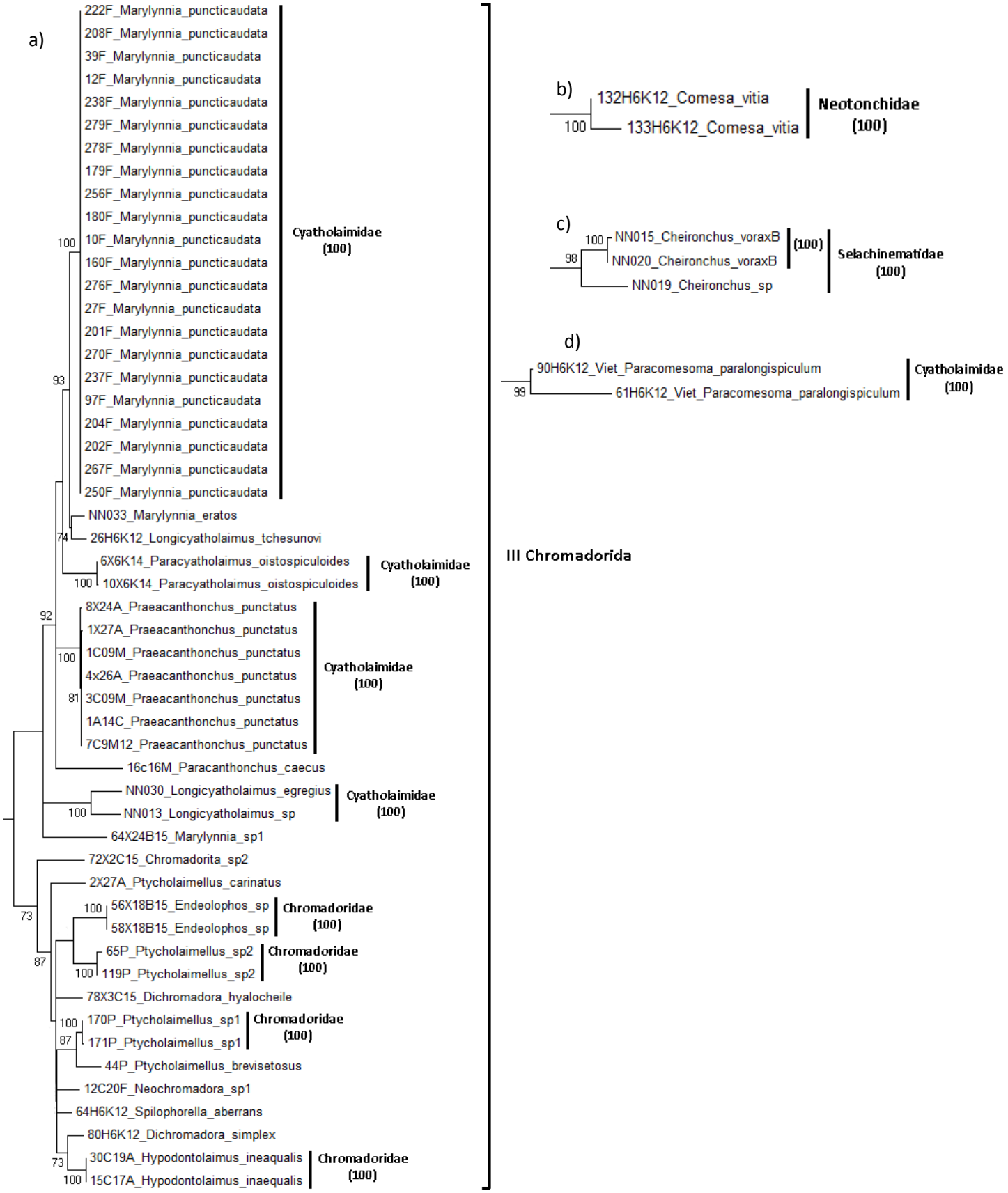


Figure 10. The most important clusters of the ML trees for specimens within the order Chromadorida, with only one strongly supported cluster for 18S within this order. COI: a-c, 18S: d. Bootstrap values (BV) of the corresponding NJ tree are given between brackets. Clusters that are supported by a BV of at least 95 for both trees (ML and NJ) are indicated in by a black line. Clusters with specimens belonging to the same order are indicated by a square bracket. BV's smaller than 50 are not shown.

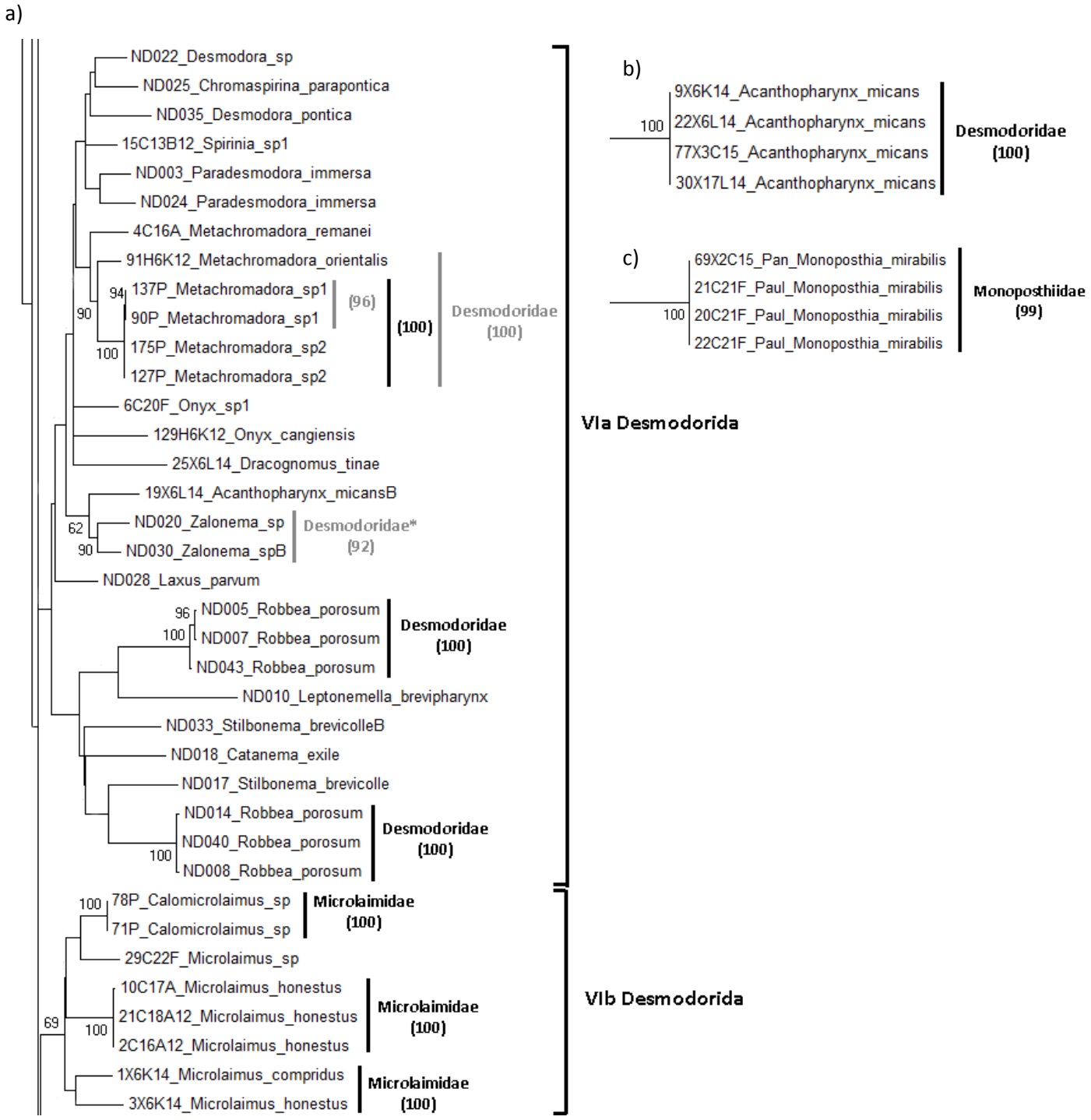


Figure 11. The most important clusters of the ML trees for specimens within the order Desmodorida, with only one strongly supported cluster for 18S within this order. COI: a-b, 18S: c. Bootstrap values (BV) of the corresponding NJ tree are given between brackets. Clusters that are supported by a BV of at least 95 for both trees (ML and NJ) are indicated in by a black line. Clusters that are supported by a BV of at least 90 are indicated by a grey line. Clusters with specimens belonging to the same order are indicated by a square bracket. BV's smaller than 50 are not shown.

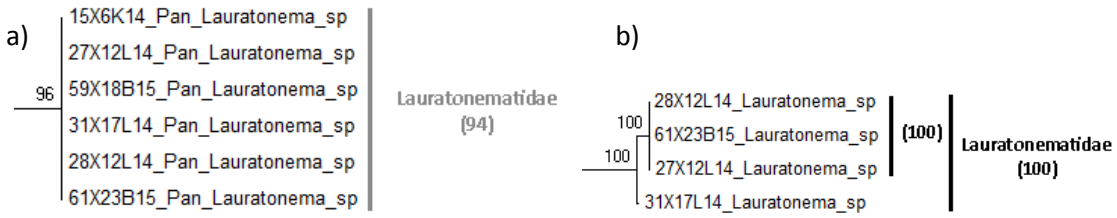


Figure 12. The only well supported cluster of the 18S ML tree (a), and strongly supported cluster of the COI tree (b) for specimens within the order Trefusiida. Bootstrap values of the corresponding NJ tree are given between brackets. Clusters that are supported by a bootstrap value (BV) of at least 95 for both trees (ML and NJ) are indicated in by a black line. Clusters that are supported by a BV of at least 90 are indicated by a grey line. BV's smaller than 50 are not shown.

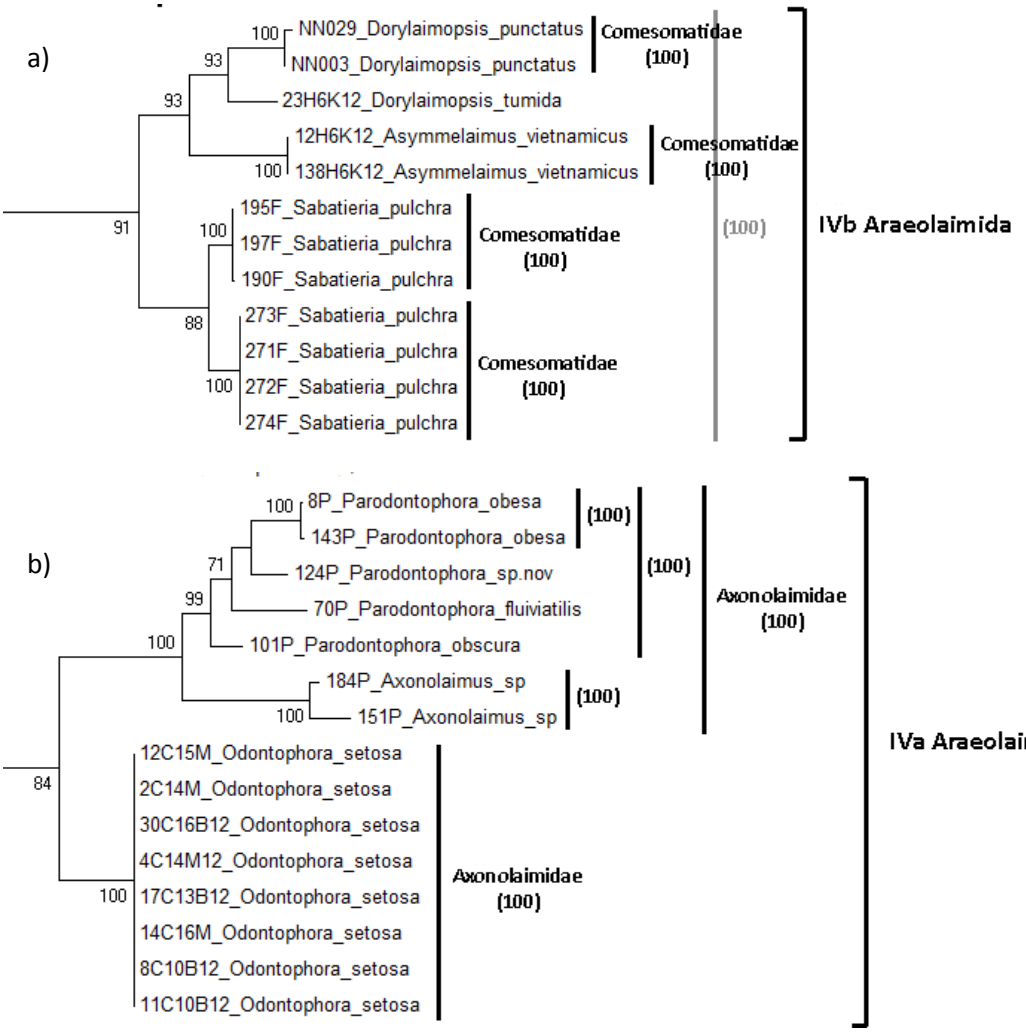


Figure 13. The only strongly supported clusters of the COI ML tree for the order Araeolaimida. No well supported clusters for this order were found for 18S. Bootstrap values (BV) of the corresponding NJ tree are given between brackets. Clusters that are supported by a BV of at least 95 for both trees (ML and NJ) are indicated in by a black line. Clusters that are supported by a BV of at least 90 are indicated by a grey line. Clusters with specimens belonging to the same order are indicated by a square bracket. BV's smaller than 50 are not shown.

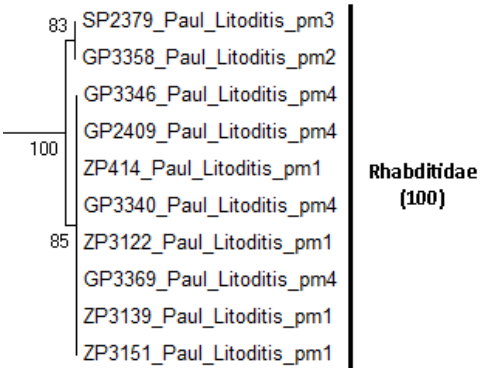


Figure 14. The only strongly supported cluster of the 18S ML tree for specimens within the order Rhabditida. No well supported clusters for this order were found for COI. Bootstrap values (BV) of the corresponding NJ tree are

given between brackets. Clusters that are supported by a BV of at least 95 for both trees (ML and NJ) are indicated in by a black line. BV's smaller than 50 are not shown.

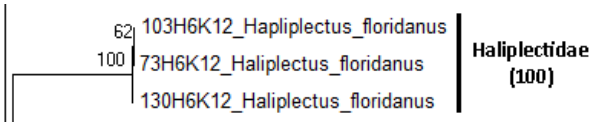


Figure 15. The only well supported cluster of the COI ML tree for specimens within the order Plectida. No well supported clusters for this order were found for 18S. Bootstrap values (BV) of the corresponding NJ tree are given between brackets. Clusters that are supported by a BV of at least 95 for both trees (ML and NJ) are indicated in by a black line. BV's smaller than 50 are not shown.

Cryptic species

We found an intraspecific distance value of more than 0.05 for 20 species. We were able to identify 13 of them as potential cryptic species, based on our three "double evidence" rules: 1) the intraspecific distance value was high for both COI and 18S (sequence divergence of one gene through chance can be ruled out), 2) the compared specimens having a high distance value came from different regions (divergence through isolation), 3) the high distance values showed consistent patterns that divided the specimens in clear groups (the same small values within and large values between groups). The species identified as potentially containing cryptic species were divided into clusters. When there was no morphological difference based on the vouchers, the clusters were considered 'true cryptic species'. This was the case for *Stilbonema brevicolle*, *Acanthopharynx micans*, *Mesacanthion* sp., *Zalonema* sp. and *Eurystomina ornata*. For *Theristus pertenuis* and *Oncholaimus campylocercoides*, we could not say with enough certainty whether there was a morphological difference based on the vouchers. Several clusters turned out to be another species than the one they were named to be. The specimens in cluster 2 of *Terschellingia longicaudata* were correctly identified, but the specimens in cluster 1 belong to a different species within the genus, which we could not identify with certainty. The specimens in cluster 1 and 2 of *Sphaerolaimus maeoticus* are correctly identified and show no morphological difference, but the specimens in cluster 3 are another species within the genus, which we could not identify with certainty. For *Anoplostoma sunderbanae*, there was no voucher available for specimen 107H6K12. The other two specimens, 60H6K12 and 66H6K12, are identified as another genus based on the vouchers, most likely *Theristus* and *Linhystera* respectively. However, they cluster nicely within the genus for both genes, so we suspect that this "misidentification" is caused by the wrong vouchers linked to these specimens. For the last three species, *Cheironchus vorax*, *Paracomesoma dubium* and *Theristus flevensis*, vouchers for at least one of the clusters were missing, so we could not make any conclusions. A full overview can be found in Table 5.

Species	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Highest value	Conclusion
Theristus pertenuis (2)	143F, 185F, 193F	6C9M12			0,2772	NC
Stilbonema brevicolle (3)	ND017	ND033			0,2417	NMD
Acanthopharynx micans (3)	9X6K14, 22X6L14, 30X17L14, 77X3C15	19X6L14, 33X17L14			0,2394	NMD
Oncholaimus campylocercoides (1, 3)	142F, 149F, 167F, 175F, 203F, 207F, 220F, 225F, 226F, 235F, 254F, 255F	132F, 177F	16X6L14, 17X6L14	29X17L14	0,2374	NC/MV (4)
Mesacanthion sp (1)	21X6L14, 87X20C15, 89X20C15	38X19L14			0,2172	NMD
Anoplostoma sunderbanae (3)	60H6K12, 107H6K12	66H6K12			0,1807	WID*
Terschellingia	60P, 85P	161P,			0,1805	WID

longicaudata (3)		27H6K12				
Sphaerolaimus maeoticus (1, 3)	01H6K12, 3H6K12, 13H6K12, 16H6K12	24H6K12, 29H6K12	16P, 145P, 150P		0,1705	NMD (1-2)/ WID (1,2-3)
Zalonema sp (3)	ND020	ND030			0,1517	NMD
Eurystomina ornata (3)	54X16B15, 60X18B15, 84X9C15	53X16B15, 86X20C15			0,1035	NMD
Cheironchus vorax (2)	143H6K12, 144H6K12	NN020, NN015			0,2306	MV
Paracomesoma dubium (2)	87F	141H6K12			0,1692	MV
Theristus flevensis (2)	1A, 9A, 10A	81H6K12			0,0765	MV

Table 5. The species identified as potentially containing multiple cryptic species, showing which specimens (given as their unique specimen code) belong to which cluster. Between brackets in the species column, the rule or combination of rules used is given: 1) the intraspecific distance value was high for both COI and 18S, 2) the compared specimens having a high distance value came from different regions, 3) the high distance values showed consistent patterns that divided the specimens in clear groups. Highest value column: shows the highest intraspecific distance value, COI for the first ten, 18S for the last three. NC = not certain, NMD = no morphological difference, MV = missing voucher(s), WID = wrong identification. Between brackets in conclusion column, the number of the cluster the conclusion applies for is given. *normal clustering within genus, probably wrong vouchers linked to these specimens.

Seven other species are suspected to contain multiple cryptic species, as they showed remarkably high intraspecific distance values, but for these we did not have enough evidence to apply one of the previous rules. Because there was still too much uncertainty, we left their names unchanged in the dataset. *Robbea porosum, Axonolaimus* sp, *Paradesmodora campbelli* and *Bolbonema brevicollis* contain possible 'true cryptic species', as no morphological difference was found between clusters. For *Sabatieria pulchra* we could not make conclusions based on the vouchers with enough certainty. For *Parodontophora quadristicha* and *Dorylaimopsis tumida*, vouchers were missing for at least one of the clusters. An overview is given in Table 6.

Species	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Highest value**	**Conclusion**
Robbea porosum	ND005, ND007, ND0043	ND008, ND0014, ND0040			0,2883	NMD
Sabatieria pulchra	271F, 273F, 274F	190F, 195F	(272F)	(197F)	0,2288	NC
Axonolaimus sp	151P	184P			0,1505	NMD
Paradesmodora campbelli	14A	3A			0,2021	NMD
Parodontophora quadristicha	93H6K12	125H6K12			0,1991	MV
Dorylaimopsis tumida	22H6K12, 23H6K12, 28H6K12, 37H6K12, 38H6K12, 48H6K12	114H6K12, 115H6K12, 116H6K12			0,1852	MV
Bolbonema brevicollis	44A	37A			0,1195	NMD

Table 6. Species that are suspected to consist of multiple cryptic species, with the suspected clusters shown. Highest value column: shows the highest intraspecific distance value, COI for the first three, 18S for the last four. Specimens between brackets did not have an clear affinity or difference with a cluster, and were therefore set apart. NC = not certain, NMD = no morphological difference, MV = missing voucher(s), WID = wrong identification.

4.3. Answering the meiofauna paradox

The total number of species that we had a 18S and/or COI sequence for is given in Table 7. Only a small number of these species were shared between locations (Table 8 and Fig. 16). Based on 18S and COI sequences respectively, 20 and 6 species were shared by two locations and 215 and 135 species respectively were unique to their location. No species were shared by more than two locations.

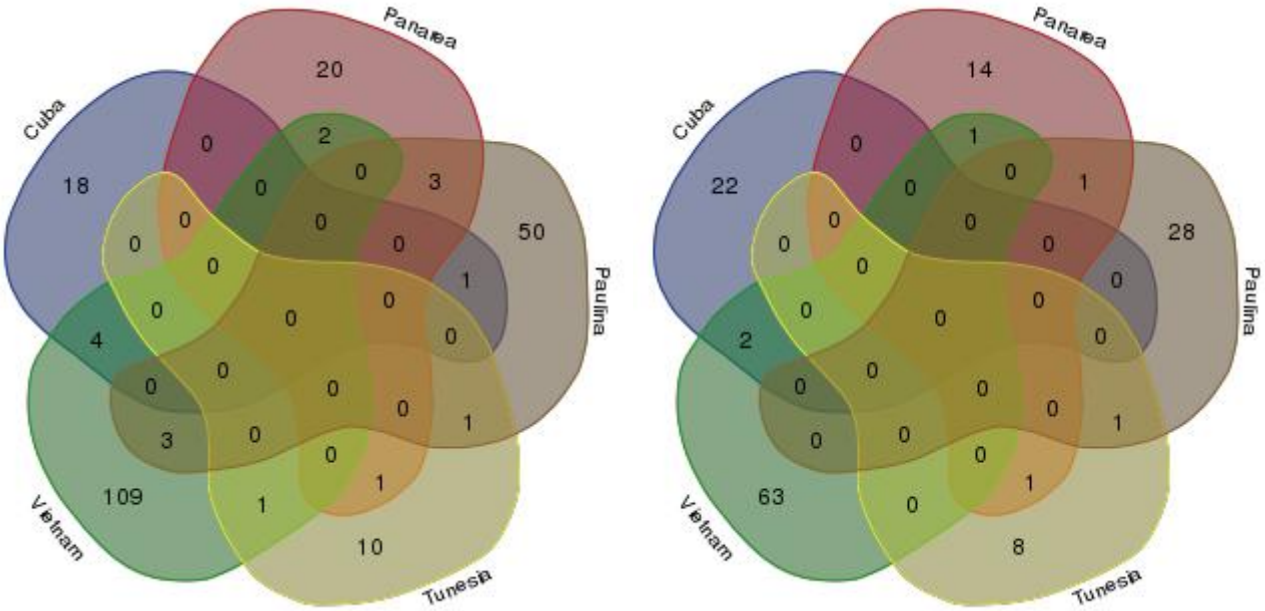


Figure 16. Venn diagrams for 18S (left) and COI (right). Each location is given in its own color. Numbers on the outer edges (single color) are the number of species unique to their corresponding location (e.g. 18 species for Cuba for 18S). The number in an overlapping region is the number of species shared by the two corresponding locations (e.g. 4 species shared by Cuba and Vietnam for 18S), see also Table 8 . Numbers for Papua New Guinea (4 species shared) are not shown, as five is the maximum number of groups to create the diagram, and no COI sequences were available for this location.

Location	Species with 18S sequence	Species with COI sequence	Species with both 18S and COI sequence
Cuba	23	24	18
Panarea	26	17	16
Paulina	58	30	18
Tunesia	13	10	5
Vietnam	119	67	55
Papua New Guinea	16	-	-

Table 7. The number of species for which a 18S and/or COI sequence is available, shown for each location. There were no COI sequences for Papua New Guinea.

	18S		COI	
Locations	**Number**	**Species**	**Number**	**Species**
Cuba, Paulina	1	*Desmodora pontica**	0	-
Cuba, Vietnam	4	*Daptonema* sp, *Gomphionema* sp, *Longicyatholaimus* sp, *Cheironchus vorax*	2	*Daptonema sp*, *Desmodora* sp
Panarea, Paulina	3	*Oncholaimus* sp, *Monoposthia mirabilis**, *Microlaimus honestus**	1	*Microlaimus honestus**
Panarea, Tunesia	1	*Oncholaimus campylocercoides**	1	*Oncholaimus campylocercoides**
Panarea, Vietnam	2	*Daptonema* sp3, *Oxystomina_sp*	1	*Daptonema* sp3
Paulina, Tunesia	1	*Linhomoeus* sp	1	*Theristus pertenuis**
Paulina, Vietnam	3	*Dichromadora* sp1, *Daptonema* sp1, *Anoplostoma sp2*	0	-
PNG, Paulina	3	*Onyx* sp1, *Bathyeurystomina* sp1, *Paracanthonchus caecus**	0	-
Tunesia, Vietnam	1	*Paracomesoma dubium*	0	-
PNG, Vietnam	1	*Theristus flevensis**	0	-
Total	20		6	

Table 8. The species that are shared between locations, for 18S and COI seperately. Number = number of species shared by the two locations, Species = the name(s) of the shared species, PNG = Papua New Guinea; no COI sequences available for this location. *Only species indicated with an asterisk were suitable for further analysis.

When excluding species that were contamination, wrongly identified, had no voucher available or that could not be identified for certain based on the vouchers, however, we were left with only seven species suitable for inter-location comparison, indicated with an asterisk in Table 8. Of these seven species, we found three species that showed high intra-specific values: *Oncholaimus campylocercoides*, *Paracomesoma dubium*, and *Theristus flevensis* (see results 4.2, Table 5). However, from none of these we could conclude if they contained different cryptic species for different locations. For *O. campylocercoides*, we could not say with enough certainty whether there was a morphological difference based on the vouchers. For the other two species, the vouchers for at least one of the clusters were missing. The four species for which we did not find high intraspecific distance values were *Desmodora pontica*, *Microlaimus honestus*, *Monoposthia mirabilis* and *Theristus pertenuis*. For each of these species, there was only one specimen for at least one of the two locations. Comparing with the specimen(s) from the other locations yielded a maximum intraspecific value of respectively 0,0359; 0,0243; 0,0043 and 0,0030 for the four species. The values of the first three species were for 18S, the last one for COI. The values for *D. pontica* and *M. honestus* are arguably high enough to suggest the presence of cryptic species, but we could not check this using the vouchers, as these were missing for at least one of the locations in both cases. The values for *M. mirabilis* and *T. pertenuis* are not high enough to consider them as potentially containing different (cryptic) species. The other five species possibly contain cryptic species, but without support or counterevidence from the vouchers, we cannot make a reliable conclusion.

4.4. Improving identification of meiofaunal communities

The number of sequences generated by the Ion Torrent was rather consistent across replicates, but of a different order of magnitude between primer sets (Table 9). Both COI primer sets resulted in far less sequences than the one for 18S, which yielded ten thousands per replicate, but the JB3-JB5 primer set yielded only very few sequences (a few dozen to a few hundred) .

Replicate	**G18S4-22R (18S)**	**JB2-JB5GED (COI)**	**JB3-JB5 (COI)**
MockA1	27539	2517	345
MockA2	40136	2933	51
MockB1	46664	2563	153
MockB2	50766	3212	48

Table 9. The community replicates (A or B) and the PCR replicates (1 or 2) and the number of sequences yielded for each primer set, after filtering the raw reads.

The artificial community contained 50 species (from 56 specimens), in varying abundance (see Appendix 10.5). We had at least one sequence available for 46 and 24 of these species for respectively 18S and COI. For 19 species, we had a sequence for both markers. For 5 species we had no sequence available for either of the genes, making them not identifiable by our reference database: *Bathylaimus assimilis, Diplolaimelloides oschei, Microlaimus punctulatus, Theristus ensifer* and *Theristus* sp2. Together they represented a theoretical 8,65% of the individuals in our mock community.

The sequences obtained from the Ion Torrent were clustered in OTU's, using different similarity thresholds. This resulted in different amounts of OTU's found, with an overall pattern of the higher the similarity level that was used, the more OTU's were found. For each of the similarity levels, a different portion of the OTU's was identified as a nematode by comparing against the reference collection (see Table 10). The two similarity percentages that had the highest percent of OTU's identified as nematodes for each primer set were further analysed: 99 and 99,5% for 18S and 94,38 and 95% for both COI primer sets. Silva referencing (18S) was only further analysed for 99%, as this was the highest percentage for which we had a reference file available.

A change in minimum cluster size for OTU picking to 1 (instead of 3) did not yield better results. The number OTU's found was 4647 for OTU picking on 99% similarity with minimum cluster size 1, very close to the number found with the alternative setting: 4655. The amount of nematodes identified was 60,36% and 60,56% respectively. Assigning taxonomy to the OTU's yielded the best results with the default sequence similarity value of 90%. All further results given were obtained using the default values for both parameters.

Gene (reference)	**Similarity**	**Nematodes**	**Unassigned**	**Total number of OTU's**	**Percent nematodes**
18S (own ref)	96	2171	2549	4720	46,00 %
18S (own ref)	97	3079	2503	5582	55,16 %
18S (own ref)	99	2805	1842	4647	60,36 %
18S (own ref)	99,5	4110	2217	6327	64,96 %
18S (Silva)	99	3018*	1602	4703	64,17 %
COI: JB2	93	3	745	748	0,40 %
COI: JB2	94,38**	6	887	893	0,67 %
COI: JB2	95	7	956	963	0,73 %
COI: JB2	96,66	6	1010	1016	0,59 %
COI: JB3	93	7	29	36	19,44 %
COI: JB3	94,38**	10	32	42	23,81 %
COI: JB3	95	12	31	43	27,91 %
COI: JB3	96,66	11	41	52	21,15 %

Table 10. Results of identifying OTU's with QIIME, using UCLUST. Gene (reference) = the barcoding gene (and forward primer for COI) used, with the database that was used as reference ("own ref" = reference database built for this study, "Silva" = the Silva rRNA database). Similarity = the percent similarity used for OTU picking with QIIME, nematodes = number of OTU's identified as nematodes, unassigned = number of OTU's not identified, total number of OTU's = the total number of different OTU's found over all replicas, percent nematodes = the percent nematodes identified on the total amount of OTU's. *Silva referencing also yielded 83 non-nematode taxonomy assignments, not included in this number. **94,38% is the COI threshold value found in this study.

Assigning taxonomy had varying success. Our own reference database identified around 70% of the artificial community based on 18S, but only 7,7-24% for COI. Using the Silva database as a reference for 18S (at 99% similarity) resulted in higher identification success using UCLUST, but lower success using BLAST as opposed to using our own reference database. For COI, the BLAST algorithm showed a higher identification success than UCLUST in all cases. The identification success for the OTU's of the different primer sets and algorithms is listed in Table 11 (see also Fig. 17 - 21).

		UCLUST		BLAST	
Gene (reference)	**Similarity**	**Identified (%)**	**Not identified (%)**	**Identified (%)**	**Not identified (%)**
18S (own ref)	99	70,19 (29/8)	29,81	69,23 (33/24)	30,77
18S (own ref)	99,5	73,08 (23/7)	26,92	69,23 (33/26)	30,77
18S (Silva)	99	93,27 (23/22*)	6,73	61,54 (25/56*)	38,46
COI: JB2	94,38**	7,69 (3/0)	92,31	22,12 (7/9)	77,88
COI: JB2	95	9,62 (4/0)	90,38	24,04 (11/8)	75,96
COI: JB3	94,38**	8,65 (2/0)	91,35	19,23 (6/4)	80,77
COI: JB3	95	8,65 (2/0)	91,35	14,42 (5/4)	85,58

Table 11. Results of assigning species level taxonomy to the OTU's. Gene (reference) = the barcoding gene (and forward primer for COI) used, with the database that was used as reference ("own ref" = reference database built for this study, "Silva" = the Silva rRNA database). Similarity = the percent similarity used for OTU picking with QIIME. Numbers given are the percent of the true mock community identified or not identified by UCLUST or BLAST. Numbers between brackets are the number of species that were correctly identified and the number of species wrongly identified ("identified", but not put in the mock community). *Silva referencing yielded 19 and 43 non-nematode taxonomy assignments for UCLUST and BLAST respectively, not included in this number. **94,38% is the COI threshold value found in this study.

Looking only so far as the genus level of the identifications for 18S resulted in even higher identification success, were the success of using our own database approached that when using the Silva database. These results are given in Table 12.

		UCLUST		BLAST	
Gene (reference)	**Similarity**	**Identified (%)**	**Not identified (%)**	**Identified (%)**	**Not identified (%)**
18S (own ref)	99	83,65 (25/6)	16,35	85,58 (30/11)	14,42
18S (own ref)	99,5	87,50 (26/5)	12,50	85,58 (30/12)	14,42
18S (Silva)	99	81,73 (20/15*)	18,27	87,50 (28/42*)	12,50

Table 12. Results of assigning genus level taxonomy to the OTU's for 18S. Database that was used as reference :"own ref" = reference database built for this study, "Silva" = the Silva rRNA database. Similarity = the percent similarity used for OTU picking with QIIME. Numbers given are the percent of the true mock community identified or not identified by UCLUST or BLAST. Numbers between brackets are the number of genera that were correctly identified and the number of genera wrongly identified ("identified", but not put in the mock community). *Silva referencing yielded 18 and 41 non-nematode taxonomy assignments for UCLUST and BLAST respectively, not included in this number.

The graph for JB2 using UCLUST is not shown because the scant proportion identified is not visible. The species level graphs are only given for 99,5% similarity 18S OTU picking, because those for 99% are

very similar, and the BLAST graph for the Silva referencing is too large to fit on a single page. The OTU composition of the mock community replicates differed drastically from composition of the original artificial community. In all cases, the BLAST algorithm left relatively less OTU's unidentified. However, it also proposed more identifications, visible as a more lengthy legend next to the graph (see Fig. 17 – 19, 21), of which more incorrect ones. In all cases, the BLAST algorithm gave more incorrect taxonomy assignments than the more conservative UCLUST method (see Table 11). This can be compensated for when we identify only to the genus level (see Table 12). Here we see that the number of correct taxa does not change much, but the number of incorrect assignments lowers and even drops to less than half for 18S referencing against our own database. For COI, there were only a few species identified, so here there was no noteworthy change compared to the species level.

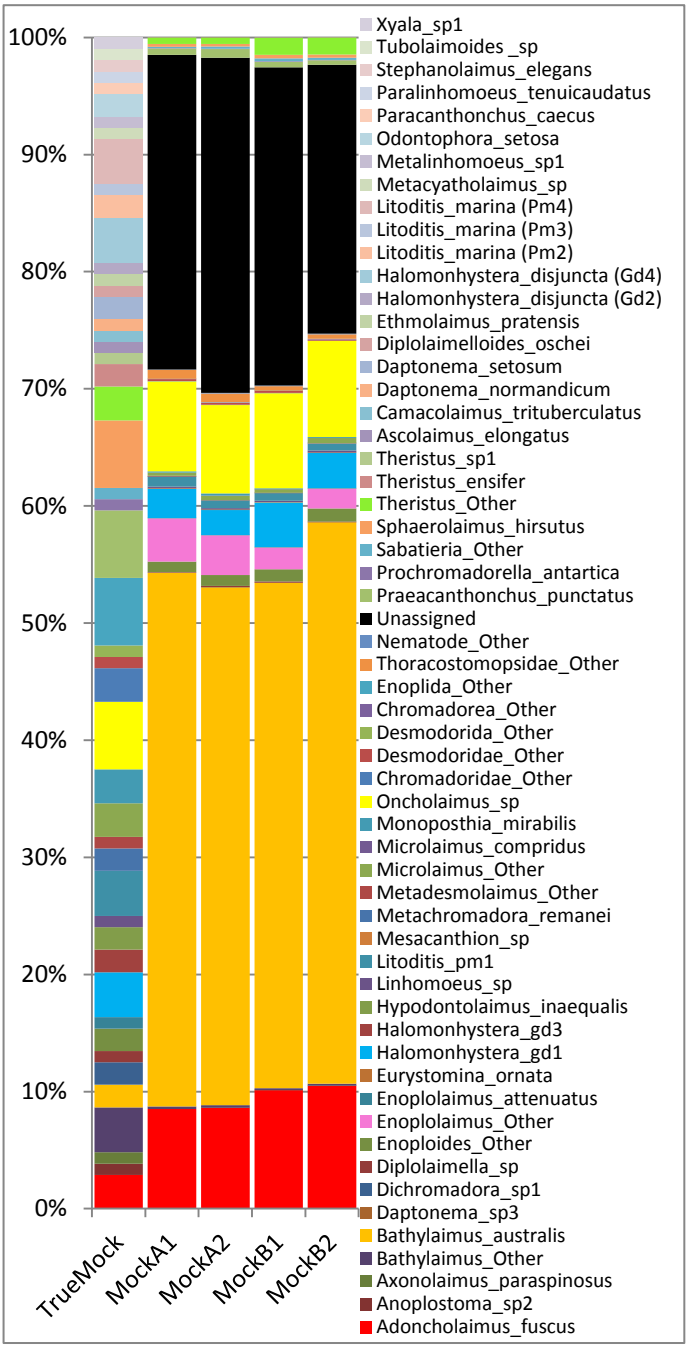
Identification of the mock community based on 18S was strongly biased towards *Bathylaimus*, identified as *B. australis* (and a fraction of an unidentified *Bathylaimus* species) by UCLUST, and as *B. denticaudatus* (and a fraction of *B. ignavus* and *B. sp)* by BLAST when referencing our own database. When using the Silva database, *B. assimilis* and a second unidentified species within the genus were assigned. The mock community only contained *B. australis* and *B. assimilis*, but only the former was identifiable by our own database. This makes the identifications by Silva referencing and by UCLUST using our own database correct. The three different *Bathylaimus* species identified by the BLAST algorithm were all incorrect. In all cases, *Bathylaimus* alone counts for more than 45% of the OTU's, whereas the mock community theoretically only contained 5,77% *Bathylaimus* DNA. *Adoncholaimus fuscus* and a not to species level identified *Oncholaimus* each took another 10% in all cases for 18S, while the mock community only contained 2,88% and 5,77% DNA extract for these species respectively. *Sphaerolaimus hirsutus* also theoretically represented 5,77% of the mock community, but was not identified based on 18S when referencing our own database, and only for less than 1% when using the Silva database. Multiple species with low abundance (having only one "individual" in the mock community, e.g. *Linhomoeus* sp.) were also identified, but are not or barely visible on the graphs because of their low abundance in the mock replicates.

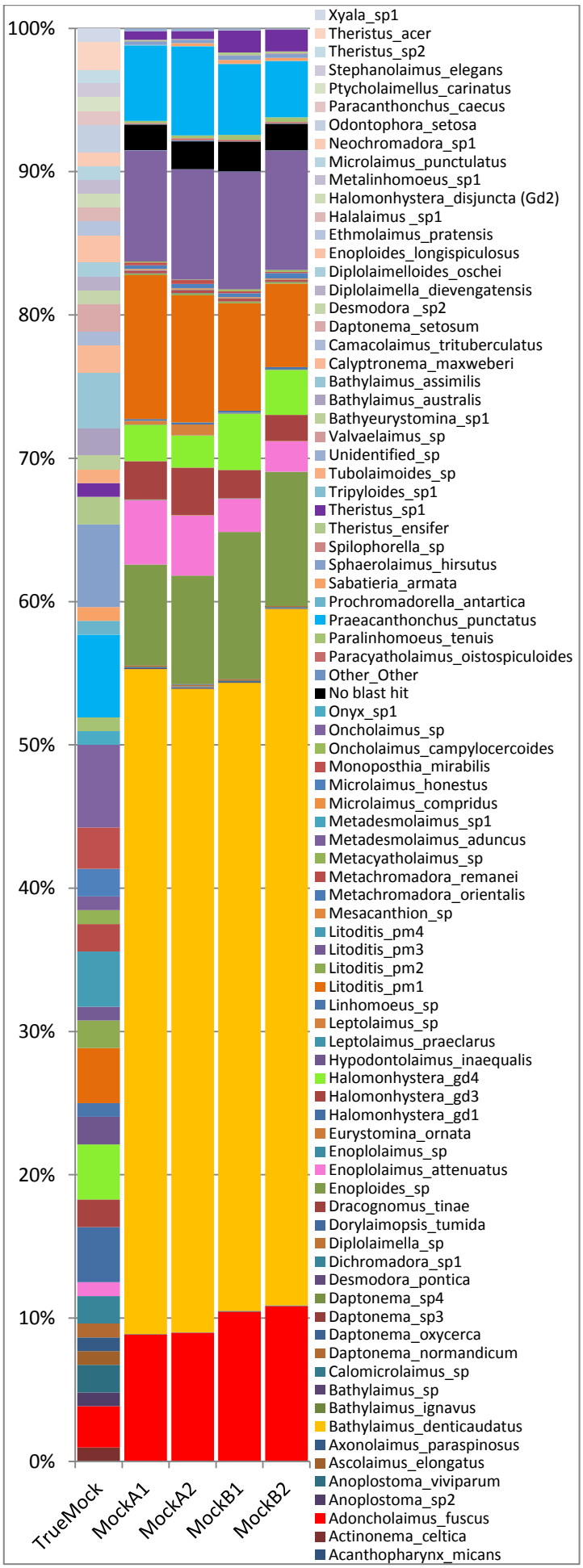
Identification based on COI identified only a few species, but for the UCLUST algorithm, these were all correct for both primer sets. The A2 and B2 mock replicates for the JB3-JB5 primer set only yielded 51 and 48 OTU's respectively. Considering there should be 50 species present in the mock communities and only 24 of them are identifiable by our COI reference database, it is not surprising that only a few of them were found. Yet, only two species identified by UCLUST, *Theristus* (no species) and *Adoncholaimus fuscus* for OTU picking with our threshold (94,38%), and *Sphaerolaimus hirsutus* and *Daptonema setosum* for 95%, is a disappointing result. For the BLAST algorithm, six and five species were identified respectively. There is however a remarkable difference between the mock replicates (figure ...): the part of the replicate OTU's identified as *A. fuscus* varies from a meagre 3% in A1 to roughly 35% in B2. For the JB2-JB5GED primer set, the OTU's identified using UCLUST were barely visible on the graph. Using the BLAST method, they counted for less than 5% of the total OTU's for each replicate.

b)

Figure 17. The species level graphs for 18S OTU picking using a 99.5% similarity threshold and our own database as reference. Bars show abundance of each species (OTU) as relative proportion of the total community. a) using the UCLUST algorithm for taxonomy assignment, b) using the BLAST algorithm. The left bar in each graph show the composition of the original artificial community ("TrueMock"), the other four bars show the OTU composition of each artificial community replicate (MockA1-B2). The proportion of OTU's for which no match was found are coloured black, other colours may vary but are shown in the legend.

a)





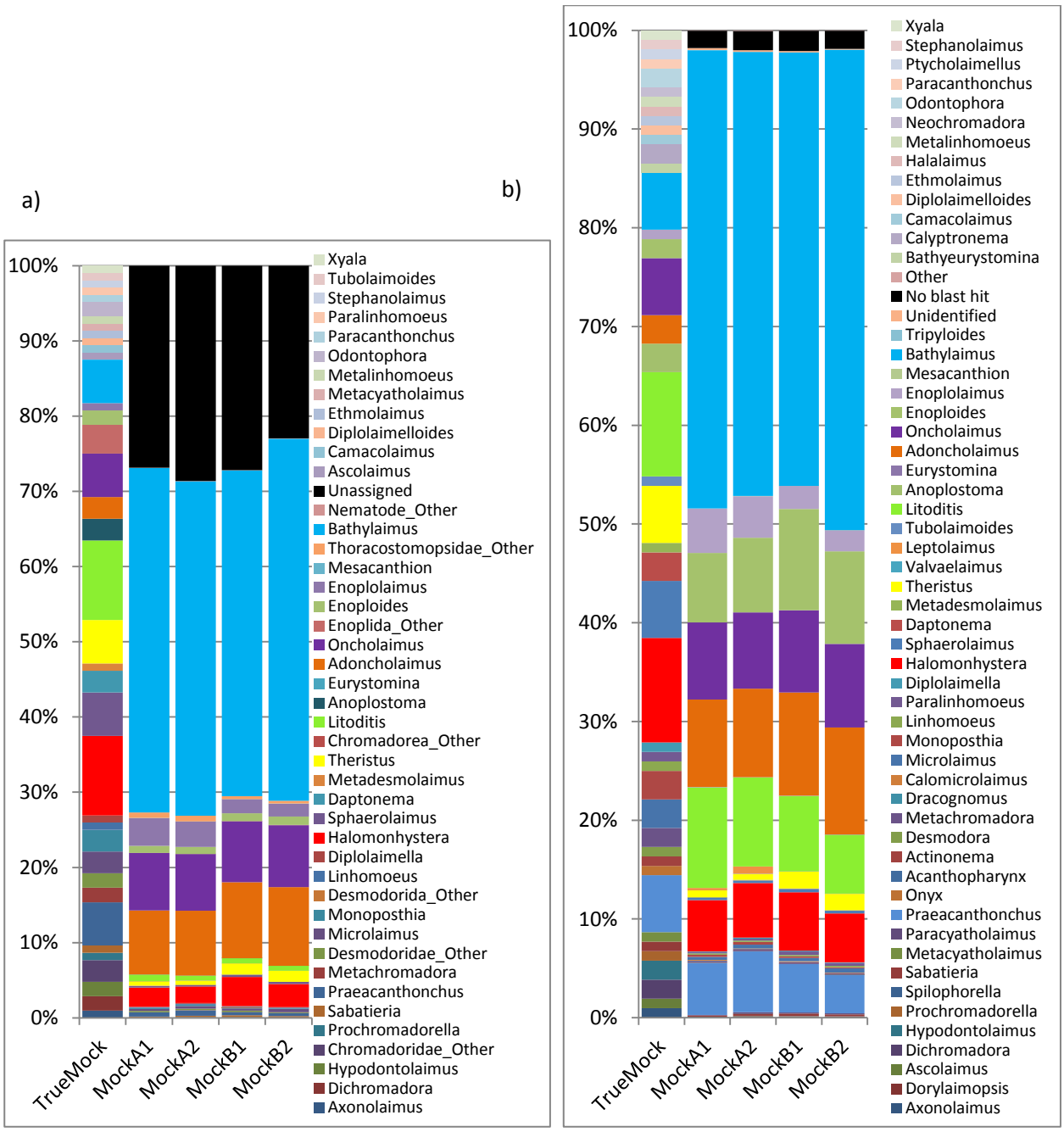
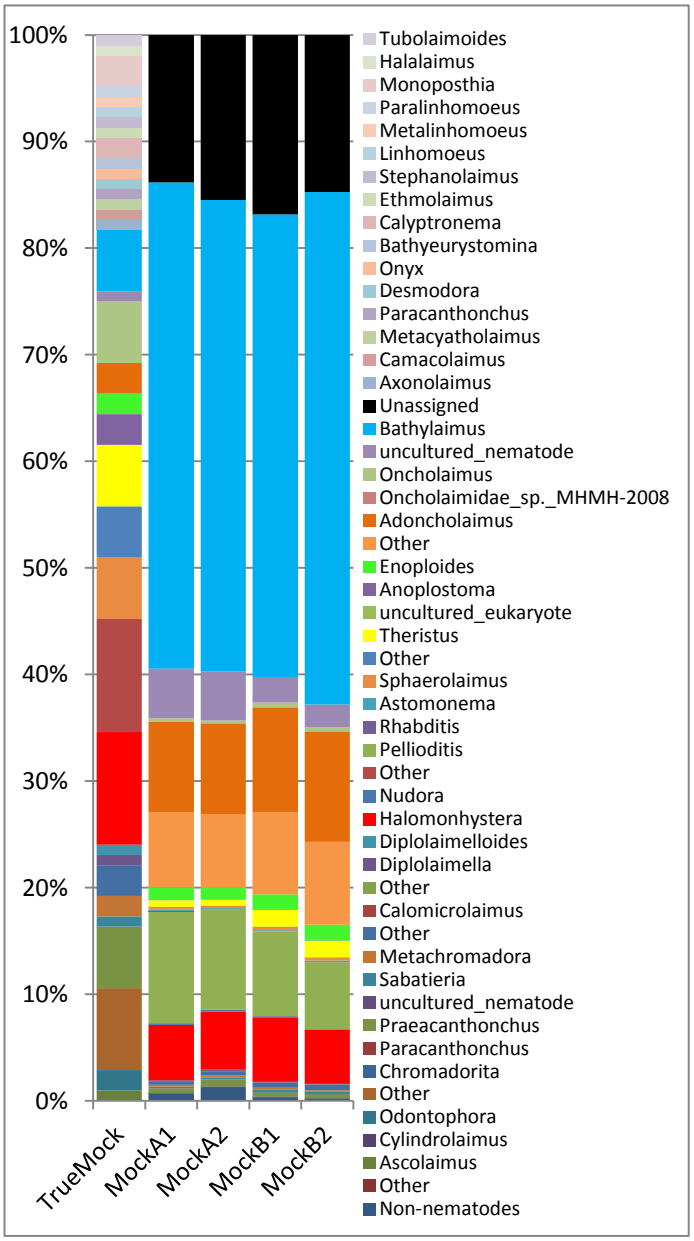


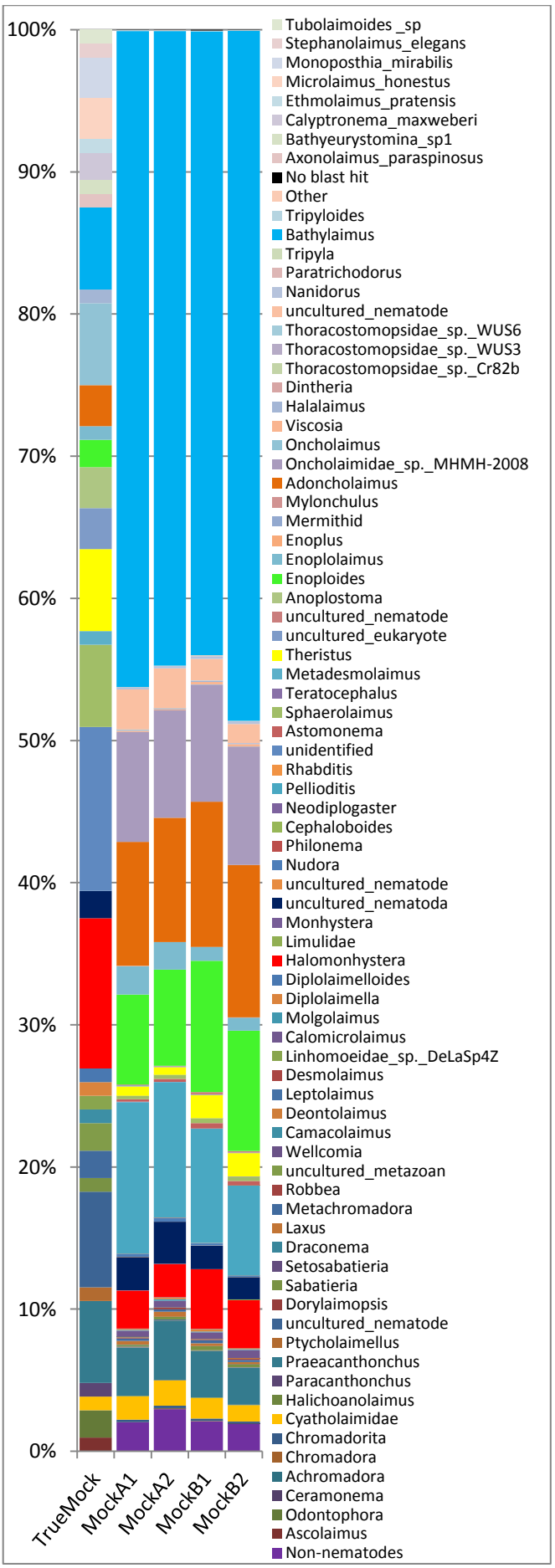
Figure 18. The genus level graphs for 18S OTU picking using a 99.5% similarity threshold and our own database as reference. Bars show abundance of each genus (OTU) as relative proportion of the total community. a) using the UCLUST algorithm for taxonomy assignment, b) using the BLAST algorithm. The left bar in each graph show the composition of the original artificial community ("TrueMock"), the other four bars show the OTU composition of each artificial community replicate (MockA1-B2). The proportion of OTU's for which no match was found are coloured black, other colours may vary but are shown in the legend.

b)

Figure 19. The genus level graphs for 18S OTU picking using a 99% similarity threshold and the Silva database as reference. Bars show abundance of each genus (OTU) as relative proportion of the total community. a) using the UCLUST algorithm for taxonomy assignment, b) using the BLAST algorithm. The left bar in each graph show the composition of the original artificial community ("TrueMock"), the other four bars show the OTU composition of each artificial community replicate (MockA1-B2). The proportion of OTU's for which no match was found are coloured black, other colours may vary but are shown in the legend.

a)





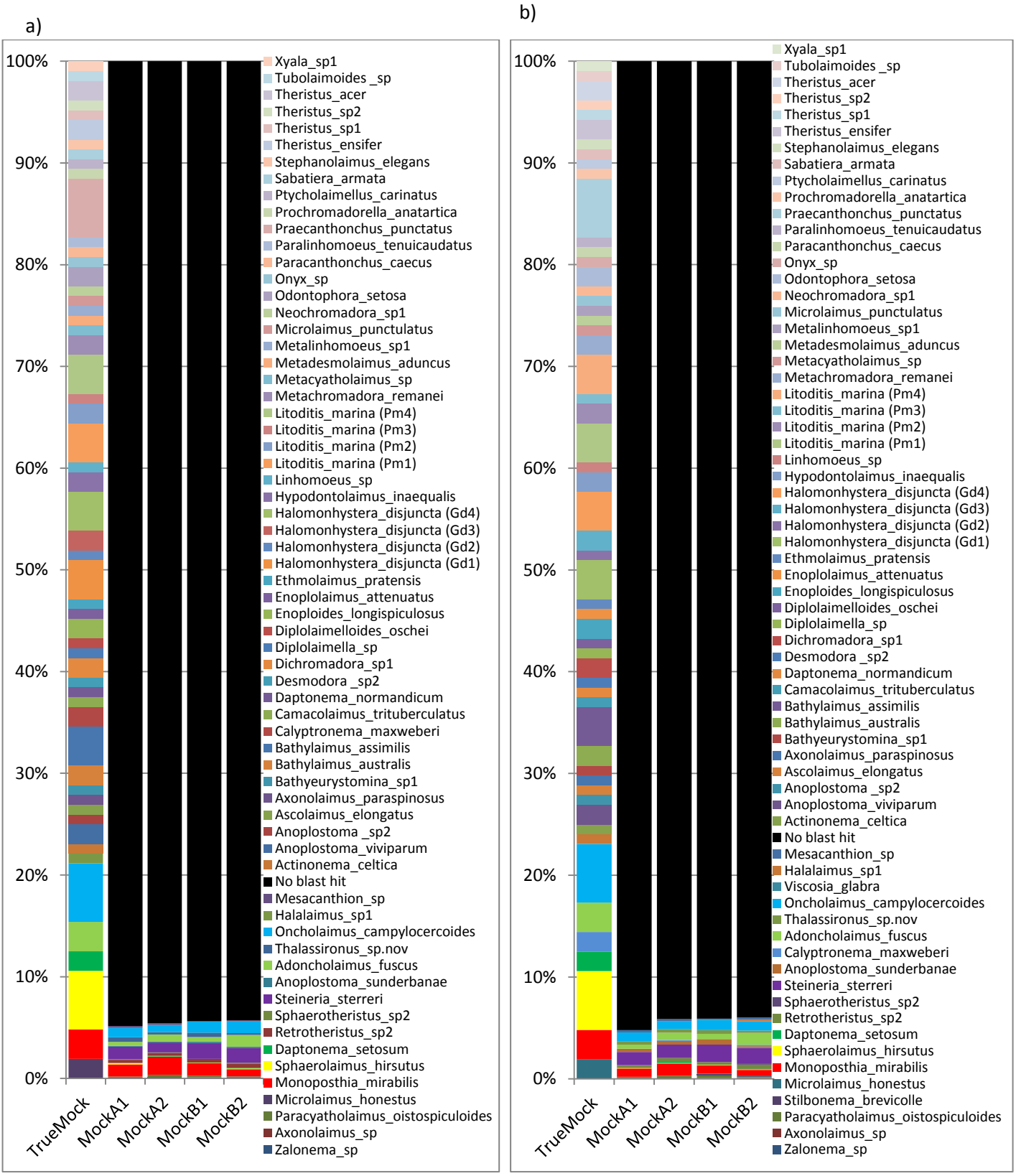


Figure 20. The species level graph for (JB2-JB5GED) for OTU picking using (a) a 94,38% similarity threshold and (b) a 95% similarity threshold and our own database as reference. The BLAST algorithm was used for taxonomy assignment. The results of the UCLUST algorithm were not visible on the graph (not shown here). Bars show abundance of each species (OTU) as relative proportion of the total community. The left bar in each graph show the composition of the original artificial community ("TrueMock"), the other four bars show the OTU composition of each artificial community replicate (MockA1-B2). The proportion of OTU's for which no match was found are coloured black, other colours may vary but are shown in the legend.

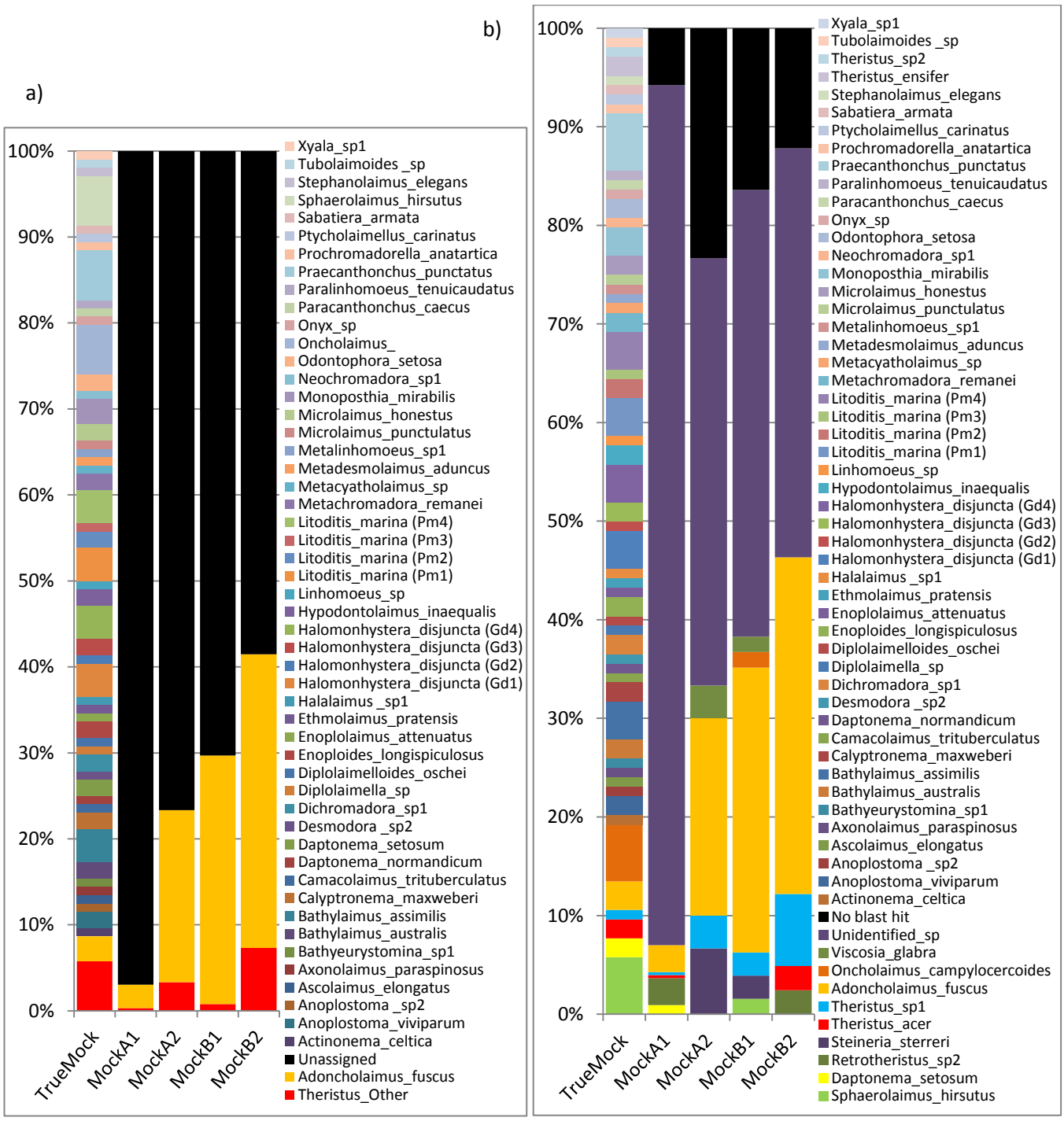


Figure 21. The species level graphs for COI (JB3-JB5) OTU picking using a 94,38% similarity threshold and our own database as reference. Bars show abundance of each species (OTU) as relative proportion of the total community. a) using the UCLUST algorithm for taxonomy assignment, b) using the BLAST algorithm. The left bar in each graph show the composition of the original artificial community ("TrueMock"), the other four bars show the OTU composition of each artificial community replicate (MockA1-B2). The proportion of OTU's for which no match was found are coloured black, other colours may vary but are shown in the legend.

5. Discussion

5.1. Distance and threshold values

The histograms for both 18S and COI reflect the expected pattern (see 4.1, Fig. 2). The relatively large overlap of intra- and interspecific distances in the 18S histogram illustrates the low resolution for identification at the species level, attributed to this barcoding gene (De Ley, et al., 2005). The COI histogram on the other hand shows a clear barcoding gap, with little overlap between the two distributions, supporting its suitability for distinguishing species. The maximum intraspecific values however, 0,1939 and 0,2883 for 18S and COI respectively, are higher than expected. This is probably largely due to the remaining suspected cryptic species that are still classified as a single species (see results 4.2). *Adhoc* found a threshold value of 0,0562 for COI, which is close to the value of 0,05 often used for COI barcoding (Derycke, Vanaverbeke, et al., 2010). We see that for 18S the lowest distance value (0,00597) other than zero still had a relative error (RE) of 0,247. This means that a quarter of all identifications done using this distance value would be wrong. We suspect that one of the main reasons for this high error probability is that there are still cryptic species in the used dataset. These probably should be considered different species based on the found distance values. However, since we did not have enough evidence to support our suspicions, we had to leave them classified as one species. They could have interfered with *Adhoc's* calculations. The large overlap in intra- and interspecific distances could also prevent the program from finding a threshold value for 18S, again suggesting that this gene is not very suitable for distinguishing species. In an effort to circumvent this problem, we made the 18S calculations separately for six major families. But even when looking at the family level, we saw this problem of overlap returning. Only for Oncholaimidae and Oxystominidae we found a significant threshold value of respectively 0,0038 and 0,0016 (see 4.1, Fig. 4). This is very low, but the former is still the double of the latter. The histograms of the six families do not all approach the form of the total histogram for 18S, although this is probably an artifact of sample size. The families with the lowest number of specimens, Oxystominidae and Sphaerolaimidae show the most distorted graph, while the one with the largest sample size, Xyalidae, seems to approach the total 18S graph the best, together with Chromadoridae.

We see in the graph of Oxystominidae that there are three cases (the first value other than zero is 0,0031) with a intraspecific value above this threshold, and yet the RE = 0. One explanation for this is that we set the "ambiguous" option to "correct". This means that when *Adhoc* finds two or more sequences as best match, if at least one of them is correct, the program considers the identification as correct. Another likely explanation is that the two combinations resulting in these higher values did not come up in *Adhoc* as best matches. We suspect the latter is the case here. We do want to note that *Adhoc* uses linear regression to find a threshold value, which might not be ideal. When looking at a large sample size, as for the complete COI dataset, we see that the relation between distance and relative error is not a linear one (see 4.1, Fig. 2b). For a smaller sample size, a clustering of some values can cause the regression line to shift, resulting in the incorrect threshold value to be calculated. An extreme case of this problem was *Adhoc*'s regression graph for Sphaerolaimidae for 18S, which was completely distorted because many of the values clustered to the left. This resulted in a threshold value of 0,1186 for this family, which is a very unlikely value for the 18S.

5.2. Pinpointing cryptic diversity

Cryptic species are relatively common and well known problem in species identification, especially for microscopic groups such as nematodes that are difficult to identify morphologically (Pfenninger & Schwenk, 2007). Giving the number of specimens in our database, we expected to find at least some cryptic species in our database. For twenty species, we found remarkably high intraspecific distance

values. When looking closer, we found that these species consisted of minimally two clusters, with high inter-cluster and low intra-cluster p-distances. For some species, like *Mesacantion* sp. from Panarea, the first was undeniably high (0,2172 for COI and 0,07 for 18S) and the latter approached zero. But not all cases were this crystal clear; we often did not have enough specimens with both genes sequenced and could therefore not make a well supported decision this way. To cope with this problem, we searched for consistent patterns in the distance values. For example: for *Oncholaimus campylocercoides*, we did not have a COI sequence of specimen 16X6L14. It did however show exactly the same divergence patterns as specimen 17X6L14, of which we had both genes. Based on this consistent pattern, we assumed that the two specimens belonged to the same cryptic species cluster (our third rule). Following this, we could assume that de COI divergence pattern for 17X6L14 would represent that of 16X6L14. For some species, for example *Theristus pertenuis*, we saw that different clusters represented groups of specimens sampled from different regions. This seems intuitively logical considering they live in apparently isolated regions, but we will come back on this in 5.3.

Detailed morphological studies may find diagnostic characters between cryptic species (Apolonio Silva De Oliveira, et al., 2012; Sudhaus & Kiontke, 2007), but we do not expect to find such characters based on voucher photos. The vouchers were used to exclude the possibility that the high intraspecific distance values were caused by incorrect identification. Five species were concluded to contain 'true cryptic species', when no morphological difference was found based on the vouchers. This was the case for *Stilbonema brevicolle* (which has also been found by its source study (Armenteros, et al., 2014)), *Acanthopharynx micans, Mesacanthion* sp.*, Zalonema* sp. and *Eurystomina ornata.* Cryptic diversity on a local scale has been shown for nematodes (Apolonio Silva De Oliveira, et al., 2012; Derycke, Backeljau, et al., 2007; Derycke, De Ley, et al., 2010). A field experiment by Derycke et al. (2007) surveyed the genetic diversity of *Litoditis marina* (present in our database), a species with a high reproductive capacity and short generation time that enables one or a few gravid females to establish populations. This species can colonize patches of a kilometre away within ten days. mtDNA haplotypes that were rare in the source population were abundant in these distant patches, suggesting that founder effects and genetic bottlenecks structured these populations. Such effects might also be the cause of the high intraspecific distances of the five true cryptic species that we found. For other species we found high distance values for, we could not say with enough certainty whether there was a morphological difference based on the vouchers, although *Robbea porosum* was also suggested to be cryptic by its source study (Armenteros, et al., 2014). In other cases, there were no vouchers available for at least one of the clusters. The vouchers did not only prove their value for identifying cryptic species, but also as a control on misidentifications. For example, the specimens in cluster 2 of *Terschellingia longicaudata* were correctly identified, but the specimens in cluster 1 belonged to a different species within the genus, which we could not identify with certainty. However, we need to be careful with blindly relying on the vouchers. Two *Anoplostoma sunderbanae* specimens, 60H6K12 and 66H6K12, were identified as another genus based on the vouchers, most likely *Theristus* and *Linhystera* respectively, but they clustered nicely within the genus for both genes. We suspect that this "misidentification" was caused by the wrong vouchers linked to these specimens. Needless to say, as the vouchers are meant as a control, such errors compromise the reliability of the whole database. It is possible that other similar errors, that we were not aware of, exist on our database. However, if the name linked to the sequence is correct, as is the case for the *A. sunderbanae* specimens, it should not have interfered with our results. Thanks to the vouchers, a bug in the '18S fasta merger' script (see Appendix 10.8.1) was discovered that caused a mismatch between voucher codes and specimen names, which could then be corrected. Because of these reasons, we would like to emphasize the importance of good vouchers. We had clear vouchers available that perfectly showed all characters of the nematode necessary for morphological identification, with all details and different levels of focus showing both different internal and external structures. Unfortunately, we also had vouchers that were not clear enough to be usable for identification, which is exactly their purpose. If the photos are taken with care, though, or even a short movie is used that can be paused on every desired character and focus, they

guarantee the usability of their corresponding sequences for countless future studies and identifications. Giving the still limited availability of such vouchered sequences in public databases, this is a crucial part of molecular work to continue to improve knowledge and identification techniques.

When we indicated the cryptic species clusters as such in the alignment and made a new ML and NJ tree, we saw that they nicely clustered, in agreement with our divisions. Marking these specimens as different cryptic species, we could also take a considerable number of high intraspecific values out of the equation. The histograms and further calculations should therefore be more reliable. There are still some high values left because of species that we suspect also consist of cryptic species clusters, but for which we did not have the evidence to decide this with enough certainty. However, as mentioned in 5.1, this likely could have had an influence on calculations of the threshold values. This underlines the importance of identifying cryptic species, to improve the robustness of both DNA barcoding and nematode taxonomy as a whole.

Phylogenetic analysis

Delimiting species in species complexes using nuclear and mitochondrial gene trees is a well-suited approach for nematodes (Byron J. Adams, 1998). Our ML trees resulted in 16 and 59 strongly supported clusters for 18S and COI respectively. We expected to find more clusters for COI, as it has a better resolution at the species level. For 18S, we did not expect many species clusters to be supported by high bootstrap values. The low resolution at species level would cause more closely related species to switch places in the replicate (bootstrapping) trees, causing the bootstrap values to drop. However, we did expect to find more clusters resolving higher level taxonomy for 18S (Blaxter, et al., 1998; Meldal, et al., 2007). We obtained a nice, large cluster for the Enoplida (see Fig. 6 Ib) and Desmodorida (see Fig. 11 VIa) and some moderately large ones for the Monhysterida (see Fig. 8 IIa and IIb), but their inner nodes were not well supported (BV often lower than 50). Only the smaller clusters within, these larger ones were strongly supported. These often consisted of different species, but mostly within the same genus and always within the same family. The inner nodes of both overall trees were not well supported.

The Tripyloididae (Enoplida) cluster (see Fig. 6e) contained one specimen of *Valvaelaimus* sp., with voucher number NN004. However, when we searched this specimen in the COI ML tree, it was named *Sphaerolaimus* sp. Both of these not belonging to the Enoplida. We did not have a voucher for this specimen, but a BLASTx suggested *Bathylaimus* or *Tripyloides* to be the correct genus, matching the specimens it clusters with. A similar case was found for COI, within the Monhysterida (see 4.2: phylogenetic analysis). One particularly interesting case was for *Zalonema* sp. When taxonomy for each genus in our database was looked up in the World Database of Free-Living Marine Nematodes (NeMys: http://nemys.ugent.be/) in October 2015, to compose the file to add the information to the database, the taxonomy of *Zalonema* sp. was unresolved. The two specimens we have in our database clustered within the Desmodorida clade, close to members of the Desmodoridae. When we looked this species up again in the WoRMS database, *Zalonema* was now considered part of the Desmodoridae. Indeed, the page had been updated roughly two months earlier, on February 13 2016, and now matched our results.

Not only did we find much more strongly supported clusters of the same species for COI, but it also yielded more larger clusters of the same family, even up to clusters of the order level, even though these were not well supported. However, the higher level clusters were also not well supported for 18S. This again proves the usefulness of COI (Derycke, Vanaverbeke, et al., 2010; Hebert, et al., 2003; Hellberg, et al., 2002). The trees we obtained for both genes thus suggest that COI would be more suited to resolve phylogenetic relationships of marine nematodes, both on species level and on higher taxonomic levels, than 18S.

5.3. Answering the meifauna paradox

Cosmopolitan species have been reported in small organisms (less than 1 millimeter in length), including nematodes (Kiontke, et al., 2011; Zauner, et al., 2007). Passive dispersal of nematodes has been shown to occur through the ballast water of ships and hydrodynamic forces (Boeckner, et al., 2009; Gingold, et al., 2011; Radziejewska, et al., 2006). The small body size of nematodes seems to be a limitation to active dispersal over large distances, even though they can actively move in the sediment or swim in the water column (Jensen, 1981; Schratzberger, et al., 2004). We found 20 and 6 species to be shared between two locations for respectively 18S and COI. We did not find any species that were shared by more than two locations, which was against our expectations considering the size of our database. We found high intraspecific distance values for three species, but when checking the vouchers, we often found them missing or we were not able confirm the specimens to be the same species with enough certainty based on the vouchers. This was disappointing, as we had suspicions of these species, but we were not able to conclude anything without evidence from the vouchers. We did however find surprisingly low distance values for two species. *Monoposthia mirabilis* and *Theristus pertenuis* had a maximum intraspecific distance value of only 0,0043 and 0,0030 respectively, the former for 18S, the latter for COI. This was surprising, considering that the specimens of different locations of both *M. mirabilis* and *T. pertenuis* live in regions more than 2000 km apart, separated by the landmass of the European continent. This finding, for species as small as nematodes with presumably low dispersal capacities, is a perfect example of the meiofauna paradox and seem to confirm it. However, the other five species possibly contained cryptic species that could provide counterevidence, but without support the vouchers, we could not make a well-founded conclusion. Nevertheless, we found cryptic diversity for five species (see 5.2) within the same location, and patchily distributed genetic diversity on a local scale has been shown for nematodes (Derycke, Van Vynckt, et al., 2007). This implies that cryptic diversity between regions, when isolation is of a much larger scale, should also exist.

5.4. Improving identification of meiofaunal communities

The number of sequences that were yielded for each mock community replicate stayed in the same order of magnitude for each primer set, but varied drastically between primer sets. The 18S primer set (G18S4-22R) yielded ten thousands of sequences. The two COI primer sets resulted in a lot less: a few thousands for JB2-JB5GED and only 48-345 for JB3-JB5. This illustrates the known amplification problem for COI (Derycke, Vanaverbeke, et al., 2010).

Of the 50 different species the artificial community contained, we had sequences in our database for 46 and 24 species for 18S and COI respectively. Theoretically, this means that we should be able to identify almost the complete community for 18S and half of it for COI. In all cases, the BLAST algorithm, identifying down to species level, left relatively less OTU's unidentified, but also proposed much more incorrect ones than the more conservative UCLUST method (see Table 11). The latter can limit identification to a certain taxonomic level when further identification is unsure. When BLAST would assign the correct genus level, but the wrong species, this is considered an incorrect identification. UCLUST can refrain from naming the species in such case, only identifying the genus, and score a correct identification. Consequently, the high rate of incorrect identifications of the BLAST algorithm can partially be compensated for when we look only to the genus level (see Table 12). Here we see that the number of correct taxa does not change much, but the number of incorrect assignments lowers and even drops to less than half for 18S referencing against our own database. However, the OTU composition of the mock community replicates differed drastically from the composition of the original artificial community (Fig. 17 - 19).

The OTU picking on 99,5% similarity resulted in the most species being identified on both species and genus level (73% and 87,5% respectively for UCLUST and 61,5% and 85,5% for BLAST), suggesting

the barcoding threshold should be found around a distance value of 0,005. Identification of the mock community based on 18S was strongly biased towards *Bathylaimus*, but only the identifications by Silva referencing and by UCLUST using our own database proposed the correct species. The three different *Bathylaimus* species identified by the BLAST algorithm were all incorrect. In all cases, *Bathylaimus* alone counts for more than 45% of the OTU's, whereas the mock community theoretically only contained 5,77% *Bathylaimus* DNA. Other discrepancies of the mock replicate composition compared to the original artificial community, discussed in 4.4, were not as strong. A small bias towards one or a few sequences in the first cycles of the PCR reaction can be exponentially enlarged in the final PCR product, completely distorting the relative sequence abundances of the DNA extract. This could cause the bias towards *Bathylaimus*, but it has also been shown that the number of 18S copies can vary within a genome and at least sixfold between species (Bik, et al., 2013). This, combined with the PCR bias, could certainly cause the observed mock replicate composition. This kind of corruption of the original relative abundances be an annoyance in qualitative metabarcoding, but it proves a huge problem for quantitative metabarcoding and might even leave it unviable for 18S.

The results from COI were rather disappointing. Less than 10% of the artificial community was identified for both primer sets using UCLUST. The results using the BLAST algorithm were higher: 14-19% for JB3-JB5 and 22-24% for JB2-JB5GED. Even though UCLUST identified only a few species, they were all correct each time. The BLAST algorithm identified more species correctly, but also gave more than half as many incorrect identifications. The results for the two best threshold values, 0,0562 and 0,05, were similar, with the latter doing even better than the former, suggesting 0,05 to still be the most suitable threshold value (Derycke, Vanaverbeke, et al., 2010). The share of total OTU's that were identified was higher for the JB3-JB5 primer set than for the JB2-JB5GED set, but differed between mock replicates from roughly 3% in A1 to more than 40% in B2. This is probably partly because differences in relative abundances of sequences by chance have a big impact when the total amount of OTU's is small. For A1, there were 345 sequences, but for B2 only 48. Another reason might be the difficult amplification of COI (Kanfra, 2015), resulting a slight discrepancy of the sequences amplified in the first phases of the PCR to be greatly enlarged in the final product. For the JB2-JB5GED primer set, the OTU's identified using UCLUST were barely visible on the graph. Using the BLAST method, they counted for less than 5% of the total OTU's for each replicate. Metabarcoding identification success for COI can be as low as 7% (Cowart, et al., 2015; Leray & Knowlton, 2015). But we suspect that a few species dominate the mock replicates, just like *Bathylaimus* for 18S, and that we do not have a COI sequence for them in our database, leaving them unidentified. Overall, 18S had a higher identification success than COI, but this does not necessarily mean that the former is a better metabarcoding marker. The higher success of 18S can be a result of the more elaborate reference sequence collection and better primer matching.

6. Conclusion

Metabarcoding resulted in the most nematode identifications when a similarity threshold of 99,5% was used. This suggests that an intraspecific distance threshold value of 0,005 is more suitable than a higher one. We did not find an overall threshold value for 18S using *Adhoc*. However, the two values found for Oncholaimidae and Oxystominidae, respectively 0,0038 and 0,0016, were even smaller than 0,005. Together with the optimal similarity threshold for metabarcoding, 99,5%, this suggests that the threshold value for 18S barcoding is probably not higher than 0,005. However, no matter how small the value used, the lack of resolution on the species level (supported by our 18S histogram and the phylogenetic analysis) seems to suggest that it is not very suitable to use on the species level. For COI, we found a threshold value of 0,0562 to be the best threshold value for our database. This is close to the commonly used value of 0,05, and results from our metabarcoding analysis found these two to give the best results, with the latter even giving slightly better results.

The meiofauna paradox seems to be supported by the low intraspecific distance values of specimens for different locations for *M. mirabilis* and *T. pertenuis.* However, we also found three species with remarkably high such values that could counter this support, but we were not able to confirm that they were the same species based on the vouchers. Consequently, we will refrain from making a conclusion on the subject because of lacking evidence. Yet, we found cryptic diversity within a species in a single region for *Stilbonema brevicolle*, *Acanthopharynx micans*, *Mesacanthion* sp., *Zalonema* sp. and *Eurystomina ornata,* and this has been reported for other nematode species (Apolonio Silva De Oliveira, et al., 2012; Derycke, Backeljau, et al., 2007; Derycke, et al., 2008; Derycke, Van Vynckt, et al., 2007). Intuitively, this implies that cryptic diversity between regions, when isolation is of a much larger scale, should also exist. Future research might be able to provide us the answer.

Or results indicate that quantitative metabarcoding might not be possible for 18S, but a decent amount of species and a large amount of genera in our artificial community could be identified qualitativly. Considering the low resolution of 18S on the species level and the better result when only identifying to genus level, it seems that qualitative metabarcoding based on 18S should probably be limited to the genus level. COI is fit to use as a barcode down to species level, but the lacking reference sequence collection and low amplification success still pose a problem. A combination of both 18S and COI seems the most ideal, as they can compensate for each other's shortcomings.

A barcoding gene without any downsides to its use is yet to be found. A clear gap in the distributions of intra- and interspecific distances implies higher interspecific values, which in turn implies an overall more variable sequence. Creating universal primers with a high amplification success for less conservative sequences proves a challenge (Creer, et al., 2010; Derycke, Vanaverbeke, et al., 2010). Something else that needs to be considered is that species are not static, immutable units. They are dynamic and continuously appearing, transforming and disappearing, something that Darwin already acknowledged in his book *On the Origin of Species*, published in 1859. This means that there will always be lineages (species) transitioning into new lineages (species). Not only do species concepts conflict in this transition zone, the so-called "grey zone" (De Queiroz, 2007), but these would also fall in between the intra- and interspecific distance distributions. If they fall close to the barcode threshold value, such ambiguous cases can cause some of the specimens to be considered the same species while others would be considered different species, however ideal the barcode used may be.

We therefore conclude that the idea of finding one superior barcoding gene, that readily amplifies and that can identify every species is a search for the Holy Grail. Barcoding is a valuable addition or cost-effective alternative for species identification, especially when morphological identification proves difficult. A high throughput system like the QIIME pipeline is very promising for qualitative metabarcoding, but it is in desperate need of a high-quality database, including appropriate vouchers, and an algorithm that finds a good compromise between the number of species identified and a low error rate. There will always be species left unnoticed because of lacking amplification success and because there are still many undiscovered species, let alone the described species still absent from the reference databases. We are convinced of the possibilities of DNA barcoding and metabarcoding, but it will need to combine multiple markers to compensate for their respective flaws, and it will need to go hand in hand with traditional, morphology based taxonomy by using high-quality vouchers.

7. Summary

7.1. English summary

Nematodes account for 80-90% of all metazoans on Earth, yet only a fraction of the estimated more than one million species are formally known and described. Free-living nematodes are dominant in both density (10^5-10^7 individuals per m²) and diversity (> 10 species per cm²) in marine sediments. They fulfil important ecological roles, are a high quality food source for higher trophic levels and influence the composition of lower trophic groups. Their small body size likely limits active dispersal over large distances, but many species seem to be cosmopolitan, with their distribution spanning the globe. This has been called the "meiofauna paradox". These microscopic metazoans possess little diagnostic characters, which makes morphological identification very time consuming. The presence of cryptic species (different species that are classified as a single species because they are morphologically indistinguishable), enlarges this problem even further, and it might cause nematode diversity to be seriously underestimated.

Molecular tools provide a promising solution to the troublesome morphological identification, both in terms of speed and reliability, and have been rapidly advancing. DNA barcoding has become a popular identification tool. One or more genes, often the small subunit ribosomal DNA (18S) or the mitochondrial cytochrome oxidase c subunit 1 (COI), are sequenced and compared against sequences from identified specimens (reference sequences). If a close enough match is found, the unknown specimen can be considered the same species as the one linked to the matched reference sequence. However, DNA barcoding still has several weak points. Whether or not the specimens are considered the same species, depends on the similarity percentage that is used as threshold (and be considered a 'match'). The choice of this similarity percentage in turn depends on the presence of a 'barcoding gap', a high degree of separation between intraspecific and interspecific distance distributions (variability in base pair composition of sequences of respectively the same or different species). A reference database is also needed that includes the DNA sequences linked to morphologically identified species. Sequences in public databases such as Genbank are often not identified to species level and lack vouchers, containing images or videos of the specimen, to verify if the identification later on. Their focus remains on 18S, and provide little COI sequences for marine nematodes.

18S rDNA has been traditionally used because of the availability of universal nematode primers, thanks to the conserved flanking regions of the sequence, and its phylogenetic resolution at genus and higher taxon levels. Its major downside is that it lacks resolution at the species. To compensate for this shortcoming, it can be used in combination with the mitochondrial COI gene. Except for some taxa (e.g. Anthozoa, Porifera…), the COI gene has been proven efficient in identifying Metazoan species. However, nuclear copies of the COI gene ("numts"), often inactive and rapidly mutating, may cause an overestimation of the taxonomic diversity. Amplification success for free-living nematode is known to give problems, but primers developed specifically for this group perform better and allow a range of nematode species to be amplified.

In this project, we have build a marine nematode reference database containing species all across the phylum that are identified, vouchered and sequenced for 18S and/or COI, from six different regions around the globe, using custom Python scripts. We have calculated the intra-interspecific distance gap (the 'barcoding gap') for both genes based on p-distance, and calculated a threshold distance value for species identification. This allowed us to identify presumably cosmopolitan "species" and track down cryptic species. Finally, we will test the applicability of both our database and the calculated threshold values for identification of nematode communities. This will be done using a metagenetic approach, with an artificial community with known species that will be compared against our database.

The database contains 586 specimens and 756 sequences (461 for 18S and 295 for COI), including representatives of 115 genera from 37 families. Intraspecific distances ranged from 0-0,1939 and from 0-0,2883 for 18S and COI respectively. Interspecific distances ranged from 0-0,3820 and from 0,0025-0,5455 for 18S and COI respectively. Using linear regression, the R package *Adhoc* found a significant threshold value of 0,0562 for COI, with a relative error (RE; the number of incorrect identifications divided by the total number of identifications) of less than 0,05. For 18S, we could not find a significant value. Here the lowest distance value (0,00597) other than zero still had a RE of 0,247. Looking at the six best represented families separately for 18S, we only found a significant threshold value of 0,0038 and 0,0016 for respectively Oncholaimidae and Oxystominidae.

Phylogenetic analysis using a maximum likelihood (ML) tree and a neighbor-joining (NJ) tree resulted in 16 clusters to be strongly supported (bootstrap value (BV) ≥ 95) and five well supported (BV ≥ 90) for 18S. For COI, 59 clusters were strongly supported (BV ≥ 95), three were well supported (BV ≥ 90). Not only were more strongly supported clusters found for COI as opposed to 18S, but specimens of the same genus, family or order more often clustered together without any specimens not belonging to the group (outsiders). For 20 species, we found remarkably high intraspecific distance values (> 0,05). 13 were identified as potential cryptic species, and five of them, *Stilbonema brevicolle, Acanthopharynx micans, Mesacanthion* sp., *Zalonema* sp. and *Eurystomina ornata* were found to contain 'true cryptic species', when no morphological difference was found based on the vouchers.

Based on 18S and COI sequences respectively, 20 and 6 species were shared by two locations and 215 and 135 species respectively were unique to their location. No species were shared by more than two locations. When excluding species that were contamination, wrongly identified, had no voucher available or that could not be identified for certain based on the vouchers, however, we were left with only seven species suitable for inter-location comparison. Of these seven species, we found three species that showed high intra-specific values: *Oncholaimus campylocercoides, Paracomesoma dubium,* and *Theristus flevensis*. However, from none of these we could conclude if they contained different cryptic species for different locations because vouchers were missing or not clear enough to make a conclusion. The values for *M. mirabilis* and *T. pertenuis,* respectively 0,0043 and 0,0030, were not high enough to consider them as potentially containing different (cryptic) species, which seems to support the meiofauna paradox. However, the five previously mentioned species for which we did find high values could have countered this. Nevertheless, we found cryptic diversity within a species in a single region for *Stilbonema brevicolle, Acanthopharynx micans, Mesacanthion* sp., *Zalonema* sp. and *Eurystomina ornata,* and this has been reported for other nematode species. Intuitively, this implies that cryptic diversity between regions, when isolation is of a much larger scale, should also exist.

Metabarcoding the artificial community using the QIIME pipeline yielded very different results for the two genes. When referencing our own database, OTU picking (clustering similar sequences in the sample as one artificial 'species') on 99,5% similarity for 18S (G18S4-22R primer set) resulted in the most species being identified on both species and genus level (73% and 87,5% respectively for the UCLUST algorithm and 61,5% and 85,5% for BLAST). Together with the previously found threshold values for Oncholaimidae and Oxystominidae that were even lower, this seems to suggest that a threshold value for 18S barcoding is probably not higher than 0,005. The results from COI were rather disappointing. Less than 10% of the artificial community was identified for both primer sets using UCLUST. The results using the BLAST algorithm were higher: 14-19% for JB3-JB5 and 22-24% for the JB2-JB5GED primer set. Even though UCLUST identified only a few species, they were all correct each time. The BLAST algorithm identified more species correctly, but also gave more than half as many incorrect identifications. The results for the two best threshold values, 0,0562 and 0,05, were similar, with the latter doing even better than the former, suggesting 0,05 to still be the most suitable threshold value for COI. The share of total OTU's that were identified for the JB3-JB5 primer set differed between mock replicates

from roughly 3% in replicate A1 to more than 40% in B2. For the JB2-JB5GED primer set, the OTU's identified using UCLUST were barely visible on the graph. Using the BLAST method, they counted for less than 5% of the total OTU's for each replicate.

Identification of the mock community based on 18S was strongly biased towards *Bathylaimus*, but only the identifications by Silva referencing and by UCLUST using our own database proposed the correct species. The three different *Bathylaimus* species identified by the BLAST algorithm were all incorrect. In all cases, *Bathylaimus* alone counts for more than 45% of the OTU's, whereas the mock community theoretically only contained 5,77% *Bathylaimus* DNA. Other discrepancies of the mock replicate composition compared to the original artificial community were not as strong, but some species were only present in very low abundance in the mock replicates. A small bias towards one or a few sequences in the first cycles of the PCR reaction can be exponentially enlarged in the final PCR product, completely distorting the relative sequence abundances of the DNA extract. This could cause the bias towards *Bathylaimus*, but it has also been shown that the number of 18S copies can vary within a genome and at least sixfold between species. This, combined with the PCR bias, could certainly cause the observed mock replicate composition. This kind of corruption of the original relative abundances be an annoyance in qualitative metabarcoding, but it proves a huge problem for quatitative metabarcoding and might even leave it unviable for 18S.

For COI, we suspect that a few species dominate the mock replicates, just like *Bathylaimus* for 18S, and that we do not have a COI sequence for them in our database, leaving them unidentified. Overall, 18S had a higher identification success than COI, but this does not necessarily mean that the former is a better metabarcoding marker. The higher success of 18S can be a result of the more elaborate reference sequence collection and better primer matching, and limiting identification to genus level yields better results. COI is fit to use as a barcode down to species level, but the lacking reference sequence collection and low amplification success still pose a problem.

Barcoding is a valuable addition or cost-effective alternative for morphological species identification, especially when morphological identification proves difficult. However, there will always be species left unnoticed because of lacking amplification success and because there are still many undiscovered species, let alone the described species still absent from the reference databases. We are convinced of the possibilities of DNA barcoding and metabarcoding, but it will need to combine multiple markers to compensate for their respective flaws, and it will need to go hand in hand with traditional, morphology based taxonomy by using high-quality vouchers.

7.2. Dutch summary

Nematoden omvatten 80-90% van alle Metazoa op Aarde, maar slechts een fractie van de meer dan één miljoen soorten zijn gekend en formeel beschreven. Ze zijn dominant in zowel densiteit (10^5-10^7 individuen per m^2) als diversiteit (> 10 soorten per cm^2) in marine sedimenten. Ze vervullen belangrijke ecologische rollen, zijn een hoogwaardige voedingsbron voor hogere trofische niveau's en beïnvloeden de samenstelling van lagere trofische groepen. Hun kleine lichaamsgrootte beperkt waarschijnlijk hun actieve verspreiding over lange afstanden, maar veel soorten lijken kosmopoliet te zijn, met een schijnbaar wereldwijde verspreiding. Dit wordt de "meiofauna paradox" genoemd. Deze microscopische Metazoa hebben weinig diagnostische kenmerken, wat morfologische identificatie heel tijdsintensief maakt. De aanwezigheid van cryptische soorten (verschillende soorten die ingedeeld worden als één soort omdat ze morfologisch niet te onderscheiden zijn), vergroten dit probleem nog verder, en kunnen voor een serieuze onderschatting van nematodendiversiteit zorgen.

Moleculaire technieken zijn een veelbelovende oplossing voor de problematische morfologische identificatie, zowel in termen van snelheid als betrouwbaarheid. DNA barcoding is een populaire identificatiemethode geworden. Eén of meer genen, vaak het small subunit ribosomaal DNA (18S) of het mitochondriaal cytochroom oxidase c subunit 1 (COI), worden gesequeneerd en vergeleken met sequenties van geïdentificeerde specimens (referentiesequenties). Als een voldoende goede match gevonden wordt, kan het onbekende specimen beschouwd worden als dezelfde soort als die van de overeenkomende referentiesequentie. DNA barcoding heeft echter nog steeds enkele zwakke punten. Of de specimens al dan niet als dezelfde soort worden beschouwd, hangt af van hoe sterk de sequenties overeenkomen (uitgedrukt in percent als 'similarity percentage') en wat als grens van minimale overeenkomst genomen wordt om als match beschouwd te worden. De keuze van dit similarity percentage hang op zijn beurt weer af van de aanwezigheid van een 'barcoding gap', een sterke graad van scheiding tussen de intraspecifieke en interspecifieke distance distributies (de variatie in basenpaarsamenstelling van sequenties van respectievelijk dezelfde of andere soorten). Er is ook een referentiedatabase nodig die DNA sequenties bevat die gelinkt zijn aan een morfologisch geïdentificeerde soorten. Sequenties in publieke databases zoals Genbank zijn vaak niet geïdentificeerd tot op soortsniveau en missen vouchers (die foto's of video's van het specimen bevatten) om later de identificatie na te kunnen gaan. Ze zijn bovendien vooral gericht op 18S, en bieden weinig COI sequenties voor marine nematoden.

Traditioneel wordt 18S rDNA gebruikt omdat er universele primers voor beschikbaar zijn, dankzij de sterk geconserveerde naastliggende regio's van de sequentie, en voor zijn goede fylogenetische resolutie op genus- en hoger niveau. Het grootste nadeel is de slechtere resolutie op soortniveau. Om dit te compenseren, kan het in combinatie met het mitochondriaal COI gen gebruikt worden. Met uitzondering van sommige taxa (vb. Anthozoa en Porifera) heeft COI zijn efficiëntie bewezen in het identificeren van Metazoa-soorten. Nucleaire kopieën van het COI gen ("numts"), die vaak inactief zijn en snel muteren, kunnen echter voor een overschatting van de taxonomische diversiteit zorgen. Amplificatiesucces voor vrijlevende nematoden kan problemen geven, maar primers die speciaal voor deze groep zijn ontwikkeld doen het beter en laten toe om een verscheidenheid aan nematodensoorten te amplificeren.

In dit project bouwen we met zelfgeschreven Python scripts een referentiedatabase van mariene nematoden, die soorten bevat van over de hele stam die geïdentificeerd, gevouchered en gesequeneerd werden voor 18S en/of COI, uit zes verschillende delen van de wereld. We hebben de intra-interspecific distance gap (the 'barcoding gap') gebaseerd op p-distance berekend, alsook de drempelwaarde (threshold value) voor soortidentificatie. Zo kunnen we de schijnbaar kosmopoliete 'soorten' identificeren en cryptische soorten opzoeken. Tenslotte hebben we ook de toepasbaarheid van onze database en de drempelwaarden getest voor identificatie van nematodengemeenschappen, door gebruik te maken van een

kunstmatige gemeenschap en deze metagenetisch (het hele staal in één keer) met onze database te vergelijken.

De database bevat 586 specimens en 756 sequenties (461 voor 18S en 295 voor COI), met vertegenwoordigers van 115 genera uit 37 families. Intraspecfieke distances gingen van 0-0,1939 en van 0-0,2883 voor 18S en COI respectievelijk. Interspecifieke distances gingen van 0-0,3820 en van 0,0025-0,5455 voor 18S en COI respectievelijk. Het R pakket *Adhoc* vond door middel van lineaire regressie een significante drempelwaarde van 0,0562 voor COI, met een relatieve error (RE; het aantal verkeerde identificaties gedeeld door het totaal aantal identificaties) van minder dan 0,05. Voor 18S konden we geen significante waarde vinden. De laagste distance waarde (0,00597) die niet nul was had nog steeds een RE van 0,247. Wanneer we voor 18S naar de zes best vertegenwoordigde families keken, vonden we alleen een significante drempelwaarde van 0,0038 en 0,0016 voor respectievelijk Oncholaimidae en Oxystominidae.

Voor de fylogenetische analyse maakten we gebruik van een maximum likelihood (ML) boom en een neighbor-joining (NJ) boom. Deze resulteerden in 16 sterk ondersteunde (bootstrap waarde (BW) $\geq$ 95) en vijf goed ondersteunde (BW $\geq$ 90) clusters voor 18S en 59 sterk ondersteunde en drie goed ondersteunde clusters voor COI. Er waren niet alleen meer sterk ondersteunde clusters voor COI in vergelijking met 18S, maar specimens van hetzelfde genus, familie of orde clusterden vaker samen zonder enige soort die niet tot deze groep behoorde ('outsiders'). Voor 20 soorten vonden we opvallend hoge intraspecifieke distance waarden (>0,05). 13 hiervan werden geïdentificeerd als potentiële cryptische soorten, en vijf (*Stilbonema brevicolle, Acanthopharynx micans, Mesacanthion* sp., *Zalonema* sp. en *Eurystomina ornata*) bevatten 'echte cryptische soorten', waarvoor we geen morfologisch verschil vonden op basis van de vouchers.

In totaal werden er 20 en 6 soorten gevonden in onze database met respectievelijk een 18S of COI sequentie die voorkwamen op twee locaties. Na uitsluiten van soorten die contaminatie waren, geen voucher hadden of die niet met zekerheid konden worden geïdentificeerd op basis van de vouchers, bleven er nog zeven soorten over waarvoor we de sequenties konden vergelijken tussen locaties. Van deze zeven soorten vonden we er drie die hoge intraspecifieke distancewaarden hadden: *Oncholaimus campylocercoides, Paracomesoma dubium*, en *Theristus flevensis*. Van geen van deze soorten konden we echter besluiten of ze cryptische soorten bevatten, omdat de vouchers ontbraken of niet duidelijk genoeg waren. De waarden van *M. mirabilis* en *T. pertenuis*, respectively 0,0043 en 0,0030, waren niet hoog genoeg om ze te overwegen als cryptische soorten. Dit lijkt de meiofauna paradox te ondersteunen, maar de hoge waarden van eerdergenoemde vijf soorten hadden dit kunnen tegenspreken als we ze als cryptische soorten hadden kunnen besluiten. Daarnaast vonden we cryptische diversiteit binnen een locatie voor *Stilbonema brevicolle, Acanthopharynx micans, Mesacanthion* sp., *Zalonema* sp. and *Eurystomina ornata,* en het is ook al gerapporteerd voor andere nematodensoorten. Als logisch gevolg zou cryptische diversiteit tussen regio's, door isolatie op een veel grotere schaal, dus waarschijnlijk ook moeten bestaan.

De resultaten van de metabarcoding identificatie van de kunstmatige gemeenschap gaf sterk verschillende resultaten voor de twee genen. Wanneer we onze eigen database als referentie gebruikten, werden er voor 18S (G18S4-22R primer set) het meest soorten geïdentificeerd met OTU clustering (clusteren van gelijkaardige sequenties in het staal als één kunstmatige 'soort') op 99,5% overeenkomst op zowel soort- als genusniveau (73% en 87,5% respectievelijk voor het UCLUST algoritme en 61,5% en 85,5% voor BLAST). Samen met de eerder gevonden drempelwaarden voor Oncholaimidae en Oxystominidae, lijkt het erop dat de drempelwaarde voor 18S barcoding niet hoger moet gezocht worden dan 0,005. Voor COI werd minder dan 10% van de kunstmatige gemeenschap werd met UCLUST geïdentificeerd voor elk van de primer sets. De resultaten voor het BLAST algoritme waren hoger: 14-19% voor JB3-JB5 en 22-24% voor de JB2-JB5GED primer set. Hoewel UCLUST slechts een paar soorten kon identificeren, waren ze

wel allemaal juist. Het BLAST algoritme identificeerde meer soorten correct, maar gaf ook meer dan de helft zoveel verkeerde identificaties. De resultaten voor de twee beste drempelwaarden, 0,0562 en 0,05, waren gelijkaardig. De laatste deed het zelfs nog iets beter dan de eerste, wat erop wijst dat 0,05 nog steeds de meest geschikte drempelwaarde voor COI zou zijn. Het deel OTU's die geïdentificeerd werden voor de JB3-JB5 primer set verschilden tussen de replicaten van de kunstmatige gemeenschap van rond de 3% in tot meer dan 40%. Voor de JB2-J5GED primer set waren de OTU's geïdentificeerd met UCLUST amper zichtbaar op de grafiek. BLAST identificeerde minder dan 5% van de OTU's.

Identificatie van de kunstmatige gemeenschap op basis van 18S was sterk afwijkend naar *Bathylaimus*, maar alleen de Silva database als referentie en onze eigen referentiedatabase met UCLUST gaven de juiste soorten. De drie *Bathylaimus* soorten die met BLAST geïdentificeerd werden waren alle drie fout. In alle gevallen maakte *Bathylaimus* meer dan 45% van het aantal OTU's uit, terwijl de kunstmatige gemeenschap slechts 5,77% *Bathylaimus* DNA bevatte. Andere discrepanties in de samenstelling van de replicaten waren niet zo uitgesproken, maar enkele soorten waren in slechts erg lage abundanties aanwezig. Een klein verschil in abundantie van een paar sequenties in het begin van de PCR reactie kan exponentieel uitvergroot worden in het PCR product, en de originele abundanties compleet vervormen. Het is ook aangetoond dat het aantal 18S kopieën kan variëren binnen een genoom, en minstens in zesvoud tussen soorten. Samen met de PCR afwijking kan dit de geobserveerde samenstelling van de replicaten veroorzaken. Dit soort vervormingen van de originele relatieve abundanties kan vervelend zijn in kwalitatieve metabarcoding, maar het is een enorm en misschien zelfs onoverkomelijk probleem voor kwalitatieve metabarcoding.

Voor COI vermoeden we dat een paar soorten de replicaten domineren, net zoals *Bathylaimus* voor 18S, en dat ze ongeïdentificeerd blijven omdat we er geen COI sequentie voor hebben in onze database. Over het algemeen had 18S meer succes dan COI voor metabarcoding identificatie dan COI, maar dat kan het resultaat zijn van een uitgebreidere referentiecollectie en betere primers. Bovendien gaf beperking tot het genusniveau ook betere resultaten. COI is geschikt als barcode tot op soortniveau, maar de gebrekkige referentiecollectie en de lage amplificatie zijn nog steeds een probleem. Barcoding is een waardevolle aanvulling of rendabel alternatief voor morfologische soortidentificatie. Er zullen echter altijd soorten onopgemerkt blijven door slechte amplificatie en omdat er nog steeds vele soorten onontdekt zijn, laat staan de beschreven soorten die nog steeds ontbreken in referentiedatabases. Wij zijn overtuigd van de mogelijkheden van DNA barcoding en metabarcoding, maar het zal meerdere genen moeten combineren om voor elkaars zwakke punten te compenseren, en het zal hand in hand moeten gaan met traditionele morfologie gebaseerde taxonomie, door gebruik te maken van hoogwaardige vouchers.

8. Acknowledgments

I would like to wholeheartedly thank dr. Sofie Derycke for providing this exciting thesis subject, giving me the chance to work on the fascinating subject that barcoding is, and even allowing me to do some scripting. Thank you for your contagious enthusiasm and all help and feedback during the project, even when I kept bumping into QIIME errors. It was a delight to work with you. I would like to thank dr. Katja Guilini, for providing me feedback and pointing out typos in the draft, and for sharing her time, expertise and patience when reviewing the vouchers. I would also like to thank Daniel Apolônio Silva de Oliveira for teaching me how to isolate, prepare and voucher nematodes and Annelien Rigaux for her guidance during the following molecular work. I would also like to express my gratitude to Maickel Armenteros and his co-researchers, Xorla Kanfra, Anouk D'Hont, Christopher Oche Eche, Fehmi Boufahja, Xuan Phuong Nguyen Thi and all the researchers from around the world who did all the hard work on obtaining the sequences that I was able to use in this study. Without you, this project would not have been possible. A sincere thank you to you all!

9. References

Adams, B. J. (1998). Species Concepts and the Evolutionary Paradigm in Modem Nematology. *Journal of Nematology, 30*(1), 1-21.

Adams, B. J. (2001). The species delimitation uncertainty principle. *Journal of Nematology, 33*(4), 153.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410. doi: http://dx.doi.org/10.1016/S0022-2836(05)80360-2

Apolonio Silva De Oliveira, D., Decraemer, W., Holovachov, O., Burr, J., Tandingan De Ley, I., De Ley, P., Moens, T., & Derycke, S. (2012). An integrative approach to characterize cryptic species in the Thoracostoma trachygaster Hope, 1967 complex (Nematoda: Leptosomatidae). *Zoological Journal of the Linnean Society, 164*(1), 18-35.

Armenteros, M., Rojas-Corzo, A., Ruiz-Abierno, A., Derycke, S., Backeljau, T., & Decraemer, W. (2014). Systematics and DNA barcoding of free-living marine nematodes with emphasis on tropical desmodorids using nuclear SSU rDNA and mitochondrial COI sequences. *Nematology, 16*(8), 979-989.

Bhadury, P., Austen, M. C., Bilton, D. T., Lambshead, P. D., Rogers, A. D., & Smerdon, G. R. (2006). Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. *Marine Ecology Progress Series, 320*, 1-9.

Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K., Meier, R., Winker, K., Ingram, K. K., & Das, I. (2007). Cryptic species as a window on diversity and conservation. *Trends in Ecology & Evolution, 22*(3), 148-155.

Bik, H. M., Fournier, D., Sung, W., Bergeron, R. D., & Thomas, W. K. (2013). Intra-genomic variation in the ribosomal repeats of nematodes. *PloS one, 8*(10), e78230.

Blaxter, M. L., De Ley, P., Garey, J. R., Liu, L. X., Scheldeman, P., Vierstraete, A., Vanfleteren, J. R., Mackey, L. Y., Dorris, M., Frisse, L. M., Vida, J. T., & Thomas, W. K. (1998). A molecular evolutionary framework for the phylum Nematoda. *Nature, 392*(6671), 71-75. doi: 10.1038/32160

Boeckner, M. J., Sharma, J., & Proctor, H. (2009). Revisiting the meiofauna paradox: dispersal and colonization of nematodes and other meiofaunal organisms in low-and high-energy environments. *Hydrobiologia, 624*(1), 91-106.

Bowles, J., Blair, D., & McManus, D. P. (1992). Genetic variants within the genus Echinococcus identified by mitochondrial DNA sequencing. *Molecular and biochemical parasitology, 54*(2), 165-173.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., & Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. [10.1038/nmeth.f.303]. *Nat Meth, 7*(5), 335-336. doi: http://www.nature.com/nmeth/journal/v7/n5/suppinfo/nmeth.f.303_S1.html

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution, 17*(4), 540-552.

Collins, R. A., Boykin, L. M., Cruickshank, R. H., & Armstrong, K. F. (2012). Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. *Methods in Ecology and Evolution, 3*(3), 457-465.

Cowart, D. A., Pinheiro, M., Mouchel, O., Maguer, M., Grall, J., Miné, J., & Arnaud-Haond, S. (2015). Metabarcoding is powerful yet still blind: a comparative analysis of morphological and molecular surveys of seagrass communities. *PloS one, 10*(2), e0117562.

Creer, S., Fonseca, V., Porazinska, D., GIBLIN-DAVIS, R., Sung, W., Power, D., Packer, M., Carvalho, G., Blaxter, M., & Lambshead, P. (2010). Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Molecular Ecology, 19*(s1), 4-20.

D'Hont, A. (2014). *Ocean Acidification: A Meiofaunal Perspective.* Department of Biology – Marine Biology. Unpublished master's thesis, University of Ghent. Master of Science in Marine Biodiversity and Conservation.

De Ley, P., De Ley, I. T., Morris, K., Abebe, E., Mundo-Ocampo, M., Yoder, M., Heras, J., Waumann, D., Rocha-Olivares, A., & Burr, A. J. (2005). An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360*(1462), 1945-1958.

De Mesel, I., Derycke, S., Moens, T., Van der Gucht, K., Vincx, M., & Swings, J. (2004). Top-down impact of bacterivorous nematodes on the bacterial community structure: a microcosm study. *Environmental Microbiology, 6*(7), 733-744.

De Queiroz, K. (2007). Species concepts and species delimitation. *Syst Biol, 56*(6), 879-886. doi: 10.1080/10635150701701083

Derycke, S., Backeljau, T., Vlaeminck, C., Vierstraete, A., Vanfleteren, J., Vincx, M., & Moens, T. (2007). Spatiotemporal analysis of population genetic structure in Geomonhystera disjuncta (Nematoda, Monhysteridae) reveals high levels of molecular diversity. *Marine Biology, 151*(5), 1799-1812.

Derycke, S., De Ley, P., Tandingan De Ley, I., Holovachov, O., Rigaux, A., & Moens, T. (2010). Linking DNA sequences to morphology: cryptic diversity and population genetic structure in the marine nematode Thoracostoma trachygaster (Nematoda, Leptosomatidae). *Zoologica Scripta, 39*(3), 276-289.

Derycke, S., Remerie, T., Backeljau, T., Vierstraete, A., Vanfleteren, J., Vincx, M., & Moens, T. (2008). Phylogeography of the Rhabditis (Pellioditis) marina species complex: evidence for long-distance dispersal, and for range expansions and restricted gene flow in the northeast Atlantic. *Molecular ecology, 17*(14), 3306-3322.

Derycke, S., Van Vynckt, R., Vanoverbeke, J., Vincx, M., & Moens, T. (2007). Colonization patterns of Nematoda on decomposing algae in the estuarine environment: Community assembly and genetic structure of the dominant species Pellioditis marina. *Limnology and Oceanography, 52*(3), 992-1001.

Derycke, S., Vanaverbeke, J., Rigaux, A., Backeljau, T., & Moens, T. (2010). Exploring the use of cytochrome oxidase c subunit 1 (COI) for DNA barcoding of free-living marine nematodes. *PLoS One, 5*(10), e13716.

Eche, C. O. (2012). *Identification of marine nematode communities through DNA barcoding: do mitochondrial COI sequences outperform the ribosomal 18S gene?* Department of Biology. Unpublished master's thesis, University of Ghent. Master of Science in Nematology.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics, 26*(19), 2460-2461. doi: 10.1093/bioinformatics/btq461

Floyd, R., Abebe, E., Papert, A., & Blaxter, M. (2002). Molecular barcodes for soil nematode identification. *Molecular ecology, 11*(4), 839-850.

Giere, O. (2008). *Meiobenthology: the microscopic motile fauna of aquatic sediments*: Springer Science & Business Media.

Gingold, R., Ibarra-Obando, S. E., & Rocha-Olivares, A. (2011). Spatial aggregation patterns of free-living marine nematodes in contrasting sandy beach micro-habitats. *Journal of the Marine Biological Association of the United Kingdom, 91*(03), 615-622. doi: doi:10.1017/S0025315410001128

Hamels, I., Moens, T., Muylaert, K., & Vyverman, W. (2001). Trophic interactions between ciliates and nematodes from an intertidal flat. *Aquatic Microbial Ecology, 26*(1), 61-72.

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences, 270*(1512), 313-321. doi: 10.1098/rspb.2002.2218

Heip, C., Vincx, M., & Vranken, G. (1985). *The ecology of marine nematodes*: Aberdeen University Press.

Hellberg, M. E., Burton, R. S., Neigel, J. E., & Palumbi, S. R. (2002). Genetic assessment of connectivity among marine populations. *Bulletin of marine science, 70*(Supplement 1), 273-290.

Holterman, M., van der Wurff, A., van den Elsen, S., van Megen, H., Bongers, T., Holovachov, O., Bakker, J., & Helder, J. (2006). Phylum-Wide Analysis of SSU rDNA Reveals Deep Phylogenetic Relationships among Nematodes and Accelerated Evolution toward Crown Clades. *Molecular Biology and Evolution, 23*(9), 1792-1800. doi: 10.1093/molbev/msl044

Jensen, P. (1981). Phyto-chemical sensitivity and swimming behavior of the free-living marine nematode Chromadorita tenuis. *Marine Ecology Progress Series*(2).

Kanfra, X. (2015). *Success of CO1 and 18S sequences for species identification of marine nematodes.* Master of Science in Nematology Master's, University of Ghent, Unpublished master's thesis.
Kiontke, K. C., Félix, M.-A., Ailion, M., Rockman, M. V., Braendle, C., Pénigault, J.-B., & Fitch, D. H. (2011). A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits. *BMC Evolutionary Biology, 11*(1), 339.
Lallias, D., Hiddink, J. G., Fonseca, V. G., Gaspar, J. M., Sung, W., Neill, S. P., Barnes, N., Ferrero, T., Hall, N., Lambshead, P. J. D., Packer, M., Thomas, W. K., & Creer, S. (2015). Environmental metabarcoding reveals heterogeneous drivers of microbial eukaryote diversity in contrasting estuarine ecosystems. [Original Article]. *ISME J, 9*(5), 1208-1221. doi: 10.1038/ismej.2014.213
Lawton, J. H., Bignell, D., Bolton, B., Bloemers, G., Eggleton, P., Hammond, P., Hodda, M., Holt, R., Larsen, T., & Mawdsley, N. (1998). Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forest. *Nature, 391*(6662), 72-76.
Leduc, D. (2009). Description of Oncholaimus moanae sp. nov.(Nematoda: Oncholaimidae), with notes on feeding ecology based on isotopic and fatty acid composition. *Journal of the Marine Biological Association of the United Kingdom, 89*(02), 337-344.
Leray, M., & Knowlton, N. (2015). DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proceedings of the National Academy of Sciences, 112*(7), 2076-2081.
Meier, R., Shiyang, K., Vaidya, G., & Ng, P. K. (2006). DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic biology, 55*(5), 715-728.
Meldal, B. H., Debenham, N. J., De Ley, P., De Ley, I. T., Vanfleteren, J. R., Vierstraete, A. R., Bert, W., Borgonie, G., Moens, T., & Tyler, P. A. (2007). An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa. *Molecular phylogenetics and evolution, 42*(3), 622-636.
Nguyen Thi, X. P. (2014). *The free living marine nematodes of the Tien Yen Estuary, Vietnam: An investigation on biodiversity using morphological and molecular methods.* Department of Biology. Unpublished master's thesis, University of Ghent. Master of Science in Nematology.
Pfenninger, M., & Schwenk, K. (2007). Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. *BMC evolutionary biology, 7*(1), 1.
Porazinska, D. L., GIBLIN-DAVIS, R. M., Faller, L., Farmerie, W., Kanzaki, N., Morris, K., Powers, T. O., Tucker, A. E., Sung, W., & Thomas, W. K. (2009). Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. *Molecular ecology resources, 9*(6), 1439-1450.
Powers, T. (2004). Nematode molecular diagnostics: from bands to barcodes. *Annu. Rev. Phytopathol., 42*, 367-383.
Radziejewska, T., Gruszka, P., & Rokicka-Praxmajer, J. (2006). A home away from home: a meiobenthic assemblage in a ship's ballast water tank sediment. *Oceanologia, 48*(S).
Ratnasingham, S., & Hebert, P. D. N. (2007). bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes, 7*(3), 355-364. doi: 10.1111/j.1471-8286.2007.01678.x
Schratzberger, M., Whomersley, P., Warr, K., Bolam, S., & Rees, H. (2004). Colonisation of various types of sediment by estuarine nematodes via lateral infaunal migration: a laboratory study. *Marine Biology, 145*(1), 69-78.
Sonet, G., Jordaens, K., Nagy, Z. T., Breman, F. C., De Meyer, M., Backeljau, T., & Virgilio, M. (2013). Adhoc: an R package to calculate ad hoc distance thresholds for DNA barcoding identification. *ZooKeys*(365), 329-336. doi: 10.3897/zookeys.365.6034
Song, H., Buhay, J. E., Whiting, M. F., & Crandall, K. A. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences, 105*(36), 13486-13491. doi: 10.1073/pnas.0803076105
Srivathsan, A., & Meier, R. (2012). On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. *Cladistics, 28*(2), 190-194. doi: 10.1111/j.1096-0031.2011.00370.x
Sudhaus, W., & Kiontke, K. (2007). Comparison of the cryptic nematode species Caenorhabditis brenneri sp. n. and C. remanei (Nematoda: Rhabditidae) with the stem species pattern of the Caenorhabditis Elegans group. *Zootaxa, 1456*, 45-62.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*, *30*(12), 2725-2729. doi: 10.1093/molbev/mst197

Van Regenmortel, M. H. V. (2010). Logical puzzles and scientific controversies: The nature of species, viruses and living organisms. *Systematic and Applied Microbiology*, *33*(1), 1-6. doi: http://dx.doi.org/10.1016/j.syapm.2009.11.001

Villesen, P. (2007). FaBox: an online toolbox for fasta sequences. *Molecular Ecology Notes*, *7*(6), 965-968. doi: 10.1111/j.1471-8286.2007.01821.x

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*, *74*(11), 5088-5090.

Xue, Q. (2013). *Nematode diversity of soil and litter in Mount Hamiguitan, the Philippines, with morphological and phylogenetic analysis of selected species.* Department of Biology. Unpublished master's thesis, University of Ghent. Master of Science in Nematology.

Zauner, H., Mayer, W. E., Herrmann, M., Weller, A., Erwig, M., & Sommer, R. J. (2007). Distinct patterns of genetic variation in Pristionchus pacificus and Caenorhabditis elegans, two partially selfing nematodes with cosmopolitan distribution. *Molecular ecology*, *16*(6), 1267-1280.

10. Appendix

10.1. Taxonomic list

Taxonomy	Number of 18S sequences	Number of COI sequences	Total number of specimens
Chromadorida	**106**	**67**	**141**
Axonolaimidae	**1**	**0**	**1**
Ascolaimus	1	0	1
Chromadoridae	**27**	**14**	**30**
Dichromadora	9	2	9
Endeolophos	2	2	3
Hypodontolaimus	3	2	3
Neochromadora	1	1	1
Prochromadorella	1	0	1
Ptycholaimellus	7	6	9
Rhips	1	0	1
Spilophorella	3	1	3
Comesomatidae	**19**	**7**	**25**
Sabatieria	19	7	25
Cyatholaimidae	**40**	**38**	**65**
Longicyatholaimus	4	3	4
Marylynnia	5	24	25
Metacyatholaimus	1	1	1
Paracanthonchus	8	1	9
Paracomesoma	13	0	13
Paracyatholaimus	3	2	3
Praeacanthonchus	6	7	10
Desmodoridae	**4**	**2**	**4**
Onyx	4	2	4
Ethmolaimidae	**1**	**0**	**1**
Ethmolaimus	1	0	1
Neotonchidae	**5**	**3**	**5**
Comesa	2	2	2
Gomphionema	3	1	3
Selachinematidae	**9**	**3**	**10**

Cheironchus	4	3	5
Halichoanolaimus	4	0	4
Latronema	1	0	1
Enoplida	**104**	**75**	**137**
Anoplostomatidae	**12**	**4**	**13**
Anoplostoma	12	4	13
Anticomidae	**1**	**1**	**1**
Cephalanticoma	1	1	1
Enchelidiidae	**9**	**9**	**13**
Bathyeurystomina	2	0	2
Calyptronema	1	4	4
Eurystomina	5	4	5
Polygastrophora	1	1	2
Enoplidae	**5**	**4**	**7**
Adoncholaimus	3	3	4
Enoplus	1	1	2
Oxyonchus	1	0	1
Ironidae	**7**	**5**	**9**
Ironus	1	1	1
Thalassironus	0	1	1
Trissonchulus	6	3	7
Leptosomatidae	**2**	**2**	**2**
Synonchus	2	2	2
Oncholaimidae	**31**	**25**	**47**
Metoncholaimus	1	0	1
Meyersia	5	0	5
Oncholaimellus	2	2	2
Oncholaimus	15	17	29
Viscosia	8	6	10
Oxystominidae	**16**	**13**	**21**
Halalaimus	7	7	11
Litinium	4	3	5
Oxystomina	5	3	5
Phanodermatidae	**1**	**1**	**1**
Micoletzkyia	1	1	1
Thoracostomopsidae	**8**	**5**	**8**
Enoploides	2	0	2
Enoplolaimus	2	1	2
Mesacanthion	4	4	4
Tripyloididae	**12**	**6**	**15**
Bathylaimus	8	4	10
Tripyloides	4	2	5
Monhysterida	**105**	**74**	**139**
Linhomoeidae	**25**	**9**	**32**
Linhomoeus	2	0	2
Megadesmolaimus	0	1	1
Metalinhomoeus	2	1	3
Paralinhomoeus	1	0	1
Terschellingia	20	7	25
Monhysteridae	**12**	**0**	**12**
Diplolaimella	1	0	1
Halomonhystera	11	0	11
Sphaerolaimidae	**15**	**19**	**26**
Parasphaerolaimus	6	2	7
Sphaerolaimus	9	17	19
Xyalidae	**53**	**46**	**69**

Daptonema	19	18	25
Metadesmolaimus	2	0	2
Paramonhystera	1	0	1
Paramonohystera	3	4	4
Retrotheristus	0	1	1
Sphaerotheristus	9	10	12
Steineria	2	2	2
Theristus	14	9	18
Trichotheristus	0	1	1
Valvaelaimus	1	1	1
Xyala	2	0	2
Overall total	**315**	**216**	**417**

Appendix 1: a list of all families and genera in the three best represented orders, giving for each the number of 18S sequences, the number of COI sequences and the total number of specimens.

10.2. Excluded specimens

Specimen	**Gene**	**Argument**
81P_Viet_Sphaerotheristus_sp5	18S	Fragment too short
29C19A_Paul_Unidentified_sp	18S	Unidentified
37C22F_Paul_Unidentified_sp	18S/ COI	Unidentified
3C23A_Paul_Anoplostoma_sp2	18S	Uncertainty in identification
3C19F_Paul_Bathyeurystomina_sp1	18S	Wrong identification
NN024_Cuba_Daptonema_sp	18S/ COI	No voucher available
33C22F_Paul_Daptonema_sp1	18S	Wrong identification
85X9C15_Pan_Daptonema_sp3	18S/ COI	No voucher available
NN026_Cuba_Gomphionema_sp	18S	No voucher available
98F_Tun_Linhomoeus_sp	18S	No voucher available
NN013_Cuba_Longicyatholaimus_sp	18S	No voucher available
41X19L14_Pan_Oncholaimus_sp	18S	No vouchers available for comparison
6C20F_Paul_Onyx_sp1	18S	Uncertainty in identification
176F_Tun_Viscosia_sp	18S	No voucher available
114H6K12_Viet_Dorylaimopsis_tumida	18S	Contamination
115H6K12_Viet_Dorylaimopsis_tumida	18S	Contamination
116H6K12_Viet_Dorylaimopsis_tumida	18S	Contamination
120H6K12_Viet_Halichoanolaimus_dolichurus	18S	Contamination
125H6K12_Viet_Parodontophora_quadristicha	18S	Contamination
141H6K12_Viet_Paracomesoma_dubium	18S	Contamination
143H6K12_Viet_Cheironchus_vorax	18S	Contamination
144H6K12_Viet_Cheironchus_vorax	18S	Contamination
140F_Tun_Unidentified_sp	COI	Unidentified
50H6K12_Viet_Desmodora_sp	COI	No voucher available

Appendix 2: a list of all specimens excluded from analysis, with the sequence that was excluded (gene) and the reason for exclusion (argument).

10.3. Species having both gene sequences available

Location	**Number of species**	**Species having both sequences**
Cuba	18	*Catanema exile, Longicyatholaimus* sp, *Cheironchus vorax, Cheironchus* sp, *Calyptronema* sp, *Oncholaimellus* sp, *Longicyatholaimus egregius,*

		Zalonema sp, *Viscosia viscosia, Paradesmodora immersa, Dorylaimopsis punctatus, Laxus parvum, Stilbonema brevicolle, Robbea porosum, Cephalanticoma* sp, *Steineria sterreri, Leptonemella brevipharynx, Daptonema* sp
Panarea	16	*Eurystomina ornata, Daptonema* sp3, *Lauratonema* sp, *Synonchus* sp, *Oxystomina* sp, *Mesacanthion* sp, *Acanthopharynx micans, Leptepsilonema* sp, *Paracyatholaimus oistospiculoides, Microlaimus compridus, Dracognomus tinae, Dichromadora hyalocheile, Endeolophos* sp, *Chromadorita* sp2, *Oncholaimus campylocercoides, Marylynnia* sp1
Paulina	18	*Onyx* sp1, *Theristus* sp1, *Stephanolaimus elegans, Adoncholaimus fuscus, Odontophora setosa, Microlaimus* sp, *Bathylaimus australis, Anoplostoma viviparum, Sphaerolaimus hirsutus, Daptonema* sp1, *Enoplolaimus attenuatus, Praeacanthonchus punctatus, Metachromadora remanei, Neochromadora* sp1, *Theristus acer, Daptonema normandicum, Paracanthonchus caecus, Hypodontolaimus inaequalis*
Tunesia	5	*Marylynnia puncticaudata, Theristus pertenuis, Oncholaimus campylocercoides, Sabatieria pulchra, Oncholaimellus mediterraneus,*
Vietnam	55	*Calomicrolaimus* sp, *Oxystomina affinis, Trissonchulus* sp1, *Halalaimus* sp2, *Sphaerotheristus* sp4, *Gomphionema parvam, Terschellingia elegans, Sphaerotheristus* sp1, *Ptycholaimellus* sp2, *Viscosia* sp3, *Steineria vietnamica, Anoplostoma sunderbanae, Daptonema* sp4, *Litinium* sp, *Desmoscolex koloensis, Theristus* sp, *Dorylaimopsis tumida, Paramonohystera megacephala, Parodontophora* sp.nov, *Onyx cangiensis, Sphaerolaimus maeoticus, Parodontophora obscura, Ironus* sp, *Asymmelaimus vietnamicus, Daptonema* sp, *Metachromadora* sp1, *Sphaerotheristus* sp, *Axonolaimus* sp, *Sphaerotheristus* sp5, *Pseudolella* sp, *Metachromadora orientalis, Spilophorella aberrans, Haliplectus floridanus, Daptonema* sp3, *Campylaimus gerlachi, Terschellingia longicaudata, Theristus flevensis, Viscosia* sp1, *Sphaerotheristus* sp3, *Bathylaimus ignavus, Halalaimus* sp3, *Ptycholaimellus brevisetosus, Ptycholaimellus* sp1, *Longicyatholaimus tchesunovi, Parodontophora obesa, Tripyloides* sp1, *Desmodora* sp, *Halalaimus gracilis, Trissonchulus* sp2, *Dichromadora simplex, Parasphaerolaimus* sp1, *Metachromadora* sp2, *Haliplectus dorsalis, Daptonema* sp2, *Comesa vitia*

Appendix 3 : the species for which we had both an 18S and a COI sequence available, listed per location. There were no COI sequences available for Papua New Guinea. In total, 112 species had both sequences.

10.4. Specimen couples with an interspecific distance value of zero for 18S

Specimen A	Specimen B
#26X12L14_Enoplolaimus_sp	#39X19L14_Enoplus_sp
#38X19L14_Mesacanthion_spB	#37X19L14_Synonchus_sp
#16X6L14_Oncholaimus_campylocercoidesC	#36X18L14_Synonchus_sp
#13C20F_Dichromadora_sp1	#78X3C15_Dichromadora_hyalocheile
#2C9M_Theristus_acer	#2A7E_Theristus_sp1
#9C20F_Dichromadora_sp1	#78X3C15_Dichromadora_hyalocheile
#ZP414_Litoditis_pm1	#GP2409_Litoditis_pm4
#12C20F_Neochromadora_sp1	#78X3C15_Dichromadora_hyalocheile
#GG2336_Halomonhystera_gd4	#OG2264_Halomonhystera_gd1
#GG2336_Halomonhystera_gd4	#OG2273_Halomonhystera_gd1
#GP3340_Litoditis_pm4	#ZP414_Litoditis_pm1
#4C9M_Theristus_acer	#2A7E_Theristus_sp1
#GP3346_Litoditis_pm4	#ZP414_Litoditis_pm1
#21C18A_Microlaimus_cyatholaimoides	#2C16A_Microlaimus_honestus
#ZP3122_Litoditis_pm1	#GP2409_Litoditis_pm4

#ZP3122_Litoditis_pm1	#GP3340_Litoditis_pm4
#ZP3122_Litoditis_pm1	#GP3346_Litoditis_pm4
#GG2320_Halomonhystera_gd4	#OG2264_Halomonhystera_gd1
#GG2320_Halomonhystera_gd4	#OG2273_Halomonhystera_gd1
#OG2261_Halomonhystera_gd1	#GG2336_Halomonhystera_gd4
#OG2261_Halomonhystera_gd1	#GG2320_Halomonhystera_gd4
#GP3369_Litoditis_pm4	#ZP414_Litoditis_pm1
#GP3369_Litoditis_pm4	#ZP3122_Litoditis_pm1
#1A30E_Enoploides_sp	#1C16A_Enoploides_longispiculosis
#1C19F_Dichromadora_sp1	#78X3C15_Dichromadora_hyalocheile
#40C20A_Tubolaimoides_sp	#89X20C15_Mesacanthion_sp
#40C20A_Tubolaimoides_sp	#26X12L14_Enoplolaimus_sp
#40C20A_Tubolaimoides_sp	#78X3C15_Dichromadora_hyalocheile
#ZP3139_Litoditis_pm1	#GP2409_Litoditis_pm4
#ZP3139_Litoditis_pm1	#GP3340_Litoditis_pm4
#ZP3139_Litoditis_pm1	#GP3346_Litoditis_pm4
#ZP3139_Litoditis_pm1	#GP3369_Litoditis_pm4
#ZP3151_Litoditis_pm1	#GP2409_Litoditis_pm4
#ZP3151_Litoditis_pm1	#GP3340_Litoditis_pm4
#ZP3151_Litoditis_pm1	#GP3346_Litoditis_pm4
#ZP3151_Litoditis_pm1	#GP3369_Litoditis_pm4
#GG2339_Halomonhystera_gd4	#OG2264_Halomonhystera_gd1
#GG2339_Halomonhystera_gd4	#OG2273_Halomonhystera_gd1
#GG2339_Halomonhystera_gd4	#OG2261_Halomonhystera_gd1
#GG2332_Halomonhystera_gd4	#OG2264_Halomonhystera_gd1
#GG2332_Halomonhystera_gd4	#OG2273_Halomonhystera_gd1
#GG2332_Halomonhystera_gd4	#OG2261_Halomonhystera_gd1
#87F_Paracomesoma_dubium	#40C20A_Tubolaimoides_sp
#231F_Sabatieria_pulchra	#40C20A_Tubolaimoides_sp
#156F_Sabatieria_punctata	#40C20A_Tubolaimoides_sp
#156F_Sabatieria_punctata	#231F_Sabatieria_pulchra
#70F_Paracomesoma_affdubium	#40C20A_Tubolaimoides_sp
#70F_Paracomesoma_affdubium	#87F_Paracomesoma_dubium
#97F_Marylynnia_puncticaudata	#40C20A_Tubolaimoides_sp
#191F_Viscosia_franzii	#40C20A_Tubolaimoides_sp
#15F_Paracomesoma_affdubium	#40C20A_Tubolaimoides_sp
#15F_Paracomesoma_affdubium	#87F_Paracomesoma_dubium
#202F_Marylynnia_puncticaudata	#40C20A_Tubolaimoides_sp
#190F_Sabatieria_pulchra	#40C20A_Tubolaimoides_sp
#190F_Sabatieria_pulchra	#156F_Sabatieria_punctata
#7F_Marylynnia_puncticaudata	#40C20A_Tubolaimoides_sp
#196F_Oncholaimellus_mediterraneus	#40C20A_Tubolaimoides_sp
#300F_Sabatieria_pulchra	#40C20A_Tubolaimoides_sp
#300F_Sabatieria_pulchra	#156F_Sabatieria_punctata
#201F_Marylynnia_puncticaudata	#40C20A_Tubolaimoides_sp
#178P_Sphaerotheristus_sp4	#68P_Sphaerotheristus_sp5
#85P_Terschellingia_longicaudata	#115P_Terschellingia_elegans
#72P_Sphaerotheristus_sp3	#68P_Sphaerotheristus_sp5
#72P_Sphaerotheristus_sp3	#178P_Sphaerotheristus_sp4
#176P_Parasphaerolaimus_sp1	#85H6K12_Parasphaerolaimus_sp

#98P_Sphaerotheristus_sp3	#68P_Sphaerotheristus_sp5
#98P_Sphaerotheristus_sp3	#178P_Sphaerotheristus_sp4
#60P_Terschellingia_longicaudata	#115P_Terschellingia_elegans
#99P_Sphaerotheristus_sp3	#68P_Sphaerotheristus_sp5
#99P_Sphaerotheristus_sp3	#178P_Sphaerotheristus_sp4
#197P_Terschellingia_eleganss	#115P_Terschellingia_elegans
#197P_Terschellingia_eleganss	#85P_Terschellingia_longicaudata
#197P_Terschellingia_eleganss	#60P_Terschellingia_longicaudata
#156P_Tripyloides_sp2	#149P_Tripyloides_sp
#70P_Parodontophora_fluviatilis	#74P_Parodontophora_obscura
#127P_Metachromadora_sp2	#138P_Metachromadora_sp1
#127P_Metachromadora_sp2	#137P_Metachromadora_sp1
#189P_Daptonema_sp5	#192P_Daptonema_sp3
#189P_Daptonema_sp5	#191P_Daptonema_sp3
#189P_Daptonema_sp5	#86P_Daptonema_sp3
#189P_Daptonema_sp5	#190P_Daptonema_sp3
#30P_Tripyloides_sp1	#149P_Tripyloides_sp
#30P_Tripyloides_sp1	#156P_Tripyloides_sp2
#46P_Parodontophora_sp2	#143P_Parodontophora_obesa
#120P_Parasphaerolaimus_sp1	#85H6K12_Parasphaerolaimus_sp
#65P_Ptycholaimellus_sp2	#100P_Ptycholaimellus_sp3
#175P_Metachromadora_sp2	#138P_Metachromadora_sp1
#175P_Metachromadora_sp2	#137P_Metachromadora_sp1
#169P_Parasphaerolaimus_sp1	#85H6K12_Parasphaerolaimus_sp
#123P_Parasphaerolaimus_sp1	#85H6K12_Parasphaerolaimus_sp
#90P_Metachromadora_sp1	#127P_Metachromadora_sp2
#90P_Metachromadora_sp1	#175P_Metachromadora_sp2
#44P_Ptycholaimellus_brevisetosus	#170P_Ptycholaimellus_sp1
#171P_Ptycholaimellus_sp1	#44P_Ptycholaimellus_brevisetosus
#161P_Terschellingia_longicaudataB	#159P_Terschellingia_sp.nov
#161P_Terschellingia_longicaudataB	#154P_Terschellingia_sp.nov
#161P_Terschellingia_longicaudataB	#24P_Terschellingia_sp.nov
#161P_Terschellingia_longicaudataB	#26P_Terschellingia_sp.nov
#161P_Terschellingia_longicaudataB	#186P_Terschellingia_sp.nov
#161P_Terschellingia_longicaudataB	#158P_Terschellingia_sp.nov
#2A_Paracanthonchus_caecus	#40C20A_Tubolaimoides_sp
#8A_Paracanthonchus_caecus	#40C20A_Tubolaimoides_sp
#20A_Paracanthonchus_sp2	#40C20A_Tubolaimoides_sp
#22A_Meyersia_sp2	#15A_Meyersia_sp1
#25A_Paracanthonchus_sp2	#40C20A_Tubolaimoides_sp
#26A_Paracanthonchus_sp2	#40C20A_Tubolaimoides_sp
#31A_Meyersia_sp2	#15A_Meyersia_sp1
#34A_Axonolaimus_paraponticus	#40C20A_Tubolaimoides_sp

Appendix 4: pairwise specimen comparisons (each time the one on the left compared to the one on the right) that yielded an interspecific distance value of zero for 18S.

10.5. Artificial community composition

Family	**Genus**	**Species**	**N**	**18S**	**COI**
Anoplostomatidae	*Anoplostoma*	sp2	1	3	0
Anoplostomatidae	*Anoplostoma*	*viviparum*	2	1	1
Axonolaimidae	*Ascolaimus*	*elongatus*	1	1	0

Axonolaimidae	*Axonolaimus*	*paraspinosus*	1	2	0
Axonolaimidae	*Odontophora*	*setosa*	2	1	8
Camacolaimidae	*Camacolaimus*	*trituberculatus*	1	1	0
Chromadoridae	*Actinonema*	*celtica*	1	1	0
Chromadoridae	*Dichromadora*	sp1	2	3	0
Chromadoridae	*Hypodontolaimus*	*inaequalis*	2	3	2
Chromadoridae	*Neochromadora*	sp1	1	1	1
Chromadoridae	*Prochromadorella*	*anatartica*	1	1	0
Chromadoridae	*Ptycholaimellus*	*carinatus*	1	0	1
Comesomatidae	*Sabatiera*	*armata*	1	1	0
Cyatholaimidae	*Metacyatholaimus*	sp	1	1	1
Cyatholaimidae	*Paracanthonchus*	*caecus*	1	5	1
Cyatholaimidae	*Praecanthonchus*	*punctatus*	6	6	7
Desmodoridae	*Desmodora*	sp2	1	4	0
Desmodoridae	*Metachromadora*	*remanei*	2	1	1
Desmodoridae	*Onyx*	sp	1	1	0
Encheliidae	*Bathyeurystomina*	sp1	1	2	0
Encheliidae	*Calyptronema*	*maxweberi*	2	0	3
Ethmolaimidae	*Ethmolaimus*	*pratensis*	1	1	0
Leptolaimidae	*Stephanolaimus*	*elegans*	1	1	1
Linhomoeidae	*Linhomoeus*	sp	1	2	0
Linhomoeidae	*Metalinhomoeus*	sp1	1	0	1
Linhomoeidae	*Paralinhomoeus*	*tenuicaudatus*	1	1	0
Microlaimidae	*Microlaimus*	*honestus*	2	3	4
Microlaimidae	*Microlaimus*	*punctulatus*	1	0	0
Monhysteridae	*Diplolaimella*	sp	1	1	0
Monhysteridae	*Diplolaimelloides*	*oschei*	1	0	0
Monhysteridae	*Halomonhystera*	*disjuncta* (Gd1)	4	3	0
Monhysteridae	*Halomonhystera*	*disjuncta* (Gd2)	1	1	0
Monhysteridae	*Halomonhystera*	*disjuncta* (Gd3)	2	3	0
Monhysteridae	*Halomonhystera*	*disjuncta* (Gd4)	4	4	0
Monoposthiidae	*Monoposthia*	*mirabilis*	3	4	2
Oncholaimidae	*Adoncholaimus*	*fuscus*	3	3	3
Oncholaimidae	*Oncholaimus**		6	15	17
Oxystominidae	*Halalaimus*	sp1	1	0	1
Rhabditidae	*Litoditis*	*marina* (Pm1)	4	4	0
Rhabditidae	*Litoditis*	*marina* (Pm2)	2	1	0
Rhabditidae	*Litoditis*	*marina* (Pm3)	1	1	0
Rhabditidae	*Litoditis*	*marina* (Pm4)	4	4	0
Sphaerolaimidae	*Sphaerolaimus*	*hirsutus*	6	2	9
Thoracostomopsidae	*Enoploides*	*longispiculosus*	2	1	0
Thoracostomopsidae	*Enoplolaimus*	*attenuatus*	1	1	1
Tripyloididae	*Bathylaimus*	*assimilis*	4	0	0
Tripyloididae	*Bathylaimus*	*australis*	2	2	3
Tubolaimoididae	*Tubolaimoides*	sp	1	1	0
Xyalidae	*Daptonema*	*normandicum*	1	1	2
Xyalidae	*Daptonema*	*setosum*	2	0	4
Xyalidae	*Metadesmolaimus*	*aduncus*	1	1	0
Xyalidae	*Theristus*	*ensifer*	2	0	0
Xyalidae	*Theristus*	sp1	1	1	1
Xyalidae	*Theristus*	sp2	1	0	0
Xyalidae	*Theristus*	*acer*	2	2	2
Xyalidae	*Xyala*	sp1	1	2	0
Total number of species			50	46	24

Apppendix 5: a list of all species present in the artificial community. N= relative abundance (by using equivalent volumes of DNA extract), representing the theoretical number of individuals present in the mock community; 18S = the number of 18S sequences present in our database for this species; COI = the number of 18S sequences present in our database for this species. Total number of species shows for "N" the number of species present in the mock community and for "18S" and "COI" the number of species identifiable by at least one sequence in our database. * Oncholaimus was not identified to species level; identification as any member of the genus was considered correct.

10.6. QIIME commands

The commands used for the QIIME bioinformatics pipeline per script, per gene:

Using own reference database (example given for 18S)

Note: "-m blast" is added to the assign_taxonomy.py script to use the BLAST algorithm instead of the default UCLUST method

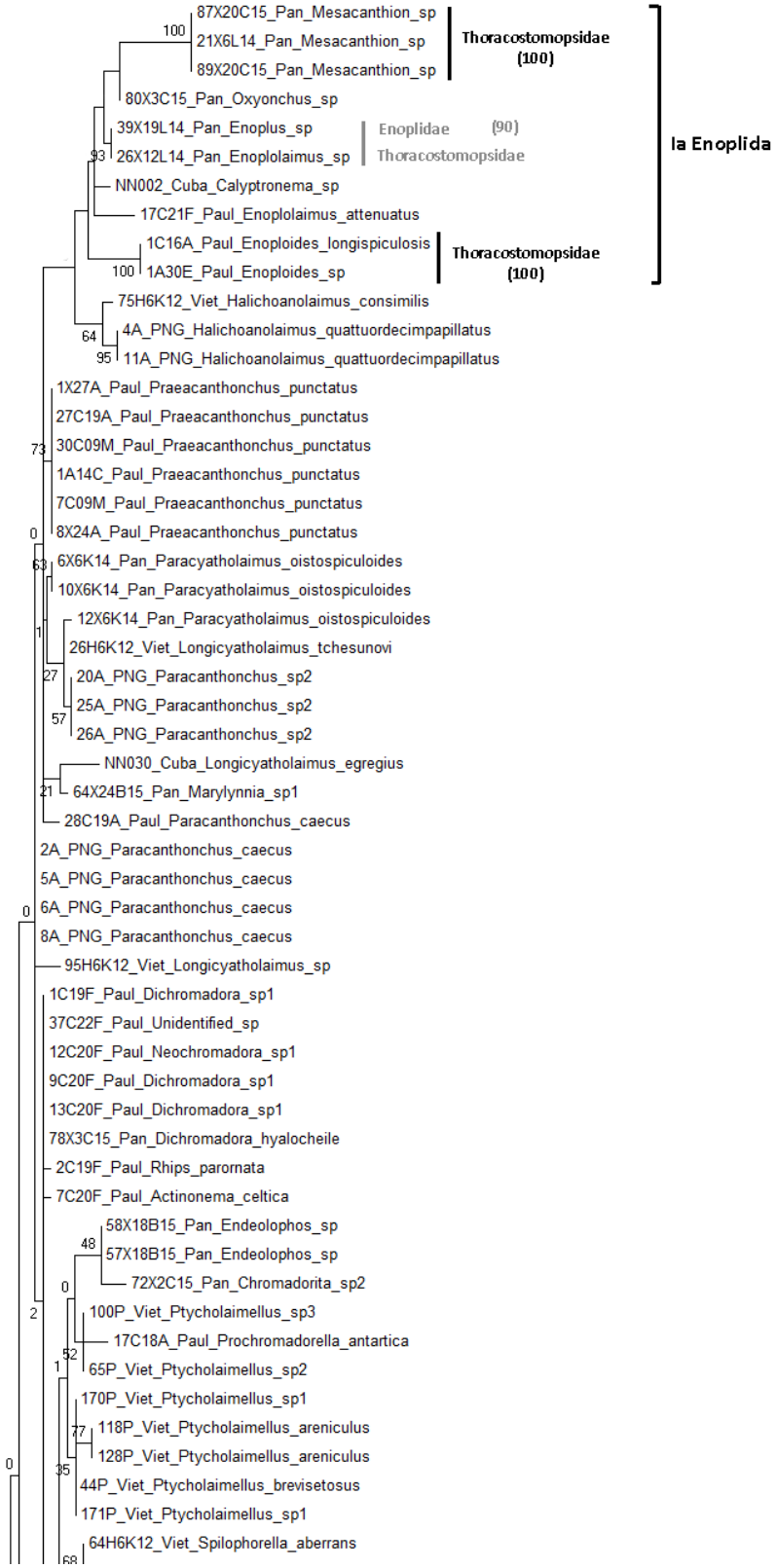
```
pick_open_reference_otus.py -i '/home/qiime/Desktop/Shared_Folder/denoise18S-chim.fasta' -r
'/home/qiime/Desktop/Shared_Folder/All18SFastasMerged.fasta' -o
'/home/qiime/Desktop/Shared_Folder/otu_picking18S' -s 0.01 -p '/home/qiime/Desktop/parameters_18S'

assign_taxonomy.py -i '/home/qiime/Desktop/Shared_Folder/otu_picking18S/rep_set.fna' -r
'/home/qiime/Desktop/Shared_Folder/All18SFastasMerged.fasta' -t
'/home/qiime/Desktop/Shared_Folder/IDtoTax.txt' -o
'/home/qiime/Desktop/Shared_Folder/uclust_assigned_taxonomy'

biom add-metadata -i '/home/qiime/Desktop/Shared_Folder/otu_picking18S/otu_table_mc2.biom' -o
'/home/qiime/Desktop/Shared_Folder/otu_picking18S/otu_table_mc2_tax.biom' --observation-metadata-fp
'/home/qiime/Desktop/Shared_Folder/uclust_assigned_taxonomy/rep_set_tax_assignments.txt' --observation-
header OTUID,taxonomy --sc-separated taxonomy

biom convert -i '/home/qiime/Desktop/Shared_Folder/otu_picking18S/otu_table_mc2_tax.biom' -o
'/home/qiime/Desktop/Shared_Folder/otu_picking18S/otu_table_mc2_tax.txt' --table-type="OTU table" --to-tsv
--header-key taxonomy

summarize_taxa_through_plots.py -i
'/home/qiime/Desktop/Shared_Folder/otu_picking18S/otu_table_mc2_tax.biom' -o
'/home/qiime/Desktop/Shared_Folder/taxa_summary'
```

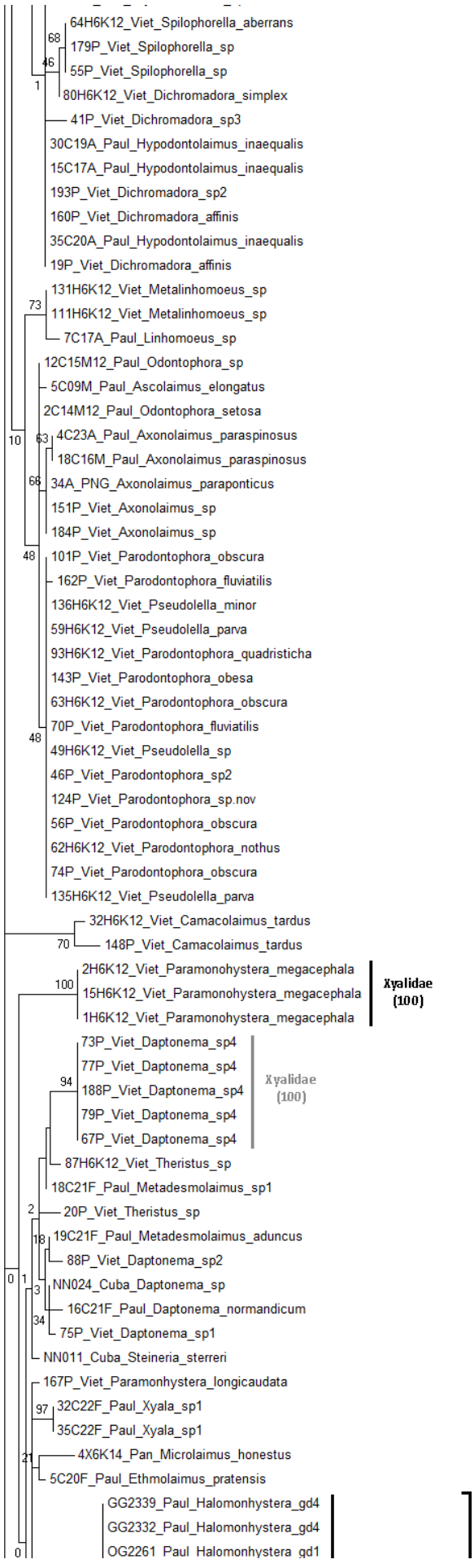
Using the Silva database as reference (18S)

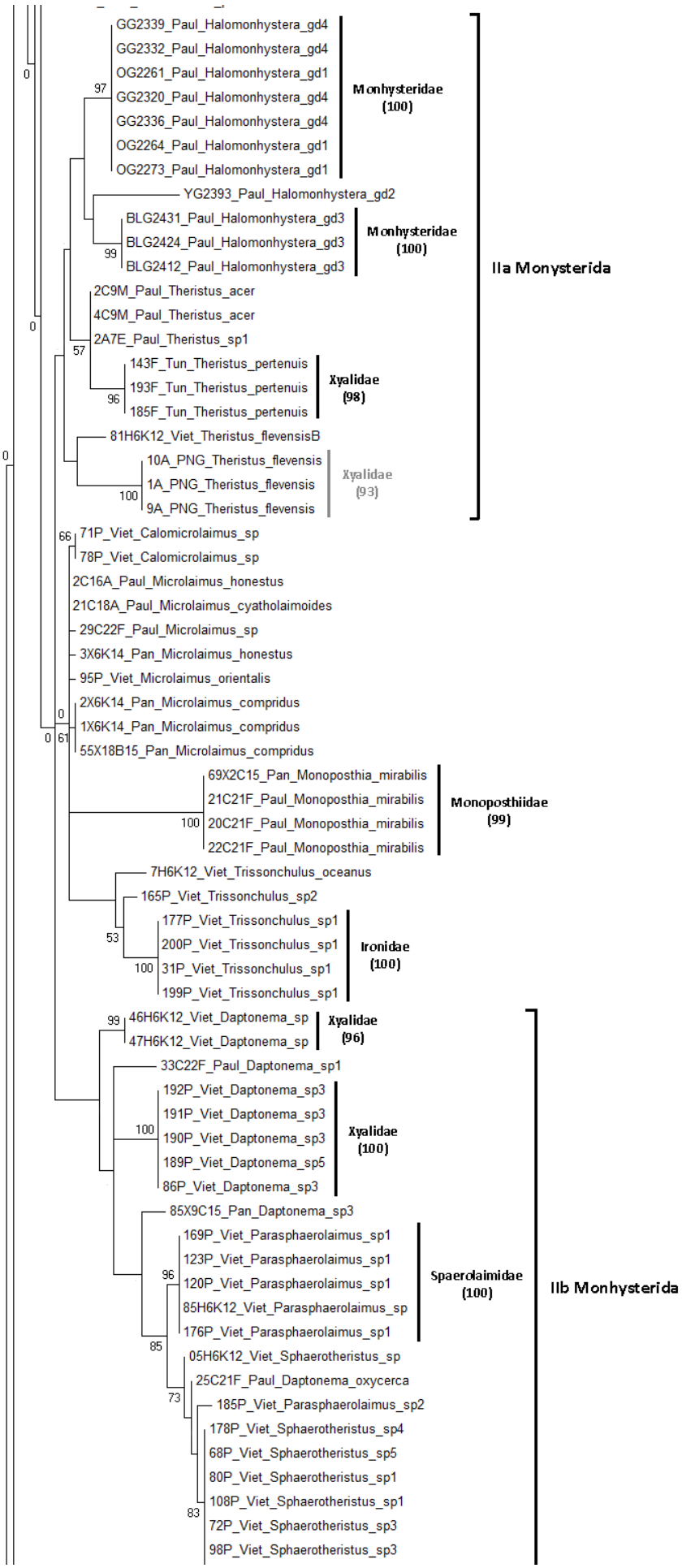
Note: "summarize_taxa:level 3,4,5,6,7,8" is added to the parameter file for the summarize_taxa_through_plots.py script, to ensure the script to summarize down to species level.

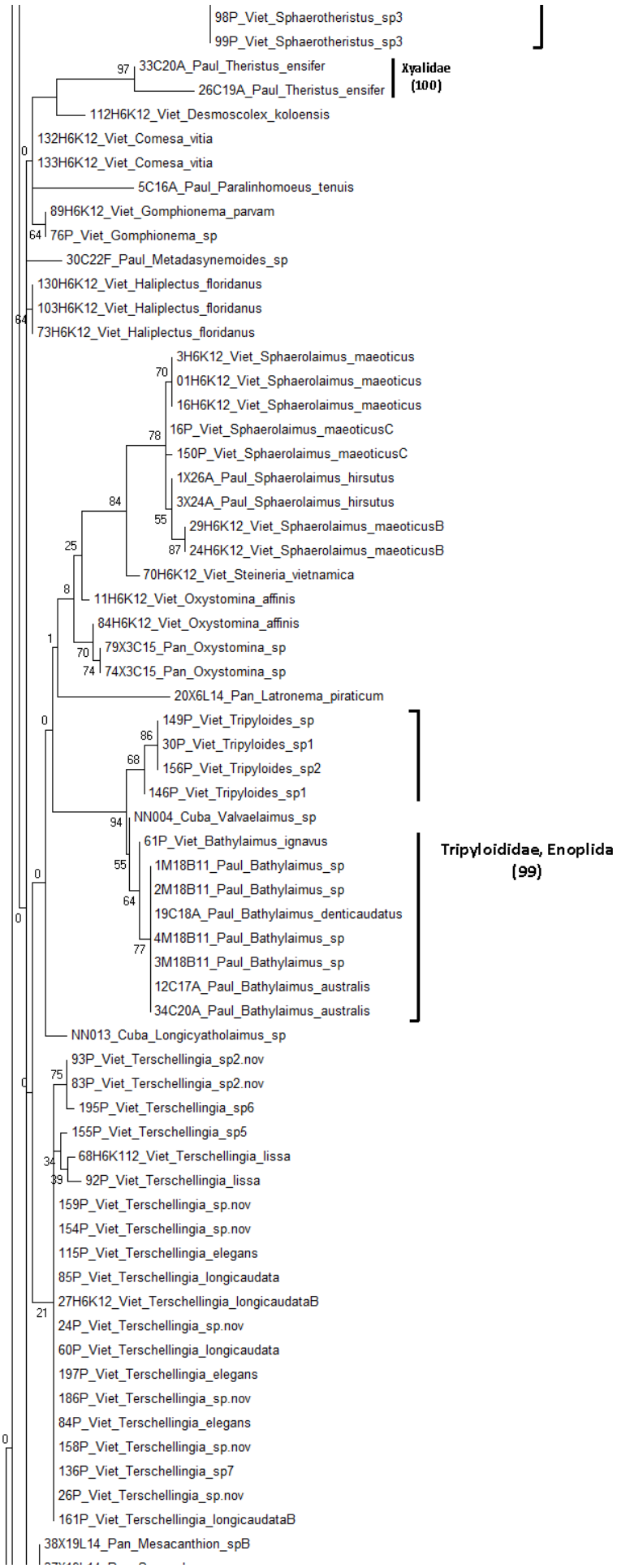
```
pick_otus.py -i /home/qiime/Desktop/Shared_Folder/denoise18S-chim.fasta -r
/home/qiime/Desktop/Shared_Folder/99_Silva_111_rep_set.fasta -o
/home/qiime/Desktop/Shared_Folder/otu_picking18S/step1_otus -m uclust_ref --minsize 3 --similarity 0.99 --
enable_rev_strand_match --suppress_new_clusters

assign_taxonomy.py -i '/home/qiime/Desktop/Shared_Folder/otu_picking18S/rep_set.fna' -r
'/home/qiime/Desktop/Shared_Folder/99_Silva_111_rep_set_euk.fasta' -t
'/home/qiime/Desktop/Shared_Folder/99_Silva_111_taxa_map_euks.txt' -o
'/home/qiime/Desktop/Shared_Folder/uclust_assigned_taxonomy'

biom add-metadata -i '/home/qiime/Desktop/Shared_Folder/otu_picking18S/otu_table_mc2.biom' -o
'/home/qiime/Desktop/Shared_Folder/otu_picking18S/otu_table_mc2_tax.biom' --observation-metadata-fp
'/home/qiime/Desktop/Shared_Folder/uclust_assigned_taxonomy/rep_set_tax_assignments.txt' --observation-
header OTUID,taxonomy --sc-separated taxonomy

biom convert -i '/home/qiime/Desktop/Shared_Folder/otu_picking18S/otu_table_mc2_tax.biom' -o
'/home/qiime/Desktop/Shared_Folder/otu_picking18S/otu_table_mc2_tax.txt' --table-type="OTU table" --to-tsv
--header-key taxonomy

summarize_taxa_through_plots.py -i
'/home/qiime/Desktop/Shared_Folder/otu_picking18S/otu_table_mc2_tax.biom' -o
'/home/qiime/Desktop/Shared_Folder/taxa_summary' –p
'/home/qiime/Desktop/Shared_Folder/Parameter_file_Silva'
```

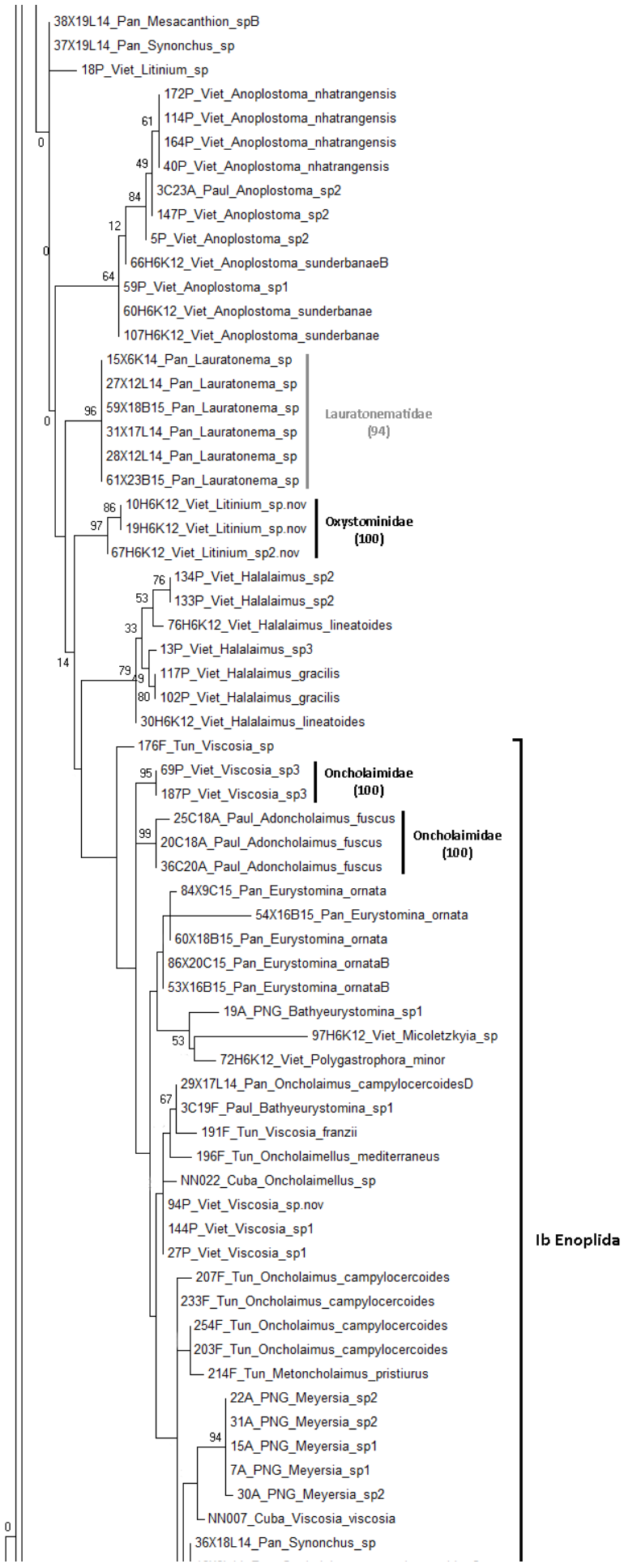
10.7. Complete maximum likelihood trees



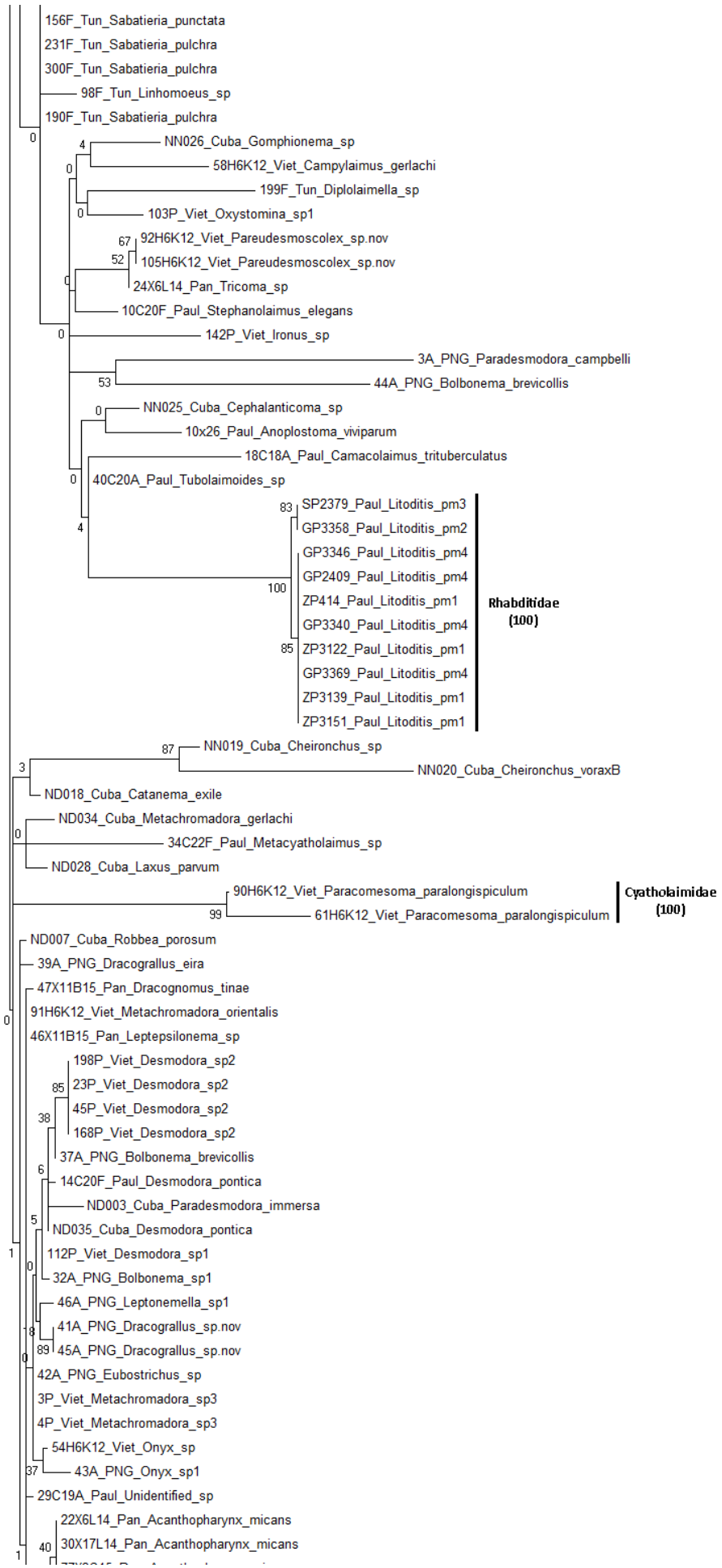

64H6K12_Viet_Spilophorella_aberrans
179P_Viet_Spilophorella_sp
55P_Viet_Spilophorella_sp
80H6K12_Viet_Dichromadora_simplex
41P_Viet_Dichromadora_sp3
30C19A_Paul_Hypodontolaimus_inaequalis
15C17A_Paul_Hypodontolaimus_inaequalis
193P_Viet_Dichromadora_sp2
160P_Viet_Dichromadora_affinis
35C20A_Paul_Hypodontolaimus_inaequalis
19P_Viet_Dichromadora_affinis
131H6K12_Viet_Metalinhomoeus_sp
111H6K12_Viet_Metalinhomoeus_sp
7C17A_Paul_Linhomoeus_sp
12C15M12_Paul_Odontophora_sp
5C09M_Paul_Ascolaimus_elongatus
2C14M12_Paul_Odontophora_setosa
4C23A_Paul_Axonolaimus_paraspinosus
18C16M_Paul_Axonolaimus_paraspinosus
34A_PNG_Axonolaimus_paraponticus
151P_Viet_Axonolaimus_sp
184P_Viet_Axonolaimus_sp
101P_Viet_Parodontophora_obscura
162P_Viet_Parodontophora_fluviatilis
136H6K12_Viet_Pseudolella_minor
59H6K12_Viet_Pseudolella_parva
93H6K12_Viet_Parodontophora_quadristicha
143P_Viet_Parodontophora_obesa
63H6K12_Viet_Parodontophora_obscura
70P_Viet_Parodontophora_fluviatilis
49H6K12_Viet_Pseudolella_sp
46P_Viet_Parodontophora_sp2
124P_Viet_Parodontophora_sp.nov
56P_Viet_Parodontophora_obscura
62H6K12_Viet_Parodontophora_nothus
74P_Viet_Parodontophora_obscura
135H6K12_Viet_Pseudolella_parva
32H6K12_Viet_Camacolaimus_tardus
148P_Viet_Camacolaimus_tardus
2H6K12_Viet_Paramonohystera_megacephala
15H6K12_Viet_Paramonohystera_megacephala
1H6K12_Viet_Paramonohystera_megacephala
Xyalidae (100)
73P_Viet_Daptonema_sp4
77P_Viet_Daptonema_sp4
188P_Viet_Daptonema_sp4
79P_Viet_Daptonema_sp4
67P_Viet_Daptonema_sp4
Xyalidae (100)
87H6K12_Viet_Theristus_sp
18C21F_Paul_Metadesmolaimus_sp1
20P_Viet_Theristus_sp
19C21F_Paul_Metadesmolaimus_aduncus
88P_Viet_Daptonema_sp2
NN024_Cuba_Daptonema_sp
16C21F_Paul_Daptonema_normandicum
75P_Viet_Daptonema_sp1
NN011_Cuba_Steineria_sterreri
167P_Viet_Paramonhystera_longicaudata
32C22F_Paul_Xyala_sp1
35C22F_Paul_Xyala_sp1
4X6K14_Pan_Microlaimus_honestus
5C20F_Paul_Ethmolaimus_pratensis
GG2339_Paul_Halomonhystera_gd4
GG2332_Paul_Halomonhystera_gd4
OG2261_Paul_Halomonhystera_gd1

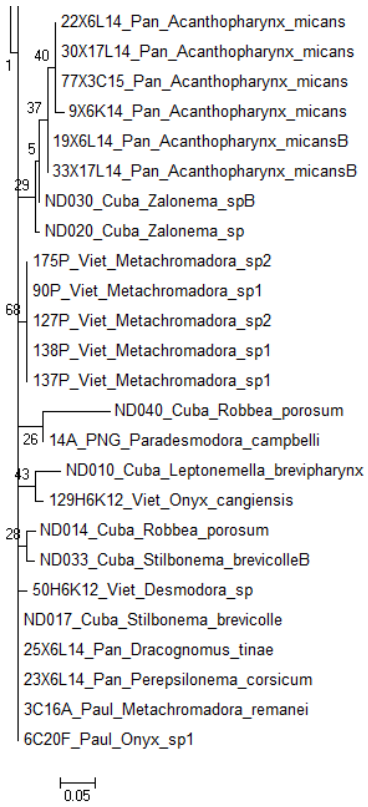

GG2339_Paul_Halomonhystera_gd4
GG2332_Paul_Halomonhystera_gd4
OG2261_Paul_Halomonhystera_gd1
GG2320_Paul_Halomonhystera_gd4
GG2336_Paul_Halomonhystera_gd4
OG2264_Paul_Halomonhystera_gd1
OG2273_Paul_Halomonhystera_gd1
Monhysteridae (100)
YG2393_Paul_Halomonhystera_gd2
BLG2431_Paul_Halomonhystera_gd3
BLG2424_Paul_Halomonhystera_gd3
BLG2412_Paul_Halomonhystera_gd3
Monhysteridae (100)
IIa Monysterida
2C9M_Paul_Theristus_acer
4C9M_Paul_Theristus_acer
2A7E_Paul_Theristus_sp1
143F_Tun_Theristus_pertenuis
193F_Tun_Theristus_pertenuis
185F_Tun_Theristus_pertenuis
Xyalidae (98)
81H6K12_Viet_Theristus_flevensisB
10A_PNG_Theristus_flevensis
1A_PNG_Theristus_flevensis
9A_PNG_Theristus_flevensis
Xyalidae (93)
71P_Viet_Calomicrolaimus_sp
78P_Viet_Calomicrolaimus_sp
2C16A_Paul_Microlaimus_honestus
21C18A_Paul_Microlaimus_cyatholaimoides
29C22F_Paul_Microlaimus_sp
3X6K14_Pan_Microlaimus_honestus
95P_Viet_Microlaimus_orientalis
2X6K14_Pan_Microlaimus_compridus
1X6K14_Pan_Microlaimus_compridus
55X18B15_Pan_Microlaimus_compridus
69X2C15_Pan_Monoposthia_mirabilis
21C21F_Paul_Monoposthia_mirabilis
20C21F_Paul_Monoposthia_mirabilis
22C21F_Paul_Monoposthia_mirabilis
Monoposthiidae (99)
7H6K12_Viet_Trissonchulus_oceanus
165P_Viet_Trissonchulus_sp2
177P_Viet_Trissonchulus_sp1
200P_Viet_Trissonchulus_sp1
31P_Viet_Trissonchulus_sp1
199P_Viet_Trissonchulus_sp1
Ironidae (100)
46H6K12_Viet_Daptonema_sp
47H6K12_Viet_Daptonema_sp
Xyalidae (96)
33C22F_Paul_Daptonema_sp1
192P_Viet_Daptonema_sp3
191P_Viet_Daptonema_sp3
190P_Viet_Daptonema_sp3
189P_Viet_Daptonema_sp5
86P_Viet_Daptonema_sp3
Xyalidae (100)
85X9C15_Pan_Daptonema_sp3
169P_Viet_Parasphaerolaimus_sp1
123P_Viet_Parasphaerolaimus_sp1
120P_Viet_Parasphaerolaimus_sp1
85H6K12_Viet_Parasphaerolaimus_sp
176P_Viet_Parasphaerolaimus_sp1
Spaerolaimidae (100)
05H6K12_Viet_Sphaerotheristus_sp
25C21F_Paul_Daptonema_oxycerca
185P_Viet_Parasphaerolaimus_sp2
178P_Viet_Sphaerotheristus_sp4
68P_Viet_Sphaerotheristus_sp5
80P_Viet_Sphaerotheristus_sp1
108P_Viet_Sphaerotheristus_sp1
72P_Viet_Sphaerotheristus_sp3
98P_Viet_Sphaerotheristus_sp3
IIb Monhysterida


98P_Viet_Sphaerotheristus_sp3
99P_Viet_Sphaerotheristus_sp3
97
33C20A_Paul_Theristus_ensifer
26C19A_Paul_Theristus_ensifer
Xyalidae
(100)
112H6K12_Viet_Desmoscolex_koloensis
132H6K12_Viet_Comesa_vitia
133H6K12_Viet_Comesa_vitia
5C16A_Paul_Paralinhomoeus_tenuis
89H6K12_Viet_Gomphionema_parvam
64
76P_Viet_Gomphionema_sp
30C22F_Paul_Metadasynemoides_sp
130H6K12_Viet_Haliplectus_floridanus
103H6K12_Viet_Haliplectus_floridanus
64
73H6K12_Viet_Haliplectus_floridanus
3H6K12_Viet_Sphaerolaimus_maeoticus
70
01H6K12_Viet_Sphaerolaimus_maeoticus
16H6K12_Viet_Sphaerolaimus_maeoticus
78
16P_Viet_Sphaerolaimus_maeoticusC
150P_Viet_Sphaerolaimus_maeoticusC
1X26A_Paul_Sphaerolaimus_hirsutus
84
3X24A_Paul_Sphaerolaimus_hirsutus
55
29H6K12_Viet_Sphaerolaimus_maeoticusB
87
24H6K12_Viet_Sphaerolaimus_maeoticusB
25
70H6K12_Viet_Steineria_vietnamica
8
11H6K12_Viet_Oxystomina_affinis
84H6K12_Viet_Oxystomina_affinis
1
70
79X3C15_Pan_Oxystomina_sp
74
74X3C15_Pan_Oxystomina_sp
20X6L14_Pan_Latronema_piraticum
0
149P_Viet_Tripyloides_sp
86
30P_Viet_Tripyloides_sp1
68
156P_Viet_Tripyloides_sp2
146P_Viet_Tripyloides_sp1
94
NN004_Cuba_Valvaelaimus_sp
61P_Viet_Bathylaimus_ignavus
55
1M18B11_Paul_Bathylaimus_sp
2M18B11_Paul_Bathylaimus_sp
64
19C18A_Paul_Bathylaimus_denticaudatus
4M18B11_Paul_Bathylaimus_sp
77
3M18B11_Paul_Bathylaimus_sp
12C17A_Paul_Bathylaimus_australis
34C20A_Paul_Bathylaimus_australis
Tripyloididae, Enoplida
(99)
NN013_Cuba_Longicyatholaimus_sp
93P_Viet_Terschellingia_sp2.nov
75
83P_Viet_Terschellingia_sp2.nov
195P_Viet_Terschellingia_sp6
155P_Viet_Terschellingia_sp5
34
68H6K112_Viet_Terschellingia_lissa
39
92P_Viet_Terschellingia_lissa
159P_Viet_Terschellingia_sp.nov
154P_Viet_Terschellingia_sp.nov
115P_Viet_Terschellingia_elegans
85P_Viet_Terschellingia_longicaudata
27H6K12_Viet_Terschellingia_longicaudataB
21
24P_Viet_Terschellingia_sp.nov
60P_Viet_Terschellingia_longicaudata
197P_Viet_Terschellingia_elegans
186P_Viet_Terschellingia_sp.nov
84P_Viet_Terschellingia_elegans
158P_Viet_Terschellingia_sp.nov
136P_Viet_Terschellingia_sp7
26P_Viet_Terschellingia_sp.nov
161P_Viet_Terschellingia_longicaudataB
38X19L14_Pan_Mesacanthion_spB

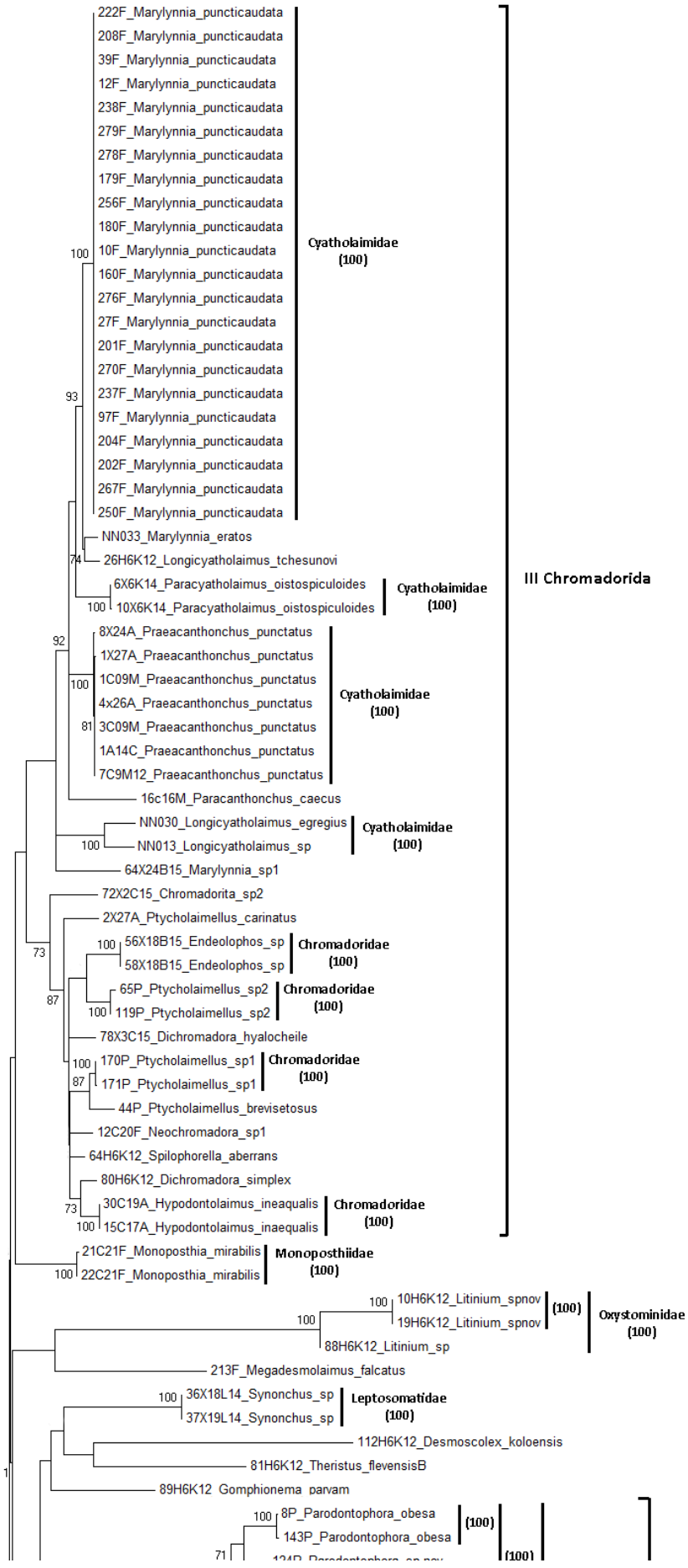

38X19L14_Pan_Mesacanthion_spB
37X19L14_Pan_Synonchus_sp
18P_Viet_Litinium_sp
172P_Viet_Anoplostoma_nhatrangensis
114P_Viet_Anoplostoma_nhatrangensis
164P_Viet_Anoplostoma_nhatrangensis
40P_Viet_Anoplostoma_nhatrangensis
3C23A_Paul_Anoplostoma_sp2
147P_Viet_Anoplostoma_sp2
5P_Viet_Anoplostoma_sp2
66H6K12_Viet_Anoplostoma_sunderbanaeB
59P_Viet_Anoplostoma_sp1
60H6K12_Viet_Anoplostoma_sunderbanae
107H6K12_Viet_Anoplostoma_sunderbanae
15X6K14_Pan_Lauratonema_sp
27X12L14_Pan_Lauratonema_sp
59X18B15_Pan_Lauratonema_sp
31X17L14_Pan_Lauratonema_sp
28X12L14_Pan_Lauratonema_sp
61X23B15_Pan_Lauratonema_sp
Lauratonematidae (94)
10H6K12_Viet_Litinium_sp.nov
19H6K12_Viet_Litinium_sp.nov
67H6K12_Viet_Litinium_sp2.nov
Oxystominidae (100)
134P_Viet_Halalaimus_sp2
133P_Viet_Halalaimus_sp2
76H6K12_Viet_Halalaimus_lineatoides
13P_Viet_Halalaimus_sp3
117P_Viet_Halalaimus_gracilis
102P_Viet_Halalaimus_gracilis
30H6K12_Viet_Halalaimus_lineatoides
176F_Tun_Viscosia_sp
69P_Viet_Viscosia_sp3
187P_Viet_Viscosia_sp3
Oncholaimidae (100)
25C18A_Paul_Adoncholaimus_fuscus
20C18A_Paul_Adoncholaimus_fuscus
36C20A_Paul_Adoncholaimus_fuscus
Oncholaimidae (100)
84X9C15_Pan_Eurystomina_ornata
54X16B15_Pan_Eurystomina_ornata
60X18B15_Pan_Eurystomina_ornata
86X20C15_Pan_Eurystomina_ornataB
53X16B15_Pan_Eurystomina_ornataB
19A_PNG_Bathyeurystomina_sp1
97H6K12_Viet_Micoletzkyia_sp
72H6K12_Viet_Polygastrophora_minor
29X17L14_Pan_Oncholaimus_campylocercoidesD
3C19F_Paul_Bathyeurystomina_sp1
191F_Tun_Viscosia_franzii
196F_Tun_Oncholaimellus_mediterraneus
NN022_Cuba_Oncholaimellus_sp
94P_Viet_Viscosia_sp.nov
144P_Viet_Viscosia_sp1
27P_Viet_Viscosia_sp1
207F_Tun_Oncholaimus_campylocercoides
233F_Tun_Oncholaimus_campylocercoides
254F_Tun_Oncholaimus_campylocercoides
203F_Tun_Oncholaimus_campylocercoides
214F_Tun_Metoncholaimus_pristiurus
22A_PNG_Meyersia_sp2
31A_PNG_Meyersia_sp2
15A_PNG_Meyersia_sp1
7A_PNG_Meyersia_sp1
30A_PNG_Meyersia_sp2
NN007_Cuba_Viscosia_viscosia
36X18L14_Pan_Synonchus_sp
Ib Enoplida

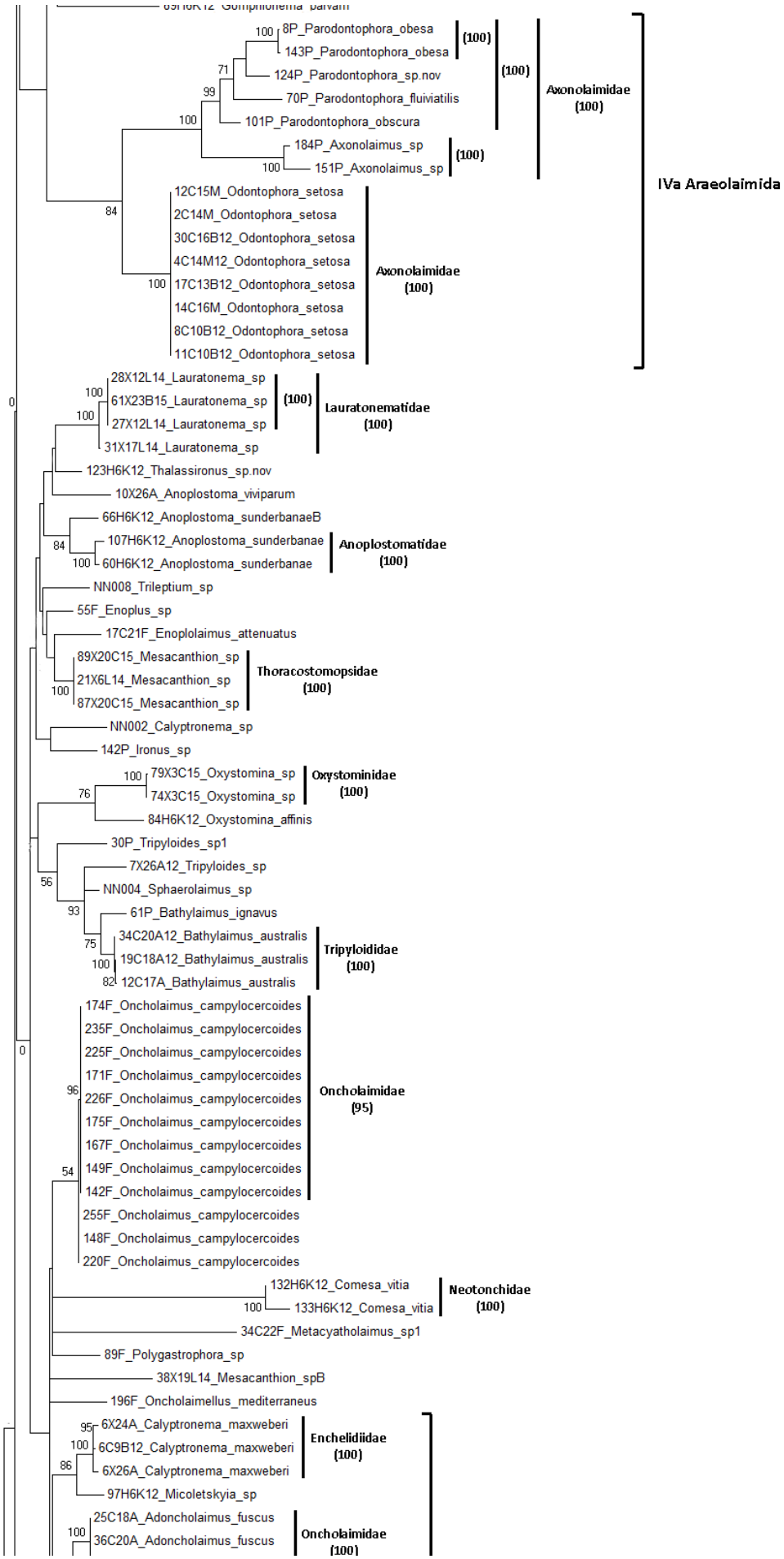

36X18L14_Pan_Synonchus_sp
16X6L14_Pan_Oncholaimus_campylocercoidesC
41X19L14_Pan_Oncholaimus_sp
17X6L14_Pan_Oncholaimus_campylocercoidesC
69A22C12_Paul_Oncholaimus_sp
65A22C12_Paul_Oncholaimus_sp
39A22_Paul_Oncholaimus_sp
85A22C12_Paul_Oncholaimus_sp
64A22C12_Paul_Oncholaimus_sp
86A22C12_Paul_Oncholaimus_sp
68A22C12_Paul_Oncholaimus_sp
NN003_Cuba_Dorylaimopsis_punctatus
102H6K12_Viet_Haliplectus_dorsalis
25H6K12_Viet_Sabatieria_sp
28C21F_Paul_Sabatieria_propissina
15C20F_Paul_Sabatieria_armata
86H6K12_Viet_Sabatieria_praedatrix
104H6K12_Viet_Sabatieria_praedatrix
31H6K12_Viet_Sabatieria_doancanhi
23H6K12_Viet_Dorylaimopsis_tumida
37H6K12_Viet_Dorylaimopsis_tumida
48H6K12_Viet_Dorylaimopsis_tumida
22H6K12_Viet_Dorylaimopsis_tumida
1P_Viet_Sabatieria_sp1
55H6K12_Viet_Sabatieria_foetida
17P_Viet_Sabatieria_sp1
2P_Viet_Sabatieria_sp1
141P_Viet_Sabatieria_sp1
13X6K14_Pan_Leptolaimus_sp
134H6K12_Viet_Leptolaimus_praeclarus
09H6K12_Viet_Asymmelaimus_vietnamicus
45H6K12_Viet_Paracomesoma_sp
44H6K12_Viet_Paracomesoma_sp
33H6K12_Viet_Paracomesoma_sp
41H6K12_Viet_Paracomesoma_sp
43H6K12_Viet_Paracomesoma_sp
42H6K12_Viet_Paracomesoma_sp
78H6K12_Viet_Paracomesoma_lisum
37P_Viet_Hopperia_sp
69H6K12_Viet_Laimella_sp1.nov
35H6K12_Viet_Hopperia_dolichura
56H6K12_Viet_Hopperia_dolichura
28H6K12_Viet_Dorylaimopsis_tumida
82P_Viet_Sabatieria_sp2
38H6K12_Viet_Dorylaimopsis_tumida
NN028_Cuba_Sabatieria_migrans
196P_Viet_Sabatieria_sp2
87P_Viet_Sabatieria_sp2
20H6K12_Viet_Asymmelaimus_vietnamicus
138H6K12_Viet_Asymmelaimus_vietnamicus
14H6K124R2_Viet_Asymmelaimus_vietnamicus
39H6K12_Viet_Asymmelaimus_vietnamicus
08H6K12_Viet_Asymmelaimus_vietnamicus
12H6K12_Viet_Asymmelaimus_vietnamicus
36H6K12_Viet_Asymmelaimus_vietnamicus
97F_Tun_Marylynnia_puncticaudata
202F_Tun_Marylynnia_puncticaudata
201F_Tun_Marylynnia_puncticaudata
7F_Tun_Marylynnia_puncticaudata
70F_Tun_Paracomesoma_affdubium
87F_Tun_Paracomesoma_dubium
15F_Tun_Paracomesoma_affdubium
156F_Tun_Sabatieria_punctata
231F_Tun_Sabatieria_pulchra

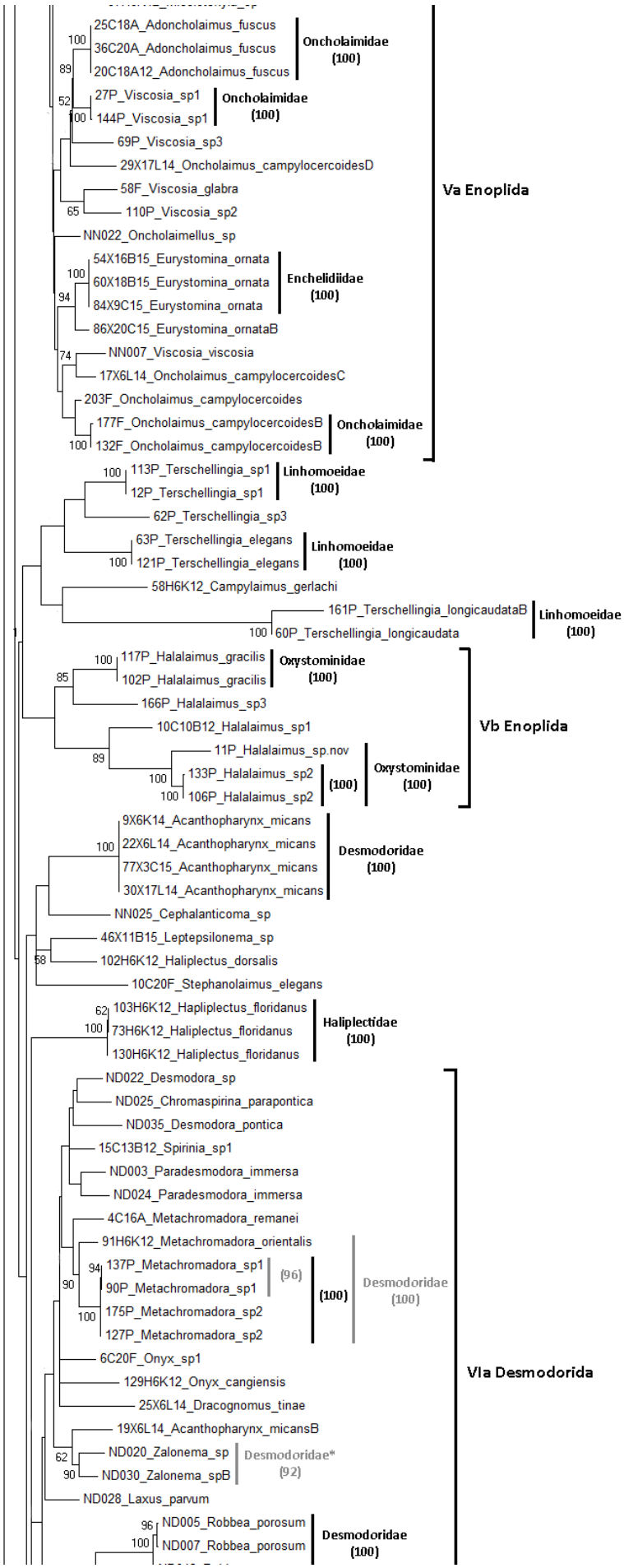

156F_Tun_Sabatieria_punctata
231F_Tun_Sabatieria_pulchra
300F_Tun_Sabatieria_pulchra
98F_Tun_Linhomoeus_sp
190F_Tun_Sabatieria_pulchra
NN026_Cuba_Gomphionema_sp
58H6K12_Viet_Campylaimus_gerlachi
199F_Tun_Diplolaimella_sp
103P_Viet_Oxystomina_sp1
92H6K12_Viet_Pareudesmoscolex_sp.nov
105H6K12_Viet_Pareudesmoscolex_sp.nov
24X6L14_Pan_Tricoma_sp
10C20F_Paul_Stephanolaimus_elegans
142P_Viet_Ironus_sp
3A_PNG_Paradesmodora_campbelli
44A_PNG_Bolbonema_brevicollis
NN025_Cuba_Cephalanticoma_sp
10x26_Paul_Anoplostoma_viviparum
18C18A_Paul_Camacolaimus_trituberculatus
40C20A_Paul_Tubolaimoides_sp
SP2379_Paul_Litoditis_pm3
GP3358_Paul_Litoditis_pm2
GP3346_Paul_Litoditis_pm4
GP2409_Paul_Litoditis_pm4
ZP414_Paul_Litoditis_pm1
GP3340_Paul_Litoditis_pm4
ZP3122_Paul_Litoditis_pm1
GP3369_Paul_Litoditis_pm4
ZP3139_Paul_Litoditis_pm1
ZP3151_Paul_Litoditis_pm1
Rhabditidae
(100)
NN019_Cuba_Cheironchus_sp
NN020_Cuba_Cheironchus_voraxB
ND018_Cuba_Catanema_exile
ND034_Cuba_Metachromadora_gerlachi
34C22F_Paul_Metacyatholaimus_sp
ND028_Cuba_Laxus_parvum
90H6K12_Viet_Paracomesoma_paralongispiculum
61H6K12_Viet_Paracomesoma_paralongispiculum
Cyatholaimidae
(100)
ND007_Cuba_Robbea_porosum
39A_PNG_Dracograllus_eira
47X11B15_Pan_Dracognomus_tinae
91H6K12_Viet_Metachromadora_orientalis
46X11B15_Pan_Leptepsilonema_sp
198P_Viet_Desmodora_sp2
23P_Viet_Desmodora_sp2
45P_Viet_Desmodora_sp2
168P_Viet_Desmodora_sp2
37A_PNG_Bolbonema_brevicollis
14C20F_Paul_Desmodora_pontica
ND003_Cuba_Paradesmodora_immersa
ND035_Cuba_Desmodora_pontica
112P_Viet_Desmodora_sp1
32A_PNG_Bolbonema_sp1
46A_PNG_Leptonemella_sp1
41A_PNG_Dracograllus_sp.nov
45A_PNG_Dracograllus_sp.nov
42A_PNG_Eubostrichus_sp
3P_Viet_Metachromadora_sp3
4P_Viet_Metachromadora_sp3
54H6K12_Viet_Onyx_sp
43A_PNG_Onyx_sp1
29C19A_Paul_Unidentified_sp
22X6L14_Pan_Acanthopharynx_micans
30X17L14_Pan_Acanthopharynx_micans

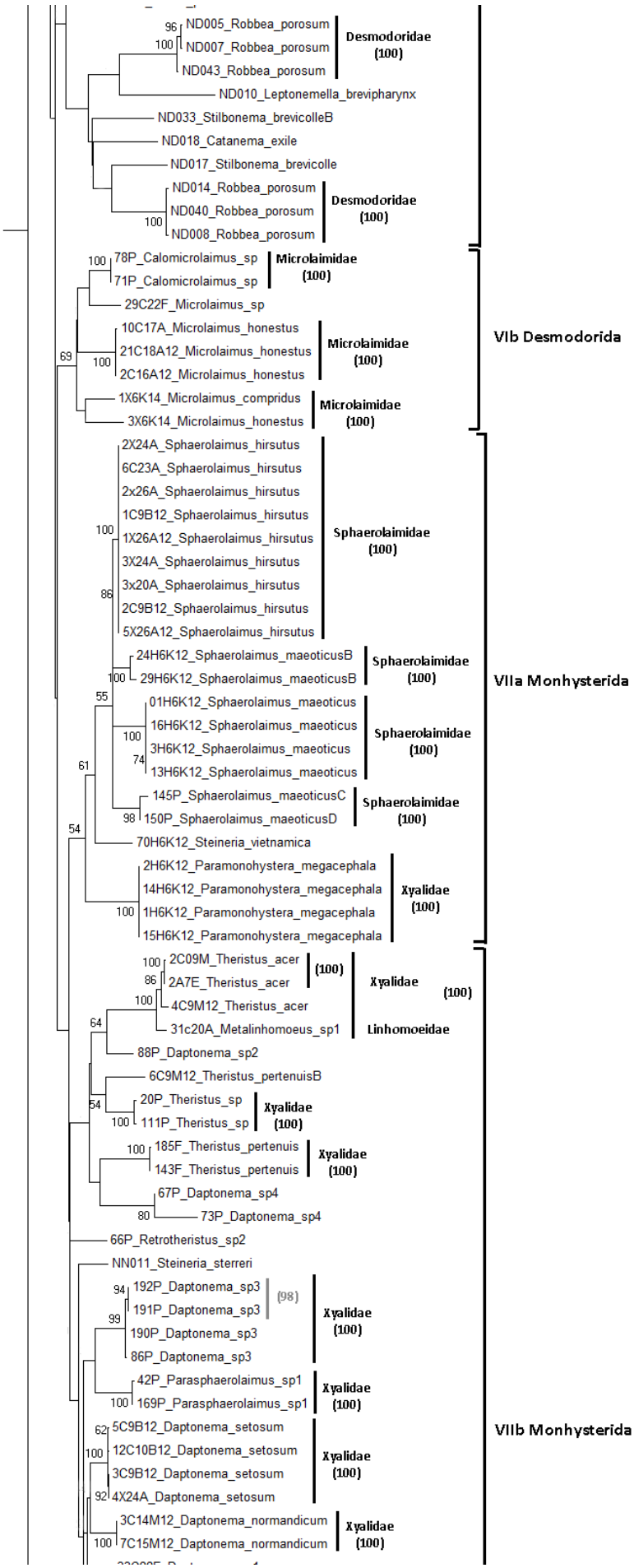


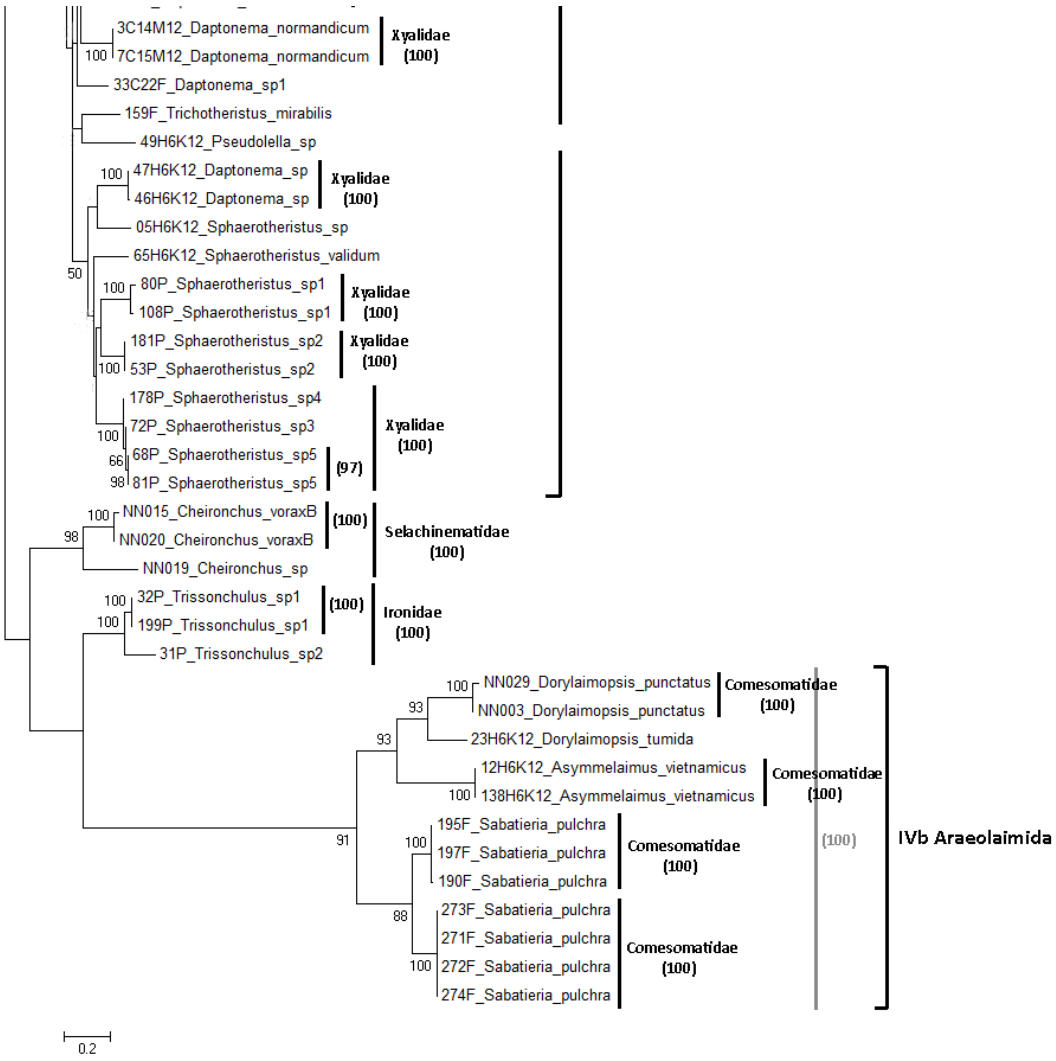
Appendix 7a: maximum likelihood tree for 18S with bootstrap values (BV) of at least 50, and BV of the neighbor joining tree between brackets. Clusters that are supported by a BV of at least 95 for both trees (ML and NJ) are indicated in by a black line. Clusters that are supported by a BV of at least 90 for both trees are indicated by a grey line. Clusters with specimens belonging to the same order are indicated by a square bracket.


222F_Marylynnia_puncticaudata
208F_Marylynnia_puncticaudata
39F_Marylynnia_puncticaudata
12F_Marylynnia_puncticaudata
238F_Marylynnia_puncticaudata
279F_Marylynnia_puncticaudata
278F_Marylynnia_puncticaudata
179F_Marylynnia_puncticaudata
256F_Marylynnia_puncticaudata
180F_Marylynnia_puncticaudata
10F_Marylynnia_puncticaudata
160F_Marylynnia_puncticaudata
276F_Marylynnia_puncticaudata
27F_Marylynnia_puncticaudata
201F_Marylynnia_puncticaudata
270F_Marylynnia_puncticaudata
237F_Marylynnia_puncticaudata
97F_Marylynnia_puncticaudata
204F_Marylynnia_puncticaudata
202F_Marylynnia_puncticaudata
267F_Marylynnia_puncticaudata
250F_Marylynnia_puncticaudata
Cyatholaimidae (100)
NN033_Marylynnia_eratos
26H6K12_Longicyatholaimus_tchesunovi
6X6K14_Paracyatholaimus_oistospiculoides
10X6K14_Paracyatholaimus_oistospiculoides
Cyatholaimidae (100)
III Chromadorida
8X24A_Praeacanthonchus_punctatus
1X27A_Praeacanthonchus_punctatus
1C09M_Praeacanthonchus_punctatus
4x26A_Praeacanthonchus_punctatus
3C09M_Praeacanthonchus_punctatus
1A14C_Praeacanthonchus_punctatus
7C9M12_Praeacanthonchus_punctatus
Cyatholaimidae (100)
16c16M_Paracanthonchus_caecus
NN030_Longicyatholaimus_egregius
NN013_Longicyatholaimus_sp
Cyatholaimidae (100)
64X24B15_Marylynnia_sp1
72X2C15_Chromadorita_sp2
2X27A_Ptycholaimellus_carinatus
56X18B15_Endeolophos_sp
58X18B15_Endeolophos_sp
Chromadoridae (100)
65P_Ptycholaimellus_sp2
119P_Ptycholaimellus_sp2
Chromadoridae (100)
78X3C15_Dichromadora_hyalocheile
170P_Ptycholaimellus_sp1
171P_Ptycholaimellus_sp1
Chromadoridae (100)
44P_Ptycholaimellus_brevisetosus
12C20F_Neochromadora_sp1
64H6K12_Spilophorella_aberrans
80H6K12_Dichromadora_simplex
30C19A_Hypodontolaimus_ineaqualis
15C17A_Hypodontolaimus_inaequalis
Chromadoridae (100)
21C21F_Monoposthia_mirabilis
22C21F_Monoposthia_mirabilis
Monoposthiidae (100)
10H6K12_Litinium_spnov
19H6K12_Litinium_spnov
(100)
88H6K12_Litinium_sp
Oxystominidae (100)
213F_Megadesmolaimus_falcatus
36X18L14_Synonchus_sp
37X19L14_Synonchus_sp
Leptosomatidae (100)
112H6K12_Desmoscolex_koloensis
81H6K12_TheristusflevensisB
89H6K12_Gomphionema_parvam
8P_Parodontophora_obesa
143P_Parodontophora_obesa
(100)
(100)


8P_Parodontophora_obesa
143P_Parodontophora_obesa
124P_Parodontophora_sp.nov
70P_Parodontophora_fluviatilis
101P_Parodontophora_obscura
184P_Axonolaimus_sp
151P_Axonolaimus_sp
(100)
(100)
Axonolaimidae
(100)
IVa Araeolaimida
12C15M_Odontophora_setosa
2C14M_Odontophora_setosa
30C16B12_Odontophora_setosa
4C14M12_Odontophora_setosa
17C13B12_Odontophora_setosa
14C16M_Odontophora_setosa
8C10B12_Odontophora_setosa
11C10B12_Odontophora_setosa
Axonolaimidae
(100)
28X12L14_Lauratonema_sp
61X23B15_Lauratonema_sp
27X12L14_Lauratonema_sp
31X17L14_Lauratonema_sp
(100)
Lauratonematidae
(100)
123H6K12_Thalassironus_sp.nov
10X26A_Anoplostoma_viviparum
66H6K12_Anoplostoma_sunderbanaeB
107H6K12_Anoplostoma_sunderbanae
60H6K12_Anoplostoma_sunderbanae
Anoplostomatidae
(100)
NN008_Trileptium_sp
55F_Enoplus_sp
17C21F_Enoplolaimus_attenuatus
89X20C15_Mesacanthion_sp
21X6L14_Mesacanthion_sp
87X20C15_Mesacanthion_sp
Thoracostomopsidae
(100)
NN002_Calyptronema_sp
142P_Ironus_sp
79X3C15_Oxystomina_sp
74X3C15_Oxystomina_sp
84H6K12_Oxystomina_affinis
Oxystominidae
(100)
30P_Tripyloides_sp1
7X26A12_Tripyloides_sp
NN004_Sphaerolaimus_sp
61P_Bathylaimus_ignavus
34C20A12_Bathylaimus_australis
19C18A12_Bathylaimus_australis
12C17A_Bathylaimus_australis
Tripyloididae
(100)
174F_Oncholaimus_campylocercoides
235F_Oncholaimus_campylocercoides
225F_Oncholaimus_campylocercoides
171F_Oncholaimus_campylocercoides
226F_Oncholaimus_campylocercoides
175F_Oncholaimus_campylocercoides
167F_Oncholaimus_campylocercoides
149F_Oncholaimus_campylocercoides
142F_Oncholaimus_campylocercoides
255F_Oncholaimus_campylocercoides
148F_Oncholaimus_campylocercoides
220F_Oncholaimus_campylocercoides
Oncholaimidae
(95)
132H6K12_Comesa_vitia
133H6K12_Comesa_vitia
Neotonchidae
(100)
34C22F_Metacyatholaimus_sp1
89F_Polygastrophora_sp
38X19L14_Mesacanthion_spB
196F_Oncholaimellus_mediterraneus
6X24A_Calyptronema_maxweberi
6C9B12_Calyptronema_maxweberi
6X26A_Calyptronema_maxweberi
Enchelidiidae
(100)
97H6K12_Micoletskyia_sp
25C18A_Adoncholaimus_fuscus
36C20A_Adoncholaimus_fuscus
Oncholaimidae
(100)


25C18A_Adoncholaimus_fuscus
36C20A_Adoncholaimus_fuscus
20C18A12_Adoncholaimus_fuscus
Oncholaimidae (100)
27P_Viscosia_sp1
144P_Viscosia_sp1
Oncholaimidae (100)
69P_Viscosia_sp3
29X17L14_Oncholaimus_campylocercoidesD
58F_Viscosia_glabra
110P_Viscosia_sp2
NN022_Oncholaimellus_sp
54X16B15_Eurystomina_ornata
60X18B15_Eurystomina_ornata
84X9C15_Eurystomina_ornata
Enchelidiidae (100)
86X20C15_Eurystomina_ornataB
NN007_Viscosia_viscosia
17X6L14_Oncholaimus_campylocercoidesC
203F_Oncholaimus_campylocercoides
177F_Oncholaimus_campylocercoidesB
132F_Oncholaimus_campylocercoidesB
Oncholaimidae (100)
Va Enoplida
113P_Terschellingia_sp1
12P_Terschellingia_sp1
Linhomoeidae (100)
62P_Terschellingia_sp3
63P_Terschellingia_elegans
121P_Terschellingia_elegans
Linhomoeidae (100)
58H6K12_Campylaimus_gerlachi
161P_Terschellingia_longicaudataB
60P_Terschellingia_longicaudata
Linhomoeidae (100)
117P_Halalaimus_gracilis
102P_Halalaimus_gracilis
Oxystominidae (100)
166P_Halalaimus_sp3
10C10B12_Halalaimus_sp1
11P_Halalaimus_sp.nov
133P_Halalaimus_sp2
106P_Halalaimus_sp2
(100)
Oxystominidae (100)
Vb Enoplida
9X6K14_Acanthopharynx_micans
22X6L14_Acanthopharynx_micans
77X3C15_Acanthopharynx_micans
30X17L14_Acanthopharynx_micans
Desmodoridae (100)
NN025_Cephalanticoma_sp
46X11B15_Leptepsilonema_sp
102H6K12_Haliplectus_dorsalis
10C20F_Stephanolaimus_elegans
103H6K12_Hapliplectus_floridanus
73H6K12_Haliplectus_floridanus
130H6K12_Haliplectus_floridanus
Haliplectidae (100)
ND022_Desmodora_sp
ND025_Chromaspirina_parapontica
ND035_Desmodora_pontica
15C13B12_Spirinia_sp1
ND003_Paradesmodora_immersa
ND024_Paradesmodora_immersa
4C16A_Metachromadora_remanei
91H6K12_Metachromadora_orientalis
137P_Metachromadora_sp1
90P_Metachromadora_sp1
(96)
175P_Metachromadora_sp2
127P_Metachromadora_sp2
(100)
Desmodoridae (100)
6C20F_Onyx_sp1
129H6K12_Onyx_cangiensis
25X6L14_Dracognomus_tinae
19X6L14_Acanthopharynx_micansB
ND020_Zalonema_sp
ND030_Zalonema_spB
Desmodoridae* (92)
ND028_Laxus_parvum
ND005_Robbea_porosum
ND007_Robbea_porosum
Desmodoridae (100)
VIa Desmodorida


ND005_Robbea_porosum
ND007_Robbea_porosum
ND043_Robbea_porosum
Desmodoridae (100)
ND010_Leptonemella_brevipharynx
ND033_Stilbonema_brevicolleB
ND018_Catanema_exile
ND017_Stilbonema_brevicolle
ND014_Robbea_porosum
ND040_Robbea_porosum
ND008_Robbea_porosum
Desmodoridae (100)
78P_Calomicrolaimus_sp
71P_Calomicrolaimus_sp
Microlaimidae (100)
29C22F_Microlaimus_sp
10C17A_Microlaimus_honestus
21C18A12_Microlaimus_honestus
2C16A12_Microlaimus_honestus
Microlaimidae (100)
VIb Desmodorida
1X6K14_Microlaimus_compridus
3X6K14_Microlaimus_honestus
Microlaimidae (100)
2X24A_Sphaerolaimus_hirsutus
6C23A_Sphaerolaimus_hirsutus
2x26A_Sphaerolaimus_hirsutus
1C9B12_Sphaerolaimus_hirsutus
1X26A12_Sphaerolaimus_hirsutus
3X24A_Sphaerolaimus_hirsutus
3x20A_Sphaerolaimus_hirsutus
2C9B12_Sphaerolaimus_hirsutus
5X26A12_Sphaerolaimus_hirsutus
Sphaerolaimidae (100)
24H6K12_Sphaerolaimus_maeoticusB
29H6K12_Sphaerolaimus_maeoticusB
Sphaerolaimidae (100)
VIIa Monhysterida
01H6K12_Sphaerolaimus_maeoticus
16H6K12_Sphaerolaimus_maeoticus
3H6K12_Sphaerolaimus_maeoticus
13H6K12_Sphaerolaimus_maeoticus
Sphaerolaimidae (100)
145P_Sphaerolaimus_maeoticusC
150P_Sphaerolaimus_maeoticusD
Sphaerolaimidae (100)
70H6K12_Steineria_vietnamica
2H6K12_Paramonohystera_megacephala
14H6K12_Paramonohystera_megacephala
1H6K12_Paramonohystera_megacephala
15H6K12_Paramonohystera_megacephala
Xyalidae (100)
2C09M_Theristus_acer
2A7E_Theristus_acer
(100)
4C9M12_Theristus_acer
31c20A_Metalinhomoeus_sp1
Xyalidae (100)
Linhomoeidae
88P_Daptonema_sp2
6C9M12_Theristus_pertenuisB
20P_Theristus_sp
111P_Theristus_sp
Xyalidae (100)
185F_Theristus_pertenuis
143F_Theristus_pertenuis
Xyalidae (100)
67P_Daptonema_sp4
73P_Daptonema_sp4
66P_Retrotheristus_sp2
NN011_Steineria_sterreri
192P_Daptonema_sp3
191P_Daptonema_sp3
(98)
190P_Daptonema_sp3
86P_Daptonema_sp3
Xyalidae (100)
42P_Parasphaerolaimus_sp1
169P_Parasphaerolaimus_sp1
Xyalidae (100)
5C9B12_Daptonema_setosum
12C10B12_Daptonema_setosum
3C9B12_Daptonema_setosum
4X24A_Daptonema_setosum
Xyalidae (100)
VIIb Monhysterida
3C14M12_Daptonema_normandicum
7C15M12_Daptonema_normandicum
Xyalidae (100)



Appendix 7b: maximum likelihood tree for COI with bootstrap values (BV) of at least 50, and BV of the neighbor joining tree between brackets. Clusters that are supported by a BV of at least 95 for both trees (ML and NJ) are indicated in by a black line. Clusters that are supported by a BV of at least 90 for both trees are indicated by a grey line. Clusters with specimens belonging to the same order are indicated by a square bracket.

10.8. Python scripts

An overview of all 5 custom Python scripts that were written specifically for this study, with an example input at the end of the script.

10.8.1. 18S fasta merger (to correct all separate 18S FASTA files and write them to one total FASTA file. "COI fasta merger" script is similar)

```
def fastamerge(Cubafile, Panfile, Paulfile, Tunfile, Vietfile, PNGfile, outfile):
    """
    #Cuba: Genus_species_code
    #Panarea: code_Genus_species[_speciesnr] (+code replace)
    #Paulina: code_Genus_species (but lots of exceptions)
    #Tunesia: code_code_Genus_species[_species] (+code replace)
    #Vietnam: Genus_species[_letter/number/"nov"]_code (!sequence on multiple lines!)
    #Papua-New-Guinea: Genus_species (code to be added from separate file)

    #Standaardlayout to Code_location_Genus_species
    #Gaps in sequence are deleted

    #Location ID resp. Cu, Pan, Paul, Tun, Viet, PNG
    """
```

```
    #1
    #Create dictionary for Cuba fasta: CubaDic['Genus_species_code'] = sequence
    CubaDic = {}
    reader = open(Cubafile, 'r')
    line = reader.readline()
    while line:
        if '>' in line:
            title = line.strip('\n')
            title = title.strip('>')
        else:
            if line == '\n':
                pass
            else:
                sequence = line.strip('\n')
                sequence = sequence.replace('-', '')                #Remove all gaps from sequence
                CubaDic[title] = sequence
        line = reader.readline()
    reader.close()

    #Set title layout to Code_Location_Genus_species + filter unidentified species (all to "sp", because
some are named 'sp.') and write to outfile
    writer = open(outfile, 'w')
    for t in CubaDic:
        genus, species, code = t.split('_')
        genus = genus.lower().capitalize()
        species = species.replace('.', '').lower()
        writer.write('>' + code + '_Cuba_' + genus + '_' + species + '\n')
        writer.write(CubaDic[t] + '\n')
    writer.close()


    #2
    #Create dictionary for Panarea fasta: PanDic['Code_Genus_species'] = sequence
    PanDic = {}
    reader = open(Panfile, 'r')
    line = reader.readline()
    while line:
        if '>' in line:
            line = line.replace(' ', '_')                #Replace all spaces in title with underscores
            title = line.strip('\n')
            title = title.strip('>')
        else:
            sequence = line.strip('\n')
            sequence = sequence.replace('-', '')                #Remove all gaps from sequence
            PanDic[title] = sequence
        line = reader.readline()
    reader.close()

    #Read the file with all PCR and DNA codes and put those in dictionary
    CodeconvDic = {}
    reader = open('PanSpecies PCR and DNA codes.txt', 'r')
    line = reader.readline()
    line = reader.readline()
    while line:
        line = line.strip('\n')
        PCR, DNA = line.split('\t')
        CodeconvDic[PCR] = DNA
        line = reader.readline()
    reader.close()

    #Set title layout to Code_Location_Genus_species + filter unidentified species and write to outfile
    writer = open(outfile, 'a')
    for t in PanDic:
        code, genus, species = t.split('_', 2)
        code = code.strip('Xo')
        code = CodeconvDic[code]                    #Replace PCR code with specimen DNA number
        genus = genus.lower().capitalize()
        species = species.replace('_', '').lower()
        writer.write('>' + code + '_Pan_' + genus + '_' + species + '\n')
        writer.write(PanDic[t] + '\n')
    writer.close()


    #3
    #Define a test to check if a sequence title contains 3 parts (code_Genus_species). If not, correct.
    def Paulstandardcheck(intitle):
```

```
        """Check if Paulina sequence title contains 3 parts (code_Genus_species). If not, correct by
adding 'sp' as species."""
        if intitle.count('_') < 2:              #If there's only one underscore, layout is Genus_code
            code, genus = t.split('_')          #Split genus and code and add 'sp' in between
            outtitle = code + '_' + genus + '_sp'
        else:
            outtitle = intitle                  #If there are multiple underscores, leave input sequence
title unchanged
        return outtitle

    #Create dictionary for Paulina fasta: PaulDic['code_Genus_species'] = sequence
    PaulDic = {}
    reader = open(Paulfile, 'r')
    line = reader.readline()
    while line:
        if '>' in line:
            title = line.strip('\n')
            title = title.strip('>')
        else:
            sequence = line.strip('\n')
            sequence = sequence.replace('-', '')                #Remove all gaps from sequence
            PaulDic[title] = sequence
        line = reader.readline()
    reader.close()

    #Set title layout to Code_Location_Genus_species + filter unidentified species and write to outfile
    writer = open(outfile, 'a')
    for t in PaulDic:
        nt = Paulstandardcheck(t)        #Check layout sequence title with previous function
        code, genus, species = nt.split('_', 2)

        if '_' in species:
            part1, *part2 = species.split('_')
            if genus == 'reverse' or genus == 'forward':    #Put the shifted genus and species name back
                genus, species = part1, part2
            else:
                if len(part2) == 1:         #Deletes haplotype letter (Species_H)
                    species = part1
        if len(species) == 1:             #Repalces haplotype letter by 'sp'
            species = 'sp'

        if 'pm' in genus.lower() and len(genus) == 3:                                        #rename
Pellioditis genus to Litoditis, keep cryptic species
            species, genus = genus.lower().capitalize(), 'Litoditis'

        if genus == 'Halomonhystera' and '_' in species:                                     #Remove
haplotype consisting of multiple characters from this genus
            species = part1

        genus = genus.lower().capitalize()
        species = species.lower()
        writer.write('>' + code + '_Paul_' + genus + '_' + species + '\n')
        writer.write(PaulDic[t] + '\n')
    writer.close()


    #4
    #Define a test to check if a sequence title contains 3 parts (code_Genus_species). If not, correct.
    def Tunstandardcheck(intitle):
        """Check if Tunesia sequence title contains 3 parts (code_Genus_species). If not, correct by
adding 'sp' as species."""
        if intitle.count('_') < 2:              #If there's only one underscore, layout is Genus_code
            code, genus = t.split('_')          #Split genus and code and add 'sp' in between
            outtitle = code + '_' + genus + '_sp'
        else:
            outtitle = intitle                  #If there are multiple underscores, leave input sequence
title unchanged
        return outtitle

    #Create dictionary for Tunesia fasta: TunDic['code_Genus_species'] = sequence
    TunDic = {}
    reader = open(Tunfile, 'r')
    line = reader.readline()
    while line:
        if '>' in line:
            title = line.strip('\n')
            title = title.strip('>')
```

```
            title = title.replace('_', '.', 1)
        else:
            sequence = line.strip('\n')
            sequence = sequence.replace('-', '')                #Remove all gaps from sequence
            TunDic[title] = sequence
        line = reader.readline()
    reader.close()

    #Read the file with all PCR and DNA codes and put those in dictionary
    CodeconvDic = {}
    reader = open('TunSpecies PCR and DNA codes.txt', 'r')
    line = reader.readline()
    line = reader.readline()
    while line:
        line = line.strip('\n')
        PCR, DNA = line.split('\t')
        PCR = PCR.replace(',', '.')
        CodeconvDic[PCR] = DNA
        line = reader.readline()
    reader.close()

    #Set title layout to Code_Location_Genus_species + filter unidentified species and write to outfile
    writer = open(outfile, 'a')
    for t in TunDic:
        nt = Tunstandardcheck(t)        #Check layout sequence title with previous function
        code, genus, species = nt.split('_', 2)
        code = CodeconvDic[code]                    #Replace PCR code with specimen DNA number
        genus = genus.lower().capitalize()
        species = species.replace('_', '').lower()
        if genus == 'Unknown':
            genus = 'Unidentified'
        writer.write('>' + code + 'F_Tun_' + genus + '_' + species + '\n')
        writer.write(TunDic[t] + '\n')
    writer.close()


    #5
    #Define a test to check if a sequence title contains 3 parts (Genus_species_code). If not, correct.
    def Vietstandardcheck(intitle):
        """Check if Vietnam sequence title contains 3 parts (Genus_species_code). If not, correct by
adding 'sp' as species."""
        if intitle.count('_') < 2:              #If there's only one underscore, layout is Genus_code
            genus, code = t.split('_')          #Split genus and code and add 'sp' in between
            outtitle = genus + '_sp_' + code
        else:
            outtitle = intitle                  #If there are multiple underscores, leave input sequence
title unchanged
        return outtitle

    #Create dictionary for Vietnam fasta: VietDic['Genus_species_code'] = sequence
    VietDic = {}
    reader = open(Vietfile, 'r')
    sequence = ''
    line = reader.readline()
    title = line.strip('\n')
    title = title.strip('>')
    line = reader.readline()
    while line:
        if '>' in line:
            VietDic[title] = sequence.strip(' ')
            sequence = ''
            title = line.strip('\n')
            title = title.strip('>')
            if '__' in title:
                title = title.replace('__', '_')        #Correct for double underscores
        else:
            line = line.replace(' ', '')                #Remove all spaces from sequence part
            line = line.replace('-', '')                #Remove all gaps from sequence part
            sequence = sequence + line.strip('\n')
        line = reader.readline()
    reader.close()

    #Set title layout to Code_Location_Genus_species + filter unidentified species (all to "sp", because
some are named 'sp.') and write to outfile
    writer = open(outfile, 'a')
    for t in VietDic:
        nt = Vietstandardcheck(t)        #Check layout sequence title with previous function
```

```
        genus, species, code = nt.split('_', 2)

        if '_' in code:
            part1, part2 = code.split('_')
            code = part2            #Code is always last part of title, so take last part as code, and
leave out potential haplotype

        #Correct for species novae
            if ('nov' in species and (len(species) < 5)) or ('nov' in part1 and (len(part1) < 5)):
#Check length to avoid full species name containing coincidential 'n' or 'nov'
                if any(char.isdigit() for char in species):                              #Correct for
"sp_n", "sp_nov" or "nov_sp". Standardize to sp.nov but keep species number.
                    for c in species:
                        if c.isdigit():
                            species = 'sp' + c + '.nov'
                elif any(char.isdigit() for char in part1):
                    for c in part1:
                        if c.isdigit():
                            species = 'sp' + c + '.nov'
                else:
                    species = 'sp.nov'
            if part1 == 'n':
                species = 'sp.nov'
        if 'spn' in species and len(species) < 5:        #Correct for "spn". Standardize to sp.nov but
keep species number. Check length to avoid full species name containing coincidential 'spn'
            if species[-1].isdigit():                    #Assuming there will not be 10 or more different
unidentified sp. nov of the same genus
                species = 'sp' + species[-1] + '.nov'
            else:
                species = 'sp.nov'

        genus = genus.lower().capitalize()
        species = species.strip('.').lower()
        writer.write('>' + code + '_Viet_' + genus + '_' + species + '\n')
        writer.write(VietDic[t] + '\n')
    writer.close()


    #6
    #Create list for PNG fasta with items ['Code_Genus_species', sequence]
    PNGlist = []
    reader = open(PNGfile, 'r')
    line = reader.readline()
    while line:
        if '>' in line:
            title = line.replace(' ', '_')               #Replace all spaces in title with underscores
        else:
            sequence = line.strip('\n')
            sequence = sequence.replace('-', '')               #Remove all gaps from sequence
            PNGlist.append([title, sequence])
        line = reader.readline()
    reader.close()

    #Read the file with all voucher codes and put those in list
    CodeList = []
    reader = open(codefile, 'r')
    line = reader.readline()
    while line:
        line = line.strip('\n')
        CodeList.append(line)
        line = reader.readline()
    reader.close()

    #Set title layout to Code_Location_Genus_species + write to outfile
    writer = open(outfile, 'w')
    codenr = 0
    for i in PNGlist:
        code = CodeList[codenr]             #Take code from code list
        firstp, genus, species, *endp = i[0].split('_', 3)
        endp = endp[0]
        if '_' in endp:
            part1, part2 = endp.split('_')
            species = species + part1
        if species == 'n.sp':            #Rename "n.sp" to "sp.nov", conform with standard layout
            species = 'sp.nov'
        genus = genus.lower().capitalize()
        species = species.strip('.').lower()
```

```
        writer.write('>' + code + 'A_PNG_' + genus + '_' + species + '\n')
        writer.write(i[1] + '\n')
        codenr += 1
    writer.close()

fastamerge('Cuba_18S_all_27_sequences.txt', 'Pan_18S_IDCORRECT.fas', 'Paulina_FINAL_18S_sofie4.fas',
'Tunesia_18S.fas', 'Vietnam_18S-Tien Yen-Can Gio-gb.fas', 'PNG_A.Dhondt_18S.fas',
'All18SFastasMerged.fasta')
```

10.8.2. **Table builder** (to write the database)

```
def tablebuilder(Ribinfile, COIinfile, Taxfile, outfile):
    """
    Creates a table (named as preferred in argument "outfile"), readable in excel, that displays the
following columns:
    - Voucher code
    - Class
    - Order
    - Family
    - Genus
    - Species
    - 18S checkbox
    - COI checkbox
    - 18S sequence
    - COI sequence
    - Location
    read from the information in two input fasta files (18S and COI) and the taxonomy list ("taxfile").

    Note: 18S is renamed "Rib" from "Ribosomal RNA" because numbers cannot be used in item names in
Python.
    """

    #Define function to convert fasta file information to dictionary
    def fastadicconv(infasta):
        """
        Reads the information in the fasta file and puts it in a working-friendly dictionary with layout
dictionary[Code_location] = [[genus, species, sequence].
        """
        Dic = {}
        reader = open(infasta, 'r')
        line = reader.readline()
        while line:
            if '>' in line:
                title = line.strip('\n')
                title = title.strip('>')
            else:
                if line == '\n':                #Correction for possible blank line
                    pass
                else:
                    sequence = line.strip('\n')
                    sequence = sequence.replace('-', '')                #Remove all gaps from sequence
                    code, loc, genus, species = title.split('_', 3)
                    Dic[code+'_'+loc] = [genus, species, sequence]
            line = reader.readline()
        reader.close()
        return Dic

    #Read the information from the two fasta files and convert it to dictionary using previous function
    RibDic = fastadicconv(Ribinfile)
    COIDic = fastadicconv(COIinfile)

    #Define function to look up higher taxonomy of the genus in the taxonomy list
    def findtaxonomy(ingen):
        """
        Searches input genus in given taxonomy list and returns
        """
        reader = open(Taxfile, 'r')
        line = reader.readline()
        while line:
            gen, htax = line.split('\t', 1)
            if gen == ingen:
                fam, ord, cl = htax.split('\t')
                fam = fam.strip('\n')
                cl = cl.strip('\n')
                return cl, ord, fam
```

```
            line = reader.readline()
            gen = ''
        reader.close()
        assert gen != '', "Genus not found in taxonomy list." #Give assertionerror if the genus is not in
the list


    #Write the information in the dictionaries to the output file in tab-delimited table format
    writer = open(outfile, 'w')
    writer.write('Code' + '\t' + 'Class' + '\t' + 'Order' + '\t' + 'Family' + '\t' + 'Genus' + '\t' +
'Species' + '\t' + '18S' + '\t' + 'COI' + '\t' + '18S sequence' + '\t' + 'COI sequence' + '\t' +
'Location' + '\n')    #Write the header of the table with column names
    for item in RibDic:
        code, loc = item.split('_')
        COIseq = ''
        COIcheck = ''
        if item in COIDic:      #Check if the specimen is present in the COI dictionary. If so, add the
sequence to the line and check the 'COI' checkbox column with an 'x'
            COIseq = COIDic[item][2]
            COIcheck = 'x'
        cl, ord, fam = findtaxonomy(RibDic[item][0])   #Get higher taxonomy using previous "findtaxonomy"
function
        writer.write(code + '\t' + cl + '\t' + ord + '\t' + fam + '\t' + RibDic[item][0] + '\t' +
RibDic[item][1] + '\t' + 'x' + '\t' + COIcheck + '\t' + RibDic[item][2] + '\t' + COIseq + '\t' + loc +
'\n')

    for item in COIDic:
        if item not in RibDic:
            code, loc = item.split('_')
            cl, ord, fam = findtaxonomy(COIDic[item][0])   #Get higher taxonomy using previous
"findtaxonomy" function
            writer.write(code + '\t' + cl + '\t' + ord + '\t' + fam + '\t' + COIDic[item][0] + '\t' +
COIDic[item][1] + '\t' + '' + '\t' + 'x' + '\t' + '' + '\t' + COIDic[item][2] + '\t' + loc + '\n')
    writer.close()

tablebuilder('All18SFastasMerged.fasta', 'AllCOIFastasMerged.fasta', 'NemTaxonomy.txt',
'DatabaseTableTest3.txt')
```

10.8.3. **Adhoc converter** (changes sequence labels for use in Adhoc)

```
def ahconv(infile, outfile, location):
    """
    Changes the sequence label format from ">Code_(Location_)Genus_species" to
">Genus_species_Code(_Location)",
    to be suitable to use with R Adhoc package.
    Third argument needs to be given to indicate if there is a location present in the sequence label.
Give "y" (yes) if there is
    a location in the sequence label. Give "n" (no) if there is not. This way the location can be
recognized in the label and the
    right path can be chosen.
    Sequences are copied unchanged.
    """
    reader = open(infile, 'r')
    line = reader.readline()
    writer = open(outfile, 'w')
    while line:
        if '>' in line:
            title = line.strip('\n')
            title = title.strip('>')
            if location == 'y':
                code, loc, name = title.split('_', 2)           #Alternative 2 script lines when location
in sequence labels
                writer.write('>' + name + '_' + code + '_' + loc +'\n')
            if location == 'n':
                code, name = title.split('_', 1)               #Alternative 2 script lines when no
location in sequence labels
                writer.write('>' + name + '_' + code + '\n')
        else:
            writer.write(line)
        line = reader.readline()

    reader.close()
    writer.close()
```

```
ahconv("Xyalidae_extracted_sequences_18S.fas", "Xyalidae_extracted_sequences_18S_Adhoc.fas", "n")
```

10.8.4. **TaxonomyMapper** (creates “ID to Taxonomy Mapping File” for use in QIIME)

```
def taxmap(tablefile, outfile):
    """
    Reads the database table (.txt) file and creates a list of all specimens in the database and their
taxonomy,
    for usage as ID to Taxonomy Mapping File in QIIME. Layout:
    Code    Pylum; Class; Order; Family; Genus; species
    """
    reader = open(tablefile, 'r')
    line = reader.readline()
    line = reader.readline()    #Skip first line of database with column headers
    writer = open(outfile, 'w')
    while line:
        code, cl, order, fam, gen, sp, rib, COI, ribseq, COIseq, loc = line.split("\t")     #Assign the
right name to all components of row (specimen)
        loc = loc.strip('\n')
        seqtitle = code + '_' + loc + '_' + gen + '_' + sp      #Reconstruct the sequence labels as used
in the total fasta files
        writer.write(seqtitle + '\t' + 'Nematoda' + ';' + cl + ';' + order + ';' + fam + ';' + gen + ';' +
sp + '\n')   #Write layout as requered for use in QIIME
        line = reader.readline()

    reader.close()
    writer.close()

taxmap("DatabaseTableTest3.txt", "IDtoTax.txt")
```

10.8.5. **Species sorter** (creates a list of all species present per location, without duplicates)

```
def speciessort(infile, outfile):
    """
    Reads total fasta file and gives output file with all species listed per location. Species names
listed are unique. All duplicates are filtered prior to writing output file.
    """

    reader = open(infile, 'r')
    line = reader.readline()
    ploc = 'Cuba'
    spset = set()
    while line:
        if '>' in line:
            title = line.strip('\n')
            title = title.strip('>')
            code, loc, name = title.split('_', 2)
            if loc == ploc:
                spset.add(name)
            else:
                writer = open(outfile, 'a')
                writer.write('\n' + ploc + '\n')
                for i in spset:
                    writer.write(i + '\n')
                ploc = loc
                spset.clear()
                code, loc, name = title.split('_', 2)
                spset.add(name)
        line = reader.readline()

    writer = open(outfile, 'a')
    writer.write('\n' + ploc + '\n')
    for i in spset:
        writer.write(i + '\n')
    reader.close()
    writer.close()


speciessort('All18SFastasMerged Final.fasta', 'Specieslist18S.txt')
```