Database, 2025, 1–10
DOI: https://doi.org/10.1093/database/baaf057
Original article

A novel taxonomic database for eukaryotic mitochondrial cytochrome oxidase subunit I gene (eKOI), with a focus on protists diversity

Rubén González-Miguéns [1,*], Àlex Gàlvez-Morante[1], Margarita Skamnelou[1], Meritxell Antó[1], Elena Casacuberta[1], Daniel J. Richter[1], Enrique Lara[2], Daniel Vaulot [3,4], Javier del Campo [1], Iñaki Ruiz-Trillo [1,5]

[1]Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), 08003 Barcelona, Spain
[2]Real Jardín Botánico de Madrid (RJB-CSIC), 28014 Madrid, Spain
[3]Sorbonne Université, CNRS, UMR7144, Station Biologique de Roscoff, 29680 Roscoff, France
[4]Department of Biosciences, University of Oslo, PO Box 1066 Blindern, 0316 Oslo, Norway
[5]ICREA, 08010 Barcelona, Spain

*Corresponding author. Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), 08003 Barcelona, Spain. E-mail: ruben.miguens@ibe.upf-csic.es

Citation details: González-Miguéns, R., Gàlvez-Morante, À., Skamnelou, M., *et al.* A novel taxonomic database for eukaryotic mitochondrial cytochrome oxidase subunit I gene (eKOI), with a focus on protists diversity. Database (2025) Vol. 2025: article ID baaf057; DOI: https://doi.org/10.1093/database/baaf057

Abstract

Metabarcoding has emerged as a robust method for assessing biodiversity patterns by retrieving environmental DNA directly from ecosystems. While the 18S rRNA gene is the primary genetic marker used for broad eukaryotic metabarcoding, it has limitations in resolving lower taxonomic levels. A potential alternative is the mitochondrial cytochrome oxidase subunit I (COI) gene because it offers resolution at the species level. However, the COI gene lacks a comprehensive, curated taxonomically informed database including protists. To address this gap, we introduce eKOI, a novel, curated COI gene database designed to enhance the taxonomic annotation for protists that can be used for COI-based metabarcoding. eKOI integrates data from GenBank and mitochondrial genomes, followed by extensive manual curation to eliminate redundancies and contaminants, recovering 15 947 sequences within 80 eukaryotic phyla. We validated the use of eKOI by reannotating several COI metabarcoding datasets, revealing previously unidentified protist biodiversity and demonstrating the database utility for community-level analyses.

Introduction

Metabarcoding has emerged as a powerful tool in the last two decades [1], allowing researchers to comprehend biodiversity patterns without the biases of traditional sampling methods [2]. Under this approach, DNA is directly retrieved from the environment (eDNA), allowing the characterization of microbial communities without the need for isolation or culture-dependent approaches [3,4]. Furthermore, its affordability nature has broadened application across diverse biological disciplines. For example, metabarcoding has provided novel insights into biogeographical [5,6] and ecological [7,8] patterns. However, the success of metabarcoding hinges on well-curated and comprehensive reference taxonomic databases to accurately annotate the sequenced eDNA.

Ribosomal genes, particularly the 18S rRNA gene (18S), are the most widely used genetic markers for species delimitation and phylogenetic inference within eukaryotes, including protists. Their widespread use is due to their universality (they are present in all living beings), the presence of both conserved and hypervariable regions that facilitate a phylogenetic resolution at various taxonomic levels, and the availability of different generalist and taxon-specific primer sets. As a result, several large taxonomic databases have been generated, such as PR^2 [9] or SILVA [10], which are essential for accurate taxonomic annotation in metabarcoding studies.

However, the 18S has limitations in resolving taxonomy at the intraspecies levels, due to its highly conserved nature [11,12]. To overcome this limitation, researchers have explored more divergent noncoding regions within and between ribosomal genes, such as the internal transcribed spacer [13], or protein-coding genes like the ribulose-bisphosphate carboxylase gene (rbcL) [14,15]. While these alternative markers offer improved taxonomic resolution appropriate for species delimitation, taxonomic databases for these markers often exhibit bias towards specific groups, such as fungi or diatoms [16,17]; although recent databases like EUKARYOME [18] aim to cover the full breadth of eukaryotic diversity.

The mitochondrial cytochrome oxidase subunit I (COI) gene has been proposed as the barcode gene for species delimitation in metazoans [19]. The COI gene has also started to be applied to the taxonomy and systematics of protists, including groups such as testate amoebae [20,21], foraminifera [22], coccolithophores [23], and diatoms [24]. This has led to the creation of several reference COI taxonomic databases, like BOLD [25], CO-ARBitrator [26], and MIDORI2 [27]. However, these existing databases are predominantly biased

Received 28 March 2025; Revised 17 June 2025; Accepted 5 August 2025

towards metazoans, and often lack curated and nonredundant sequences for other eukaryotic groups. Additionally, they may lack standardized taxonomic ranks that are commonly used for eukaryotes [28]. These limitations hinder the effective application of the COI gene in community-level studies of eukaryotes, particularly those including protists, and obstruct taxonomic integration across different molecular marker databases, such as the widely used PR^2 for 18S rRNA gene [9].

To address these limitations, we introduce eKOI (environmental eukaryotic cytochrome oxidase subunit I) (version 1.0), a novel and curated eukaryote-wide database encompassing 80 phyla of the mitochondrial COI gene. This new reference database aims to overcome the limitations of existing COI taxonomic databases at the protist community-level, analogous to what PR^2 represents for 18S. This will facilitate accurate taxonomic annotation of metabarcoding sequences, as well as comparisons among different taxonomic databases derived from other molecular markers. To create this dataset, we combined COI gene data extracted from GenBank and the complete COI gene obtained from publicly available mitochondrial genomes. A thorough and manual curation process was implemented to eliminate redundant sequences, identify potential contaminants, and correct taxonomic annotation errors. We evaluated eKOI by taxonomically reannotating various COI-based metabarcoding studies, resulting in the identification of previously unidentified diversity. Finally, we further validated the new taxonomic annotations for these studies by constructing phylogenetic trees using sequences from the eKOI database, thereby confirming the large amount of protist biodiversity previously uncharacterized with the COI gene.

Materials and methods

GenBank database sequences downloading and curation

To construct the eKOI taxonomic database, we initially retrieved the sequences from the mitochondrial gene 'Cytochrome oxidase subunit I' (COI) from GenBank (Fig. S1). We established keywords to search for each taxonomic group in the 'NCBI taxonomy browser': '(("X"[Organism] OR "X"[All Fields]) AND co1[All Fields]) OR (("X"[Organism] OR "X"[All Fields]) AND cox1[All Fields]) OR (("X"[Organism] OR X[All Fields]) AND coi[All Fields]) OR (("X"[Organism] OR X[All Fields]) AND cytochrome oxidase[All Fields] AND subunit[All Fields] AND 1[All Fields]) OR (("X"[Organism] OR "X"[All Fields]) AND cytochrome oxidase[All Fields] AND subunit[All Fields]) OR (("X"[Organism] OR "X"[All Fields]) AND coxi[All Fields])', where 'X' denotes the name of each major taxonomic group of 'NCBI taxonomic browser' within 'Eukaryota'. The files were downloaded in 'INSDSeq XML' format, obtaining about 4 million sequences. The resultant sequences, grouped by taxonomic groups, mainly phyla, were processed using a custom script 1_sequences_procesing.py. This script eliminated the sequences that were duplicated, smaller than 200 bp and larger than 3000 bp. Next, to reduce the redundancy and the total number of sequences, clusters were created based on similarity percentages using vsearch ver. 2.14.1 [29], selecting a representative sequence for each cluster. The similarity percentage was established at 97%, except for Arthropoda, Chordata, and Mollusca for which it was set at 90%, aiming to balance the number of sequences per taxonomic group in the final eKOI database. This limits identifications to low taxonomic levels such as species or genus. Chimeric sequence detection was performed using vsearch ver. 2.14.1 '*de novo*' [29] per phylum and the final database, using default settings. Lastly, a 'fasta' file was generated for each taxonomic group containing the sequences with the taxonomy string defined in GenBank.

Alignments were generated for each taxonomic group, using MAFFT version 7.490 [30], using default parameters. Finally, manual curation of the resulting sequences was performed using the software Geneious Prime (version 2019.0.4), removing divergent sequences that may be errors or taxonomic misclassifications.

Mitochondrial genome database curation and integration with GenBank database

The resulting curated sequences retrieved from GenBank were combined with the mitochondrial genome of public databases, such as GenBank and Zenodo. The COI gene was extracted from complete mitochondrial genomes present in GenBank based on the sequence annotations. Some resulting sequences contained exons and introns. It has been demonstrated that some introns can result from nuclear pseudogenes [31]. Therefore, we first tested whether the introns were potential pseudogenes or chimeric sequences. Additionally, the presence of pseudogenes was confirmed through alignments, verifying whether there were divergent sequences. To accomplish this, two datasets were generated from the mitochondrial genomes: one with the entire COI gene including introns and exons, and another one containing only the coding region, the exons.

Alignments were performed using MAFFT version 7.490 [30], using default parameters, aiming to determine if the sequences from the GenBank database or the newly amplified sequences contained the intron region. The script 2_percentage_identity_graphic.py graphically represents the percentage of identity for each position in an alignment. Once the introns were confirmed to be potential pseudogenes or contaminations, the curated GenBank sequences were combined with the coding regions of the COI gene extracted from the mitochondrial genomes (see results section 'Testing the presence of introns within COI gene'). The final fasta files and alignments for each taxonomic group are available from the Supplementary data. Finally, the alignments without introns were used to curate the sequences with a wrong taxonomy following the 'curation process' in EukRef [32], generating phylogenetic reference trees and alignments per phylum.

Taxonomy path curation

One of the limitations of current molecular databases for the taxonomic annotation of eDNA sequences is how to handle the variable taxonomic ranks across clades of eukaryotes. Even though higher taxonomic ranks, such as order or class, lack comparable evolutionary context, in terms of divergence time or evolutionary history in general, many computational tools require a fixed number of ranks across taxa.

To achieve this, the names of the sequences were extracted into a CSV file using the script 3_fasta_name_extraction.py. This file was manually curated to correct the taxonomy of

Table 1. Metabarcoding studies based on the COI molecular marker that has been reanalyzed with our new eKOI dataset, together with the primers used in each case and the environment

ID	Paper	Primer R	Primer F	Environment
1	[57]	ArCOIR [20]	LCO [40]	Freshwater and soil
2	[33]	ArCOIR [20]	LCO [40]	Freshwater and marine
3	[58]	HCO [40]	LCO [40]	Soil
4	[44]	jgHCO2198 [41]	mlCOIintF-XT [43]	Marine
5	[59]	jgHCO2198 [41]	mlCOIintF [42]	Marine
6	[60]	jgHCO2198 [41]	mlCOIintF [42]	Marine
7	[46]	jgHCO2198 [41]	mlCOIintF [42]	Marine
8	[61]	jgHCO2198 [41]	mlCOIintF [42]	Freshwater
9	[62]	jgHCO2198 [41]	mlCOIintF [42]	Marine
10	[63]	jgHCO2198 [41]	mlCOIintF [42]	Freshwater
11	[64]	jgHCO2198 [41]	mlCOIintF [42]	Marine
12	[65]	EPTDr2n [66]	fwhF2 [67]	Freshwater
13	[68]	HCO [40]	LCO [40]	Freshwater
14	[69]	jgHCO2198 [41]	mlCOIintF [42]	Marine
15	[45]	jgHCO2198 [41]	mlCOIintF [42]	Marine

each sequence. The aim was to ensure that all sequences have 'homologous' taxonomic categories among them at each taxonomic level. We used the taxonomic levels proposed by PR^2 version 5.0 (released in 2023) [9], comprising nine levels: 'domain; supergroup; division; subdivision; class; order; family; genus; species'. We also generated another dataset, to which we manually added the level 'phylum' between 'subdivision' and 'class' since phylum is one of the most widely used taxonomic ranks in metazoans. Once each CSV file was manually curated, the names of the sequences in the fasta file were substituted by the accession identifier unique to each sequence, using the script 4_fasta_name_substitution.py. The final eKOI taxonomic database version 1.0 is available from the supplementary data (file 'eKOI_ver1.fasta').

Testing the eKOI database accuracy and comparing with other taxonomic databases

To test the accuracy of the eKOI database (version 1.0), 15 metabarcoding studies based on COI were selected (see Table 1 for details on metabarcoding studies and primers used). The raw data of these studies was reanalysed using the protocol described in [33], using the dada2 R package version 1.32 [34]. Subsequently, the resulting amplicon sequence variants (ASVs) smaller than 100 bp were discarded, as taxonomic annotations for fragments of such small size are typically unreliable.

Taxonomic annotation was performed using the eKOI and MIDORI2 [27] databases (accessed 27 October 2024). For this purpose, a custom script 5_taxonomic_assignation.py was employed. This script generates an independent folder for each fasta file present in the script's directory. Within each folder, an Excel file is generated containing the taxonomic annotation information for each ASV using vsearch usearch_global command version 2.14.1 [29]. ASVs with lower than 84% similarity to reference sequences were not considered. This threshold was established based on Amoebozoa from the order Arcellinida and metazoans COI pairwise distance [19,33]. However, this threshold should be applied with caution, as specific studies are needed to refine and validate it across eukaryotic phyla. Subsequently, a fasta file is generated for each desired taxonomic level. In this case, we selected phylum. The resulting ASVs taxonomically annotated for each phylum with eKOI can be downloaded from '3_ASV_metabarcoding' in the supplementary data.

To graphically represent the diversity of each taxonomic group and test the accuracy of taxonomic annotations using eKOI, we generated phylogenetic trees for each taxonomic group, focusing on protists. To reduce the number of sequences for graphical representation, the ASVs were grouped into operational taxonomic units (OTUs) based on a similarity percentage of 97% using a custom script 6_cluster_OTU.py. Alignments of the resulting OTUs were performed using MAFFT version 7.490 [30], with default parameters, along with sequences from the eKOI database of the same taxonomic groups, and at least two sequences from sister groups as outgroups. Tree topologies and node supports were evaluated using maximum likelihood with IQTREE2 version 2.0, where the best substitution models were selected with ModelFinder [35]. Node supports were assessed with 10 000 ultrafast bootstrap replicate approximations [36]. The resulting trees were graphically edited using the R package ggtree version 3.12 [37].

For some groups (Apicomplexa, Cercozoa, Filasterea, and Heterolobosea), reconstructions of ancestral habitat characters were performed to illustrate the potential of the eKOI database. For that purpose, we used the phytools R package version 2.3 [38]. We employed the function *make.simmap* [39] to generate 500 stochastic character maps from our dataset, under the equal rates model. These sets of stochastic maps were then summarized using the function *densityMap*, plotting the posterior probability of being in each state across all the edges and nodes of the tree. For the graphical representation of the distribution maps of the new eDNA sequences, we used the maps R package version 3.4.2.

Once taxonomic annotations were made, the different samples were grouped by study (Table 1) and then by sampled environment. From these groups, the mean ASV taxonomic annotation values were obtained across all samples using the script 7_taxonomic_assignation_mean.py. This script generates a CSV with the percentage of the number of ASV assigned to each phylum. This was performed for the eKOI and MIDORI2 databases ('5_COI_databses_comparison' in supplementary data). Barplots were generated in R using ggplot2 version 3.5.1, selecting the phyla with at least 1% of ASV taxonomically assigned in each study.

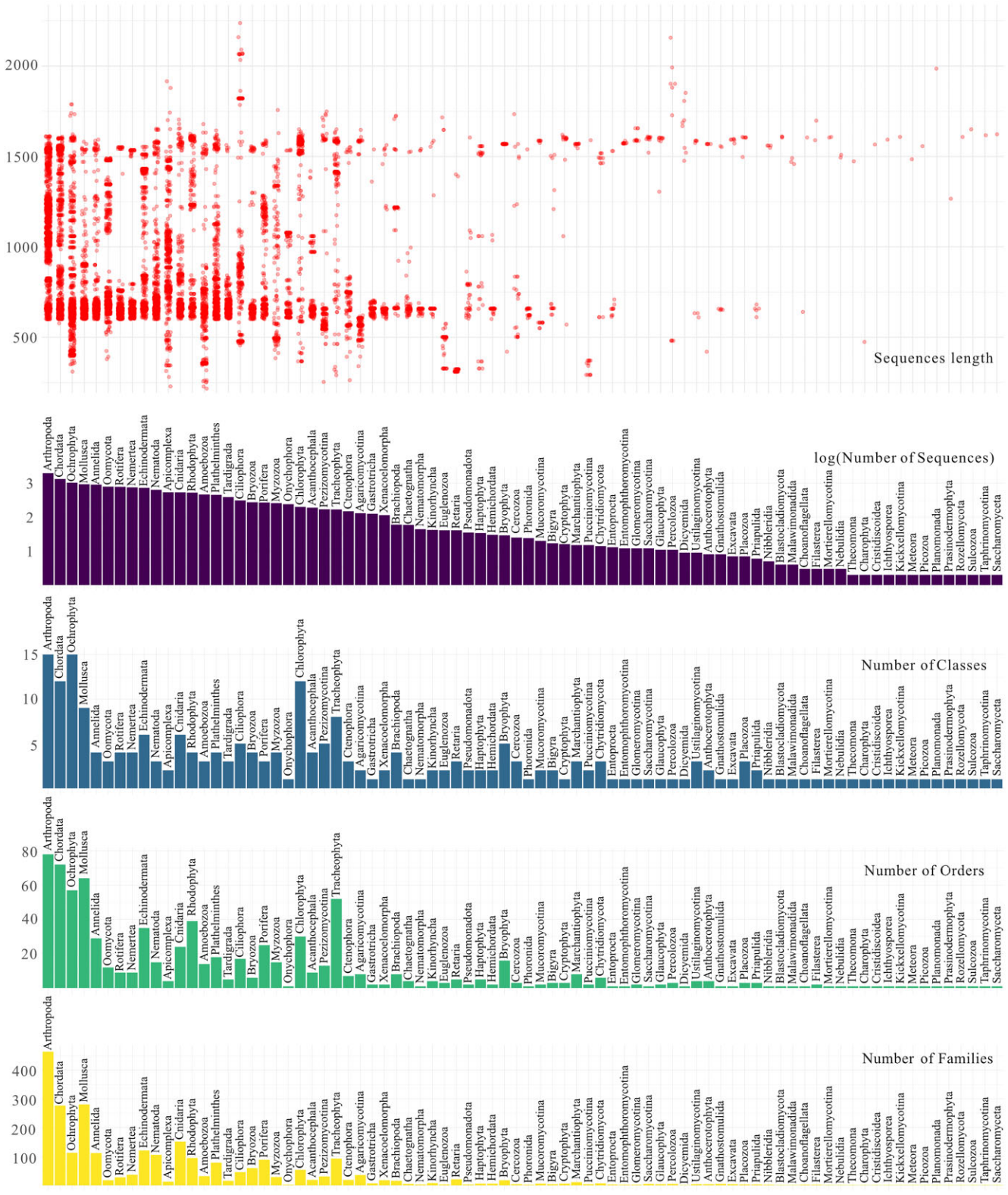

Figure 1. Graphs representing the number of families (yellow), orders (green), classes (blue), and total sequences (violet) per phylum present in the eKOI database. Red circles represent the size, in base pairs, of each sequence included in each phylum.

Results

eKOI database

The final eKOI database includes 15 947 sequences, representing 80 eukaryotic phyla, 231 classes, 796 orders, and 2646 families of eukaryotes (Fig. 1 and 'eKOI_taxonomy.csv' in supplementary data). Therefore, eKOI expands protist taxonomic coverage compared to other COI databases and incorporates eukaryotic groups absent in existing databases, such as Picozoa, Nibbleridia, or Rozellomycota (see Tables S1 and

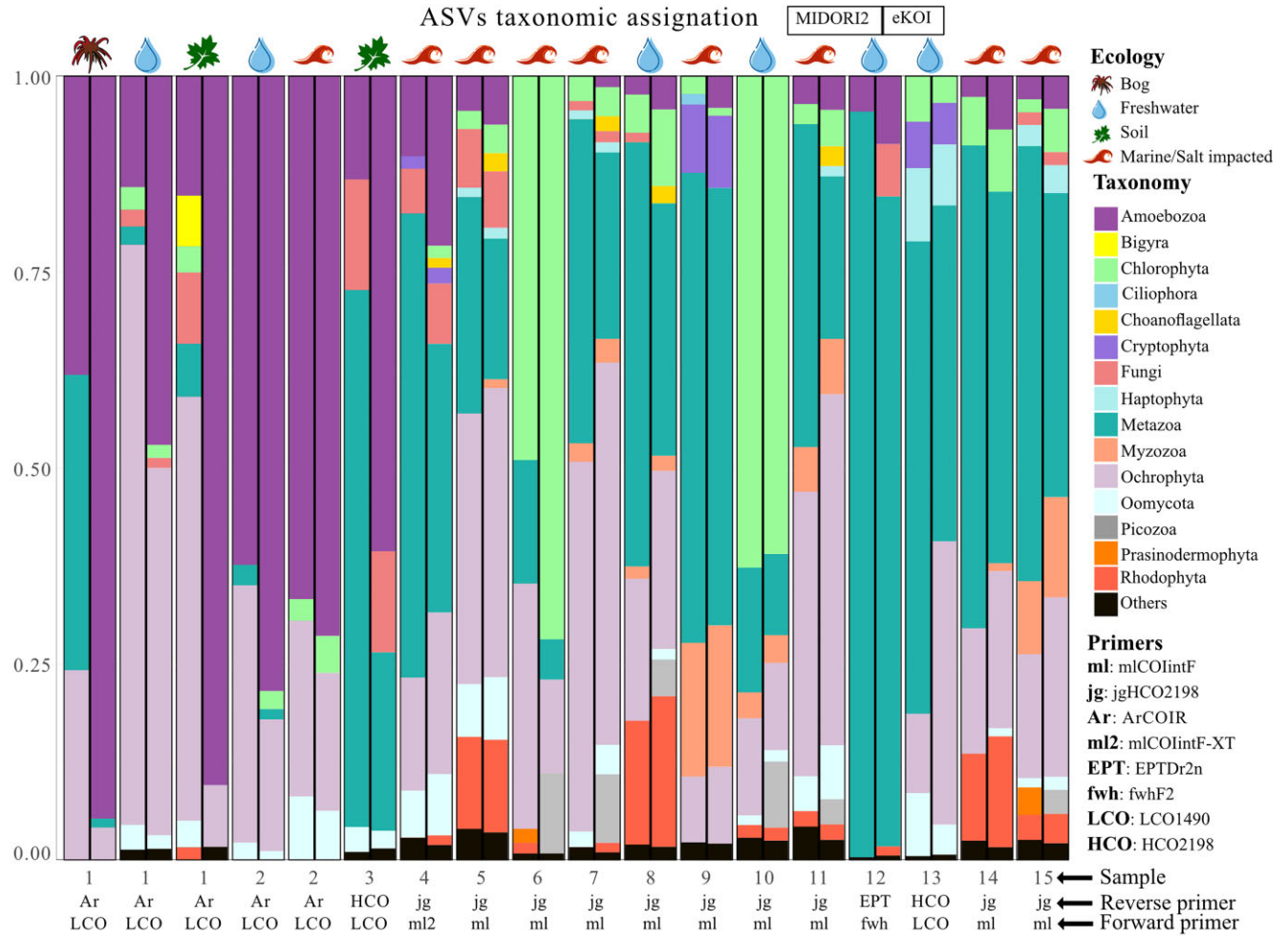

Figure 2. Bar chart representing the proportion of ASVs, relative to the total taxonomically classified, to each phylum in the MIDORI2 (left barplots) eKOI database (right barplots) from 15 metabarcoding studies (Table 1). Due to the large number of phyla present in each study, only the phyla with at least 1% of ratio per study are represented (the rest of the other phyla are represented as 'others'). The metazoan phyla were grouped as 'Metazoa' and fungi phyla as 'Fungi'. The environment type of each study is represented by different symbols. The pair of primers used in each study is also presented.

S2 in supplementary data). Almost every eukaryotic phylum in the eKOI database is represented by at least one complete COI gene sequence. eKOI predominantly contains sequences of two lengths: ˜600 base pairs and 1600 base pairs (Fig. 1). The ∼600 bp sequences correspond to fragments amplified using the HCO and LCO primer pair [40], while the ∼1600 bp sequences represent the complete COI gene obtained from mitochondrial genomes. The coding region (exons) of the COI gene exhibits relatively stable length across eukaryotic phyla.

To validate taxonomic annotations and the presence of introns, we constructed two alignments one with and one without introns (see the section 'Materials and Methods', 'Mitochondrial Genome Database Curation and Integration with GenBank Database'). In the alignment including introns, some GenBank sequences displayed large gaps due to the absence of introns (Fig. S2). Sequences with introns were considered as potential pseudogenes and were removed [31]. The lack of highly divergent sequences and long branches in the phylogenetic trees, supports the taxonomy accuracy of the eKOI database (supplementary data folder '2_alignment_eKOI').

Applying the eKOI database in metabarcoding studies

To test the effectiveness and accuracy of the eKOI database to taxonomically assign eukaryotic COI eDNA diversity, we analysed 15 metabarcoding studies (Table 1 and Fig. 2). The prevalent use of primers jgHCO2198 [41], mlCOIintF [42], and mlCOIintF-XT [43] is due to their ability to amplify a diverse range of eukaryotic taxonomic groups. The variation in the proportion of phyla within similar ecosystem types, using identical primers, is linked to differences in sample substrates, such as plankton [44,45] versus sediments [46] for example. Across all studies, eKOI identified significant eukaryotic microbial diversity, including Amoebozoa, Chlorophyta, Choanoflagellata, and Picozoa, which constituted a substantial portion of ASVs (Fig. 2 and 'eKOI_MIDORI2_comparison.xlsx' in supplementary data). Notably, this diversity revealed a substantial underestimation of protist diversity and highlighted previously overlooked taxa, such as choanoflagellates and Picozoa (Fig. 3 and Fig. S3), which were not recovered using MIDORI2 (Fig. 2 and 'eKOI_MIDORI2_comparison.xlsx' in supplementary data).

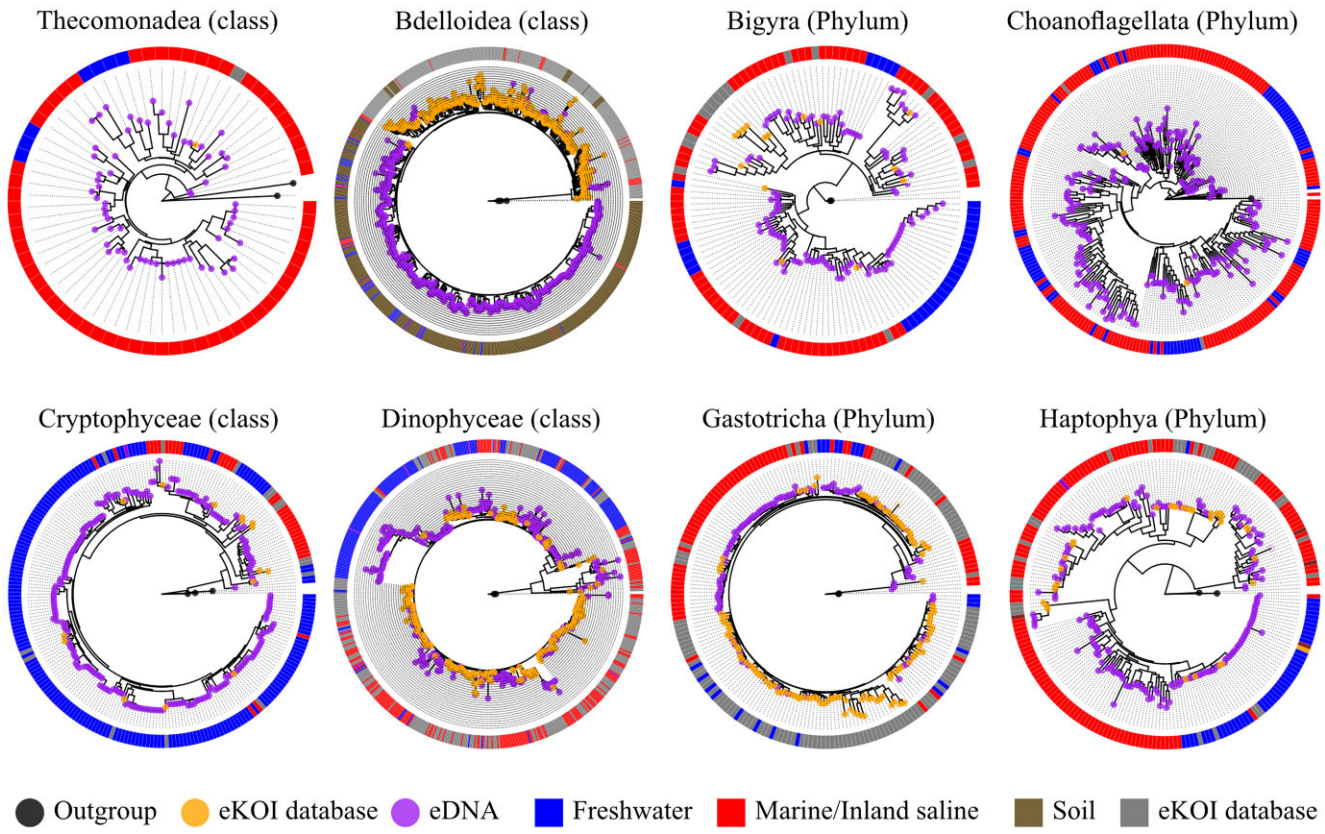

Figure 3. Phylogenetic trees obtained by combining sequences from the eKOI database of different taxonomic groups (orange dot), with outgroups (grey dot) and eDNA OTUs (purple dot). OTUs were obtained from similarity clustering of taxonomically reannotated ASVs from metabarcoding studies (Table 1). The circle surrounding each phylogenetic tree represents the environment of each OTU, sequences from the eKOI database are represented in grey, with the ecology not represented to reflect the newly characterized biodiversity for each taxonomic group.

Instead, MIDORI2 database assigned a higher representation of metazoans, unlike eKOI (Fig. S4).

The utility of ASVs derived from eDNA samples extends beyond characterizing novel molecular biodiversity. For instance, we examined the Cyphoderiidae (Rhizaria), a group in which ecological transitions across the salinity barrier have influenced its evolutionary history [47]. While these transitions were primarily characterized using the 18S marker, the COI database lacked marine lineages that were exclusively obtained with the nuclear marker. The integration of new COI ASVs (Table 1), taxonomically classified as Cyphoderiidae using the eKOI database, enabled the characterization of marine lineages that had been previously identified with 18S, like the 'marine clade 1' [47]. This increase in the Cyphoderiidae COI sequences enhances the robustness of characterizing ecological transitions between marine and freshwater environments (Fig. 4), allowing for consistent ecological and phylogenetic patterns using both nuclear and mitochondrial markers. This approach can be applied to other difficult-to-sample microorganisms, such as free-living or parasitic taxonomic groups; for example, Conoidasida (Apicomplexa) and Filasterea (Fig. 4).

Discussion

The novel eKOI database fills a crucial taxonomic gap present for protist COI, allowing an integration of taxonomic annotation between 18S rRNA and COI metabarcoding studies. The inclusion of curated protist sequences mitigates certain limitations present in existing COI databases that primarily focus on metazoans. Reducing redundant sequences improves the application of eKOI in community-level studies of eukaryotes; however, this limits its utility for species-level assignments of metazoans. For these specific assignments, specialized databases such as BOLD [48] or MIDORI2 [27] and others focused on specific taxonomic groups, like insects [49], metazoans [26], or zooplankton [50] remain valuable for better taxonomic resolution within those clades.

The taxonomic curation and standardization of eKOI with databases such as PR^2, which focuses on the 18S ribosomal gene, enables comparison of taxonomic annotation results. Currently, most community-level [51] or protist-focused [52] metabarcoding studies rely on 18S rRNA. However, a notable limitation of 18S rRNA is its phylogenetic resolution at lower taxonomic levels, due to its slow mutation rate compared to mitochondrial genes. While effective for inferring relationships at deeper taxonomic scales, its utility diminishes for species or intraspecific level distinctions [12,33], with some exceptions [53,54]. To solve this problem, fast-evolving genes such as COI provide species-level resolution [21,55], but can present challenges for characterizing novel divergent eukaryotic lineages due to high sequence divergence from existing database entries. The eKOI database comprises sequences for the entire COI gene for nearly all known eukaryotic phyla (Fig. 1), but lacks some phyla,

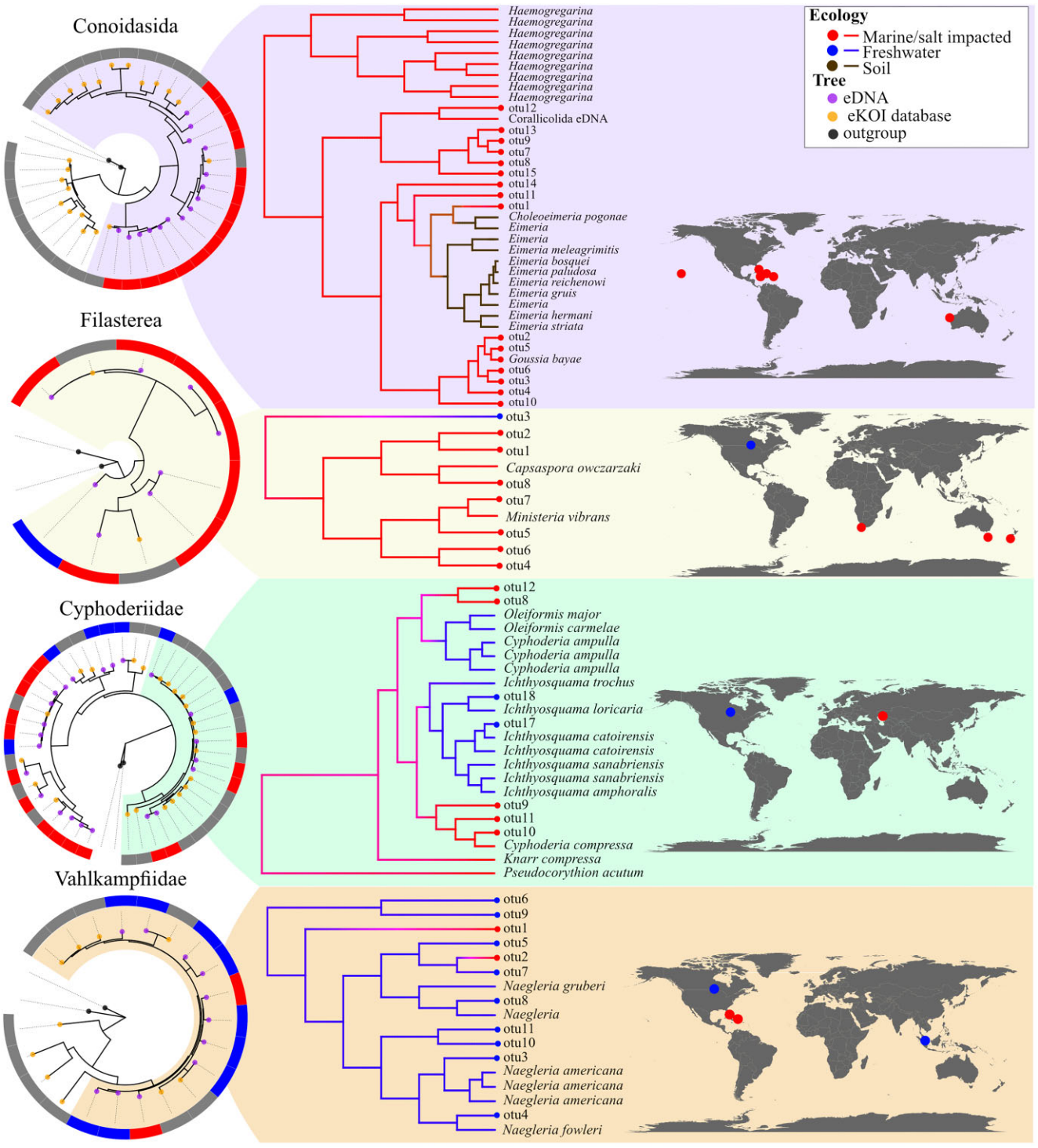

Figure 4. Phylogenetic trees obtained by combining sequences from the eKOI database of each taxonomic group (orange circles) and with outgroups (black circles). The eDNA OTUs (purple circles) were calculated and generated independently for each taxonomic group, with the IDs shared across groups. The circle surrounding each phylogenetic tree represents the habitat of each OTU, sequences from the eKOI database are represented in grey, to reflect the newly characterized biodiversity for each taxonomic group. From each taxonomic group, a part of the tree was selected from which the ancestral habitat was reconstructed, as an example. We also represented with dots on the map (blue for freshwater, brown for soil, and red for marine) the origin of the different OTUs.

like Dimorpha, Hemimastigophora, or Telonemia. Therefore, we propose that combining these independent nuclear and mitochondrial molecular markers could be ideal to uncover hidden patterns that may not be detected when relying on a single marker alone. Integrating mitochondrial eKOI (COI) and nuclear PR^2 (18S) databases will provide new perspectives for hypothesis testing using eDNA.

Another key aim of eKOI is to facilitate the development of taxonomic-specific primers, similar to what PR^2 [56] allows for 18S. This will allow applications of metabarcod-

ing to reach beyond community-level analyses, enabling targeted protocols for specific taxonomic groups. Two examples are the use of COI metabarcoding for Arcellinida [33] and Foraminifera [11]. This enables applied studies, such as the development of bioindicators [57] and ecological assessments [22], to be focused on specific taxonomic groups using the COI gene. Furthermore, the integration of eKOI with eDNA-derived ASVs could offer a valuable tool for inferring biogeographical, diversification, or ecological patterns in future metabarcoding studies. While Fig. 4 illustrates this potential, the current sample size is too limited for robust inferences. These examples illustrate the potential of developing specific metabarcoding protocols targeting understudied protist groups, often hindered by sampling difficulties or the impossibility to culture some lineages.

Overall, the eKOI database represents a significant advancement for COI-based metabarcoding, particularly in the realm of protist diversity. eKOI provides a curated, comprehensive, and eukaryote-wide resource with a focus on previously underrepresented eukaryotic groups, not only addressing limitations of existing COI resources biased towards metazoans but also facilitating direct comparisons with 18S rRNA datasets. This integration paves the way for a deeper understanding of eukaryotic community structure and function in environmental samples. The potential for developing targeted primers and uncovering hidden biogeographical and ecological patterns further solidifies eKOI as a valuable tool for future research in protist diversity and ecology. Moreover, the integration of eKOI with the next version of the widely used PR^2 database will enhance its long-term viability and facilitate regular updates, thereby increasing its applicability in future studies.

Acknowledgements

The views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. We would also like to thank the Multicellgenome Lab team for their valuable discussions and support.

Supplementary data

Supplementary data is available at *Database* online.

Conflict of interest: The authors declare no competing interests.

Funding

This project has been funded by the European Research Council (ERC, MISSINGRELATIVES, 101097659) and the European Union's Horizon 2020 research and innovation programme (grant agreement number 949745). We also acknowledge support from the Departament de Recerca i Universitats de la Generalitat de Catalunya (exp. 2021 SGR 00751), Juan de la cierva MICIU/AEI/10.13039/501100011033 JDC2023-050439-I funded by the Spanish Ministry of Science and Innovation, and by PIE-202120E047 and PID2023-152955NA-I00- Conexiones-Life.

Data availability

The resulting eKOI taxonomic database, scripts, and raw data, such as alignments and phylogenetic trees generated in this study have been deposited in GitHub (https://github.com/rubenmiguens/eKOI_taxonomy_database.git) and figshare (10.1101/2024.12.05.626972), in the supplementary materials and PR^2 database.

References

1. Rondon MR, August PR, Bettermann AD *et al.* Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* 2000;**66**:2541–47. https://doi.org/10.1128/AEM.66.6.2541-2547.2000
2. Taberlet P, Coissac E, Hajibabaei M *et al.* Environmental DNA. *Mol Ecol* 2012;**21**:1789–93. https://doi.org/10.1111/j.1365-294X.2012.05542.x
3. Ogram A, Sayler GS, Barkay T. The extraction and purification of microbial DNA from sediments. *J Microbiol Methods* 1987;**7**:57–66. https://doi.org/10.1016/0167-7012(87)90025-X
4. Pawlowski J, Bonin A, Boyer F *et al.* Environmental DNA for biomonitoring. *Mol Ecol* 2021;**30**:2931–36. https://doi.org/10.1111/mec.16023
5. de Vargas C, Audic S, Henry N *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* 2015;**348**:1261605. https://doi.org/10.1126/science.1261605
6. Vernette C, Henry N, Lecubin J *et al.* The Ocean Barcode Atlas: a web service to explore the biodiversity and biogeography of marine organisms. *Mol Ecol Resour* 2021;**21**:1347–58. https://doi.org/10.1111/1755-0998.13322
7. Burki F, Sandin MM, Jamy M. Diversity and ecology of protists revealed by metabarcoding. *Curr Biol* 2021;**31**:R1267–80. https://doi.org/10.1016/j.cub.2021.07.066
8. Deiner K, Bik HM, Mächler E *et al.* Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol Ecol* 2017;**26**:5872–95. https://doi.org/10.1111/mec.14350
9. Guillou L, Bachar D, Audic S *et al.* The Protist Ribosomal Reference database (PR^2): a catalog of unicellular eukaryote small subunit rRNA sequences with curated taxonomy. *Nucleic Acids Res* 2013;**41**:D597–604. https://doi.org/10.1093/nar/gks1160
10. Quast C, Pruesse E, Yilmaz P *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;**41**:D590–96. https://doi.org/10.1093/nar/gks1219
11. Girard EB, Langerak A, Jompa J *et al.* Mitochondrial cytochrome oxidase subunit 1: a promising molecular marker for species identification in foraminifera. *Front Mar Sci* 2022;**9**. https://doi.org/10.3389/fmars.2022.809659
12. Lara E, Singer D, Geisen S. Discrepancies between prokaryotes and eukaryotes need to be considered in soil DNA-based studies. *Environ Microbiol* 2022;**24**:11.
13. Op De Beeck M, Lievens B, Busschaert P *et al.* Comparison and validation of some ITS primer pairs useful for fungal metabarcoding studies. *PLoS One* 2014;**9**:e97629. https://doi.org/10.1371/journal.pone.0097629
14. Apothéloz-Perret-Gentil L, Bouchez A, Cordier T *et al.* Monitoring the ecological status of rivers with diatom eDNA metabarcoding: a comparison of taxonomic markers and analytical approaches for the inference of a molecular diatom index. *Mol Ecol* 2021;**30**:2959–68. https://doi.org/10.1111/mec.15646
15. Pichard SL, Campbell L, Paul JH. Diversity of the ribulose bisphosphate carboxylase/oxygenase form I gene (rbcL) in natural phytoplankton communities. *Appl Environ Microbiol* 1997;**63**:3600–3606. https://doi.org/10.1128/aem.63.9.3600-3606.1997
16. Bell KL, Loeffler VM, Brosi BJ. An rbcL reference library to aid in the identification of plant species mixtures by DNA metabarcod-

ing. *Appl Plant Sci* 2017;**5**:1600110. https://doi.org/10.3732/apps.1600110
17. Větrovský T, Morais D, Kohout P *et al.* GlobalFungi, a global database of fungal occurrences from high-throughput-sequencing metabarcoding studies. *Sci Data* 2020;**7**:228.
18. Tedersoo L, Hosseyni Moghaddam MS, Mikryukov V *et al.* EUKARYOME: the rRNA gene reference database for identification of all eukaryotes. *Database* 2024;**2024**:baae043. https://doi.org/10.1093/database/baae043
19. Hebert PDN, Ratnasingham S, de Waard JR. Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proc R Soc Lond B* 2003;**270**. https://doi.org/10.1098/rsbl.2003.0025
20. González-Miguéns R, Soler-Zamora C, Villar-Depablo M *et al.* Multiple convergences in the evolutionary history of the testate amoeba family Arcellidae (Amoebozoa: arcellinida: sphaerothecina): when the ecology rules the morphology. *Zool J Linnean Soc* 2022;**194**:1044–71. https://doi.org/10.1093/zoolinnean/zlab074
21. Kosakyan A, Lahr DJG, Mulot M *et al.* Phylogenetic reconstruction based on COI reshuffles the taxonomy of hyalosphenid shelled (testate) amoebae and reveals the convoluted evolution of shell plate shapes. *Cladistics* 2016;**32**:606–23. https://doi.org/10.1111/cla.12167
22. Girard EB, Macher J-N, Jompa J *et al.* COI metabarcoding of large benthic Foraminifera: method validation for application in ecological studies. *Ecol Evol* 2022;**12**:e9549. https://doi.org/10.1002/ece3.9549
23. Hagino K, Bendif EM, Young JR *et al.* New evidence for morphological and genetic variation in the cosmopolitan coccolithophore *Emiliania huxleyi* (prymnesiophyceae) from the cox1b-atp4 genes. *J Phycol* 2011;**47**:1164–76. https://doi.org/10.1111/j.1529-8817.2011.01053.x
24. Moniz MBJ, Kaczmarska I. Barcoding diatoms: is there a good marker? *Mol Ecol Resour* 2009;**9**:65–74. https://doi.org/10.1111/j.1755-0998.2009.02633.x
25. Ratnasingham S, Hebert PDN. bold: the Barcode of Life Data System (http://www.barcodinglife.org). *Mol Ecol Notes* 2007;**7**:355–64. https://doi.org/10.1111/j.1471-8286.2007.01678.x
26. Heller P, Casaletto J, Ruiz G *et al.* A database of metazoan cytochrome c oxidase subunit I gene sequences derived from GenBank with CO-ARBitrator. *Sci Data* 2018;**5**:180156. https://doi.org/10.1038/sdata.2018.156
27. Leray M, Knowlton N, Machida RJ. MIDORI2: a collection of quality controlled, preformatted, and regularly updated reference databases for taxonomic assignment of eukaryotic mitochondrial sequences. *Environ DNA* 2022;**4**:894–907. https://doi.org/10.1002/edn3.303
28. Burki F, Roger AJ, Brown MW *et al.* The new tree of eukaryotes. *Trends Ecol Evol* 2020;**35**:43–55.
29. Rognes T, Flouri T, Nichols B *et al.* VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 2016;**4**:e2584. https://doi.org/10.7717/peerj.2584
30. Katoh K, Misawa K, Kuma K *et al.* MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;**30**:3059–66. https://doi.org/10.1093/nar/gkf436
31. Andújar C, Creedy TJ, Arribas P *et al.* Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. *Mol Ecol Resour* 2021;**21**:1772–87.
32. del Campo J, Kolisko M, Boscaro V *et al.* EukRef: phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. *PLoS Biol* 2018;**16**:e2005849.
33. González-Miguéns R, Cano E, Guillén-Oterino A *et al.* A needle in a haystack: a new metabarcoding approach to survey diversity at the species level of Arcellinida (Amoebozoa: tubulinea). *Mol Ecol Resour* 2023;**23**:1034–49. https://doi.org/10.1111/1755-0998.13771
34. Callahan BJ, McMurdie PJ, Rosen MJ *et al.* DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016;**13**:581–83. https://doi.org/10.1038/nmeth.3869
35. Kalyaanamoorthy S, Minh BQ, Wong TKF *et al.* ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017;**14**:587–89. https://doi.org/10.1038/nmeth.4285
36. Hoang DT, Chernomor O, von Haeseler A *et al.* UFBoot2: improving the Ultrafast Bootstrap approximation. *Mol Biol Evol* 2018;**35**:518–22. https://doi.org/10.1093/molbev/msx281
37. Yu G, Smith DK, Zhu H *et al.* ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017;**8**:28–36. https://doi.org/10.1111/2041-210X.12628
38. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 2012;**3**:217–23. https://doi.org/10.1111/j.2041-210X.2011.00169.x
39. Bollback JP. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinf* 2006;**7**:88. https://doi.org/10.1186/1471-2105-7-88
40. Folmer O, Black M, Hoeh W *et al.* DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotechnol* 1994;**3**:294–99.
41. Geller J, Meyer C, Parker M *et al.* Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Mol Ecol Resour* 2013;**13**:851–61. https://doi.org/10.1111/1755-0998.12138
42. Leray M, Yang JY, Meyer CP *et al.* 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. *Front Zool* 2013;**10**:34. https://doi.org/10.1186/1742-9994-10-34
43. Wangensteen OS, Palacín C, Guardiola M *et al.* DNA metabarcoding of littoral hard-bottom communities: high diversity and database gaps revealed by two molecular markers. *PeerJ* 2018;**6**:e4705. https://doi.org/10.7717/peerj.4705
44. Bakker J, Wangensteen OS, Baillie C *et al.* Biodiversity assessment of tropical shelf eukaryotic communities via pelagic eDNA metabarcoding. *Ecol Evol* 2019;**9**:14341–55. https://doi.org/10.1002/ece3.5871
45. Suter L, Polanowski AM, Clarke LJ *et al.* Capturing open ocean biodiversity: comparing environmental DNA metabarcoding to the continuous plankton recorder. *Mol Ecol* 2021;**30**:3140–57. https://doi.org/10.1111/mec.15587
46. Koziol A, Stat M, Simpson T *et al.* Environmental DNA metabarcoding studies are critically affected by substrate selection. *Mol Ecol Resour* 2019;**19**:366–76. https://doi.org/10.1111/1755-0998.12971
47. González-Miguéns R, Soler-Zamora C, Useros F *et al.* *Cyphoderia ampulla* (Cyphoderiidae: rhizaria), a tale of freshwater sailors: the causes and consequences of ecological transitions through the salinity barrier in a family of benthic protists. *Mol Ecol* 2022;**31**:2644–63. https://doi.org/10.1111/mec.16424
48. Ratnasingham S, Hebert PDN. A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS One* 2013;**8**:e66213. https://doi.org/10.1371/journal.pone.0066213
49. Magoga G, Forni G, Brunetti M *et al.* Curation of a reference database of COI sequences for insect identification through DNA metabarcoding: cOins. *Database* 2022;**2022**:baac055. https://doi.org/10.1093/database/baac055
50. Bucklin A, Peijnenburg KTCA, Kosobokova KN *et al.* Toward a global reference database of COI barcodes for marine zooplankton. *Mar Biol* 2021;**168**:78. https://doi.org/10.1007/s00227-021-03887-y
51. Oliverio AM, Geisen S, Delgado-Baquerizo M *et al.* The global-scale distributions of soil protists and their contributions to belowground systems. *Sci Adv* 2020;**6**:eaax8787. https://doi.org/10.1126/sciadv.aax8787

52. Segawa T, Matsuzaki R, Takeuchi N *et al.* Bipolar dispersal of red-snow algae. *Nat Commun* 2018;**9**:3094. https://doi.org/10.1038/s41467-018-05521-w
53. Ribeiro CG, Lopes dos Santos A, Trefault N *et al.* Arctic phytoplankton microdiversity across the marginal ice zone: subspecies vulnerability to sea-ice loss. *Element Sci Anthropocene* 2024;**12**:00109.
54. Tragin M, Vaulot D. Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. *Sci Rep* 2019;**9**:5190. https://doi.org/10.1038/s41598-019-41680-6
55. Pinseel E, Janssens SB, Verleyen E *et al.* Global radiation in a rare biosphere soil diatom. *Nat Commun* 2020;**11**:2382. https://doi.org/10.1038/s41467-020-16181-0
56. Vaulot D, Geisen S, Mahé F *et al.* pr2-primers: an 18S rRNA primer database for protists. *Mol Ecol Resour* 2022;**22**:168–79. https://doi.org/10.1111/1755-0998.13465
57. González-Miguéns R, Cano E, García-Gallo Pinto M *et al.* The voice of the little giants: Arcellinida testate amoebae in environmental DNA-based bioindication, from taxonomy free to haplotypic level. *Mol Ecol Resour* 2024;**24**:e13999. https://doi.org/10.1111/1755-0998.13999
58. Drummond AJ, Newcomb RD, Buckley TR *et al.* Evaluating a multigene environmental DNA approach for biodiversity assessment. *Gigascience* 2015;**4**:46. https://doi.org/10.1186/s13742-015-0086-1
59. Jeunen G, Knapp M, Spencer HG *et al.* Environmental DNA (eDNA) metabarcoding reveals strong discrimination among diverse marine habitats connected by water movement. *Mol Ecol Resour* 2019;**19**:426–38. https://doi.org/10.1111/1755-0998.12982
60. Rossouw EI, Landschoff J, Ndhlovu A *et al.* Detecting kelp-forest associated metazoan biodiversity with eDNA metabarcoding. *npj Biodivers* 2024;**3**:1–8. https://doi.org/10.1038/s44185-023-00033-3
61. Lim NKM, Tay YC, Srivathsan A *et al.* Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities. *R Soc Open Sci* 2016;**3**:160635. https://doi.org/10.1098/rsos.160635
62. Stat M, Huggett MJ, Bernasconi R *et al.* Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Sci Rep* 2017;**7**:12240. https://doi.org/10.1038/s41598-017-12501-5
63. Littlefair JE, Hleap JS, Palace V *et al.* Freshwater connectivity transforms spatially integrated signals of biodiversity. *Proc R Soc B Biol Sci* 2023;**290**:20230841. https://doi.org/10.1098/rspb.2023.0841
64. Jacobs-Palmer E, Gallego R, Ramón-Laca A *et al.* A halo of reduced dinoflagellate abundances in and around eelgrass beds. *PeerJ* 2020;**8**:e8869. https://doi.org/10.7717/peerj.8869
65. Brantschen J, Pellissier L, Walser J-C *et al.* Evaluation of primer pairs for eDNA-based assessment of Ephemeroptera, Plecoptera, and Trichoptera across a biogeographically diverse region. *Environ DNA* 2022;**4**:1356–68. https://doi.org/10.1002/edn3.342
66. Leese F, Sander M, Buchner D *et al.* Improved freshwater macroinvertebrate detection from environmental DNA through minimized nontarget amplification. *Environ DNA* 2021;**3**:261–76. https://doi.org/10.1002/edn3.177
67. Vamos E, Elbrecht V, Leese F. Short COI markers for freshwater macroinvertebrate metabarcoding. *Metabarcod Metagenomics* 2017;**1**:e14625. https://doi.org/10.3897/mbmg.1.14625
68. Deiner K, Walser J-C, Mächler E *et al.* Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biol Conserv* 2015;**183**:53–63. https://doi.org/10.1016/j.biocon.2014.11.018
69. Nichols PK, Timmers M, Marko PB. Hide 'n seq: direct versus indirect metabarcoding of coral reef cryptic communities. *Environ DNA* 2022;**4**:93–107. https://doi.org/10.1002/edn3.203

Received 28 March 2025; Revised 17 June 2025; Accepted 5 August 2025