Received: 28 August 2020 | Revised: 27 March 2022 | Accepted: 28 March 2022

DOI: 10.1002/edn3.303

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

WILEY

ORIGINAL ARTICLE

MIDORI2: A collection of quality controlled, preformatted, and regularly updated reference databases for taxonomic assignment of eukaryotic mitochondrial sequences

Matthieu Leray[1] | Nancy Knowlton[2] | Ryuji J. Machida[3]

[1]Smithsonian Tropical Research Institute, Balboa, Ancon, Panama

[2]National Museum of Natural History, Smithsonian Institution, Washington, District of Columbia, USA

[3]Biodiversity Research Centre, Academia Sinica, Taipei, Taiwan

Correspondence
Ryuji J. Machida, Biodiversity Research Centre, Academia Sinica, Taipei 11529, Taiwan.
Email: ryujimachida@gmail.com

Funding information
Gordon and Betty Moore Foundation, Grant/Award Number: GBMF5603; Scientific Committee on Oceanic Research; Academia Sinica; Ministry of Science and Technology, Taiwan, Grant/Award Number: 108-2611-M-001-001, 109-2611-M-001-003 and 110-2611-M-001-009

Abstract

Analysis of environmental DNA is increasingly used to characterize ecological communities, but the effectiveness of this approach depends on the accuracy of taxonomic reference databases. The MIDORI databases, first released in 2017, were built to improve accuracy for mitochondrial metazoan (animal) sequences. MIDORI has now been significantly improved and renamed MIDORI2 (available at http://www.reference-midori.info). Like MIDORI, MIDORI2 is built from GenBank and contains curated sequences of thirteen protein-coding and two ribosomal RNA mitochondrial genes. Coverage has been substantially expanded to cover all eukaryotes, including fungi, green algae and land plants, other multicellular algal groups, and diverse protist lineages. MIDORI2 also now includes not only species with full binomials, but also taxa referred to by genus with species left unspecified ("sp."). Another new feature is the updating of the databases approximately every two months with version numbers corresponding to each new GenBank release. Additional potentially erroneously annotated sequences have also been removed. Finally, the ability to export data files to BLAST+ has been added to the original ability to export preformatted data to five taxonomic assignment programs, and databases of amino acid sequences are also made available for protein-coding genes. As a technical validation, we conducted a preliminary comparison of the performance of MIDORI2 with five taxonomic assignment programs. Results suggest that BLAST+ top hits performed better for assigning *CO1* sequences than alignment-free methods based on compositional features. Comparing MIDORI2 with two other commonly used curated databases of mitochondrial sequences, CO-ARBitrator and BOLD, we show that MIDORI2 includes sequences from a broader range of metazoan and non-metazoan taxa. Overall, in many contexts, MIDORI2 offers clear advantages: a higher diversity of taxa than other databases, a variety of user-friendly features, and regular updates. MIDORI2 is particularly well-suited for environmental DNA studies that target mitochondrial genes with broad primers.

KEYWORDS
eukaryote, fungi, GenBank, metabarcoding, metazoan, mitochondrial genes, plants, protist

1 | INTRODUCTION

Accurate approaches for detecting and identifying species are essential to measure how communities and ecosystems respond to ongoing environmental changes (Blowes et al., 2019; Morecroft et al., 2019). Assessments of Bacteria and Archaea have long relied on molecular analyses of environmental DNA, but these approaches are now increasingly used for basic and applied studies of eukaryotic communities, to complement or replace traditional methods (Cahill et al., 2018; Leray & Knowlton, 2016; Machida et al., 2009; Ruppert et al., 2019; Thomsen & Willerslev, 2015). There are several reasons for this wider adoption of molecular methods for describing eukaryotic communities. Traditional methods that rely on collections, visual observations, or acoustics are invariably biased toward some taxonomic groups, and they are notably ineffective for small, difficult to identify, rare, and elusive taxa that make up the majority of life on Earth. They also often require large efforts with minimal return (i.e., comparatively few observations per unit of effort), can be environmentally destructive, are subject to observer biases, and pose challenges for comparisons across habitats and geographic regions [e.g., limited sample sizes, lack of standardization, insufficient detections of target species (Buckland & Johnston, 2017; Trimble & Aarde, 2012)]. Of course, molecular approaches for censusing species have their own biases and limitations (Aylagas et al., 2016; Kelly et al., 2019). Nevertheless, studies so far show that analyses of environmental DNA provide higher species detectability in a variety of environments (Bush et al., 2020; Machida et al., 2021; Nguyen et al., 2020; Thomsen & Sigsgaard, 2019), and new applications continue to be developed (e.g., Clare et al., 2022; Lynggaard et al., 2022).

Typically in these studies, extra-organismal DNA/RNA or samples containing multiple entire organisms are collected from the environment (eDNA/eRNA; Rodriguez-Ezpeleta et al., 2021). After the DNA/RNA is isolated, one or several variable marker genes are targeted using PCR amplification (metabarcoding) or the total gDNA/cDNA is sheared for shotgun metagenomics (metagenomics/metatranscriptomics). After sequencing on a high-throughput platform, the resulting raw reads are processed to differentiate spurious sequences from true biological variants (Antich et al., 2022; Callahan et al., 2017). These can then be clustered into groups [also referred to as molecular Operational Taxonomic Units (mOTUs) (Blaxter et al., 2005)] that best reflect species. Finally, each sequence variant or mOTU is compared to a set of reference sequences for taxonomic assignment.

The downstream interpretation of molecular surveys is highly dependent upon the completeness and accuracy of the reference databases used for taxonomic assignment. Taxonomically narrow or incomplete databases may lead to false negatives (i.e., the species is present in the sample but not reported because it is absent from the database). Mislabeled taxa in reference databases can lead to false positives and erroneous conclusions (e.g., positive detection of a species where it does not occur naturally). The research community has been well aware of these issues, and substantial efforts have been dedicated to develop taxonomically broad and error-free reference databases for the nuclear-encoded ribosomal 18S gene [e.g., Silva (Quast et al., 2013), PR2 (Guillou et al., 2013)], nuclear ribosomal internal transcribed spacer (ITS) regions [e.g., UNITE (Nilsson et al., 2019), PLANiTS (Banchi et al., 2020)], and chloroplast genes [e.g., (Rimet et al., 2019)].

Mitochondrial genes are increasingly targeted in DNA-based surveys of environmental DNA because they are highly variable between species in many taxonomic groups. Conserved primers are available for PCR-based surveys of selected metazoans, all metazoans, or all eukaryotes [e.g., Cytochrome c oxidase subunit 1 (*CO1*) (Elbrecht et al., 2019; Leray et al., 2013); Small rRNA (*srRNA*) (Machida et al., 2012); Large rRNA (*lrRNA*) (Kelly et al., 2016)]. Studies have also successfully used PCR-free mitochondrial metagenomics (mitogenomics) and metatranscriptomics to infer composition and relative biomass of complex mixed-species samples (Bista et al., 2018; Lopez et al., 2022; Machida et al., 2021).

In 2017, our group released MIDORI, a collection of quality-controlled reference databases of mitochondrial metazoan sequences, designed to facilitate and improve the accuracy of taxonomic assignments (Machida et al., 2017). MIDORI includes sequences of thirteen protein-coding genes (ATP synthase subunit 6 [*A6*] and 8 [*A8*], Cytochrome c oxidase subunit I [*CO1*], II [*CO2*] and III [*CO3*], Cytochrome b apoenzyme [*Cytb*], NADH dehydrogenase subunits 1–4 [*ND1–ND4*], 4L [*ND4L*], 5 [*ND5*] and 6 [*ND6*]), and two ribosomal RNA genes (Large and Small ribosomal subunit RNA) encoded in the mitochondrial genome. To build MIDORI, sequences were extracted from the GenBank BLAST NT database (Benson et al., 2013), entries with inaccurate annotations were removed, and the databases were preformatted for several commonly used taxonomic assignment programs [Mothur (Schloss et al., 2009), Qiime (Caporaso et al., 2010), RDP Classifier (Wang et al., 2007), Sintax (Edgar, 2016), and Spingo (Allard et al., 2015)]. Queries against MIDORI can be performed through a Webserver: http://reference-midori.info/server.php (Leray et al., 2018). Last updated in February 2018, the databases have until now remained limited to metazoan sequences.

Here, we describe a major update of the databases, MIDORI2 (http://www.reference-midori.info). First, we extended the taxonomic scope of the databases from metazoans to all eukaryote groups represented in GenBank, including fungi, green algae and land plants, other multicellular algal groups, and diverse protist lineages. Databases now also include both sequences with binomial identification and sequences without a species name ("sp."). Second, we increased the quality of the databases by removing sequences that we identified as potentially mislabeled in a previous study (Leray et al., 2019). Third, we classified sequences by gene based on the sequence similarity to the RefSeq complete mitochondrial genome sequences to avoid incorrect gene annotation. Fourth, we built a pipeline to automate updates of the databases approximately every 2 months following the release of each GenBank database. At each iteration, databases are formatted for six taxonomic assignment software programs, including for the first time BLAST+. Databases of amino acid sequences are also available for protein-coding genes.

To validate the databases and guide future users, we examined the performance of five taxonomic assignment programs to which MIDORI2 can export data [the sixth, SPINGO (Allard et al., 2015), was not tested because it performs taxonomic assignments only

at the species level]. Finally, we compared MIDORI2 to two other publicly available databases of curated mitochondrial sequences, the Barcode Of Life Database [BOLD (Ratnasingham & Hebert, 2007)] and CO-ARBitrator (Heller et al., 2018).

2 | MATERIAL AND METHODS

2.1 | Database construction

2.1.1 | Construction of gene classification reference databases

To construct MIDORI, we separated sequences into each gene based on text descriptions in GenBank flat files. However, the procedure has two disadvantages: There are potential mistakes in the text description of genes, and the procedure requires time-consuming manual checking of all text. To streamline the process in the construction of MIDORI2, we created the gene classification reference databases from RefSeq complete mitochondrial genome sequences.

For the preparation of the gene classification reference databases, we first downloaded RefSeq complete mitochondrial genome sequences (mitochondrion.1.1.genomic.fna.gz, mitochondrion.2.1.genomic.fna.gz, mitochondrion.1.genomic.gbff.gz, mitochondrion.2.genomic.gbff.gz) from the GenBank ftp site (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/). After extraction of "CDS" and "rRNA" positional information from the RefSeq flat files, sequences of the thirteen protein-coding (amino acid sequences) and two rRNA (nucleic acid sequences) regions were retrieved using E-utilities commands (EFetch; Sayers, 2010). Next, after the concatenation of those sequences, reciprocal BLAST searches (BLASTP for amino acid sequences and BLASTN for nucleic acid sequences) were performed for those fasta files. These reciprocal BLAST searches were required to remove the gene sequences with incorrect gene annotations in RefSeq. BLAST results with multiple gene hits with high significance (*e*-value lower than 1e-15) were manually checked and removed if their gene annotation was wrong. After the removal of erroneously annotated sequences, the fasta files were used as reference databases to classify the sequences into the 15 genes as described below.

2.1.2 | GenBank database preparation

First, the following domains of GenBank flat files were downloaded from the GenBank ftp site: ftp://ftp.ncbi.nlm.nih.gov/genbank/gbinv*; gbmam*; gbpln*; gbpri*; gbrod*; gbvrt*. The GenBank taxonomy file was also downloaded from the site: ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz. After concatenation of the downloaded GenBank flat files, eukaryote mitochondrial flat files were extracted from the total flat files. Next, using the positional information of the "CDS" and "rRNA," we created a multi-amino acid fasta file for protein-coding regions and a multi-nucleic acid fasta file for rRNA regions. After insertion of taxonomy headers using GenBank taxonomy, these multi-fasta files were used as query sequences in the following procedures.

2.1.3 | Classification of the sequences into genes using BLAST searches

BLASTP and BLASTN (Altschul et al., 1990; Camacho et al., 2009) were used for the classification of the protein-coding and ribosomal RNA-coding genes, respectively. We used BLASTP for protein-coding genes because of its higher sensitivity compared to BLASTN. Reference sequences for the BLAST searches were those created from complete mitochondrial RefSeqs. Query sequences for the BLAST were multi-fasta files created from GenBank flat files. The following options were used for the BLASTP: "-num_alignments 100 -word_size 3 -outfmt 7 -seg no -soft_masking false" and for the BLASTN: "-num_alignments 100 -word_size 11 -outfmt 7 -dust no -soft_masking false." The following criteria were used for the assignment of the genes: evalue <1.7E-05, bit score >42.4 for BLASTP; *e*-value <2.04E-09, bit score >63.9 for BLASTN.

2.1.4 | Formatting of the prepared reference sequences

In the original release of the databases (MIDORI), we removed sequences without binomial species descriptions, such as "cf.," "aff.," "sp.," "environment," "undescribed," "uncultured," "complex," "unclassified," "nom.," "nud." and "unidentif." Because we subsequently received several requests from users to include those sequences, we created two types of databases, with and without binomial species descriptions. Taxonomically mislabeled sequences identified in our previous study (Leray et al., 2019) were also removed from the databases before formatting files for use by taxonomic assignment software programs.

In all formats, we inserted GenBank taxonomy IDs after all taxonomic names. The reason for this insertion was to differentiate synonyms. For example, there are three taxa called Ctenophora: a phylum Ctenophora_10197, a genus of diatoms Ctenophora_1003038 and a genus of flies Ctenophora_516519. Insertion of the GenBank taxonomy ID makes each name unique, which is required for some taxonomic assignment software programs to run properly.

Length restrictions were applied to each nucleotide and amino acid sequence database as "*srRNA*: 100–2,000, *lrRNA*: 100–2,500, *A6*: 100–1,000, *A8*: 100–500, *CO1*: 100–2,000, *CO2*: 100–1,500, *CO3*: 100–1,300, *Cytb*: 100–1,500, *ND1*: 100–1,200, *ND2*: 100–1,500, *ND3*: 100–600, *ND4*: 100–2,000, *ND4L*: 100–700, *ND5*: 100–2,000, *ND6*: 100–1,500" and "*A6*: 32–334, *A8*: 32–167, *CO1*: 32–767, *CO2*: 32–500, *CO3*: 32–434, *Cytb*: 32–500, *ND1*: 32–400, *ND2*: 32–500, *ND3*: 32–200, *ND4*: 32–667, *ND4L*: 32–234, *ND5*: 32–667, *ND6*: 32–500," respectively.

In total, we formatted the reference databases for six taxonomic assignment software programs: Mothur (Schloss et al., 2009), Qiime (Caporaso et al., 2010), RDP Classifier (Wang et al., 2007), Sintax (Edgar, 2016), Spingo (Allard et al., 2015), and BLAST+ (Camacho et al.,

2009). Additionally, we also provide the original databases "RAW" with full taxonomy. Each format has two types of databases: "unique" contains all unique haplotypes associated with each species and "longest" contains a single sequence, the longest, for each species. "RAW" format also has "total," which contains all sequences. List files are listing accession numbers collapsed into each sequence both for unique and longest databases. Additionally, we provide Perl scripts, which users can use to extract target taxa from databases (all available at http://www.reference-midori.info/download.php). Because RDP Classifier cannot run with missing taxonomic ranks, we created missing taxonomy from a lower taxonomic ranking (e.g., description in class-level was missing, so it was created from order-level in the following example: >JF502242.1.7041.7724 root_1; Eukaryota_2759; Chordata_7711; class_Crocodylia_1294634; Crocodylia_1294634; Crocodylidae_8493; Crocodylus_8500; Crocodylus intermedius_184240).

The first version of MIDORI2 was built with GenBank release version 238 (downloaded in June 2020) and named MIDORI2 vGB238. Subsequent upgrades of MIDORI2 are identified with the corresponding GenBank release number.

2.2 | Technical validation

We confirmed that the databases were compatible with Mothur, QIIME, RDP Classifier, Sintax, Spingo, and BLAST+. A test of the accuracy of taxonomic assignments of some *CO1* sequences was built into the technical validation process. We removed 186 complete mitochondrial *CO1* gene sequences, one from each of the 186 taxonomic classes represented in MIDORI2, from the LONGEST_GB242 *CO1* database to use as queries.

Accuracy tests were conducted for three regions of the query sequences: (1) the complete mitochondrial *CO1*, (2) the ~658 bp *CO1* "barcode region" (Folmer et al., 1994), and (3) a ~313 bp fragment (Leray et al., 2013) within the barcode region. Before the trimming, the MAFFT server [options: --localpair--maxiterate (Katoh et al., 2019)] was used to align complete mitochondrial sequences. The query and reference databases are available on figshare (https://doi.org/10.6084/m9.figshare.19134722 and https://doi.org/10.6084/m9.figshare.19134704).

Five taxonomic assignment software programs were compared as follows: Qiime2 (parameters: feature-classifier classify-sklearn)

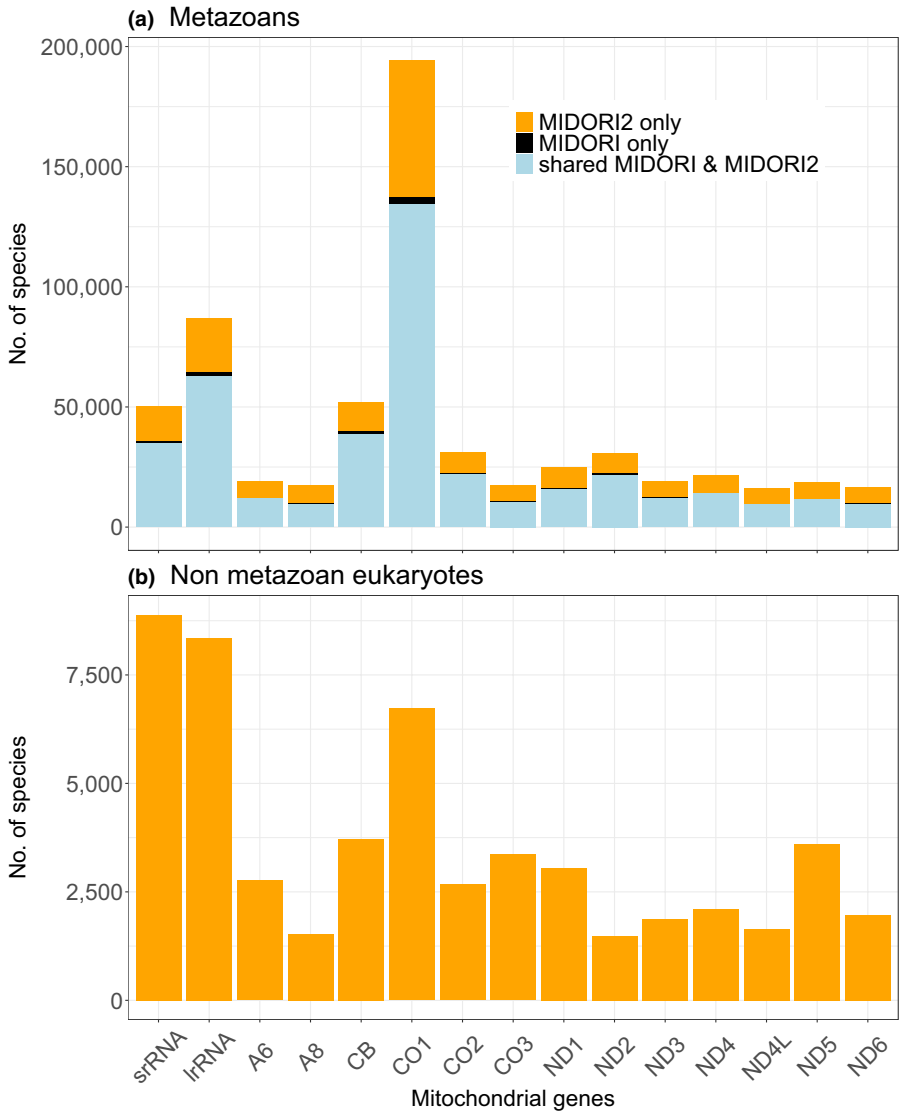

FIGURE 1 Composition of the MIDORI and MIDORI2 vGB244 databases. Number of species of (a) metazoans and (b) other eukaryotes represented in the original mitochondrial gene sequence reference databases of MIDORI (built from GenBank BLAST NT.fasta database downloaded in February 2018) and in the updated version MIDORI2 vGB244 (built from GenBank version 244, June 2021). The category "MIDORI only" identifies sequences for which the taxonomy was updated in GenBank between the release of MIDORI and the construction of MIDORI2 vGB244, or sequences that were flagged as potentially dubious by NCBI. The category "shared MIDORI & MIDORI2" indicates the number of species that are present in both databases

(Caporaso et al., 2010), Mothur (parameters: classify.seqs, cutoff = 0, ksize = 8) (Schloss et al., 2009), RDP Classifier (Wang et al., 2007), Sintax (parameters: -strand both, -sintax_cutoff 0.5) (Edgar, 2016), and BLAST+ (parameters: blastn -word_size 11 -max_target_seqs 100 -outfmt "7 sseqid evalue bitscore length nident pident"; Min *e*-value: 1e-04) (Camacho et al., 2009). For BLAST+ searches, each query sequence was assigned to the taxonomic group of the top hit. A perl program was used to extract the top hit (available on figshare, https://doi.org/10.6084/m9.figshare.19134740).

2.3 | Comparison of databases

We compared the taxonomic composition of MIDORI2 vGB244 (released in June 2021) and the latest update of the original MIDORI (BLAST NT.fasta downloaded in February 2018) by matching species, genus and family names between files. We used a similar approach to also compare the composition of MIDORI2 vGB244 to the composition of BOLD downloaded in July 2021 and to the latest release of CO-ARBitrator released in August 2019. Because the Application Programming Interface (API) of BOLD times out when searching for too many records at once, we downloaded sequences by taxonomic groups (using the modified R script bold_datapull_byGroup.R, https://osf.io/m5cgs/; see Figshare https://doi.org/10.6084/m9.figshare.19134740) for all 15 mitochondrial encoded genes. We retrieved a phylum level phylogeny of the eukaryotes represented in MIDORI2 vGB244 from the Open Tree of Life (OTL) with the rotl package (Michonneau et al., 2016) in R (R Development Core Team, 2019). Several groups of Metazoa (e.g., Xenacoelomorpha and Brachiopoda) and Fungi (e.g., Blastocladiomycota, Zoopagomycota, and Cryptomycota), which were absent from the synthetic tree of the OTL, were added manually in TreeGraph2 (Stöver & Müller, 2010) following published phylogenies (Laumer et al., 2019; McCarthy & Fitzpatrick, 2017). Several groups missing from the OTL but represented in the MIDORI2 vGB244 by only a few sequences (e.g., Prasinodermophyta, Perkinsozoa, Endomyxa, Evosea, and Haptista) could not be confidently placed in the tree.

3 | RESULTS

3.1 | Composition of the MIDORI2 vGB244 databases

For all genes combined, MIDORI2 vGB244 contains sequences belonging to a total of 247,481 species with full binomial name, 63,605 genera and 6981 families of eukaryotes. The animal barcode gene, *CO1*, is by far the best represented with sequences of 197,688 species with full binomial name, 53,385 genera and 6042 families, followed by *lrRNA* and *srRNA* with 92,949 and 57,791 species, respectively, and *Cytb* with 53,993 species (Figure 1).

About one-third of the taxa are additions: A total of 82,361 species with full binomial name, 20,694 genera and 2231 families

TABLE 1 Comparison of reference databases of curated mitochondrial sequences

	MIDORI	MIDORI2	BOLD	CO-ARBitrator
Reference	Machida et al. (2017)	Herein	Ratnasingham and Hebert (2007)	Heller et al. (2018)
Mitochondrial genes included	*CO1, CO2, CO3, srRNA (12S), lrRNA (16S), A6, A8, CytB, ND1, ND2, ND3, ND4, ND4L, ND5, ND6*	*CO1, CO2, CO3, srRNA (12S), lrRNA (16S), A6, A8, CytB, ND1, ND2, ND3, ND4, ND4L, ND5, ND6*	*CO1 (mostly), CO2, CO3, srRNA (12S), lrRNA (16S), A6, A8, CytB, ND1, ND2, ND3, ND4, ND4L, ND5, ND6*	*CO1*
Taxonomic scope	Metazoan	All eukaryotes	All eukaryotes	Metazoan
Includes sequences that lack binomial identification	No	Yes	Yes	Yes
Sequence type	Nucleotide	Nucleotide & Amino Acid	Nucleotide	Nucleotide & Amino Acid
Curation method	Sequence properties	Sequence properties & taxonomic annotations	Sequence properties & taxonomic annotations	Sequence properties
Updates	Twice	Every two months at each new release of the GenBank nucleotide database	Continuously as users make data public	Once (08-2019)
Taxonomic assignment program format	Mothur, Qiime, RDP, Sintax, Spingo	Mothur, Qiime, RDP, Sintax, Spingo, BLAST+	None	None

are new to MIDORI2 vGB244 compared with MIDORI (built from BLAST NT.fasta database downloaded in February 2018). Across all genes, most of the new records of species and genera are metazoan sequences that were made public in GenBank between February 2018 and June 2021 (62,271 species [76%], 14,377 genera [69%]), but only 637 families (29%) fall into that category. Cytochrome c oxidase subunit 1 sequences of 56,613 metazoan species previously absent from MIDORI were incorporated into MIDORI2 vGB244, which makes it the gene with the fastest growing number of taxa, followed by *lrRNA* and *srRNA* with 22,075 and 14,164 species, respectively. Other sequences that are new to MIDORI2 belong to non-metazoan eukaryotes (20,090 species [24% of additions], 6537 genera [32%], and 1618 families [73%]), which had not been included in the construction of the metazoan-specific MIDORI (Table 1). Interestingly, 4477 species, 387 genera, and 10 families were present in MIDORI but are not in MIDORI2 vGB244. They may be sequences that were flagged by NCBI as potentially mislabeled or dubious (and therefore removed from the database), or sequences for which the taxonomy was updated. Non-metazoan eukaryotic groups newly added to MIDORI2 represent half of the phyla represented in the databases (Figure 2; but note that not all non-metazoan phyla included in MIDORI2 vGB244 are included in the tree). Yet, the total species diversity for these groups is proportionately much lower (20,090 species [8.1% of total species] with full binomial names and an additional 6537 taxa identified to genus with the species name unspecified).

3.2 | Performance of taxonomic assignment software programs

We verified that the databases were compatible with taxonomic assignment programs for which they were preformatted, and we built a test of the five most commonly used software programs into the validation process to guide future users. As expected, assignment accuracy decreased at lower taxonomic levels for all assignment programs (by 20%–30% from phylum to genus; Table 2). Assignment accuracy also differed markedly among taxonomic groups. The likelihood of correctly assigning metazoan sequences, the best represented group in MIDORI2, was higher at all taxonomic ranks than the likelihood of correctly assigning non-metazoan sequences. Protist and multicellular red algae ("others" in Table 2) were <50% likely to be correctly identified at the family and genus levels with all programs. Interestingly, assignment accuracy did not notably decrease with shorter query DNA sequences, further supporting the effectiveness of short barcodes in environmental DNA studies. BLAST+ outperformed the other four programs for classification at the phylum, class, order, and family levels, whereas the RDP classifier and Mothur, which implements a version of the RDP classifier by default, provided comparable accuracy at the genus level. The Naive Bayes classifier implemented in Qiime2 performed poorly, with 10%–20% fewer sequences correctly assigned at all taxonomic ranks.

3.3 | Comparison of MIDORI2 vGB244 with CO-Arbitrator and BOLD

We compared MIDORI2 with the latest release (August 2019) of CO-ARBitrator, a reference database of metazoan *CO1* gene sequences extracted from GenBank and BOLD (Figure 3). Of the 148,578 species (not including records with "sp." as species name), 45,647 genera and 4796 families represented in CO-ARBitrator, 8.7%, 6.2% and 0.9%, respectively, are absent from MIDORI2 vGB244. On the other hand, much higher percentages of taxa are represented in MIDORI2 vGB244 (for *CO1*) but absent from CO-ARBitrator: 55,260 of the 190,957 animal species (28.9%), 8925 of the 51,758 animal genera (17.2%), and 399 of the 5150 animal families (7.7%).

We also compared MIDORI2 vGB244 with BOLD, an open access library of DNA barcodes developed at the Centre for Biodiversity Genomics in Canada. Out of a total of 4,701,487 sequences downloaded from BOLD, the vast majority (4,615,617, 98.2%) were animal *CO1* barcodes. Despite the fact that BOLD focuses on *CO1*, MIDORI2 vGB244 contains a higher diversity of *CO1* sequences: 190,957 versus 181,975 animal species with full binomial names, 51,758 versus 45,186 genera, and 5150 versus 4727 families. In general, there was more overlap between MIDORI2 vGB244 and CO-ARBitrator than between MIDORI2 and BOLD at all taxonomic levels (Figure 3). A total of 56,306 animal species represented in BOLD (30.9%) were absent from MIDORI2 vGB244, whereas only 12,881 animal species represented in CO-ARBitrator (8.6%) were absent from MIDORI2 vGB244. Similarly, 11.2% and 10.0% of the genera and families in BOLD were absent from MIDORI2 vGB244 but only 6.2% and 0.9% of the genera and families in CO-ARBitrator were absent from MIDORI2 vGB244.

A recent study identified 2215 records that are mislabeled at the order, class, or phylum levels in GenBank (Leray et al., 2019). These records were removed from MIDORI2. However, 719 and 659 of them remain present in BOLD and CO-ARBitrator, respectively.

4 | DISCUSSION

4.1 | The utility of *CO1* sequences for non-metazoan taxa

One of the major upgrades to the databases is the inclusion of non-metazoan mitochondrial sequences represented in GenBank such as those of fungi, green algae and land plants, other multicellular algal groups, and diverse protist lineages. Although they represent most of the terrestrial and marine diversity on the planet, they remain

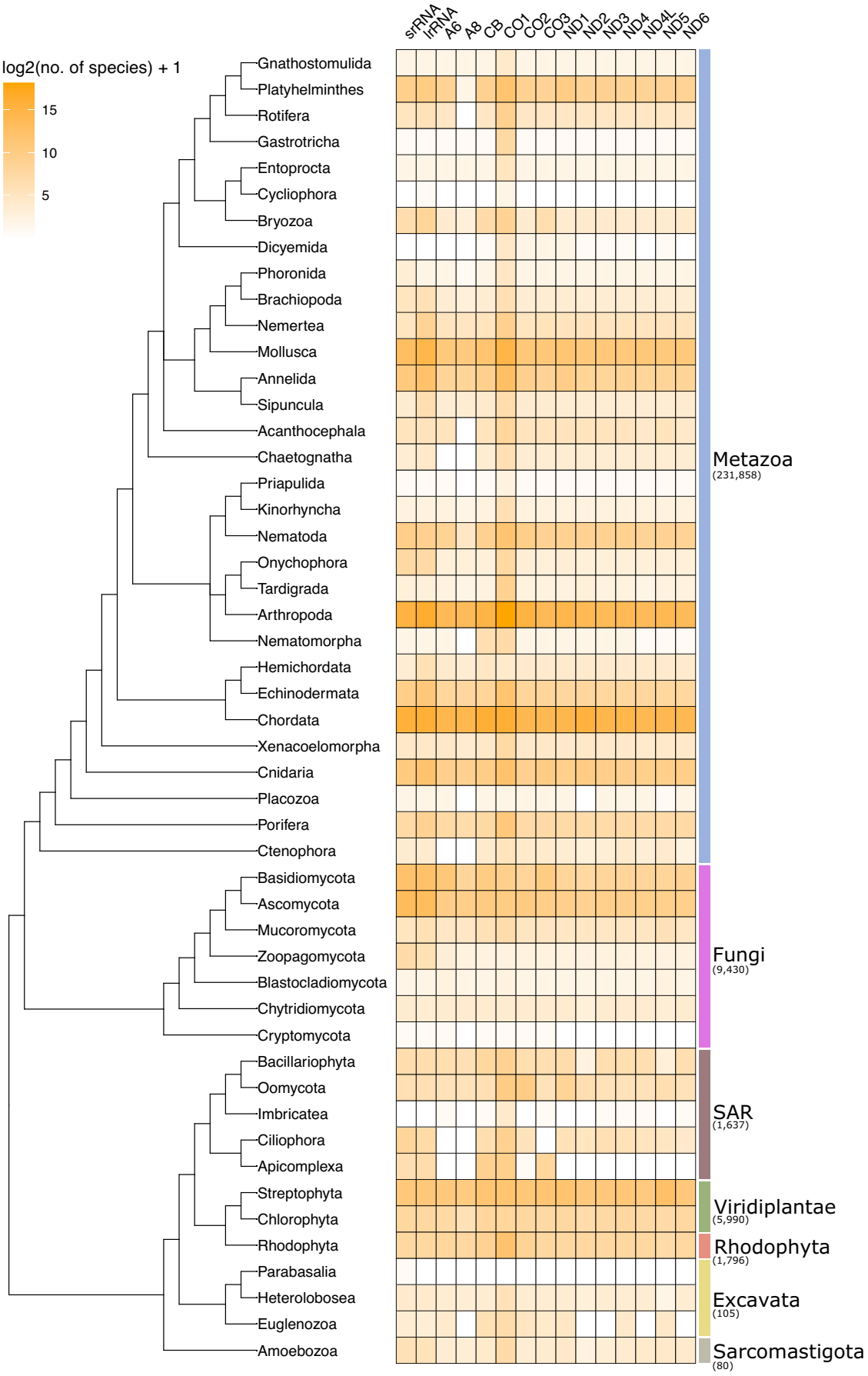

FIGURE 2 Taxonomic breakdown of mitochondrial sequences available in MIDORI2 vGB244 (built from GenBank version 244, June 2021). The phylogenetic tree was modified from the synthetic tree of the Open Tree of Life. The heatmap shows the log2 + 1 transformed number of species represented in MIDORI2 vGB244 per gene and per taxonomic group. Numbers between parentheses indicate the number of species in each group. Several taxa missing from the Open Tree of Life but represented in the MIDORI2 vGB244 by only a few sequences (e.g., Prasinodermophyta, Perkinsozoa, Endomyxa, Evosea, and Haptista) are not in the tree

largely underrepresented in databases of mitochondrial sequences. Part of this is due to lack of study, but it also reflects the fact that non-mitochondrial genes are the commonly used marker genes for well-studied non-metazoan groups. For example, chloroplast genes have become standard molecular barcodes for land plants (Hollingsworth et al., 2009) and the nuclear ribosomal transcribed spacer has been preferred for fungi and oomycetes (Schoch et al., 2012). Similarly, the marker of choice for unicellular eukaryotes has been the nuclear 18S rRNA gene because phylogenetically informative gene sequences can be amplified across diverse lineages (Pawlowski et al., 2012).

Numerous environmental surveys using metabarcoding also target non-mitochondrial genes, such as nuclear small and large ribosomal RNA genes, to take advantage of well-populated barcode databases (Guillou et al., 2013; Quast et al., 2013) that facilitate taxonomic annotations (Leray & Knowlton, 2016). Furthermore, some studies have demonstrated important features of nuclear markers: They are more abundant relative to mitochondrial markers (Jo et al., 2020; Moushomi et al., 2019) and allow the detection and identification of environmentally sensitive taxa that would otherwise have been missed with mitochondrial markers (bioindicator taxa; Seymour et al., 2020). Nonetheless, highly variable mitochondrial genes are becoming increasing popular markers for understanding trophic interactions and biogeographic patterns, and in studies that employ eDNA sequencing for biodiversity monitoring, in part because of the availability of an increasing number of effective primer sets (Gaither et al., 2022). Recent research showed that broad-range primers designed to amplify metazoan mitochondrial sequences also recover a diversity of mitochondrial sequences belonging to unicellular eukaryotes and fungi, particularly when samples contain traces of target metazoan DNA (Collins et al., 2019; Nguyen et al., 2020). These sequences are valuable data that remain poorly exploited. The inclusion of representatives of a broad range of non-metazoan groups in MIDORI2 will greatly assist the interpretation of these diverse sequence datasets.

4.2 | Comparison of taxonomic assignment programs during technical validation

Performance tests of taxonomic assignment programs showed that assignment accuracy of *CO1* sequences was overall higher at all taxonomic levels using BLAST+. BLAST+ uses a heuristic method to find close matches using an alignment. The algorithm identifies homologous sequences by initially looking for exact matches between short strings of the query sequence and strings within the database. When matches are found, BLAST+ attempts to produce an alignment in both directions. The alignment does not have to be exact, which increases the likelihood of finding close matches. On the other hand, the RDP classifier, Mothur, Sintax, and Qiime2 are alignment-free assignment methods that use the compositional features of query and reference sequences. They are based on the random sampling of short words (k-mers) and require an exact match between the query sequence and one or several sequences of the training set. Because the third codon position of protein-coding genes is highly variable between taxa, the likelihood of numerous exact matches is low, hence decreasing the assignment accuracy of alignment-free methods. Approaches based on k-mers are thus likely more suited for taxonomic assignments with ribosomal genes (e.g., Large [*16S*] and Small [*12S*] ribosomal subunit RNA) because they have longer conserved regions. Our results are consistent with previous analyses showing the robustness of BLAST+ for protein-coding genes (Hleap et al., 2021). Yet, the performance of BLAST+ was significantly lower for non-metazoan eukaryotes that are the least represented at the species level in MIDORI2. Phylogeny-based approaches may perform better for groups with limited representation in databases, as shown with nematodes (Holovachov et al., 2017). Further tests are needed to quantify differences in the accuracy of assignments between taxonomic groups, between genes and in relation to parameters used (Bik, 2021; Creedy et al., 2022). We performed the test only up to the genus level in the present study.

4.3 | Comparison of *CO1* content of MIDORI2, CO-ARBitrator and BOLD

We compared the diversity and composition of MIDORI vGB244 to the two most commonly used databases of mitochondrial sequences, CO-ARBitrator and BOLD, focusing on the *CO1* gene because it is the only gene represented in CO-ARBitrator, and *CO1* sequences represent 98.2% of the data in BOLD. We found that the vast majority of taxa included in CO-ARBitrator were also present in MIDORI2 vGB244, but MIDORI2 included many additional taxa. The lower diversity in the CO-ARBitrator database is partly explained by the public release of sequences in GenBank since CO-ARBitrator was last updated in 2019. In addition, CO-ARBitrator was built from a relatively small number of representative amino acid sequences randomly selected from 10 metazoan phyla (3055 sequences total), a subset that likely does not capture the diversity of the hypervariable *CO1* gene. While effectively rejecting non-metazoan sequences, the algorithm may fail to incorporate sequences of many species, genera

TABLE 2 Comparisons of taxonomic assignment accuracy. Percentages of correctly assigned sequences are indicated. The complete mitochondrial *CO1* region, the ~658 bp barcode region and the short ~313 bp barcode region were used as the test queries. Five programs, BLAST+, RDP, Sintax, Mothur, Qiime2, were used in this comparison

	Complete *CO1* (828–2094 bp)					*CO1* barcode region (~658 bp)	
	Total (186 seqs)	Metazoa (85 seqs)	Viridiplantae (33 seqs)	Fungi (27 seqs)	others (41 seqs)	Total (186 seqs)	Metazoa (85 seqs)
BLAST+							
Phylum	86%	92%	94%	85%	68%	86%	96%
Class	80%	86%	79%	81%	66%	80%	93%
Order	74%	82%	76%	74%	56%	75%	89%
Family	69%	80%	70%	59%	51%	70%	86%
Genus	55%	64%	58%	44%	41%	60%	74%
RDP							
Phylum	83%	91%	88%	78%	68%	82%	91%
Class	77%	84%	79%	78%	63%	75%	85%
Order	72%	81%	76%	70%	51%	72%	82%
Family	68%	80%	70%	59%	46%	68%	81%
Genus	60%	74%	58%	48%	41%	60%	73%
Sintax							
Phylum	77%	82%	82%	81%	59%	77%	86%
Class	71%	76%	76%	78%	51%	69%	79%
Order	67%	75%	73%	70%	44%	68%	78%
Family	65%	75%	73%	56%	41%	66%	78%
Genus	56%	68%	61%	41%	37%	58%	69%
Mothur							
Phylum	83%	92%	85%	78%	66%	81%	91%
Class	76%	85%	76%	78%	59%	74%	84%
Order	72%	82%	70%	70%	51%	72%	81%
Family	68%	81%	67%	59%	46%	68%	80%
Genus	60%	74%	58%	48%	41%	60%	73%
Qiime2							
Phylum	65%	73%	70%	59%	49%	65%	76%
Class	59%	62%	67%	59%	44%	61%	69%
Order	58%	61%	64%	59%	44%	58%	68%
Family	55%	61%	58%	52%	41%	55%	66%
Genus	48%	56%	45%	41%	37%	49%	62%

and families with low similarity to domains contained in the limited training set.

Our analysis also found a significantly higher number of species in MIDORI2 vGB244 than in BOLD. However, 31% of the species in BOLD were absent from MIDORI2 vGB244. The discrepancy between MIDORI2 vGB244 and BOLD could be due to three main reasons related to the fact that they do not have identical aims. First, as BOLD users make project data publicly visible to other users of the web-platform, sequences become part of the downloadable BOLD database but they do not become uploaded to GenBank, at least initially. They are therefore absent from MIDORI2 vGB244. Second, BOLD periodically mines GenBank for additional sequences, but they may exclude records missing associated metadata (e.g., collecting site, date). We were unable to find information on the process used by BOLD to mine GenBank. Third, BOLD has a built-in species delineation tool (Barcode Index Numbers; BINs) that is used, in part, to identify mislabeled sequences that our curation approach might have missed.

Unlike CO-ARBitrator and BOLD, MIDORI2 also includes curated, preformatted and regularly updated databases for mitochondrial encoded genes that have not been traditionally targeted in studies using eDNA sequencing, such as ATP synthase subunit 6 (*A6*) and 8 (*A8*), Cytochrome c oxidase subunit II (*CO2*) and III (*CO3*), Cytochrome b apoenzyme (*Cytb*) and NADH dehydrogenase

			CO1 short barcode (~313 bp)				
Viridiplantae (33 seqs)	Fungi (27 seqs)	others (41 seqs)	Total (186 seqs)	Metazoa (85 seqs)	Viridiplantae (33 seqs)	Fungi (27 seqs)	others (41 seqs)
91%	85%	61%	81%	93%	79%	85%	56%
73%	81%	59%	77%	91%	70%	81%	54%
70%	78%	49%	74%	87%	67%	78%	49%
67%	59%	46%	69%	82%	67%	63%	46%
58%	44%	44%	60%	74%	58%	44%	44%
85%	81%	61%	76%	88%	79%	81%	46%
73%	70%	59%	70%	81%	73%	70%	46%
70%	67%	54%	68%	79%	70%	67%	44%
70%	56%	49%	65%	78%	67%	56%	41%
61%	41%	46%	57%	71%	58%	41%	39%
79%	81%	56%	73%	82%	79%	74%	46%
70%	74%	54%	67%	74%	73%	70%	46%
67%	74%	51%	66%	73%	70%	70%	44%
67%	56%	49%	62%	71%	70%	56%	41%
58%	41%	44%	55%	68%	58%	37%	39%
85%	78%	61%	76%	88%	79%	81%	46%
73%	70%	59%	69%	79%	73%	70%	46%
70%	70%	54%	67%	76%	70%	67%	44%
70%	56%	49%	64%	76%	67%	56%	41%
61%	41%	46%	56%	69%	58%	41%	39%
67%	56%	46%	62%	72%	67%	56%	44%
64%	56%	44%	57%	62%	64%	56%	41%
58%	52%	41%	54%	62%	52%	56%	39%
55%	44%	41%	52%	61%	48%	48%	39%
42%	33%	39%	48%	59%	45%	37%	37%

subunits 1–4 (*ND1*–*ND4*), 4L (*ND4L*), 5 (*ND5*) and 6 (*ND6*). The availability of MIDORI2's user-friendly and regularly updated databases will promote the use of these markers in future studies as the databases are readily usable with a range of taxonomic assignment programs. The curated databases may also be used to design broad and group-specific PCR primers and to conduct in silico assessments of existing primers. The MIDORI2 databases will also be valuable resources as the field moves toward PCR-free shotgun metagenomics and metatranscriptomics, as these approaches provide data beyond *CO1* and ribosomal RNA mitochondrial genes.

5 | CONCLUSION

Overall, the MIDORI2 databases constitute an important new resource for environmental DNA studies that target mitochondrial genes. Not only do they include sequences from more taxa, they also come without obvious contaminants that are still included in BOLD and Co-ARBitrator. Databases are preformatted for a range of taxonomic assignment programs, which make them easy to integrate into sequence analysis pipelines. Automatic database updates and versioning following each new release of GenBank will ultimately

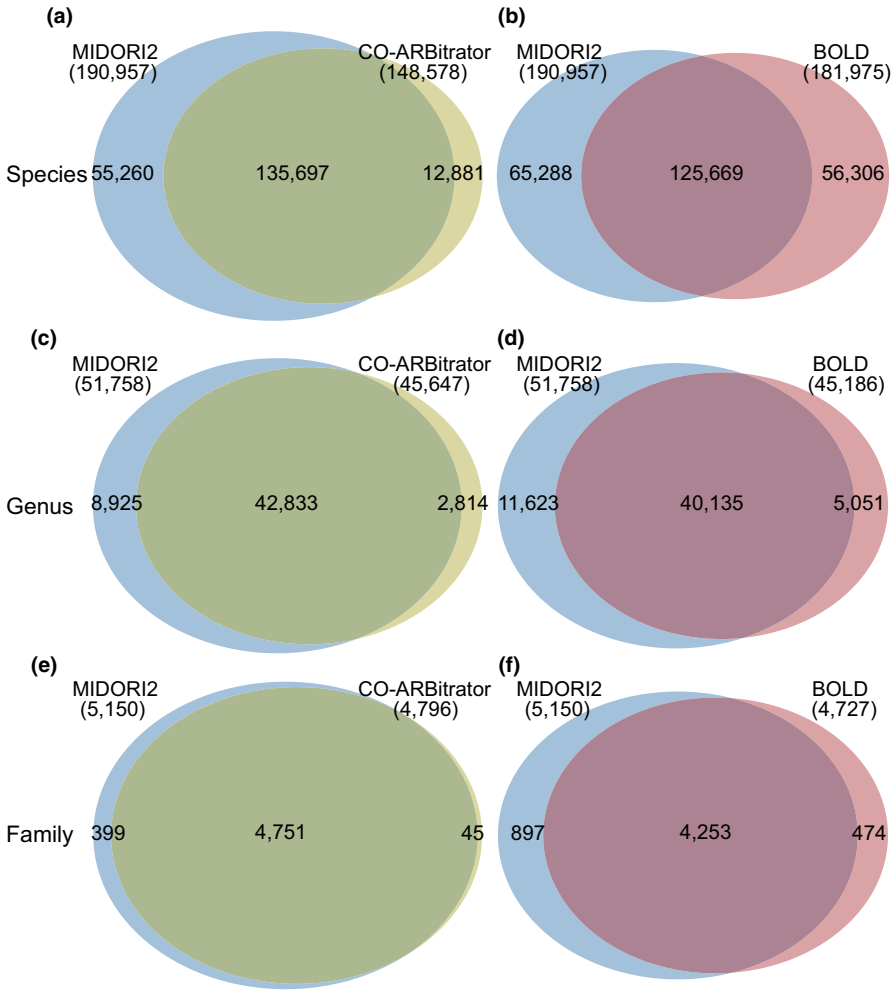

FIGURE 3 Shared diversity of metazoan *CO1* sequences among three curated databases of mitochondrial sequences. The area of overlap between circles is proportional to the number of shared species (a, b), genera (c, d) and families (e, f) between databases. The total number of species, genera and families is indicated in parenthesis. MIDORI2 vGB244 (built from GenBank version 244, June 2021), the latest version of CO-ARBitrator (released in August 2019) and BOLD (downloaded in July 2021) were compared

promote the reproducible processing of eDNA datasets (Creedy et al., 2022). The benefits of MIDORI2 extend beyond taxonomic assignments. For example, the ability to extract sequences from particular taxonomic groups will also facilitate efforts to design primers and target capture probes.

ACKNOWLEDGEMENTS

ML was funded by a grant from the Gordon and Betty Moore foundation to the Smithsonian Tropical Research Institute and the University of California Davis (https://www.moore.org/grant-detail?grantId=GBMF5603, PIs: Drs. William T. Wcislo and Jonathan A. Eisen). RJM was funded by grants from the Academia Sinica, Taiwan, the Ministry of Science and Technology, Taiwan (108-2611-M-001-001, 109-2611-M-001-003, 110-2611-M-001-009) and the Scientific Committee on Oceanic Research Working group 157.

CONFLICT OF INTEREST

The authors declared no conflicts of interest.

AUTHOR CONTRIBUTIONS

Concept and design: R.J.M., M.L., and N.K.; Data analyses: R.J.M. and M.L.; Wrote the paper: M.L., R.J.M., and N.K.

DATA AVAILABILITY STATEMENT

Scripts and datasets used in this study are available on figshare (https://doi.org/10.6084/m9.figshare.19134746, https://doi.org/10.6084/m9.figshare.19134740, https://doi.org/10.6084/m9.figshare.19134722, and https://doi.org/10.6084/m9.figshare.19134704).

ORCID

Matthieu Leray https://orcid.org/0000-0002-7327-1878
Nancy Knowlton https://orcid.org/0000-0002-4062-5502
Ryuji J. Machida https://orcid.org/0000-0003-1687-4709

REFERENCES

Allard, G., Ryan, F. J., Jeffery, I. B., & Claesson, M. J. (2015). SPINGO: A rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics*, *16*(1), 324. https://doi.org/10.1186/s12859-015-0747-1

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

Antich, A., Palacín, C., Turon, X., & Wangensteen, O. S. (2022). DnoisE: Distance denoising by entropy. An open-source parallelizable alternative for denoising sequence datasets. *PeerJ*, *10*, e12758. https://doi.org/10.7717/peerj.12758

Aylagas, E., Borja, Á., Irigoien, X., & Rodríguez-Ezpeleta, N. (2016). Benchmarking DNA metabarcoding for biodiversity-based monitoring and assessment. *Frontiers in Marine Science*, *3*, 96. https://doi.org/10.3389/fmars.2016.00096

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(D1), D36–D42. https://doi.org/10.1093/nar/gks1195

Banchi, E., Ametrano, C. G., Greco, S., Stanković, D., Muggia, L., & Pallavicini, A. (2020). PLANiTS: A curated sequence reference dataset for plant ITS DNA metabarcoding. *Database*, *2020*, baz155. https://doi.org/10.1093/database/baz155

Bik, H. M. (2021). Just keep it simple? Benchmarking the accuracy of taxonomy assignment software in metabarcoding studies. *Molecular Ecology Resources*, *21*(7), 2187–2189. https://doi.org/10.1111/1755-0998.13473

Bista, I., Carvalho, G. R., Tang, M., Walsh, K., Zhou, X., Hajibabaei, M., Shokralla, S., Seymour, M., Bradley, D., Liu, S., Christmas, M., & Creer, S. (2018). Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Molecular Ecology Resources*, *18*(5), 1020–1034. https://doi.org/10.1111/1755-0998.12888

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., & Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1462), 1935–1943. https://doi.org/10.1098/rstb.2005.1725

Blowes, S. A., Supp, S. R., Antão, L. H., Bates, A., Bruelheide, H., Chase, J. M., Moyes, F., Magurran, A., McGill, B., Myers-Smith, I. H., Winter, M., Bjorkman, A. D., Bowler, D. E., Byrnes, J. E. K., Gonzalez, A., Hines, J., Isbell, F., Jones, H. P., Navarro, L. M., … Dornelas, M. (2019). The geography of biodiversity change in marine and terrestrial assemblages. *Science*, *366*(6463), 339–345. https://doi.org/10.1126/science.aaw1620

Buckland, S. T., & Johnston, A. (2017). Monitoring the biodiversity of regions: Key principles and possible pitfalls. *Biological Conservation*, *214*, 23–34. https://doi.org/10.1016/j.biocon.2017.07.034

Bush, A., Monk, W. A., Compson, Z. G., Peters, D. L., Porter, T. M., Shokralla, S., Wright, M. T. G., Hajibabaei, M., & Baird, D. J. (2020). DNA metabarcoding reveals metacommunity dynamics in a threatened boreal wetland wilderness. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(15), 8539–8545. https://doi.org/10.1073/pnas.1918741117

Cahill, A. E., Pearman, J. K., Borja, A., Carugati, L., Carvalho, S., Danovaro, R., Dashfield, S., David, R., Féral, J.-P., Olenin, S., Šiaulys, A., Somerfield, P. J., Trayanova, A., Uyarra, M. C., & Chenuil, A. (2018). A comparative analysis of metabarcoding and morphology-based identification of benthic communities across different regional seas. *Ecology and Evolution*, *8*(17), 8908–8920. https://doi.org/10.1002/ece3.4283

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, *11*(12), 2639–2643. https://doi.org/10.1038/ismej.2017.119

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303

Clare, E. L., Economou, C. K., Bennett, F. J., Dyer, C. E., Adams, K., McRobie, B., Drinkwater, R., & Littlefair, J. E. (2022). Measuring biodiversity from DNA in the air. *Current Biology*, *32*, 693–700. https://doi.org/10.1016/j.cub.2021.11.064

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., Genner, M. J., & Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, *10*(11), 1985–2001. https://doi.org/10.1111/2041-210X.13276

Creedy, T. J., Andújar, C., Meramveliotakis, E., Noguerales, V., Overcast, I., Papadopoulou, A., Morlon, H., Vogler, A. P., Emerson, B. C., & Arribas, P. (2022). Coming of age for COI metabarcoding of whole organism community DNA: Towards bioinformatic harmonisation. *Molecular Ecology Resources*, *22*, 847–861. https://doi.org/10.1111/1755-0998.13502

Edgar, R. C. (2016). SINTAX: A simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *BioRxiv*, 074161, https://doi.org/10.1101/074161

Elbrecht, V., Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Hajibabaei, M., Wright, M., Zakharov, E. V., Hebert, P. D. N., & Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, *7*, e7745. https://doi.org/10.7717/peerj.7745

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, *3*(5), 294–299.

Gaither, M. R., DiBattista, J. D., Leray, M., & von der Heyden, S. (2022). Metabarcoding the marine environment: From single species to biogeographic patterns. *Environmental DNA*, *4*(1), 3–8. https://doi.org/10.1002/edn3.270

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., … Christen, R. (2013). The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(Database issue), D597–D604. https://doi.org/10.1093/nar/gks1160

Heller, P., Casaletto, J., Ruiz, G., & Geller, J. (2018). A database of metazoan cytochrome c oxidase subunit I gene sequences derived from GenBank with CO-ARBitrator. *Scientific Data*, *5*(1), 180156. https://doi.org/10.1038/sdata.2018.156

Hleap, J. S., Littlefair, J. E., Steinke, D., Hebert, P. D. N., & Cristescu, M. E. (2021). Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. *Molecular Ecology Resources*, *21*(7), 2190–2203. https://doi.org/10.1111/1755-0998.13407

Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M. W., Cowan, R. S., Erickson, D. L., Fazekas, A. J., Graham, S. W., James, K. E., Kim, K.-J., Kress, W. J., Schneider, H., van AlphenStahl, J., Barrett, S. C. H., van den Berg, C., Bogarin, D., … Little, D. P. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(31), 12794–12797. https://doi.org/10.1073/pnas.0905845106

Holovachov, O., Haenel, Q., Bourlat, S. J., & Jondelius, U. (2017). Taxonomy assignment approach determines the efficiency of identification of OTUs in marine nematodes. *Royal Society Open Science*, *4*(8), 170315. https://doi.org/10.1098/rsos.170315

Jo, T., Arimoto, M., Murakami, H., Masuda, R., & Minamoto, T. (2020). Estimating shedding and decay rates of environmental nuclear DNA with relation to water temperature and biomass. *Environmental DNA*, *2*(2), 140–151. https://doi.org/10.1002/edn3.51

Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: Multiple sequence alignment, interactive sequence choice and

visualization. *Briefings in Bioinformatics*, *20*(4), 1160–1166. https://doi.org/10.1093/bib/bbx108

Kelly, R. P., O'Donnell, J. L., Lowell, N. C., Shelton, A. O., Samhouri, J. F., Hennessey, S. M., Feist, B. E., & Williams, G. D. (2016). Genetic signatures of ecological diversity along an urbanization gradient. *PeerJ*, *4*, e2444. https://doi.org/10.7717/peerj.2444

Kelly, R. P., Shelton, A. O., & Gallego, R. (2019). Understanding PCR processes to draw meaningful conclusions from environmental DNA studies. *Scientific Reports*, *9*(1), 12133. https://doi.org/10.1038/s41598-019-48546-x

Laumer, C. E., Fernández, R., Lemer, S., Combosch, D., Kocot, K. M., Riesgo, A., Andrade, S. C. S., Sterrer, W., Sørensen, M. V., & Giribet, G. (2019). Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1906), 20190831. https://doi.org/10.1098/rspb.2019.0831

Leray, M., Ho, S.-L., Lin, I.-J., & Machida, R. J. (2018). MIDORI server: A webserver for taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using a curated database. *Bioinformatics*, *34*(21), 3753–3754. https://doi.org/10.1093/bioinformatics/bty454

Leray, M., & Knowlton, N. (2016). Censusing marine eukaryotic diversity in the twenty-first century. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1702), 20150331. https://doi.org/10.1098/rstb.2015.0331

Leray, M., Knowlton, N., Ho, S.-L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(45), 22651–22656. https://doi.org/10.1073/pnas.1911714116

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., & Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*(1), 34. https://doi.org/10.1186/1742-9994-10-34

Lopez, M. L. D., Lin, Y., Sato, M., Hsieh, C., Shiah, F.-K., & Machida, R. J. (2022). Using metatranscriptomics to estimate the diversity and composition of zooplankton communities. *Molecular Ecology Resources*, *22*(2), 638–652. https://doi.org/10.1111/1755-0998.13506

Lynggaard, C., Bertelsen, M. F., Jensen, C. V., Johnson, M. S., Frøslev, T. G., Olsen, M. T., & Bohmann, K. (2022). Airborne environmental DNA for terrestrial vertebrate community monitoring. *Current Biology*, *32*, 701–707. https://doi.org/10.1016/j.cub.2021.12.014

Machida, R. J., Hashiguchi, Y., Nishida, M., & Nishida, S. (2009). Zooplankton diversity analysis through single-gene sequencing of a community sample. *BMC Genomics*, *10*(1), 438. https://doi.org/10.1186/1471-2164-10-438

Machida, R. J., Kweskin, M., & Knowlton, N. (2012). PCR primers for metazoan mitochondrial 12S ribosomal DNA sequences. *PLoS ONE*, *7*(4), e35887. https://doi.org/10.1371/journal.pone.0035887

Machida, R. J., Kurihara, H., Nakajima, R., Sakamaki, T., Lin, Y.-Y., & Furusawa, K. (2021). Comparative analysis of zooplankton diversities and compositions estimated from complement DNA and genomic DNA amplicons, metatranscriptomics, and morphological identifications. *ICES Journal of Marine Science*, *78*(9), 3428–3443. https://doi.org/10.1093/icesjms/fsab084

Machida, R. J., Leray, M., Ho, S.-L., & Knowlton, N. (2017). Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Scientific Data*, *4*(1), 170027. https://doi.org/10.1038/sdata.2017.27

McCarthy, C. G. P., & Fitzpatrick, D. A. (2017). Phylogenomic reconstruction of the Oomycete phylogeny derived from 37 genomes. *mSphere*, *2*(2), e00095-17. https://doi.org/10.1128/mSphere.00095-17

Michonneau, F., Brown, J. W., & Winter, D. J. (2016). rotl: An R package to interact with the Open Tree of Life data. *Methods in Ecology and Evolution*, *7*(12), 1476–1481. https://doi.org/10.1111/2041-210X.12593

Morecroft, M. D., Duffield, S., Harley, M., Pearce-Higgins, J. W., Stevens, N., Watts, O., & Whitaker, J. (2019). Measuring the success of climate change adaptation and mitigation in terrestrial ecosystems. *Science*, *366*(6471), eaaw9256. https://doi.org/10.1126/science.aaw9256

Moushomi, R., Wilgar, G., Carvalho, G., Creer, S., & Seymour, M. (2019). Environmental DNA size sorting and degradation experiment indicates the state of *Daphnia magna* mitochondrial and nuclear eDNA is subcellular. *Scientific Reports*, *9*, 12500. https://doi.org/10.1038/s41598-019-48984-7

Nguyen, B. N., Shen, E. W., Seemann, J., Correa, A. M. S., O'Donnell, J. L., Altieri, A. H., Knowlton, N., Crandall, K. A., Egan, S. P., McMillan, W. O., & Leray, M. (2020). Environmental DNA survey captures patterns of fish and invertebrate diversity across a tropical seascape. *Scientific Reports*, *10*(1), 6729. https://doi.org/10.1038/s41598-020-63565-9

Nilsson, R. H., Larsson, K.-H., Taylor, A. F. S., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F. O., Tedersoo, L., Saar, I., Kõljalg, U., & Abarenkov, K. (2019). The UNITE database for molecular identification of fungi: Handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research*, *47*(D1), D259–D264. https://doi.org/10.1093/nar/gky1022

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., Bowser, S. S., Cepicka, I., Decelle, J., Dunthorn, M., Fiore-Donno, A. M., Gile, G. H., Holzmann, M., Jahn, R., Jirků, M., Keeling, P. J., Kostka, M., Kudryavtsev, A., Lara, E., … de Vargas, C. (2012). CBOL protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLOS Biology*, *10*(11), e1001419. https://doi.org/10.1371/journal.pbio.1001419

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(Database issue), D590–D596. https://doi.org/10.1093/nar/gks1219

R Development Core Team (2019). *R: A language and environment for statistical computing*. http://www.r-project.org

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (http://www.barcodinglife.org) *Molecular Ecology Notes*, *7*(3), 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Rimet, F., Gusev, E., Kahlert, M., Kelly, M. G., Kulikovskiy, M., Maltsev, Y., Mann, D. G., Pfannkuchen, M., Trobajo, R., Vasselon, V., Zimmermann, J., & Bouchez, A. (2019). Diat.barcode, an open-access curated barcode library for diatoms. *Scientific Reports*, *9*(1), 15116. https://doi.org/10.1038/s41598-019-51500-6

Rodriguez-Ezpeleta, N., Morissette, O., Bean, C. W., Manu, S., Banerjee, P., Lacoursière-Roussel, A., Beng, K. C., Alter, S. E., Roger, F., Holman, L. E., Stewart, K. A., Monaghan, M. T., Mauvisseau, Q., Mirimin, L., Wangensteen, O. S., Antognazza, C. M., Helyar, S. J., Boer, H., Monchamp, M.-E., … Deiner, K. (2020). Trade-offs between reducing complex terminology and producing accurate interpretations from environmental DNA: Comment on "Environmental DNA: What's behind the term?" by Pawlowski et al,Trade-offs between reducing complex terminology and producing accurate interpretations from environmental DNA: Comment on "Environmental DNA: What's behind the term?" by Pawlowski. *Molecular Ecology*, *30*, 4601–4605. https://doi.org/10.1111/mec.15942

Ruppert, K. M., Kline, R. J., & Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, *17*, e00547. https://doi.org/10.1016/j.gecco.2019.e00547

Sayers, E. (2010). A general introduction to the E-utilities. In *Entrez programming utilities help [Internet]*. National Center for Biotechnology

Information (US). https://www.ncbi.nlm.nih.gov/books/NBK25497/

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., & Consortium, F. B. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(16), 6241–6246. https://doi.org/10.1073/pnas.1117018109

Seymour, M., Edwards, F. K., Cosby, B. J., Kelly, M. G., de Bruyn, M., Carvalho, G. R., & Creer, S. (2020). Executing multi-taxa eDNA ecological assessment via traditional metrics and interactive networks. *Science of the Total Environment*, *729*, 138801. https://doi.org/10.1016/j.scitotenv.2020.138801

Stöver, B. C., & Müller, K. F. (2010). TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics*, *11*(1), 7. https://doi.org/10.1186/1471-2105-11-7

Thomsen, P. F., & Sigsgaard, E. E. (2019). Environmental DNA metabarcoding of wild flowers reveals diverse communities of terrestrial arthropods. *Ecology and Evolution*, *9*(4), 1665–1679. https://doi.org/10.1002/ece3.4809

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, *183*, 4–18. https://doi.org/10.1016/j.biocon.2014.11.019

Trimble, M. J., & van Aarde, R. J. (2012). Geographical and taxonomic biases in research on biodiversity in human-modified landscapes. *Ecosphere*, *3*(12), 1–16. https://doi.org/10.1890/ES12-00299.1

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07

How to cite this article: Leray, M., Knowlton, N., & Machida, R. J. (2022). MIDORI2: A collection of quality controlled, preformatted, and regularly updated reference databases for taxonomic assignment of eukaryotic mitochondrial sequences. *Environmental DNA*, *4*, 894–907. https://doi.org/10.1002/edn3.303