Environmental DNA

WILEY

Environmental DNA
Open Access
Dedicated to the study and use of environmental DNA for basic and applied sciences

ORIGINAL ARTICLE OPEN ACCESS

Persistent Gaps and Errors in Reference Databases Impede Ecologically Meaningful Taxonomy Assignments in 18S rRNA Studies: A Case Study of Terrestrial and Marine Nematodes

Alejandro De Santiago[1,2] | Tiago José Pereira[1,2] | Timothy John Ferrero[3] | Natalie Barnes[3] | Delphine Lallias[4] | Simon Creer[4] | Holly M. Bik[1,2]

[1]Department of Marine Sciences, University of Georgia, Athens, Georgia, USA | [2]Institute of Bioinformatics, University of Georgia, Athens, Georgia, USA | [3]Department of Zoology, The Natural History Museum, London, UK | [4]Molecular Ecology and Evolution at Bangor (MEEB), School of Environmental and Natural Sciences, Bangor University, Bangor, Gwynedd, UK

Correspondence: Holly M. Bik (hbik@uga.edu)

Received: 6 October 2024 | **Revised:** 21 February 2025 | **Accepted:** 25 February 2025

Funding: Funding for this study was provided by the North Pacific Research Board (NPRB project 1303), The Gulf of Mexico Research Initiative, and institutional startup funding from the University of Georgia to H.M.B. We also thank the Center for Conservation Biology at the University of California, Riverside, and The Shipley-Skinner Reserve—Riverside County Endowment for partially funding this project. Part of this work was funded by a NERC Post-Genomics and Proteomics Grant (Ref NE/F001266/1) and New Investigator Grant NE/E001505/1 to S.C. Research support for ADS was provided by the University of Georgia Research Foundation and the National Institute of General Medical Sciences of the National Institute of Health under award number 1T32GM142623. This work was also supported by a National Science Foundation CAREER award to HMB (DEB- 2144304).

Keywords: 18S rRNA metabarcoding | microbial eukaryotes | nematodes | reference databases | taxonomy assignment

ABSTRACT

In metabarcoding studies, Linnaean taxonomy assignments of Operational Taxonomic Units (OTUs) or Amplicon Sequence Variants (ASVs) underpin many downstream bioinformatics analyses and ecological interpretations of environmental DNA (eDNA) datasets. However, public molecular databases (i.e., SILVA, EUKARYOME, BOLD) for most microbial metazoan phyla (nematodes, tardigrades, kinorhynchs, etc.) are sparsely populated, negatively impacting our ability to assign ecologically meaningful taxonomy to these understudied groups. Additionally, the choice of bioinformatics parameters and computational algorithms can further affect the accuracy of eDNA taxonomy assignments. Here, we use two *in silico* datasets to show that taxonomy assignments using the 18S rRNA gene can be dramatically improved by curating Linnaean taxonomy strings associated with each reference sequence and closing phylogenetic gaps by improving taxon sampling. Using free-living nematodes as a case study, we applied two commonly used taxonomy assignment algorithms (BLAST+ and the QIIME2 Naïve Bayes classifier) across six iterations of the SILVA 138 reference database to evaluate the precision and accuracy of taxonomy assignments. The BLAST+ top hit with a 90% sequence similarity cutoff often returned the highest percentage of correctly assigned taxonomy at the genus level, and the QIIME2 Naïve Bayes classifier performed similarly well when paired with a reference database containing corrected taxonomy strings. Our results highlight the urgent need for phylogenetically informed expansions of public reference databases (encompassing both genomes and common gene markers), focused on poorly sampled lineages that are now robustly recovered via eDNA metabarcoding approaches. Additional taxonomy curation efforts should be applied to popular reference databases such as SILVA, and taxon sampling could be rapidly improved by more frequent incorporation of newly published GenBank sequences linked to genus- and/or species-level identifications.

1 | Introduction

Meiofauna (microbial metazoan groups < 1 mm in size, such as nematodes, tardigrades, copepods, kinorhynchs, and other "minor" phyla) are abundantly found in terrestrial and marine habitats worldwide. This group exhibits high biodiversity but suffers from difficult morphological taxonomy, often requiring a high level of expertise for accurate genus- or species-level identifications (Blaxter 2016; De Ley et al. 2005; Lambshead 1993; Miljutin et al. 2010). Historically, the small size and low biomass of meiofauna have hindered the use of whole genome approaches, which typically require High Molecular Weight (HMW) DNA, resulting in sparse molecular databases compared to larger invertebrate taxa. (Dell'Anno et al. 2015; Holovachov et al. 2017; Weigand et al. 2019). High-throughput environmental DNA (eDNA) metabarcoding approaches are now touted as a robust method for overcoming persistent gaps in our biodiversity knowledge of environmental microbiomes, including meiofaunal communities (Bik et al. 2012; Deiner et al. 2017; van der Loos and Nijland 2021), and facilitating rapid studies of species richness and phylogeographic patterns on a global scale (Holman et al. 2021; Hu et al. 2022). However, in contrast to traditional morphological studies, metabarcoding surveys inherently rely on reference databases for linking unknown environmental DNA sequences (ASVs—Amplicon Sequence Variants; OTUs—Operational Taxonomic Units) with ecologically informative taxonomic identifications (Fonseca et al. 2014; Sinniger et al. 2016). Furthermore, taxonomy assignments of eDNA sequences are strongly impacted by the choice of tools and parameters in bioinformatics workflows (De Santiago et al. 2022; Hleap et al. 2021). For meiofauna with historically poor database representation, we do not fully understand the scientific and practical implications that database composition and algorithm choice may have on metabarcoding taxonomy assignments.

Several recent studies have comprehensively reviewed and assessed the performance and accuracy of taxonomy assignment algorithms (Gardner et al. 2019; Hleap et al. 2021). These studies have primarily focused on mitochondrial loci, such as the Cytochrome Oxidase subunit I (COI) for metabarcoding studies of macroinvertebrates and vertebrates (Bourret et al. 2023; Hleap et al. 2021; Mathon et al. 2021; O'Rourke et al. 2020), as well as the 16S rRNA (Ribosomal RNA) gene targeting prokaryotic assemblages (Almeida et al. 2018; Ritari et al. 2015; Siegwald et al. 2017). Both COI and 16S rRNA benefit from robust databases stemming from historically intensive sequencing and curation efforts (e.g., of distant prokaryotic phylogenetic clades; Mukherjee et al. 2017; Thompson et al. 2017), and constrained species diversity and broad public interest (e.g., in fish and mammals; Ratnasingham and Hebert 2007; Zhang et al. 2020; Zhu et al. 2023). For most metabarcoding markers, database completeness varies by habitat and taxa, with better representation for terrestrial and freshwater invertebrates. For example, marine invertebrates in Europe are the most underrepresented group in reference databases, with only 22.1% of described species having publicly available DNA barcodes. In stark contrast, over 80% of freshwater and marine fish species have reference DNA barcodes available (Weigand et al. 2019). Furthermore, there is a growing recognition that regional reference databases are critical for accurate taxonomic assignments (Bourret et al. 2023) since non-native sister taxa frequently lead to misassignments when global reference databases are used (Gold et al. 2021). Thus, there is a critical need to conduct similar evaluations for other common metabarcoding loci, such as the 18S rRNA gene, where the performance of bioinformatics pipelines and taxonomy assignment algorithms has been comparatively understudied.

Hleap et al. (2021) defined four discrete categories of taxonomy assignment algorithms for metabarcoding studies: (1) *Sequence Similarity* tools use global or local nucleotide alignments to match ASVs/OTUs with sequences in reference databases, using percent identity cutoffs (e.g., usually 80%–99%, depending on loci) or minimum confidence values to make robust taxonomy assignments. BLAST (Altschul et al. 1997) is the most widely used algorithm in this category; (2) *Sequence Composition* approaches harness compositional features of ASVs/OTUs, such as nucleotide frequency patterns, using a model-based approach that link these profiles to specific taxonomic groups in reference databases. The Ribosomal Database Project (RDP; Wang et al. 2007) and the newer QIIME2 Naïve Bayes classifier (Bokulich et al. 2018) are two popular Naïve Bayes algorithms in this category; (3) *Phylogenetic approaches* infer taxonomy assignments by placing ASVs/OTUs within a reference phylogeny, most often using a read recruitment strategy to place short ASVs/OTUs onto branches within a pre-computed guide tree (typically built using full-length reference sequences). Algorithms in this category include pplacer (implementing Bayesian and Maximum Likelihood options; Matsen et al. 2010), the Evolutionary Placement Algorithm (EPA, a Maximum Likelihood method; Barbera et al. 2019), and HmmUFOtu (utilizing Hidden Markov Models in conjunction with pplacer/EPA algorithms; Zheng et al. 2018); (4) *Probabilistic methods* rely on statistical frameworks, such as multinomial regression, to evaluate the probability that a given ASV/OTU should be assigned to a particular taxonomic rank in a reference database. The PROTAX algorithm (Somervuo et al. 2016) is the most commonly applied approach in this category. Other software tools exist in each category, but the above represent some of the most commonly cited workflows with active development and user support (Hleap et al. 2021).

Interestingly, BLAST and the QIIME2 Naïve Bayes classifier—despite being considered more "simplistic" approaches—consistently outperform other algorithm classes and software tools and return the most accurate taxonomy assignments (Hleap et al. 2021). Sequence similarity tools, such as BLAST, perform robustly in the face of large and heterogeneous reference databases, in contrast to more complex machine learning methods (probabilistic and phylogenetic software tools) which are more sensitive to database size and taxonomic coverage. Furthermore, the phylogenetic method HmmUFOtu alarmingly places randomly generated control sequences into a taxonomic group (Hleap et al. 2021). Despite the promise of evolutionary-informed guide tree analyses, fast heuristic searches required for large phylogeny reconstruction do not yet return reliably accurate results for metabarcoding taxonomy assignments (Hleap et al. 2021). Subsequent investigations have upheld the consistently robust performance of BLAST in metabarcoding analysis pipelines (Bourret et al. 2023), and benchmarking analyses indicate that gaps in

reference databases actually represent the primary source of error in eDNA metabarcoding studies. This scenario results in false negatives where eDNA sequences remain unassigned, or eDNA sequences are "under-classified" to higher-level taxonomic groups when using Least Common Ancestor (LCA) methods (Gardner et al. 2019).

Among meiofaunal taxa, the phylum Nematoda provides an ideal case study for advancing our knowledge of how 18S rRNA database gaps and taxonomy assignment algorithms impact eDNA metabarcoding inferences in poorly studied metazoan groups. Similar to many protists and other meiofaunal groups, nematodes are poorly represented in reference databases and harbor large amounts of cryptic diversity, making it difficult to delimit species boundaries when solely using traditional morphological methods. Estimates of global nematode richness range from 50,000 up to 100 million species (Blaxter 2016; Lambshead 1993; Mokievsky and Azovsky 2002), yet this phylum suffers from a perpetual and severe taxonomic deficit, with only ~26,000 nematode species formally described (Hugot et al. 2001). Molecular databases for terrestrial groups have disproportionately benefited from an active research community surrounding model species (i.e., *Caenorhabditis elegans, Pristionchus pacificus*, etc.) and parasitic nematodes. In contrast, approximately ~5000 free-living marine nematode species have been formally described, and of that number, only around 16% represent deep-sea species (despite deep-sea habitats covering 91% of the earth's surface and generally exhibiting high biodiversity; Miljutin et al. 2010). Furthermore, the vast majority of nematode species are not linked to a corresponding DNA barcode, with marine species in particular being especially underrepresented in public sequence databases (Dell'Anno et al. 2015; Macheriotou et al. 2019). Nematodes and other meiofauna taxa have been touted as ideal bioindicators for aquatic biomonitoring, owing to their short generation times and rapid species-specific responses to environmental stress (Moens et al. 2014; Zeppilli et al. 2015). For nematodes especially, family- and genus-level taxonomic assignments convey important ecological information such as feeding groups (e.g., bacterial-feeding vs. predatory species), inference of taxon-specific responses to disturbance or environmental stress, and definitive identification of parasitic or pathogenic species (Bongers 1990; Bongers et al. 1991; Chitwood 2003; Moens et al. 2014). Metabarcoding workflows for most meiofaunal groups are still limited in providing robust low-level taxonomy without parallel morphological identifications or significant manual curation of taxonomy strings from typical bioinformatics outputs.

In this study, we aimed to comprehensively assess the status of 18S rRNA reference databases for the phylum Nematoda (using this group as a proxy for poorly-studied meiofauna) and investigate how database structure and taxonomy assignment algorithms can impact the accuracy of family- and genus-level identifications of molecular sequences. We further focused on investigating (1) whether public sequence databases exhibit uneven representation across nematode taxa, habitat, and lifestyle, (2) how bioinformatics parameters and database curation impact the accuracy of metabarcoding taxonomy assignments, including the accuracy and completeness of database-derived reference taxonomy strings, and (3) whether manual curation of reference databases and increasing representation of nematode groups improve taxonomy assignments for the phylum Nematoda, especially at the ecologically informative genus-level.

2 | Materials and Methods

2.1 | Creating the In Silico Nematode Metabarcoding Datasets

Using full-length 18S rRNA sequences generated as part of an ecological study from the Shipley Skinner Reserve, CA, USA (Pereira et al. 2024), a genetics study of the 18S rRNA gene across marine nematodes (Pereira et al. 2020), and two unpublished biodiversity studies from Bodega Bay, CA, USA, and the Northern Gulf of Alaska, we created two synthetic metabarcoding datasets representing nematode communities typically present in marine and terrestrial habitats (herein referred to as "*in silico* datasets"; Figure 1; Figure S1). To create the original full-length reference sequences, nematode specimens were extracted from bulk soil or sediment samples. Soil nematodes were isolated from ~100 g of soil using the Baermann Funnel Technique (Viglierchio and Schmitt 1983) as described in Pereira et al. (2024). Marine nematodes were isolated from sediments using a decantation-flotation method in a 2-L glass cylinder and decanted over a 63-μm sieve using sterile artificial seawater (Instant Ocean), as described in Pereira et al. (2020). Prior to DNA extraction, nematode specimens were picked under a dissecting microscope, mounted on temporary slides, and morphologically identified using a compound microscope (Pereira et al. 2020). All the sequences included in the *in silico* datasets were taxonomically identified to at least the genus level. The full-length 18S rRNA gene was amplified using three overlapping primers (Pereira et al. 2020). Sanger sequences were quality-checked and assembled using CodonCode v4.2.7 (CodonCode Corporation, LI-COR Inc.). To subsequently create the *in silico* datasets used in this study, the full-length 18S sequences were aligned and trimmed with MAAFT (Katoh and Standley 2013) using the default settings via the Geneious Prime Software v2022.0.1 (https://www.geneious.com). Full-length sequence alignments were trimmed down to the V1-V2 hypervariable region (350–400 bp) commonly used in metabarcoding studies of meiofauna. Trimming was carried out using the G18S4/F04 forward (5′-GCTTGTCTCAAAGATTAAGCC-3′) and R22 reverse (5′-GCCTGCTGCCTTCCTTGGA-3′) primers (Blaxter et al. 1998; Creer et al. 2010), by matching to these conserved regions and trimming at the 3′ end of the primers. Primers were expected to match at least 10 bp with a maximum of 3 mismatches. None of the Sanger sequences were discarded due to low sequence quality or primer mismatches.

The final marine *in silico* dataset consists of 221 (117 unique) nematode sequences, composed of 79 sequences examined in a previous study (Pereira et al. 2020) and 142 previously unpublished 18S rRNA sequences deposited in GenBank as part of this study (Table S1). The marine *in silico* dataset includes 20 families and 47 genera. Marine nematodes represent deep-sea and shallow-water taxa spanning multiple ocean basins, including the Gulf of Mexico, intertidal and offshore sites in Southern California, and continental shelf and deep-sea sites in the Alaskan Beaufort and Chukchi Seas. The terrestrial *in silico* dataset consists of 117 (96 unique) nematode sequences

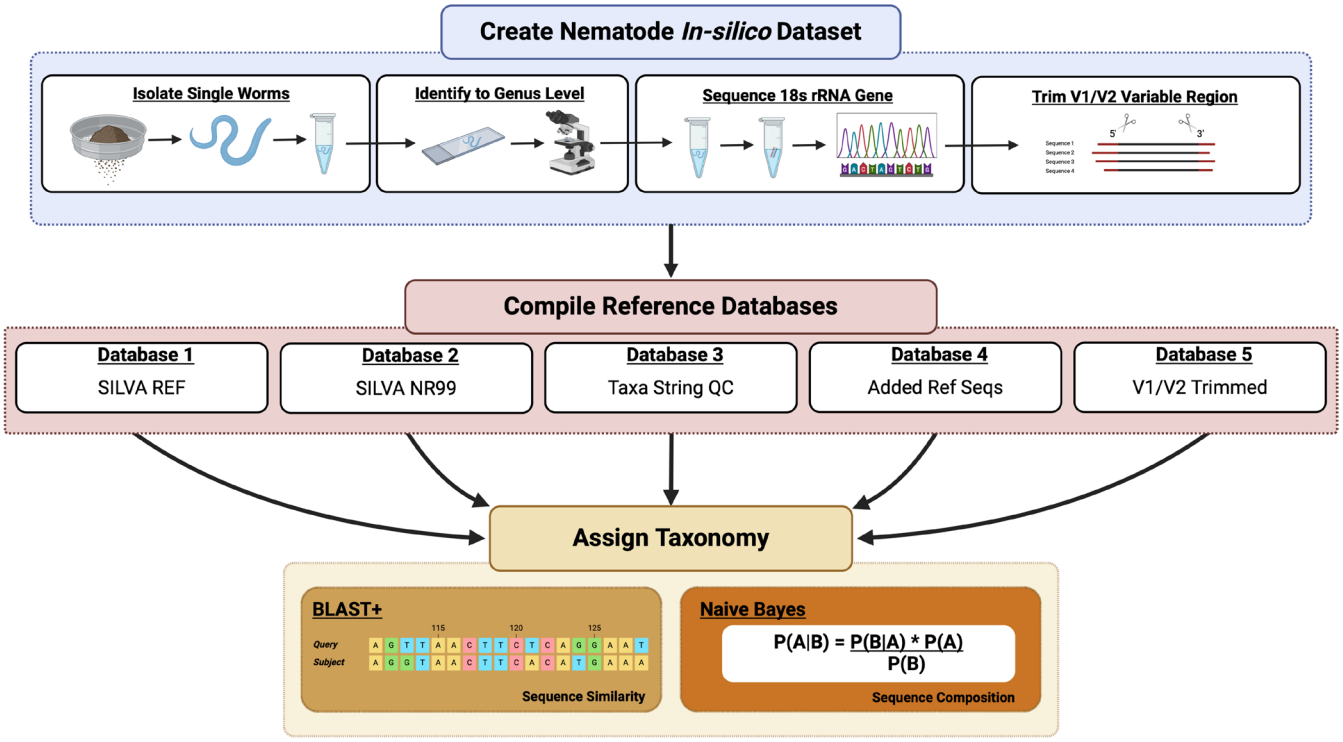

FIGURE 1 | General Workflow for this Study. Workflow diagram of (top) how the *in silico* datasets were generated, (middle) the reference datasets used in this study, and (bottom) the two methods used for taxonomy assignment. Illustration created with BioRender.com.

generated during a study of semi-arid ecosystems within the Shipley-Skinner Reserve in Southern California (Pereira et al. 2024). The terrestrial *in silico* dataset includes 17 families and 30 free-living and plant-parasitic terrestrial nematode genera (Table S1). A midpoint-rooted phylogenetic tree, to visualize the taxonomic diversity included in the *in silico* datasets, was constructed based on the V1-V2 regions of the 18S rRNA gene using the QIIME2 *align-to-tree-mafft-raxml* pipeline (Katoh and Standley 2013; Stamatakis 2014) with the default settings (Figure S1).

2.2 | Constructing the Reference Databases

In this study, we focused our assessments solely on the SILVA database (Quast et al. 2013) due to its stringent quality control of reference sequences and wide use as a reference database for both 16S and 18S rRNA metabarcoding studies (Yilmaz et al. 2013). SILVA eliminates short, chimeric, and erroneous sequences through a stringent QA/QC pipeline by employing a phylogeny-guided approach of high-quality DNA sequences and associated taxonomy strings routinely retrieved from GenBank (Quast et al. 2013). Here, we tested the effects of six iterations of the SILVA 138 database (the most recent version released in 2020) on the accuracy of taxonomy assignments on the phylum Nematoda (Figure 1; Table 1).

Database 1 (SILVA REF) is the original SILVA 138 release reference database (Quast et al. 2013). Database 2 (SILVA NR99) is the SILVA 138 release NR99, where DNA sequences are clustered at 99% sequence similarity to reduce the number of redundant sequences and overrepresentation of common taxa in the dataset (Quast et al. 2013). In our assessments, we found that databases 1 and 2 contained sequences with degenerate bases (i.e., Y, M, K, S, N, etc.) that can interfere with taxonomy assignment workflows for metabarcoding datasets. To reduce potential errors, we removed sequences with more than five degenerate bases from these databases. For Database 3 (Taxa Strings QC), we manually curate the taxonomy strings associated with nematode reference sequences in Database 2 (SILVA NR99; Figure 2). Taxonomic strings downloaded from SILVA were curated using Linnaean hierarchies from the WoRMS database (WoRMS Editorial Board 2023). The WoRMS database integrates several species databases, including the Nemys database (Nemys Editorial Board 2024) that includes taxonomic descriptions of both marine and terrestrial nematode taxa. Nematode sequences found to have a chimeric taxonomy string (34 sequences; Table S2) were removed from the SILVA NR99 database to eliminate any ambiguous or potentially incorrect sequences. Additionally, non-identifiable information was removed from the taxonomy string (i.e., "uncultured eukaryote", "environmental sample", "metagenome", etc.; 134 sequences).

For Database 4 (Added More Seq), we took the SILVA NR99 database with curated taxonomy strings and added additional marine nematode sequences to expand the number of families and genera represented in the reference database (Table S3). We did not include the sequences from our *in silico* datasets in the reference data to prevent biasing our results. Adding the same Sanger sequences from our test dataset into our reference databases would artificially inflate the number of true positives from our taxonomy assignment methods. In particular, the BLAST+ top hit methods would result in a 100% true positive rate but would perform drastically differently in real-world applications. Instead, we included 567 short (~350 bp)

TABLE 1 | Overview of database construction, formatting, and the number of nematode sequences available.

Database name	Database contents	Nematode sequences	Reference
Database 1 SILVA REF	SILVA 138 REF Database	5781	Quast et al. (2013), Yilmaz et al. (2013)
Database 2 SILVA NR99	SILVA 138 Non-Redundant Database (99% clustered OTUs)	2174	Quast et al. (2013)
Database 3 Taxa String QC	Database 2 (SILVA NR99) + manual curation of taxonomy strings	2068	Holovachov et al. (2017)
Database 4 Added Ref Seqs	Database 3 (Taxa String QC) + additional nematode reference sequences (567 added)	2635	Lallias et al. (2015), Macheriotou et al. (2019)
Database 5 V1/V2 Trimmed	Database 4 (Added Ref Seqs) + all nucleotide sequences trimmed down to V1/V2 region (~350 bp)	2618	Werner et al. (2012), Macheriotou et al. (2019)
Database 6 Closed Gaps	Database 5 (V1/V2 Trimmed) + in silico Sanger sequences (~350 bp)	2956	Pereira et al. (2020), Pereira et al. (2024)

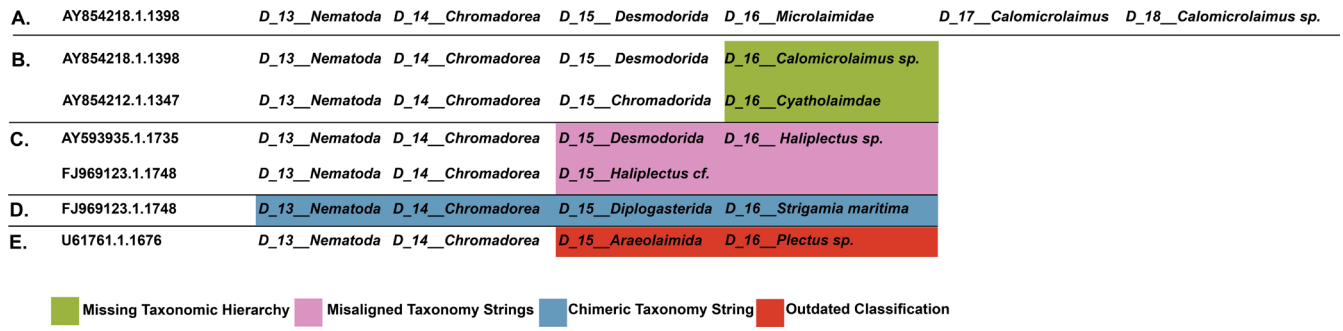

FIGURE 2 | Taxonomy String Errors in the SILVA Database. Four different types of taxonomy string errors found in the SILVA reference database that contribute to misleading taxonomy assignments for eukaryotic metabarcoding studies compared to an ideal taxonomy string: (A) an ideal taxonomy string with major Linnean taxonomic ranks, (B) Hierarchy that is missing some low-level taxonomic ranks (i.e., family or genus), (C) Sequences that represent the same genera or species, but the lower-level ranks do not align vertically due to flexibility of intermediate taxonomic ranks that SILVA accepts. This can affect methods that rely on consistent taxonomy strings, such as BLAST+ LCA and QIIME2 Naïve Bayes. (D) A chimeric taxonomic string that consists of a hybrid of Nematoda higher ranks with a species-level Arthropoda classification (*Strigamia maritima*). (E) Red sequences represent an outdated morphological taxonomic hierarchy (i.e., the genus *Plectus* is currently classified within the order Plectida based on evidence from recent molecular phylogenies).

and large (~700 bp) V1-V2 18S rRNA fragments generated as part of Macheriotou et al. (2019), Fonseca et al. (2012), and unpublished DNA sequences generated as part of Lallias et al. (2015). Thus, a total of 567 nematode sequences (483 unique sequences) were further added to the Database 4 reference dataset (Figure 3; Figure S2). After adding these sequences, 195 families and 627 genera are represented in Database 4, an increase of 3 families (e.g., Aegialoalaimidae, Lauratonematidae, and Tubolaimoididae) and 60 genera compared to the SILVA NR99 reference database (Database 2). We made the decision to include short-length sequences in order to greatly improve the taxonomic coverage of marine nematode taxa in our reference database. Genbank includes short-fragment sequences that are not included in the SILVA reference database due to stringent quality control. Only marine nematodes were added to the reference database to assess how expanding reference databases using habitat-specific short-fragment 18S rRNA sequences impact the accuracy of taxonomy assignment methods in metabarcoding studies (e.g., marine nematodes).

For Database 5 (V1/V2 Trimmed), we took Database 4 and carried out an additional step to trim down all reference sequences to the V1-V2 regions, applying the same method used to generate the *in silico* datasets for marine and terrestrial nematodes (described above). Previous studies have shown that for both 16S and 18S rRNA datasets, trimming the reference database to the region of interest can increase taxonomic resolution (Holovachov et al. 2017; Macheriotou et al. 2019; Werner et al. 2012). A total of 17 sequences that were shorter than 100 bp and did not span the primer regions (V1-V2 regions) were removed from Database 5 in order to standardize the sequence length and coverage of the gene region. To assess whether our results are largely impacted by phylogenetic gaps in the reference database or by parameterization of the taxonomy assignment methods, we generated Database 6, which

A.

Lineage Name	Number of Sequences		
	SILVA 138 Full	SILVA NR99	Added Ref Seq
Muspiceida	1	1	1
Rhabditida	3435	1110	1139
Plectida	82	49	67
Araeolaimida	58	18	106
Monhysterida	164	69	235
Isolaimida	3	2	2
Desmoscolecida	3	3	7
Desmodorida	113	69	127
Chromadorida	61	34	124
Dorylaimida	438	144	148
Mermithida	28	16	16
Mononchida	88	19	19
Trichinellida	667	241	241
Dioctophymatida	2	2	2
Triplonchida	249	75	77
Enoplida	388	189	324

B.

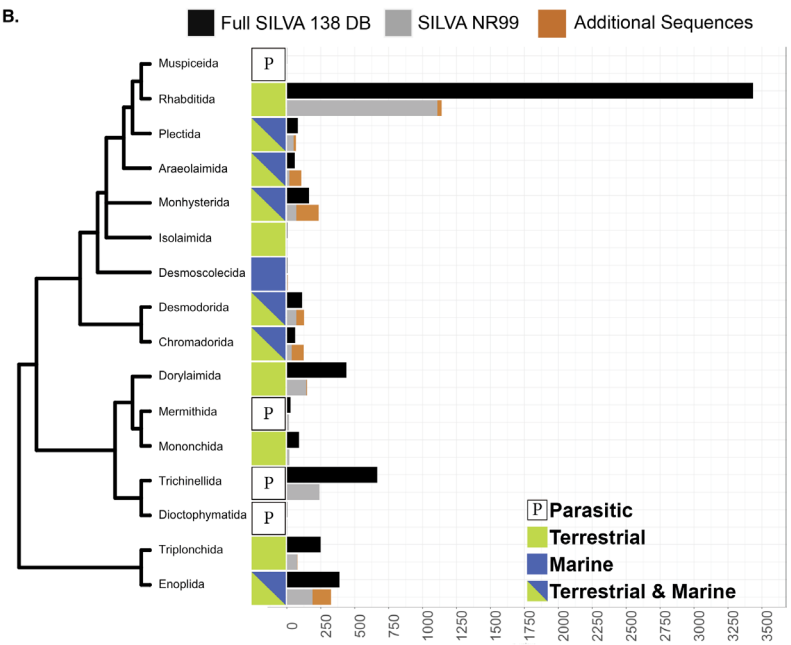

FIGURE 3 | Publicly available 18S rRNA gene sequences (> 1200 bp) linked to a nematode morphological ID. (A) Number of sequences of each major nematode lineage in each reference database. (B) The phylogenetic backbone for phylum Nematoda is derived from Ahmed and Holovachov (2021). The primary habitats of each nematode clade are denoted by colored squares (Bik et al. 2010; De Ley 2006; WoRMS Editorial Board 2023). Black bars represent sequences passing quality control parameters in the SILVA 138 REF database, and gray bars represent sequences passing quality control parameters in the SILVA non-redundant database clustered at 99% sequence identity (SILVA NR99). The minimum sequence length of public sequences in SILVA is 1200 bp, and SILVA sequences with more than five degenerate bases (i.e., Y, M, K, S, N, etc.) were removed prior to data analysis. Orange bars represent nematode partial 18S rRNA sequences not currently represented in SILVA and added as part of this study (> 350 bp, see methods for details).

includes Database 5 in addition to our *in silico* 18S rRNA sequences of marine and terrestrial nematodes.

2.3 | Assessing Nematode Representation and Taxonomy String Accuracy in the SILVA 138 18S rRNA Reference Database

The sequences and taxonomy strings in the SILVA 138 (Database 1) and the clustered SILVA 138 NR99 reference database (Database 2) were analyzed to assess the accuracy and taxon coverage of the phylum Nematoda (Figure 3; Figure S2). We manually explored the taxonomy strings associated with each nematode reference sequence and compared them to the WoRMS database to identify taxonomic errors. We identified four types of errors in the SILVA database with respect to the taxonomic string associated with nematode DNA sequences: (1) missing hierarchical taxonomic data, (2) misaligned taxonomy strings (e.g., nematodes from the same genus are missing different taxonomic ranks), (3) chimeric taxonomy strings, and (4) use of outdated taxonomic classifications (Figure 2). Error types 1 and 2 largely impact taxonomy assignment methods that heavily rely on consistent and accurately labeled data to assign a consensus taxonomy, such as the BLAST+ LCA (Least Common Ancestor) and Naïve Bayes methods. While errors 1 and 2 are minor (i.e., the sequences are correctly classified but lack taxonomic information at certain levels) and may lead to an unassigned sequence, errors 3 and 4 are considered major taxonomic errors (i.e., the sequences have the wrong taxonomic classification) and can lead to incorrect taxonomic assignments of ASV/OTU sequences (i.e., false positives).

2.4 | Assigning Taxonomy to the In Silico Datasets

The *in silico* datasets were assigned taxonomy using the BLAST+ classifier (Camacho et al. 2009) and QIIME2 Naïve Bayes classifier (Pedregosa et al. 2011) run within QIIME2 v2022.2 (Bokulich et al. 2018; Bolyen et al. 2018). For BLAST+, six scenarios were tested using a predetermined number of top hits (e.g., "N hits" parameter, which is the first N sequences with the percent similarity threshold that matches the query sequences) and percent identity parameters. The first three variations were designated "Top Hit" where the sequence query was assigned according to the first hit that matched at least the chosen percent identity threshold (e.g., 90%, 95%, and 99%) of the reference sequence. For the other two scenarios, the "*N* hits" parameter was increased (i.e., 3 and 10 hits) to determine the LCA of the query sequence. These parameters were chosen to account for two common scenarios in the SILVA database that can impact LCA methods: (1) the overrepresentation of some genera in the database and (2) some orders being represented by ≤ 3 sequences (see results). The minimum consensus (i.e., the fraction of assignments that must match to accept the taxonomy for the query sequence) for the BLAST+ LCA methods was left as the QIIME2 default (51%). Taxonomy assignment with the QIIME2 Naïve Bayes

TABLE 2 | Overview of the taxonomy assignment methods and the parameters that were used in this study.

Method name	Tool	Max accepts	Percent ID	Minimum consensus	% Confidence
BLAST.1.90	BLAST+	1	90%	—	—
BLAST.1.95	BLAST+	1	95%	—	—
BLAST.1.99	BLAST+	1	99%	—	—
BLAST.3.90	BLAST+	3	90%	51%	—
BLAST.10.95	BLAST+	10	95%	51%	—
NAÏVE.70	Naïve Bayes	—	—	—	70%
NAÏVE.80	Naïve Bayes	—	—	—	80%
NAÏVE.90	Naïve Bayes	—	—	—	90%

classifier was carried out using three different confidence value thresholds (e.g., 70%, 80%, and 90%). Every database described above was used to test how an improved reference database can impact the accuracy in assigning taxonomy to Sanger barcodes. A total of 48 comparisons were conducted in this study: six databases were each assessed using three QIIME2 Naïve Bayes scenarios and five BLAST+ scenarios. A summary of all eight taxonomy assignment methods with their respective parameters is described in Table 2.

2.5 | Evaluating the Accuracy and Precision of Taxonomy Assignment Methods and Databases

The accuracy and precision of each taxonomy assignment method and database combinations were assessed at different taxonomic ranks, viz. order, family, and genus. Each assignment was either designated as a true positive (i.e., taxonomy assignment of a sequence reflects the correct taxonomy of the query sequence at that specific taxonomic level.), a false positive (i.e., taxonomy at that specific level does not match the identity of the sequence), or a false negative (i.e., the sequence remains unassigned at that specific taxonomic level). Due to the sparse reference databases and the lack of taxonomic identification at the species level for most sequences, we did not assess the accuracy or precision of taxonomic assignments at the species level. Similar to Bourret et al. (2023), accuracy (a) was calculated as the proportion of true positives among all the sequences examined ($a = TP/(TP + FP + FN)$); precision (p) was calculated as the proportion of true positives among all the positive predictions ($p = TP/(TP + FP)$). Previous studies have shown that the accuracy of taxonomy assignments to several invertebrate groups using BLAST+ and LCA methods routinely falls below 70% (Bourret et al. 2023). Thus, we defined accuracy or precision $< 70\%$ as "low", while accuracy or precision $\geq 70\%$ was classified as "high".

2.6 | The Impact of Taxonomy Assignment Workflows on Downstream Ecological Analysis

The impact of different reference databases and taxonomy assignment methods on community composition and richness was explored in RStudio (R Core Team 2023). We generated barplots with qiime2R (Bisanz 2018) and ggplot2 (Wickham 2009) of the taxonomic groups at the order and family levels. For genus-level taxonomy barplots, only the seven most abundant genera for each reference database tested were plotted due to the high number of genera represented in our *in silico* dataset. The rest were grouped together as "other". Genus richness of each method and reference database combination was estimated. A one-sample Wilcoxon test was conducted to test whether the estimated richness for each reference database differed from the known diversity of each *in silico* dataset (marine = 48 genera; terrestrial = 30 genera).

3 | Results

3.1 | Assessing Nematode Representation and Integrity of the SILVA Reference Databases

As of 2024, the SILVA 138 REF database (Database 1) consists of 172,240 18S rRNA sequences, 5781 belonging to the phylum Nematoda (Figure 3; Figure S2). Among nematode sequences, 3435 (59.42%) belong to Rhabditida—a large order containing mostly terrestrial nematodes and some animal and plant parasitic clades (i.e., Strongylida, Tylenchomorpha, etc.). Trichinellida—an order composed of vertebrate parasitic nematodes—is represented by 667 sequences (11.53%). The least represented orders in the SILVA database (≤ 3 sequences for each order) are Dioctophymatida (animal parasite), Desmoscolecida (free-living marine), Muspiceida (animal parasite), and Isolaimida (free-living terrestrial). The SILVA NR99 clustered database reduced the number of nematode sequences to 2174, with the vast majority of sequences (e.g., 1351 sequences; 62.14% of the nematode sequences in the SILVA NR99 database) belonging to Rhabditida (49.82%) and Trichinellida (11.09%). The total number of sequences belonging to poorly represented orders in the SILVA NR99 database (i.e., Mermithida [16 sequences; 0.76%], Mononchida [19 sequences; 0.87%], and Chromadorida [34 sequences; 1.56%]) was also reduced (Figure 3; Figure S2).

Despite the phylogeny tree-guided curation of the taxonomy in the SILVA database, errors are still common (Figure 2). For example, in the SILVA NR99 database, 34 sequences (1.61%) from various non-nematode phyla were incorrectly

merged with a nematode string, resulting in chimeric taxonomy strings (Table S2). The source of this error is most likely SILVA's automated curation pipeline, which "corrects" the GenBank-derived taxonomy hierarchy according to where the sequence is placed in a guide tree phylogeny, where the true identity of the nucleotide sequences is misclassified in the GenBank record (e.g., a parasite sequence co-amplified during PCR that is incorrectly deposited as a DNA barcode for the host organism). Of the 2174 nematode sequences, all of them had some kind of taxonomic error; however, 36.2% of the nematode sequences in the SILVA NR99 database are considered major taxonomic errors (i.e., either a chimeric string or an outdated classification). Of the nematode 18S rRNA sequences, 566 (26.03%) are classified as Tylenchida, currently accepted as an infraorder (Tylenchomorpha) within Rhabditida. Additionally, SILVA NR99 incorrectly places three nematode families (e.g., Plectidae, Leptolaimidae, and Camacolaimidae, a total of 23 sequences or 1.06%) in the order Araeolaimida instead of the currently accepted Plectida.

3.2 | Evaluating the Accuracy of Taxonomy Assignments Using the SILVA REF Database as Reference (Database 1)

The SILVA REF reference database resulted in low accuracy across every taxonomic level (i.e., order, family, and genus; Figures 4 and 5). For the marine nematode dataset, the BLAST.1.90 method had the highest accuracy at the order (57.47%), family (74.21%), and genus (53.39%) levels (Tables S4–S6). Using the QIIME2 Naïve Bayes methods, most sequences remained unassigned at the family and genus ranks (66.06%–86.43%; Figure 4; Table S5). At the genus level, BLAST.10.95 had the lowest accuracy (5.43%; Table S6).

For the terrestrial nematode dataset, all methods yielded similar results at the order rank ($a = 55.56\%–58.12\%$, $p = 58.26\%–59.63\%$)–with the exception of the stringent BLAST.1.99 ($a = 24.79\%$ and $p = 44.62\%$; Table S4). Assigning taxonomy to family and genus levels results in low accuracy across all methods and parameters (Table S6). Fewer than 59% of the terrestrial nematode sequences were accurately assigned at the family level (11.11%–58.97%; Table S5). BLAST.1.90 had the highest accuracy at the family (58.97%) and genus (39.32%) levels. The QIIME2 Naïve Bayes methods consistently had low accuracy (11.11%–16.24%) but high precision (100%; Figure 4; Table S5).

3.3 | Evaluating the Accuracy of Taxonomy Assignments Using the SILVA NR99 Database as Reference (Database 2)

Clustering the SILVA database did not significantly improve the accuracy or precision at any taxonomic level. SILVA NR99 performed similarly to the SILVA REF using the different taxonomy assignment methods (Figure 5). BLAST.1.90 had the highest accuracy at the order (57.47%), family (60.63%), and genus (36.20%) levels (Tables S4–S6). The stringent BLAST.1.99 performed the worst at the order and family ranks. At the genus level, BLAST.10.95 and BLAST.3.90, resulted in 189 (85.52%) and 176 (79.64%) false negatives, respectively (i.e., unassigned sequences; Figure 4; Table S6).

Similarly, BLAST.1.99 performed the worst for order-level taxonomy assignments of the terrestrial *in silico* dataset ($a = 23.93\%$, $p = 45.16\%$), whereas the QIIME2 Naïve Bayes methods resulted in only half of the sequences being correctly assigned taxonomy ($a = 53.85\%–56.41\%$, $p = 58.41\%–63\%$). At the family and genus levels, BLAST.1.90 had the highest accuracy (Figures 4 and 5; Tables S5 and S6). At the genus level, all the QIIME2 Naïve Bayes methods failed to assign taxonomy to more than 16 terrestrial nematode sequences (i.e., 13.68% of the terrestrial dataset; Figure 4). Compared to the taxonomy assignment at the family level, the number of false positives at the genus level increased drastically.

3.4 | Evaluating the Effect of Taxonomy String Curation on Taxonomy Assignment Accuracy and Precision (Database 3)

Improving the database taxonomy strings (Taxa String QC) had the most positive impact on accurately assigning taxonomy at the order and family levels. For the marine *in silico* dataset, taxonomy string curation had a positive effect on the QIIME2 Naïve Bayes methods, with the Naïve.70 method having the highest accuracy and precision at the order ($a = 79.19\%$, $p = 95.63\%$), family ($a = 79.19\%$, $p = 95.63\%$) and genus ($a = 36.65\%$, $p = 72.32\%$) levels (Tables S4–S6). The BLAST.1.99 method produced the worst results at the order ($a = 6.79\%$, $p = 88.24\%$), family ($a = 5.43\%$, $p = 80\%$), and genus ($a = 4.52\%$, $p = 66.67\%$) levels (Tables S4–S6). Regardless of the method, most genus-level assignments were false positives (2.26%–25.24%) or false negatives (38.46%–93.21%; Figure 4; Table S6).

For the terrestrial *in silico* dataset, the QIIME2 Naïve Bayes outperformed every method at the order ($a = 99.15\%$, $p = 100\%$) and family ($a = 70.09\%–76.07\%$, $p = 95.70\%–100\%$) levels (Figure 4; Tables S4 and S5). However, the reliability of genus-level taxonomic assignment with the QIIME2 Naïve Bayes methods drastically decreased ($a = 22.22\%–30.77\%$, $p = 81.82\%–96.30\%$; Tables S6). Similar to the marine *in silico* dataset, most genus-level assignments, regardless of the method, were either false positives (0.85%–22.22%) or false negatives (35.04%–84.62%; Table S6). BLAST.10.95 performed the worst ($a = 11.11\%$, $p = 72.22\%$; Table S6).

3.5 | Evaluating the Impact of Increasing Database Representation on the Accuracy of Taxonomy Assignment Methods (Database 4)

Including 567 more nematode sequences in the reference database increased the accuracy and reduced the number of unassigned sequences at every taxonomic level (Figures 4 and 5). BLAST.3.90 was the best-performing method assigning taxonomy at the order ($a = 95.48\%$, $p = 95.91\%$) and family ($a = 92.31\%$, $p = 94.88\%$) levels for the marine *in silico* dataset (Figure 5; Tables S4 and S5). Using the stringent BLAST.1.99 method, most sequences (79.19%) remained unassigned at both the family and genus levels (Figure 4; Tables S5 and S6). QIIME2 Naïve Bayes

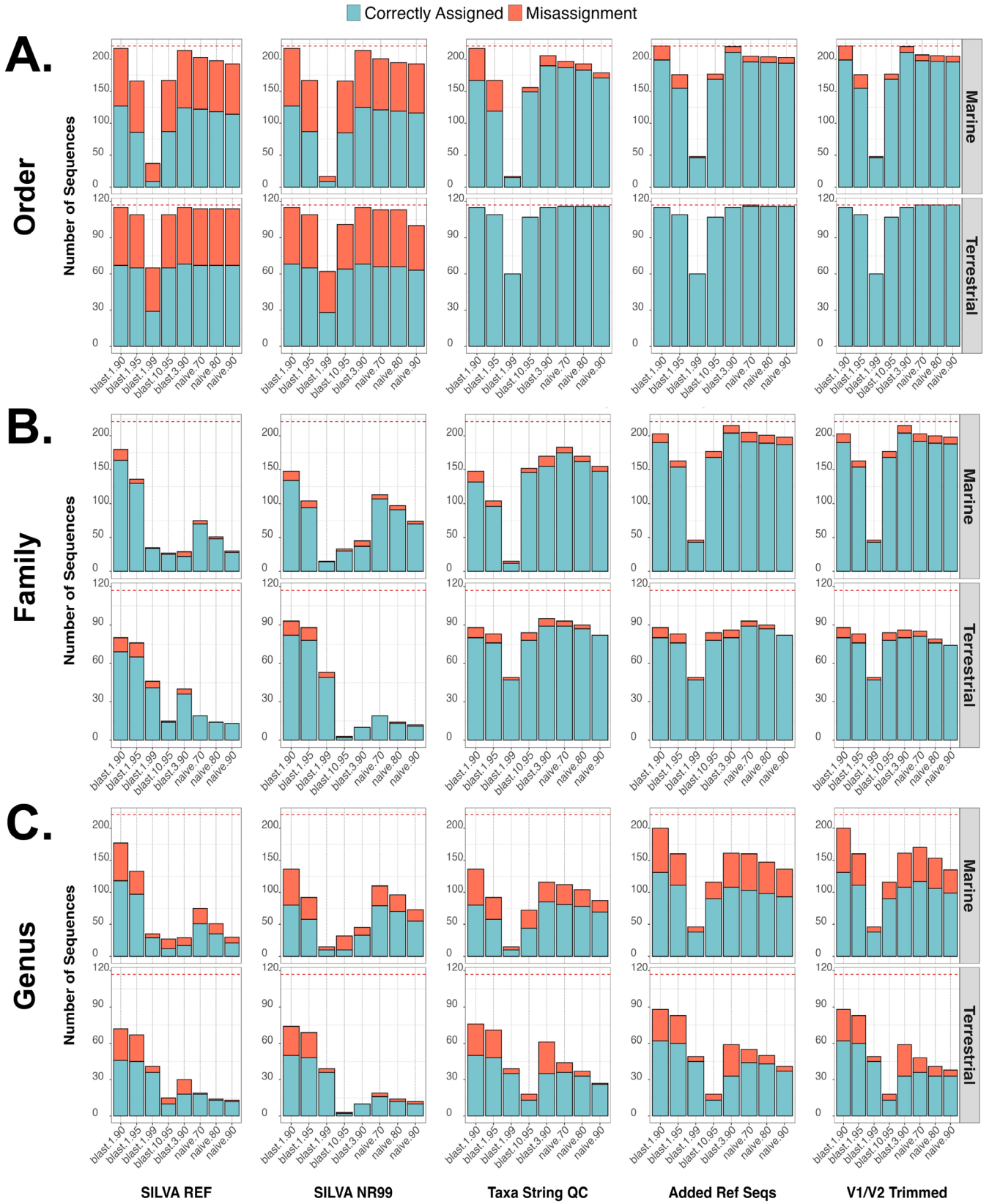

FIGURE 4 | Accuracy of nematode taxonomy assignments across databases, algorithms, and parameters. The number of correctly assigned (true positive) and misassigned reads (false positive) at the (A) order, (B) family, and (C) genus levels. The red bars are sequences that were misassigned at that particular taxonomic level. Blue bars indicate sequences that were correctly assigned to their known morphological ID (obtained via light microscopy). The red horizontal line is the number of sequences included in the *in silico* datasets (i.e., Marine: 221 total and 170 unique sequences; Terrestrial: 117 total and 96 unique sequences).

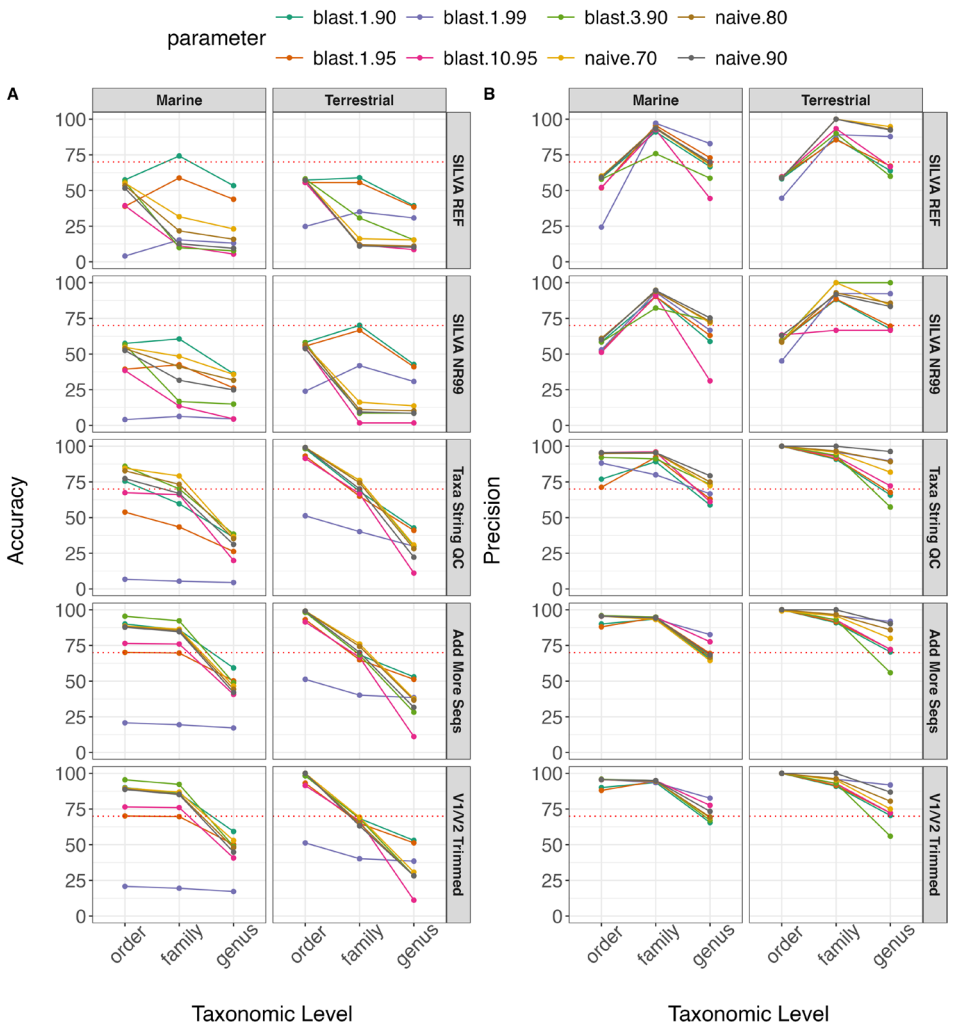

FIGURE 5 | Accuracy (A) and Precision (B) across five reference databases. Line color in each plot refers to different taxonomy assignment algorithms and parameter settings tested in the present study.

performed similarly across order and family-level taxonomy assignments (a = 84.62%–88.59%, p = 93.17%–95.65%).

For the terrestrial *in silico* dataset, QIIME2 Naïve Bayes accurately assigned order level taxonomy to 99.15% of the sequences (p = 99.15%–100%). However, the accuracy of the QIIME2 Naïve Bayes methods dropped when assigning taxonomy at the family (a = 70.09%–76.07%) and genus (a = 22.22%–30.77%) levels (Figure 5; Tables S5 and S6). The accuracy and precision of every taxonomy method drastically decreased at the genus level (Figure 5; Table S6). BLAST.1.90 and BLAST.1.95 were the best methods to assign genus-level taxonomy for both *in silico* datasets (a = 50.23%–59.28%, p = 65.50%–72.29%).

3.6 | Evaluating the Effect of Trimming on Taxonomy Assignment Accuracy and Precision (Database 5)

The accuracy and precision of the V1/V2 Trimmed Reference Database varied between the marine and terrestrial *in silico* datasets (Figures 4 and 5). However, it still outperformed the default SILVA databases (e.g., SILVA REF and SILVA NR99). For the marine *in silico* dataset, database trimming only slightly improved the accuracy and precision of taxonomy assignments. Naïve.70 performed the best at the family (a = 86.88%, p = 94.58%) level, while BLAST.1.90 performed the best at the genus level (a = 59.28%, p = 65.50%). Increasing the confidence interval of QIIME2 Naïve Bayes increased the number of false negatives at the order level (from 6.33% to 7.24%; Table S4). At the family level, the stringent BLAST.1.99 resulted in the vast majority of sequences remaining unassigned (a = 19.46%, p = 93.48%; Table S5).

For the terrestrial *in silico* dataset, Database 5 (Trimmed V1/V2) performed equally well at the order level compared to Database 4 (Added More Seq) and Database 3 (Fixed Taxonomy Strings; Figures 4 and 5). However, at the family and genus levels, the accuracy and precision of taxonomy assignment decreased compared to Database 4 (Added More Seq), which contained full-length reference sequences. At the family level, every tested method had varying levels of accuracy (40.17%–69.23%) but all had high levels of precision (90.91%–100%; Table S5). At the genus level, the BLAST.1.90 (a = 52.99%, p = 70.45%) method outperformed the QIIME2 Naïve Bayes methods (a = 28.21%–30.77%, p = 75.00%–86.84%); however, BLAST.1.90 also had the highest number of false positives (22.22%; Table S6).

3.7 | Evaluating the Effect of Closing Phylogenetic Gaps in the Reference Database (Database 6)

As expected, by closing phylogenetic gaps in reference databases (by including our *in silico* 18S rRNA sequences in Database 6), we drastically improved the accuracy and precision of taxonomy assignments across all taxonomic methods and levels. Aside from the BLAST.10.95 method, the accuracy and precision for every taxonomy assignment method were high (> 70%) for both *in silico* datasets (Figure S3B). For the marine *in silico* dataset, the BLAST+ top hit methods performed equally well at the order ($a = 98.19\%$, $p = 98.19\%$), family ($a = 98.64\%$, $p = 98.64\%$), and genus levels ($a = 90.05\%$; $p = 90.05\%$; Figure S3A; Tables S4–S6). The Naïve Bayes methods performed similarly at the order and family levels; however, at the genus level, the Naïve.70 performed the best ($a = 87.33\%$, $p = 89.35\%$; Figure S3).

For the terrestrial dataset, the BLAST+ LCA methods performed the worst at the family ($a = 73.50\%$–91.45%; $p = 94.51\%$–97.27%) and genus levels ($a = 51.28\%$–77.78%, $p = 82.19\%$–88.35%; Figure S3A; Tables S4–S6). For the marine *in-silico* dataset, the Naïve Bayes methods behaved similarly, with the Naïve.70 performing comparably to the BLAST+ methods at the family ($a = 98.29\%$, $p = 99.14\%$) and genus levels (94.87%, $p = 97.37\%$).

3.8 | Effect of Reference Databases and Taxonomy Assignment Methods on Downstream Ecological Analysis

The genus richness for both the terrestrial and marine *in silico* datasets differed from the true richness regardless of the database or method used (Figure S4). For the marine *in silico* dataset (true richness = 48 genera), the Added Ref Seq reference dataset results in the highest genus richness across every method used (median = 25.5); however, the genus richness significantly differed from the true value (p-value = 0.008). For the terrestrial *in silico* dataset (true richness = 30 genera), Database 3 (Taxa String QC) had the highest genus richness among most methods (except for BLAST.10.95 and BLAST.3.90) but was still significantly lower than the true richness (p-value = 0.015). For both *in silico* datasets, BLAST.1.90 resulted in the highest genus richness, regardless of the reference dataset used to assign taxonomy. The BLAST.1.90 performed best in predicting the genus richness of marine (richness = 36 genera) and terrestrial (richness = 26 genera) *in silico* datasets.

We further assessed the impact of the various reference databases on nematode community composition at different taxonomic ranks. For marine nematodes, the original SILVA databases (i.e., Databases 1 and 2) overestimate specific taxa, such as Monhysterida, and underestimate Araeolaimida and Plectida (Figure S5). However, the results for terrestrial nematodes were slightly different. For instance, no Araeolaimida terrestrial nematodes were included in the *in silico* dataset, but using the SILVA REF and SILVA NR99 reference databases resulted in ~15% of the sequences being assigned to Araeolaimida. Assigning taxonomy with QIIME2 Naïve Bayes and the BLAST.3.90 with Database 4 (Added Ref Seq) or Database 5 (V1/V2 Trimmed) as reference yielded the most accurate results at the family level (Figure S6) for both *in silico* datasets. Meanwhile, most methods (except for BLAST.1.90) with the SILVA REF (Database 1) and SILVA NR99 (Database 2) resulted in > 25% of unassigned sequences. For the terrestrial *in silico* dataset, the QIIME2 Naïve Bayes methods with the original SILVA databases (i.e., Databases 1 and 2) resulted in > 75% unassigned sequences at the family level. At the genus level, using the BLAST+ LCA or QIIME2 Naïve Bayes methods resulted in the majority of terrestrial nematode sequences remaining unassigned (Figure S7).

4 | Discussion

4.1 | Status of 18S rRNA Reference Databases for Phylum Nematoda

The availability of reference sequences in the SILVA database across the phylum Nematoda is a reflection of lopsided global research efforts (mainly driven by biomedical and agricultural interests), and unfortunately does not encapsulate the estimated diversity of this speciose metazoan group (Smythe et al. 2019; Ahmed and Holovachov 2021; Ahmed et al. 2022). As expected, nematode 18S rRNA reference databases are dominated by terrestrial and parasitic lineages, with an especially high representation of sequences within the clade containing the model species *C. elegans* and closely related species (Rhabditida; Figure 3). Rhabditid nematodes have four times as many reference sequences compared to the second most represented clade Trichinellida, which includes important parasites of pigs and other vertebrates that commonly cause foodborne illness in humans (Anderson 2000). Dorylaimida, a diverse group of free-living soil and plant-parasitic nematodes, is the third most well-represented clade in the SILVA database, followed by free-living marine nematodes in the early-branching clade Enoplida (Bik et al. 2010). Many free-living nematode lineages, across diverse marine ecosystems (e.g., estuaries, sandy beaches, deep sea, etc.), continue to be poorly represented in SILVA, especially for groups known to contain high marine biodiversity such as Chromadorida, Desmodorida, Monhysterida, and Araeolaimida (Ahmed et al. 2022; Moens et al. 2013; Schratzberger et al. 2019). Poor or lack of representation of typical deep-sea genera (e.g., *Manganonema* and *Acantholaimus*; Fonseca et al. 2006; Miljutina and Miljutin 2011) reflects the taxonomic gaps in molecular databases regarding nematode diversity and habitat representation, and precludes robust identifications of these groups in marine eDNA studies. Desmoscolecida, an important group of stout nematodes with body rings that encase themselves in sediment using mucus secretions, has a cosmopolitan distribution and high species diversity in diverse marine ecosystems (Decraemer 1997; Hwang et al. 2009; Vanreusel et al. 2010; Decraemer and Rho 2013) especially in deep-sea habitats (25% of the species were described from the deep sea; Decraemer and Rho 2013). However, only three representative sequences are available in the SILVA database (Schmidt-Rhaesa 2013; Fernandez-Leborans et al. 2017; Gattoni et al. 2023) despite a collection of 18S rRNA sequences recently deposited in Genbank (Pereira et al. 2020).

In terms of public database resources for nematodes, marine and terrestrial habitats have starkly different patterns of coverage across taxonomic groups. At the family level, marine clades are evenly (albeit shallowly) sampled, with at least a

handful of public DNA barcodes existing for most of the described marine lineages. In contrast, terrestrial nematode families are either very well sampled with ≥100 sequenced representatives (Rhabditidae, Trichinellidae, Diplogasteridae, Aphelenchoididae) or have minimal representation with a large number of families absent from public molecular databases (Figure S2). For example, within the terrestrial subclass Dorylaimia, approximately one-third of nematode families completely lack any DNA barcoded representative in SILVA.

Curated reference databases such as SILVA implement stringent quality control protocols to maintain high-quality and nearly full-length sequences (Ratnasingham and Hebert 2007; Quast et al. 2013). Quality control protocols (i.e., minimum sequence quality and sequence length) largely represent a trade-off between nucleotide quality and the overall representation of each taxon in the database. While stringent quality control protocols allow for the creation of a high-quality reference database, they can severely limit taxonomic coverage to fewer well-studied species and reduce the effectiveness of the database in metabarcoding studies. Currently, SILVA quality control protocols require eukaryotic sequences to be at least 1200 bp in length, have an alignment quality ≥50, and an alignment identity ≥70. Many marine nematode genera (e.g., *Pselionema, Oxyonchus, Parodontophora, Acantholaimus*) are absent in the SILVA database because their GenBank records only contain short (~300–400 bp) and partial fragments (~800 bp) of the 18S rRNA gene, which prevents them from being incorporated into SILVA and thus negatively impacts downstream eDNA taxonomy assignments. Macheriotou et al. (2019) suggested that incorporating these short and partial 18S rRNA sequences into the SILVA database has a positive impact on the accuracy of taxonomy assignment of marine nematodes. Additionally, SILVA's last major update was in 2020 (Release 138), and so any full-length 18S rRNA sequences produced in the last 4 years (e.g., from recent phylogenomic studies; Ahmed and Holovachov 2021; Ahmed et al. 2022) have not been integrated into the reference database. More frequent SILVA releases would also help to further improve eDNA taxonomy assignments, since this would allow new full-length 18S rRNA sequences to be rapidly incorporated into this popular public resource for eDNA bioinformatics workflows.

4.2 | Importance of Reference Databases on Ecological Data Analysis

These contrasting patterns of database coverage between marine (shallow but even) and terrestrial nematodes (a largely bimodal distribution) have important repercussions for the performance of taxonomy assignment algorithms (discussed further below). For eDNA studies and biomonitoring applications, the paucity (or even complete absence) of molecular data across lower taxonomic levels will continue to hinder the use of cutting-edge -Omics tools for rapid environmental assessments and management (Macheriotou et al. 2019; Tytgat et al. 2019; Gold et al. 2021; Pantó et al. 2021). Several macro- and meiofaunal indexes, such as GAMBI (genetics-based AZTI's Marine Biotic Index; Aylagas et al. 2014), SPEAR (SPEcies At Risk; Liess and Von Der Ohe 2005), and nemaSPEAR (a nematode-specific SPEAR index; Brüchner-Hüttemann et al. 2021), have been developed to assess ecosystem health and monitor pollution levels. However, biotic indexes and habitat assessments are often contingent on reliable family- and genus-level identifications of faunal assemblages (Jones 2008; Aylagas et al. 2014). For nematodes, lineages across the phylum exhibit different tolerances to organic and chemical pollutants (Ekschmitt and Korthals 2006). Certain opportunistic marine genera, such as *Sabatieria*, can tolerate anoxic conditions and high sulfide concentrations and are typically used as a proxy for disturbance and environmental stress (Franco et al. 2008; Moreno et al. 2008; Soto et al. 2017). Other genera (e.g., *Enoplus*) contain a mixture of sensitive and persistent species. For example, *E. communis*, a true marine species, is more sensitive to chemical pollution than the euryhaline *E. brevis* (Howell 1984; Vranken et al. 1991). Additionally, some studies have indicated that in soil habitats, fungal-feeding nematodes are less susceptible to heavy metal contaminants than bacterial feeders and predatory nematodes (Chauvin et al. 2020; Jiang et al. 2023). In order to harness information about such "bioindicator" species in metabarcoding studies, we need to ensure that (1) the target genera/species are well represented in public reference databases and (2) our bioinformatic taxonomy assignment pipelines are able to accurately report the presence and relative abundance of these sentinel species.

From a genomic perspective, microbial metazoan groups are among the most neglected lineages with the greatest taxonomic gaps across public sequence databases (Bik et al. 2012; Keeling and Campo 2017; Ahmed and Holovachov 2021; Ahmed et al. 2022). Bacterial and archeal databases have historically benefitted from coordinated genome sequencing efforts aimed at capturing the breadth of phylogenetic diversity, such as the "Genomic Encyclopedia of Bacteria and Archaea" project, which initially aimed to sequence 250 new prokaryote genomes and was later extended to target 1250 genomes (Kyrpides et al. 2014; Whitman et al. 2015; Hug et al. 2016; Mukherjee et al. 2017). This coordinated sequencing effort systematically expanded the phylogenetic diversity of prokaryote reference genomes by 25% and specifically targeted non-clinically relevant organisms (Mukherjee et al. 2017). For example, the MMESTP project (Keeling et al. 2014) has created high-quality transcriptomes for 40 major lineages (i.e., Ascomycetes, Ciliates, Choanoflagellates, etc.), including those that did not previously have a reference genome. However, these genomic projects have not included microbial metazoan groups. More recently, the Earth Biogenome Project (EBP; https://goat.genomehubs.org/projects/EBP) has begun a decadal initiative to generate genome sequences for all eukaryotic life (Lewin et al. 2022), but the EBP genome projects currently underway continue to be largely biased towards arthropods and chordates. Genome projects for microbial metazoans continue to be hindered by the low biomass of individual organisms, insufficient HMW DNA, and high cryptic species diversity, which prevents the pooling of individuals before sequencing.

Marine ecosystems are a reservoir for cryptic diversity, with 11%–43% of marine eukaryotes estimated to be cryptic species (Appeltans et al. 2012). Fewer than 10% of the estimated diversity of marine eukaryotic taxa with high amounts of cryptic diversity (i.e., copepods, nematodes, etc.) have been described (Derycke et al. 2010, 2016; Leray and Knowlton 2016). Linking genomic data (or at minimum, species-diagnostic DNA barcodes such as full-length 18S rRNA gene) with morphological data is

fundamental for delineating species and uncovering the hidden diversity of marine metazoans. Recent phylogenomic studies of nematodes and other meiofauna have expanded our knowledge of deep evolutionary relationships (Smythe et al. 2019; Ahmed and Holovachov 2021; Ahmed et al. 2022); however, taxonomic sampling remains limited compared to the known morphological diversity within these groups. Our results indicate that more phylogenetically representative sampling is urgently needed across microbial metazoan groups (especially nematodes), even for commonly used metabarcoding markers such as the 18S rRNA gene, where public database resources continue to be sparsely populated.

4.3 | Performance of BLAST Versus QIIME2 Feature Classifier

Overall, BLAST+ and the QIIME2 Naïve Bayes classifier returned similar taxonomy assignment results (Figures 4 and 5), with the exception of Blast top hit at 99% (BLAST.1.99) sequence similarity, which performed consistently poorly across all databases, *in silico* datasets, and taxonomic levels. This was presumably due to the lack of closely related species or sister taxa in the SILVA database for most free-living nematodes (Figure 3; Figure S2). Additionally, low 18S rRNA sequence divergence among nematode taxa and possible taxonomic misidentifications of closely related taxa in the *in silico* datasets and reference databases may contribute to the high false positive rate seen in this study. For example, in the terrestrial *in silico* dataset, the V1-V2 region of the 18S rRNA sequence belonging to *Aporcelaimellus* (Aporcelaimidae) has a high sequence similarity (> 99%) with other Dorylaimida taxa, including *Dorylaimoides* (Mydonomidae) and *Ecumenicus* (Qudsianematidae), all from distinct families. Low sequence divergence among different nematode families can impact the accuracy of taxonomy assignment methods, particularly if using the BLAST+ LCA methods. Furthermore, BLAST.1.99 performed much worse for Sanger barcodes from marine habitats across order, family, and genus levels since the handful of sequences for any given marine lineage was likely to be distantly related to the Sanger barcodes in our *in silico* dataset. Unsurprisingly, our expanded reference database containing additional nematode lineages (Added Ref Seqs) led to an increased number of correct taxonomy assignments for marine nematodes across all taxonomic ranks (Figure 4), suggesting that the targeted generation of new DNA reference sequences (e.g., full-length and partial fragments of the 18S rRNA gene obtained via Sanger sequencing) can quickly improve metabarcoding taxonomic assignments for groups with historically poor taxon sampling. Similarly, Hleap et al. (2021) showed that taxonomy assignment methods are more robust when used in conjunction with larger and more phylogenetically diverse reference databases. In contrast, the additional marine nematode reference sequences also benefitted terrestrial nematodes at the genus level; however, they did not improve the higher nematode ranks commonly found in terrestrial habitats. Therefore, habitat-specific 18S rRNA nematode sequences are also needed to improve taxonomy assignment methods for terrestrial nematodes at higher taxonomic ranks. Specific soil nematode databases have recently been created, such as NemaTaxa and NemaBase (Baker et al. 2023; Gattoni et al. 2023), in order to overcome the persistent issues with lower-level taxonomic identification of nematodes when using the SILVA database. However, such customized databases are still biased towards terrestrial and parasitic species and are unlikely to be broadly utilized by researchers working across eDNA fields.

For eukaryotic metabarcoding studies, trimming reference databases down to the target locus (~350 bp of the V1/V2 gene region of 18S rRNA) slightly improves genus-level taxonomy assignment for marine nematodes; however, it does not appear to offer any benefits for improving taxonomy assignments for terrestrial nematodes (Figure 4). Marine nematode Sanger barcodes assigned at the order and family levels were virtually equivalent to the untrimmed database. For terrestrial nematodes, the trimmed database has reduced the accuracy of the taxonomy assignments for the family and genus levels. Previous studies recommending locus-specific database trimming were focused on bacterial/archaeal taxonomy and 16S rRNA reference databases (Werner et al. 2012). Interestingly, studies focused on eukaryotic taxonomy and reference databases have yielded conflicting results. Macheriotou et al. (2019) suggested that trimming 18S rRNA reference databases can improve the accuracy of taxonomy assignment algorithms. However, Robeson et al. (2021) found that trimming the BOLD databases (COI mitochondrial marker) to a locus-specific region results in fewer unique sequences and reduces the amount of taxonomic information of marine metazoa, thus further reducing the number of lineages represented in the database. Our findings suggest that the effectiveness of this step is dependent on the database structure (gene marker, length of reference sequences, and database representation) and has variable results across different nematode lineages. The V1/V2 trimmed database performed better than the SILVA REF, SILVA NR99, and "Added More Seq" databases; however, the accuracy and precision of taxonomy assignments for the terrestrial nematodes were reduced when compared to the "Added More Seq" database. While some nematode lineages had more misassigned (i.e., *Acrobeles* and *Chromadorita*) or unassigned (i.e., *Plectus*) sequences when using the V1/V2 trimmed database than the "Added More Seq" database, other lineages (i.e., *Sabatieria*) were improved when they were assigned taxonomy using the V1/V2 Trimmed Database (Figure 6; Figure S8).

We also found that the composition and curation of reference databases had unpredictable and variable effects on the accuracy of low-level taxonomy assignments. When examining genus-level assignment patterns for BLAST top hit at 90% (Figure 6) and the QIIME2 Naïve Bayes classifier (Figure S8), no single method or database was able to maximize taxonomy assignment accuracy across a phylogenetically diverse subset of marine and terrestrial nematode lineages. For example, the addition of a single reference sequence in the genus *Plectus* resulted in 100% of *Plectus* Sanger barcodes being correctly assigned by BLAST+ top hit at 90% identity (compared to less than a quarter correctly assigned in the previous "Taxa String QC" database), but there was no change in taxonomy assignment accuracy using the QIIME2 Naïve Bayes classifier with the reference database trimmed to the primer region (where three-quarters of sequences remained unassigned in the "Taxa String QC"). In another case, clustering of the SILVA database into 99% non-redundant reference sequences notably improved the accuracy of *Sabatieria* taxonomy assignments using the QIIME2 Naïve Bayes classifier ("SILVA REF" versus "SILVA NR99" databases in Figure S8),

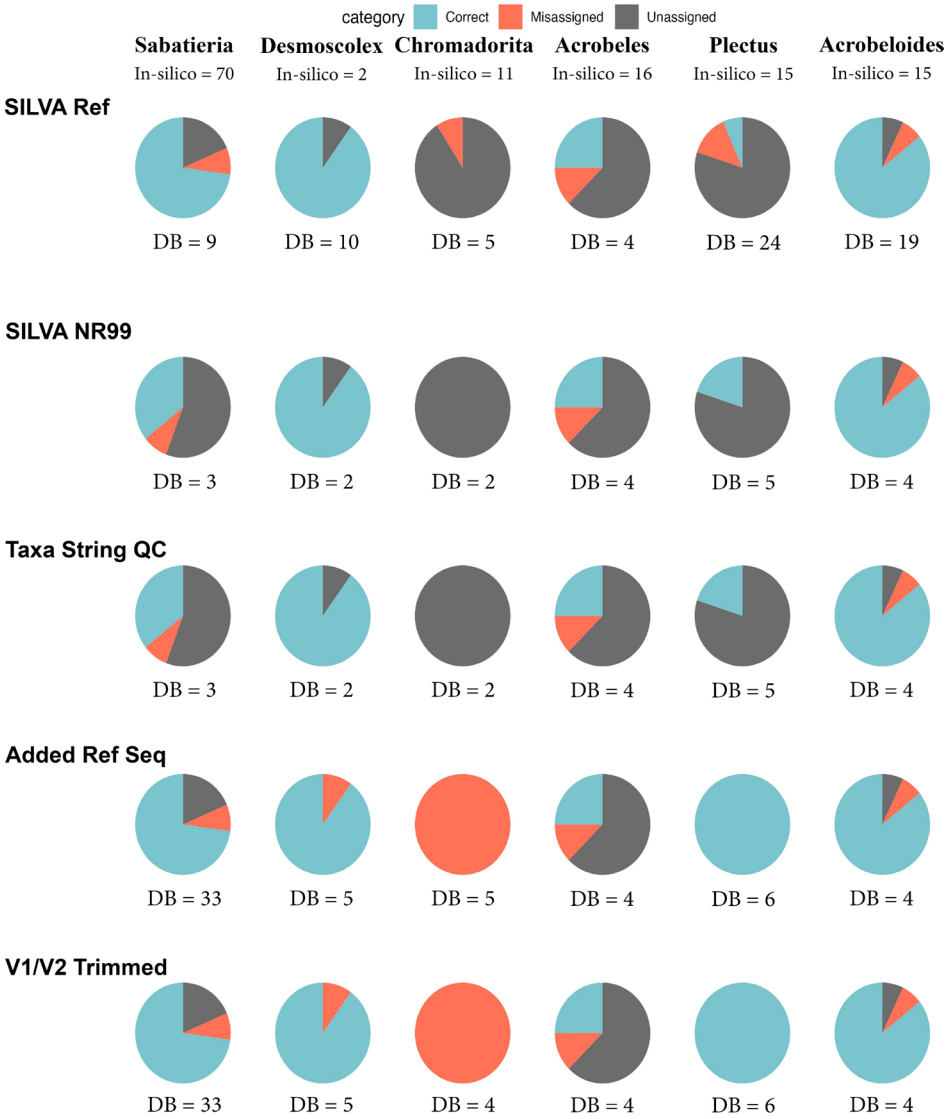

FIGURE 6 | Taxonomy assignment results across databases for five abundant nematode genera using BLAST+ top hit. Results are shown for the top three well-sampled marine (*Sabatieria, Desmoscolex, Chomadorita*) and terrestrial/freshwater (*Acrobeles, Plectus, Acrobeloides*) nematode genera. All taxonomy assignments were carried out using the BLAST top hit with a cutoff of 90% sequence similarity. Results for each genus are displayed as a vertical column, with "DB" below each pie chart representing the number of sequences from that genus in each reference database. The total number of sequence representatives from each genus in the *in silico* dataset is listed at the top.

but had the opposite effect using BLAST+ top hit at 90% identity (converting correctly assigned Sanger barcodes into misassignments; "SILVA REF" versus "SILVA NR99" databases in Figure 6). Notably, when we observed improvements in the number of correct taxonomy assignments, it was not necessarily contingent on the total number of reference sequences available in public databases. For example, the QIIME2 Naïve Bayes classifier was able to accurately assign taxonomy for two-thirds of 70 *Sabatieria* Sanger barcodes using only 3 SILVA reference sequences (SILVA NR99 in Figure S8). Similar results were observed for *Desmoscolex*, *Plectus*, and *Acrobeloides*. BLAST top hit at 90% identity (Figure 6), accurately assigned taxonomy to Sanger barcodes that are represented by a few sequences in the SILVA reference database. Our results suggest that the phylogenetic distance between Sanger barcodes and reference database lineages is one of the most important factors for obtaining accurate metabarcoding taxonomy assignments. As exemplified by Database 6 (Figure S3), closing phylogenetic gaps led to a drastic improvement in the accuracy of taxonomy assignments at every taxonomic level. Despite having an exact sequence match in Database 6, sequences are occasionally misassigned across taxonomy assignment methods. Algorithmic choices can have a slight impact on the accuracy of taxonomy assignments. For example, BLAST+ top hit assigns taxonomy using the first reference sequence it encounters that exceeds the sequence similarity threshold. Neither the overall number of database sequences nor bioinformatics parameters can overcome the real limitations of sparse phylogenetic sampling.

When choosing BLAST+ cutoffs and parameters for 18S rRNA eukaryotic metabarcoding data, our data suggests that using BLAST+ top hit with a less stringent cutoff of 90% often returns the highest percentage of Sanger barcodes with correctly assigned taxonomy. Using more stringent sequence similarity

cutoffs (95% or 99%) and/or incorporating the top 3 or 10 database hits appears to dramatically reduce the accuracy of BLAST-derived taxonomy assignments (at worst) or provides mostly equivalent performance to Blast top hit at 90% (at best; Figures 4 and 5). BLAST top hit at 90% and 95% stringency also exhibited overall higher performance than the LCA BLAST+ and QIIME2 Naïve Bayes methods for taxonomy assignments using the original SILVA REF and SILVA NR99 clustered databases, most likely due to conflicting taxonomy strings in both of these reference datasets. The QIIME2 Naïve Bayes classifier tended to outperform BLAST+ once the SILVA database was subjected to additional manual curation efforts (Taxa String QC in Figure 4, discussed further below), returning fewer misassignments than BLAST+ top hit methods even when the number of correctly assigned sequences were similar across the two methods. Misassignments at the genus level were high across both BLAST+ and QIIME2 Naïve Bayes methods, and QIIME2 Naïve Bayes was less accurate at this lowest taxonomic level, suggesting significant gaps in reference sequences for nematodes at the genus level (and a correspondingly smaller training set for the QIIME2 Naïve Bayes classifier which results in poorer performance).

4.4 | Database Curation Dramatically Improves Taxonomy Assignments

Putting aside the impact of database gaps, we found that error correction and manual curation of reference taxonomy strings (Figure 2) had the most obvious and compelling impact on the performance of metabarcoding taxonomy assignments (Figures 4 and 5). When we discarded low-quality SILVA sequences (exhibiting chimeric taxonomy strings) and ensured that taxonomic hierarchies adhering to standard Linnaean hierarchies were correctly aligned, we observed a drastic reduction in the number of misassignments at the order level and a notable increase in the number of correct assignments at the family and genus levels (Figure 4), with particular improvement seen in the QIIME2 Naïve Bayes results.

Machine learning methods (such as the QIIME2 Naïve Bayes classifier) have a training step that is reliant on accurate and consistently labeled taxonomy strings to correctly differentiate between distinct sequences in the dataset (Pedregosa et al. 2011). Inaccurate and/or conflicting strings affect the training step of these classifiers and negatively impact the accuracy of these methods. Although manually curating the taxonomy strings of a reference database is a time-intensive endeavor, it will notably improve the accuracy and precision of the taxonomy classifier (Figure 5). Similarly, BLAST+ LCA methods rely on consistent taxonomic strings to accurately summarize the top "*N*" hits in a reference database (Bokulich et al. 2018). For example, if a *Haliplectus* sp. sequence is assigned taxonomy using a top 2-hit LCA method (Figure 2C), the sequence will only be assigned to a higher-level classification (Chromadorea) due to an inconsistent labeling schema. Some bioinformatics pipelines, such as RESCRIPt (Robeson et al. 2021) and Meta-fish-lib (Collins et al. 2021), have been developed to automate the curation of reference sequence taxonomy strings; however, the developers of these much-needed workflows emphasize the need for manually checking the phylogenetic placement of molecular sequences. Furthermore, specialized expertise is required due to the historical taxonomic uncertainty of specific taxa and the frequent reclassification of meiofauna groups due to the continual expansion of phylogenetic studies.

In this study, the accuracy and precision of the QIIME2 Naïve Bayes classifier performed best with the most robust databases (Taxa String QC and Added Ref Seqs) compared to the SILVA REF and SILVA NR99 (Figure 5). This is in line with previous studies that have found a positive linear relationship between the accuracy of the QIIME2 Naïve Bayes algorithm and the size of the reference database (Richardson et al. 2017), exemplifying the need for increasing reference databases of undersampled nematode taxa. Additionally, our findings support previous studies that showed that the QIIME2 Naïve Bayes classifier with the most stringent confidence interval (90%) might decrease the error rate but also increase the proportion of unassigned sequences. Metabarcoding studies using databases with large gaps and without well-curated taxonomic strings should implement sequence similarity methods (BLAST+) using a top hit approach with a 90% cutoff and avoid using machine learning methods, such as the Naive Bayes classifier, when assigning taxonomy. In addition, assigning taxonomy for both BLAST+ and Naïve Bayes methods can be greatly improved by curating taxonomy strings, and more intensive curation efforts of these taxonomic hierarchies should ideally occur within public database resources (e.g., SILVA). Furthermore, we strongly emphasize that genus- and species-level taxonomy should be interpreted with care, as large taxonomic gaps in reference databases lead to an astonishing number of unassigned sequences or sequences assigned to incorrect lineages (and misassignment at lower taxonomic levels can occasionally occur even when exact database matches are present).

5 | Conclusions

The choice of metabarcoding locus has been shown to significantly influence the performance of taxonomy assignment algorithms (Richardson et al. 2017), suggesting that bioinformatics workflows should be evaluated independently for the genetic marker(s), reference database, and specific taxonomic group(s) being investigated in any given study. Here, we evaluated the status of public nematode reference databases and assessed the accuracy and precision of two commonly used taxonomy methods (BLAST+ and QIIME2 Naïve Bayes) on marine and terrestrial nematodes using various parameter combinations with six iterations of the SILVA database. Many free-living nematode lineages are still poorly represented in the SILVA database, impacting the ability to deduce ecologically meaningful (genus-level) taxonomy assignments for marine and soil nematodes. The publicly available SILVA taxonomy strings (in SILVA REF and SILVA NR99) are often conflicting and outdated, which hinders accurate taxonomy assignments across all taxonomic levels. Increasing the representation of diverse nematode taxa by incorporating publicly available short-length 18S rRNA sequences (250–700 bp) increases the ability to accurately assign taxonomy at the family and genus levels. Finally, trimming a full-length curated database down to the V1/V2 primer region (Database 5) results in minor taxonomy assignment improvements for marine nematodes but slightly reduces the accuracy

of assigning taxonomy to terrestrial nematodes. Our findings highlight the need for renewed efforts to produce reference DNA barcodes (linked to morphological identifications to the genus or species levels) that adequately sample the biodiversity of microbial metazoans, as well as more rapid incorporation of newly published sequences into public curated databases such as SILVA. We emphasize that nucleotide sequences and taxonomy hierarchies in well-regarded curated reference databases, such as SILVA, need to be subjected to higher quality controls to improve eDNA taxonomy assignments. Furthermore, communication between database curators and taxonomic experts is indispensable for improving the organization and completeness of Linnaean taxonomic hierarchies associated with each reference sequence.

Author Contributions

Alejandro De Santiago, Tiago José Pereira, and Holly M. Bik conceived the ideas and designed the study methodology. Alejandro De Santiago collated and analyzed the data. Alejandro De Santiago, Tiago José Pereira, and Holly M. Bik led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Acknowledgments

Funding for this study was provided by the North Pacific Research Board (NPRB project 1303), The Gulf of Mexico Research Initiative, and institutional startup funding from the University of Georgia to H.M.B. We also thank the Center for Conservation Biology at the University of California, Riverside, and The Shipley-Skinner Reserve—Riverside County Endowment for partially funding this project. Part of this work was funded by a NERC Post-Genomics and Proteomics Grant (Ref NE/F001266/1) and New Investigator Grant NE/E001505/1 to S.C. Research support for A.D.S. was provided by the University of Georgia Research Foundation and the National Institute of General Medical Sciences of the National Institute of Health under award number 1T32GM142623. We also acknowledge support from The Thames Estuary partnership, The Mersey Basin Campaign, Dr. Kerry Walsh, Prof. P. John D. Lambshead, and colleagues at the UK Environment Agency and Bangor University School of Ocean Sciences for boat support. This work was also supported by a National Science Foundation CAREER award to HMB (DEB- 2144304).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Full-length 18S rRNA sequences of marine (Accession Numbers: PV244086-PV244363) nematodes generated via Sanger sequencing in this study are deposited in Genbank. The soil and marine nematode *in silico* datasets, six databases, and all scripts used for processing and analyzing the data are available on GitHub (https://github.com/alejandrodesantiago/taxonomy_benchmarking). GOMRI Data are publicly available through the Gulf of Mexico Research Initiative Information & Data Cooperative (GRIIDC) at https://data.gulfresearchinitiative.org (doi: 10.7266/N7959G4X).

References

Ahmed, M., and O. Holovachov. 2021. "Twenty Years After De Ley and Blaxter—How Far Did we Progress in Understanding the Phylogeny of the Phylum Nematoda?" *Animals* 11, no. 12: 3479. https://doi.org/10.3390/ani11123479.

Ahmed, M., N. G. Roberts, F. Adediran, A. B. Smythe, K. M. Kocot, and O. Holovachov. 2022. "Phylogenomic Analysis of the Phylum Nematoda: Conflicts and Congruences With Morphology, 18s rRNA, and Mitogenomes." *Frontiers in Ecology and Evolution* 9: 2021. https://doi.org/10.3389/fevo.2021.769565.

Almeida, A., A. L. Mitchell, A. Tarkowska, and R. D. Finn. 2018. "Benchmarking Taxonomic Assignments Based on 16s rRNA Gene Profiling of the Microbiota From Commonly Sampled Environments." *GigaScience* 7, no. 5: giy054. https://doi.org/10.1093/gigascience/giy054.

Altschul, S. F., T. L. Madden, A. A. Schäffer, et al. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25, no. 17: 3389–3402.

Anderson, R. C. 2000. *Nematode Parasites of Vertebrates: Their Development and Transmission*. CABI.

Appeltans, W., S. T. Ahyong, G. Anderson, et al. 2012. "The Magnitude of Global Marine Species Diversity." *Current Biology: CB* 22, no. 23: 2189–2202.

Aylagas, E., A. Borja, and N. Rodríguez-Ezpeleta. 2014. "Environmental Status Assessment Using DNA Metabarcoding: Towards a Genetics Based Marine Biotic Index (gAMBI)." *PLoS One* 9, no. 3: e90529.

Baker, H. V., J. R. Ibarra Caballero, C. Gleason, et al. 2023. "NemaTaxa: A New Taxonomic Database for Analysis of Nematode Community Data." *Phytobiomes Journal* 7, no. 3: 385–391.

Barbera, P., A. M. Kozlov, L. Czech, et al. 2019. "EPA-Ng: Massively Parallel Evolutionary Placement of Genetic Sequences." *Systematic Biology* 68, no. 2: 365–369.

Bik, H. M., P. J. D. Lambshead, W. K. Thomas, and D. H. Lunt. 2010. "Moving Towards a Complete Molecular Framework of the Phylum Nematoda: A Focus on the Enoplida and Early-Branching Clades." *BMC Evolutionary Biology* 10: 353.

Bik, H. M., D. L. Porazinska, S. Creer, J. G. Caporaso, R. Knight, and W. K. Thomas. 2012. "Sequencing Our Way Towards Understanding Global Eukaryotic Biodiversity." *Trends in Ecology & Evolution* 27, no. 4: 233–243.

Bisanz, J. E. 2018. "Qiime2r: Importing qiime2 artifacts and Associated Data into r Sessions. Version 0.99." https://github.com/jbisanz/qiime2R.

Blaxter, M. 2016. "Imagining Sisyphus Happy: DNA Barcoding and the Unnamed Majority." *Philosophical Transactions of the Royal Society, B: Biological Sciences* 371, no. 1702: 20150329. https://doi.org/10.1098/rstb.2015.0329.

Blaxter, M., P. De Ley, J. R. Garey, et al. 1998. "A Molecular Evolutionary Framework for the Phylum Nematoda." *Nature* 392, no. 6671: 71–75. https://doi.org/10.1038/32160.

Bokulich, N. A., B. D. Kaehler, J. R. Rideout, et al. 2018. "Optimizing Taxonomic Classification of Marker-Gene Amplicon Sequences With QIIME 2's q2-Feature-Classifier Plugin." *Microbiome* 6, no. 1: 90.

Bolyen, E., J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. Abnet, and G. A. Al-Ghalith. 2018. "QIIME 2: Reproducible, Interactive, Scalable, and Extensible Microbiome Data Science." (No e27295v2). PeerJ. https://doi.org/10.7287/peerj.preprints.27295v2.

Bongers, T. 1990. "The Maturity Index: An Ecological Measure of Environmental Disturbance Based on Nematode Species Composition." *Oecologia* 83, no. 1: 14–19.

Bongers, T., R. Alkemade, and G. W. Yeates. 1991. "Interpretation of Disturbance-Induced Maturity Decrease in Marine Nematode Assemblages by Means of the Maturity Index." *Marine Ecology Progress Series* 76: 135–142.

Bourret, A., C. Nozères, E. Parent, and G. J. Parent. 2023. "Maximizing the Reliability and the Number of Species Assignments in Metabarcoding Studies Using a Curated Regional Library and a Public Repository." *Metabarcoding and Metagenomics* 7: e98539.

Brüchner-Hüttemann, H., S. Höss, C. Ptatscheck, M. Brinke, J. Schenk, and W. Traunspurger. 2021. "Added Value of the NemaSPEAR[%]-Index to Routinely Used Macrofauna-Based Indices for Assessing the Quality of Freshwater Sediments." *Ecological Indicators* 121: 107015.

Camacho, C., G. Coulouris, V. Avagyan, et al. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10: 421.

Chauvin, C., M. Trambolho, M. Hedde, et al. 2020. "Soil Nematodes as Indicators of Heavy Metal Pollution: A Meta-Analysis." *Open Journal of Soil Science* 10, no. 12: 579–601.

Chitwood, D. J. 2003. "Research on Plant-Parasitic Nematode Biology Conducted by the United States Department of Agriculture-Agricultural Research Service." *Pest Management Science* 59, no. 6–7: 748–753.

Collins, R. A., G. Trauzzi, K. M. Maltby, et al. 2021. "Meta-Fish-Lib: A Generalised, Dynamic DNA Reference Library Pipeline for Metabarcoding of Fishes." *Journal of Fish Biology* 99, no. 4: 1446–1454.

Creer, S., V. G. Fonseca, D. L. Porazinska, et al. 2010. "Ultrasequencing of the Meiofaunal Biosphere: Practice, Pitfalls and Promises." *Molecular Ecology* 19, no. Suppl 1: 4–20.

De Ley, P. 2006. "A Quick Tour of Nematode Diversity and the Backbone of Nematode Phylogeny." *WormBook: The Online Review of C. Elegans Biology*: 1–8.

De Ley, P., I. T. De Ley, K. Morris, et al. 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. Series B, Biological Sciences* 360, no. 1462: 1945–1958.

De Santiago, A., T. J. Pereira, S. L. Mincks, and H. M. Bik. 2022. "Dataset Complexity Impacts Both MOTU Delimitation and Biodiversity Estimates in Eukaryotic 18S rRNA Metabarcoding Studies." *Environmental DNA* 4, no. 2: 363–384.

Decraemer, W. 1997. "Origin and Evolution of the Desmoscolecida, an Aberrant Group of Nematodes." *Journal of Zoological Systematics and Evolutionary Research = Zeitschrift Fur Zoologische Systematik Und Evolutionsforschung* 15, no. 3: 232–236.

Decraemer, W. I., and H. S. Rho. 2013. "7.11 Order Desmoscolecida." In *Volume 2 Nematoda*, 351–372. De Gruyter.

Deiner, K., H. M. Bik, E. Mächler, et al. 2017. "Environmental DNA Metabarcoding: Transforming How We Survey Animal and Plant Communities." *Molecular Ecology* 26, no. 21: 5872–5895.

Dell'Anno, A., L. Carugati, C. Corinaldesi, G. Riccioni, and R. Danovaro. 2015. "Unveiling the Biodiversity of Deep-Sea Nematodes Through Metabarcoding: Are We Ready to Bypass the Classical Taxonomy?" *PLoS One* 10, no. 12: e0144928.

Derycke, S., P. De Ley, I. Tandingan Ley, O. Holovachov, A. Rigaux, and T. Moens. 2010. "Linking DNA Sequences to Morphology: Cryptic Diversity and Population Genetic Structure in the Marine Nematode *Thoracostoma trachygaster* (Nematoda, Leptosomatidae)." *Zoologica Scripta* 39, no. 3: 276–289.

Derycke, S., N. De Meester, A. Rigaux, et al. 2016. "Coexisting Cryptic Species of the Litoditis Marina Complex (Nematoda) Show Differential Resource Use and Have Distinct Microbiomes With High Intraspecific Variability." *Molecular Ecology* 25, no. 9: 2093–2110.

Ekschmitt, K., and G. W. Korthals. 2006. "Nematodes as Sentinels of Heavy Metals and Organic Toxicants in the Soil." *Journal of Nematology* 38, no. 1: 13–19.

Fernandez-Leborans, G., S. Román, and D. Martin. 2017. "Erratum to: A New Deep-Sea Suctorian-Nematode Epibiosis (Loricophrya-Tricoma) From the Blanes Submarine Canyon (NW Mediterranean)." *Microbial Ecology* 74, no. 4: 1009.

Fonseca, G., W. Decraemer, A. Vanreusel, and A. Wegener. 2006. "Taxonomy and Species Distribution of the Genus Manganonema Bussau, 1993 (Nematoda: Monhysterida)." https://www.vliz.be/imisdocs/publications/289384.pdf.

Fonseca, V. G., G. R. Carvalho, B. Nichols, et al. 2014. "Metagenetic Analysis of Patterns of Distribution and Diversity of Marine Meiobenthic Eukaryotes." *Global Ecology and Biogeography: A Journal of Macroecology* 23, no. 11: 1293–1302.

Fonseca, V. G., B. Nichols, D. Lallias, et al. 2012. "Sample Richness and Genetic Diversity as Drivers of Chimera Formation in nSSU Metagenetic Analyses." *Nucleic Acids Research* 40, no. 9: e66.

Franco, M. A., M. Steyaert, H. N. Cabral, et al. 2008. "Impact of Discards of Beam Trawl Fishing on the Nematode Community From the Tagus Estuary (Portugal)." *Marine Pollution Bulletin* 56, no. 10: 1728–1736.

Gardner, P. P., R. J. Watson, X. C. Morgan, et al. 2019. "Identifying Accurate Metagenome and Amplicon Software via a Meta-Analysis of Sequence to Taxonomy Benchmarking Studies." *PeerJ* 7: e6160.

Gattoni, K., E. M. S. Gendron, R. Sandoval-Ruiz, et al. 2023. "18S-NemaBase: Curated 18S rRNA Database of Nematode Sequences." *Journal of Nematology* 55, no. 1: 20230006.

Gold, Z., E. E. Curd, K. D. Goodwin, et al. 2021. "Improving Metabarcoding Taxonomic Assignment: A Case Study of Fishes in a Large Marine Ecosystem." *Molecular Ecology Resources* 21, no. 7: 2546–2564.

Hleap, J. S., J. E. Littlefair, D. Steinke, P. D. N. Hebert, and M. E. Cristescu. 2021. "Assessment of Current Taxonomic Assignment Strategies for Metabarcoding Eukaryotes." *Molecular Ecology Resources* 21, no. 7: 2190–2203.

Holman, L. E., M. de Bruyn, S. Creer, G. Carvalho, J. Robidart, and M. Rius. 2021. "Animals, Protists and Bacteria Share Marine Biogeographic Patterns." *Nature Ecology & Evolution* 5, no. 6: 738–746.

Holovachov, O., Q. Haenel, S. J. Bourlat, and U. Jondelius. 2017. "Taxonomy Assignment Approach Determines the Efficiency of Identification of OTUs in Marine Nematodes." *Royal Society Open Science* 4, no. 8: 170315.

Howell, R. 1984. "Acute Toxicity of Heavy Metals to Two Species of Marine Nematodes." *Marine Environmental Research* 11, no. 3: 153–161.

Hu, S. K., A. R. Smith, R. E. Anderson, et al. 2022. "Globally-Distributed Microbial Eukaryotes Exhibit Endemism at Deep-Sea Hydrothermal Vents." *Molecular Ecology* 32, no. 23: 6580–6598. https://doi.org/10.1111/mec.16745.

Hug, L. A., B. J. Baker, K. Anantharaman, et al. 2016. "A New View of the Tree of Life." *Nature Microbiology* 1: 16048.

Hugot, J.-P., P. Baujard, and S. Morand. 2001. "Biodiversity in Helminths and Nematodes as a Field of Study: An Overview." *Nematology: International Journal of Fundamental and Applied Nematological Research* 3, no. 3: 199–208.

Hwang, U. W., E. H. Choi, D. S. Kim, W. Decraemer, and C. Y. Chang. 2009. "Monophyly of the Family Desmoscolecidae (Nematoda, Demoscolecida) and Its Phylogenetic Position Inferred From 18S rDNA Sequences." *Molecules and Cells* 27, no. 5: 515–523.

Jiang, R., M. Wang, and W. Chen. 2023. "Heavy Metal Pollution Triggers a Shift From Bacteria-Based to Fungi-Based Soil Micro-Food Web: Evidence From an Abandoned Mining-Smelting Area." *Journal of Hazardous Materials* 459: 132164.

Jones, F. C. 2008. "Taxonomic Sufficiency: The Influence of Taxonomic Resolution on Freshwater Bioassessments Using Benthic Macroinvertebrates." *Environmental Review* 16: 45–69.

Katoh, K., and D. M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30, no. 4: 772–780.

Keeling, P. J., F. Burki, H. M. Wilcox, et al. 2014. "The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating

the Functional Diversity of Eukaryotic Life in the Oceans Through Transcriptome Sequencing." *PLoS Biology* 12, no. 6: e1001889.

Keeling, P. J., and J. D. Campo. 2017. "Marine Protists Are Not Just Big Bacteria." *Current Biology: CB* 27, no. 11: R541–R549.

Kyrpides, N. C., T. Woyke, J. A. Eisen, et al. 2014. "Genomic Encyclopedia of Type Strains, Phase I: The One Thousand Microbial Genomes (KMG-I) Project." *Standards in Genomic Sciences* 9, no. 3: 1278–1284.

Lallias, D., J. G. Hiddink, V. G. Fonseca, et al. 2015. "Environmental Metabarcoding Reveals Heterogeneous Drivers of Microbial Eukaryote Diversity in Contrasting Estuarine Ecosystems." *ISME Journal* 9, no. 5: 1208–1221.

Lambshead, P. J. D. 1993. "Recent Developments in Marine Benthic Biodiversity Research." *Oceanis* 19: 5–24.

Leray, M., and N. Knowlton. 2016. "Censusing Marine Eukaryotic Diversity in the Twenty-First Century." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 371, no. 1702: 20150331. https://doi.org/10.1098/rstb.2015.0331.

Lewin, H. A., S. Richards, E. Lieberman Aiden, et al. 2022. "The Earth BioGenome Project 2020: Starting the Clock." *Proceedings of the National Academy of Sciences of the United States of America* 119, no. 4: e2115635118. https://doi.org/10.1073/pnas.2115635118.

Liess, M., and P. C. Von Der Ohe. 2005. "Analyzing Effects of Pesticides on Invertebrate Communities in Streams." *Environmental Toxicology and Chemistry / SETAC* 24, no. 4: 954–965.

Macheriotou, L., K. Guilini, T. N. Bezerra, et al. 2019. "Metabarcoding Free-Living Marine Nematodes Using Curated 18S and CO1 Reference Sequence Databases for Species-Level Taxonomic Assignments." *Ecology and Evolution* 9, no. 3: 1211–1226.

Mathon, L., A. Valentini, P.-E. Guérin, et al. 2021. "Benchmarking Bioinformatic Tools for Fast and Accurate eDNA Metabarcoding Species Identification." *Molecular Ecology Resources* 21, no. 7: 2565–2579.

Matsen, F. A., R. B. Kodner, and E. V. Armbrust. 2010. "Pplacer: Linear Time Maximum-Likelihood and Bayesian Phylogenetic Placement of Sequences Onto a Fixed Reference Tree." *BMC Bioinformatics* 11: 538.

Miljutin, D. M., G. Gad, M. M. Miljutina, V. O. Mokievsky, V. Fonseca-Genevois, and A. M. Esteves. 2010. "The State of Knowledge on Deep-Sea Nematode Taxonomy: How Many Valid Species Are Known Down There?" *Marine Biodiversity: A Journal of the Senckenberg Research Institute / Senckenberg Forschungsinstitut Und Naturmudeum* 40, no. 3: 143–159.

Miljutina, M. A., and D. M. Miljutin. 2011. "Seven New and Four Known Species of the Genus Acantholaimus (Nematoda: Chromadoridae) From the Abyssal Manganese Nodule Field (Clarion-Clipperton Fracture Zone, North-Eastern Tropical Pacific)." *Helgoland Marine Research* 66, no. 3: 413–462.

Moens, T., U. Braeckman, S. Derycke, et al. 2013. "Ecology of Free-Living Marine Nematodes." In *Volume 2 Nematoda*, 109–152. De Gruyter.

Moens, T., U. Braeckman, S. Derycke, et al. 2014. "Ecology of Free-Living Marine Nematodes." In *Handbook of Zoology, Volume 2: Nematoda*, edited by A. Schmidt-Rhaesa, 109–159. De Gruyter.

Mokievsky, V., and A. Azovsky. 2002. "Re-Evaluation of Species Diversity Patterns of Free-Living Marine Nematodes." *Marine Ecology Progress Series* 238: 101–108.

Moreno, M., L. Vezzulli, V. Marin, P. Laconi, G. Albertelli, and M. Fabiano. 2008. "The Use of Meiofauna Diversity as an Indicator of Pollution in Harbours." *ICES Journal of Marine Science: Journal Du Conseil* 65, no. 8: 1428–1435.

Mukherjee, S., R. Seshadri, N. J. Varghese, et al. 2017. "1,003 Reference Genomes of Bacterial and Archaeal Isolates Expand Coverage of the Tree of Life." *Nature Biotechnology* 35, no. 7: 676–683.

Nemys Editorial Board. 2024. "Nemys, World Database of Nematodes." https://doi.org/10.48580/dglq4-4rf.

O'Rourke, D. R., N. A. Bokulich, M. A. Jusino, M. D. MacManes, and J. T. Foster. 2020. "A Total Crapshoot? Evaluating Bioinformatic Decisions in Animal Diet Metabarcoding Analyses." *Ecology and Evolution* 10, no. 18: 9721–9739.

Pantó, G., F. Pasotti, L. Macheriotou, and A. Vanreusel. 2021. "Combining Traditional Taxonomy and Metabarcoding: Assemblage Structure of Nematodes in the Shelf Sediments of the Eastern Antarctic Peninsula." *Frontiers in Marine Science* 8: 2021. https://doi.org/10.3389/fmars.2021.629706.

Pedregosa, F., G. Varoquaux, A. Gramfort, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830.

Pereira, T. J., A. De Santiago, and H. M. Bik. 2024. "Soil Properties Predict Below-Ground Community Structure, but Not Nematode Microbiome Patterns in Semi-Arid Habitats." *Molecular Ecology* 33, no. 18: e17501. https://doi.org/10.1111/mec.17501.

Pereira, T. J., A. De Santiago, T. Schuelke, S. M. Hardy, and H. M. Bik. 2020. "The Impact of Intragenomic rRNA Variation on Metabarcoding-Derived Diversity Estimates: A Case Study From Marine Nematodes." *Environmental DNA* 44: 97.

Quast, C., E. Pruesse, P. Yilmaz, et al. 2013. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41, no. Database issue: D590–D596.

R Core Team. 2023. "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing. https://www.R-project.org/.

Ratnasingham, S., and P. D. N. Hebert. 2007. "Bold: The Barcode of Life Data System." *Molecular Ecology Notes* 7, no. 3: 355–364. http://www.barcodinglife.org.

Richardson, R. T., J. Bengtsson-Palme, and R. M. Johnson. 2017. "Evaluating and Optimizing the Performance of Software Commonly Used for the Taxonomic Classification of DNA Metabarcoding Sequence Data." *Molecular Ecology Resources* 17, no. 4: 760–769.

Ritari, J., J. Salojärvi, L. Lahti, and W. M. de Vos. 2015. "Improved Taxonomic Assignment of Human Intestinal 16S rRNA Sequences by a Dedicated Reference Database." *BMC Genomics* 16: 1056.

Robeson, M. S., D. R. O'Rourke, B. D. Kaehler, et al. 2021. "RESCRIPt: Reproducible sequence taxonomy reference database management." *PLoS Computational Biology* 17, no. 11: e1009581.

Schmidt-Rhaesa, A. 2013. *Handbook of Zoology/ Handbuch Der Zoologie. Handbook of Zoology. Gastrotricha, Cycloneuralia and Gnathifera: Volume 2: Nematoda*. De Gruyter.

Schratzberger, M., M. Holterman, D. van Oevelen, and J. Helder. 2019. "A Worm's World: Ecological Flexibility Pays Off for Free-Living Nematodes in Sediments and Soils." *Bioscience* 69, no. 11: 867–876.

Siegwald, L., H. Touzet, Y. Lemoine, D. Hot, C. Audebert, and S. Caboche. 2017. "Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics." *PLoS One* 12, no. 1: e0169563.

Sinniger, F., J. Pawlowski, S. Harii, et al. 2016. "Worldwide Analysis of Sedimentary DNA Reveals Major Gaps in Taxonomic Knowledge of Deep-Sea Benthos." *Frontiers in Marine Science* 3: 2016. https://doi.org/10.3389/fmars.2016.00092.

Smythe, A. B., O. Holovachov, and K. M. Kocot. 2019. "Improved Phylogenomic Sampling of Free-Living Nematodes Enhances Resolution of Higher-Level Nematode Phylogeny." *BMC Evolutionary Biology* 19, no. 1: 121.

Somervuo, P., S. Koskela, J. Pennanen, R. Henrik Nilsson, and O. Ovaskainen. 2016. "Unbiased Probabilistic Taxonomic Classification for DNA Barcoding." *Bioinformatics* 32, no. 19: 2920–2927.

Soto, L. A., D. L. Salcedo, K. Arvizu, and A. V. Botello. 2017. "Interannual Patterns of the Large Free-Living Nematode Assemblages in the Mexican Exclusive Economic Zone, NW Gulf of Mexico After the Deepwater Horizon Oil Spill." *Ecological Indicators* 79: 371–381.

Stamatakis, A. 2014. "RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies." *Bioinformatics (Oxford, England)* 30, no. 9: 1312–1313.

Thompson, L. R., J. G. Sanders, D. McDonald, et al. 2017. "A Communal Catalogue Reveals Earth's Multiscale Microbial Diversity." *Nature* 551, no. 7681: 457–463.

Tytgat, B., D. T. Nguyen, T. X. P. Nguyen, et al. 2019. "Monitoring of Marine Nematode Communities Through 18S rRNA Metabarcoding as a Sensitive Alternative to Morphology." *Ecological Indicators* 107: 105554.

van der Loos, L. M., and R. Nijland. 2021. "Biases in Bulk: DNA Metabarcoding of Marine Communities and the Methodology Involved." *Molecular Ecology* 30, no. 13: 3270–3288.

Vanreusel, A., G. Fonseca, R. Danovaro, et al. 2010. "The Contribution of Deep-Sea Macrohabitat Heterogeneity to Global Nematode Diversity." *Marine Ecology* 31, no. 1: 6–20.

Viglierchio, D. R., and R. V. Schmitt. 1983. "On the Methodology of Nematode Extraction From Field Samples: Baermann Funnel Modifications." *Journal of Nematology* 15, no. 3: 438–444.

Vranken, G., R. Vanderhaeghen, and C. Heip. 1991. "Effects of Pollutants on Life-History Parameters of the Marine Nematode *Monhystera disjuncta*." *ICES Journal of Marine Science: Journal Du Conseil* 48, no. 3: 325–334.

Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. "Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences Into the New Bacterial Taxonomy." *Applied and Environmental Microbiology* 73, no. 16: 5261–5267.

Weigand, H., A. J. Beermann, F. Čiampor, et al. 2019. "DNA Barcode Reference Libraries for the Monitoring of Aquatic Biota in Europe: Gap-Analysis and Recommendations for Future Work." *Science of the Total Environment* 678: 499–524.

Werner, J. J., O. Koren, P. Hugenholtz, et al. 2012. "Impact of Training Sets on Classification of High-Throughput Bacterial 16s rRNA Gene Surveys." *ISME Journal* 6, no. 1: 94–103.

Whitman, W. B., T. Woyke, H.-P. Klenk, Y. Zhou, T. G. Lilburn, and B. J. Beck. 2015. "Genomic Encyclopedia of Bacterial and Archaeal Type Strains, Phase III: The Genomes of Soil and Plant-Associated and Newly Described Type Strains." *Standards in Genomic Sciences* 10: 26.

Wickham, H. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. Springer Publishing Company, Incorporated.

WoRMS Editorial Board. 2023. "World Register of Marine Species (WoRMS)." https://www.marinespecies.org.

Yilmaz, P., L. W. Parfrey, P. Yarza, et al. 2013. "The SILVA and "All-Species Living Tree Project (LTP)" Taxonomic Frameworks." *Nucleic Acids Research* 42, no. D1: D643–D648.

Zeppilli, D., J. Sarrazin, D. Leduc, et al. 2015. "Is the Meiofauna a Good Indicator for Climate Change and Anthropogenic Impacts?" *Marine Biodiversity: A Journal of the Senckenberg Research Institute/ Senckenberg Forschungsinstitut Und Naturmudeum* 45, no. 3: 505–535.

Zhang, S., J. Zhao, and M. Yao. 2020. "A Comprehensive and Comparative Evaluation of Primers for Metabarcoding eDNA From Fish." *Methods in Ecology and Evolution* 11, no. 12: 1609–1625.

Zheng, Q., C. Bartow-McKenney, J. S. Meisel, and E. A. Grice. 2018. "HmmUFOtu: An HMM and Phylogenetic Placement Based Ultra-Fast Taxonomic Assignment and OTU Picking Tool for Microbiome Amplicon Sequencing Studies." *Genome Biology* 19, no. 1: 82.

Zhu, T., Y. Sato, T. Sado, M. Miya, and W. Iwasaki. 2023. "MitoFish, MitoAnnotator, and MiFish Pipeline: Updates in 10 Years." *Molecular Biology and Evolution* 40, no. 3: msad035. https://doi.org/10.1093/molbev/msad035.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.