COInr and mkCOInr : Building and customizing a nonredundant barcoding reference database from BOLD and NCBI using a semi-automated pipeline

Emese Meglécz

Received: 20 May 2022 | Revised: 30 November 2022 | Accepted: 12 January 2023

DOI: 10.1111/1755-0998.13756

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

COInr and mkCOInr: Building and customizing a nonredundant barcoding reference database from BOLD and NCBI using a semi-automated pipeline

Emese Meglécz

Aix Marseille Univ, Avignon Univ, CNRS, IRD, IMBE, Marseille, France

Correspondence
Emese Meglécz, Aix Marseille Univ, Avignon Univ, CNRS, IRD, IMBE, Chemin de la batterie des Lions, 13007 Marseille, France.
Email: emese.meglecz@imbe.fr

Handling Editor: Simon Creer

Abstract

Reference databases with wide taxonomic coverage are greatly needed in many fields of biology, most particularly for the taxonomic assignment of metabarcoding sequences. Therefore, it is fundamental to be able to access and pool data from different primary databases. The COInr database is a freely available, easy-to-access database of COI reference sequences extracted from the BOLD and NCBI nucleotide databases. It is a comprehensive database: not limited to a taxon, a gene region or a taxonomic rank; therefore, it is a good starting point for creating custom databases. Sequences are dereplicated between databases and within taxa. Each taxon has a unique taxonomic identifier (taxID), fundamental to avoid ambiguous associations of homonyms and synonyms in the source database. TaxIDs form a coherent hierarchical system fully compatible with the NCBI taxIDs, allowing their full or ranked lineages to be created. The MKCOINR tool is a series of Perl scripts designed to download sequences from BOLD and NCBI, to build the COInr database and to customize it according to the users' needs. It is possible to select or eliminate sequences for a list of taxa, select a specific gene region, select for minimum taxonomic resolution, add new custom sequences, and format the database for BLAST, VTAM, QIIME and RDP CLASSIFIER. This is a semi-automated pipeline using command lines in a Linux environment. The COInr database can be downloaded from https://doi.org/10.5281/zenodo.6555985 and MKCOINR and its full documentation is available at https://github.com/meglecz/mkCOInr.

KEYWORDS
COI, database, download, metabarcoding, taxID, taxonomic assignment

1 | INTRODUCTION

Reference databases of particular genes or markers are used for DNA-based identification and thus have various applications in mitogenomics (Ho & Gilbert, 2010), metagenomics (Santamaria et al., 2012), phylogenetics (Khater et al., 2021; Slater-Baker et al., 2022; Vijapure et al., 2019), identification of unknown barcodes (Hebert, Cywinska, et al., 2003; Hebert, Ratnasingham, & deWaard, 2003) and most particularly for metabarcoding. The use of metabarcoding has increased dramatically in the past decade due to

technological advances, and the continuous reduction in sequencing costs has made it accessible for a wide range of studies (Slatko et al., 2018). Metabarcoding is applied mainly for biodiversity assessment, but it can be used in other fields such as studying interaction networks or understanding animal diets (Compson et al., 2020)

One of the difficulties of metabarcoding lies in the taxonomic assignment of sequences and the completeness of the underlying reference databases. Methods of taxonomic assignment can be alignment-based, relying of sequence similarities detected by BLAST (Altschul et al., 1997) or VSEARCH (Rognes et al., 2016) implemented in different software (Bokulich et al., 2018; Huson et al., 2007) or based on machine learning (Murali et al., 2018; Pedregosa et al., 2011; Wang et al., 2007). However, for all methods, the quality of the reference database is crucial (Hleap et al., 2021). Many methods are sensitive to gaps in the taxonomic coverage of the reference database (Hleap et al., 2021), and thus the creation of a reference database with the best coverage available is greatly needed.

Several different markers can be used for metabarcoding, since each of them are subject to different taxonomic biases and provide different taxonomic resolution (Ruppert et al., 2019). The most widespread markers are the ribosomal RNA (rRNA) genes (18 S, 28 S, 16 S), the Cytochrome Oxidase C subunit I (COI) gene and internal transcribed spacer sequences (ITS) (Creer et al., 2016; Porter & Hajibabaei, 2020). rRNA genes allow amplification from a wide range of taxa, and are the most widely used markers for microorganisms (Creer et al., 2016). The choice of the ideal marker is more difficult when dealing with Eukaryotes. Plants and fungal studies most often use ITS markers or rbcL, since the COI gene often contains indels of variable size and location and is not sufficiently variable in these groups. In addition, the taxonomic resolution of plant and fungal rRNA genes is relatively low (Bruns et al., 1991; Yao et al., 2010). For animals, the use of both rRNA genes and COI sequences is widespread (Creer et al., 2016). The COI marker has been proposed as the marker of choice for animals and it is one of the most widely sequenced genes (Porter & Hajibabaei, 2018b), since it is the main maker of the Barcode of Life database (Hebert, Cywinska, et al., 2003; Hebert, Ratnasingham, & deWaard, 2003). Although it has become clear that the COI gene or any of its fragment is not sufficient to differentiate species in some groups, most particularly Diptera (Meier et al., 2006; Roe & Sperling, 2007; Rubinoff et al., 2006), it is still a frequently used marker, mostly because more animal taxa have been barcoded with COI than with any other marker (Andújar et al., 2018).

Regularly updated, curated and marker-specific databases are available for ITS—UNITE (Rolf Henrik Nilsson et al., 2019), PLANTiTS (Banchi et al., 2020)—and for rRNA genes—Greengenes (DeSantis et al., 2006), SILVA (Pruesse et al., 2007). Conversely, COI sequences are deposited in two different major databases, which are not COI-specific: (i) the nucleotide database of NCBI (hereafter NCBI-nt database; Sayers et al., 2022) and their European (ENA) and Japanese equivalents (DDBJ) are generalist databases without focusing on a taxon or a gene; and (ii) the Barcoding of Life Data System (BOLD; Ratnasingham & Hebert, 2007) contains barcoding sequences of several markers, but most of the sequences are from the barcoding fragment of the COI gene. Although the overlap of data between these databases is considerable, each of them has sequences that are not found in the other database. Therefore, creating a merged database with sequences from both sources is highly desirable. Most existing COI databases are sourced from NCBI-nt (e.g., Bengtsson-Palme et al., 2018; Curd et al., 2019; Keller et al., 2020; Richardson et al., 2020) and only a few of them combine sequences from BOLD and NCBI (Arranz et al., 2020; Balech et al., 2022; Macher et al., 2017; Porter & Hajibabaei, 2018a).

A major challenge of pooling sequences from different sources into a single database is to reconcile their taxonomic lineages. This step is not trivial due to the presence of homonyms (e.g., Plecoptera is both an insect order and a moth genus), synonyms and misspellings. Therefore, the only clean solution to deal with taxon names is the use of unique taxonomic identifiers (taxID) that are connected to a nonambiguous, hierarchical system and allow the identification of the lineage for each taxon. Both the NCBI-nt and the BOLD databases use taxIDs, but the two systems are independent of each other, and thus they cannot be simply merged. Finding the equivalent taxon names and taxIDs between the two databases calls for a careful comparison of taxon names and their lineages in order to match them. However, a further complication arises from occasional inconsistencies of taxonomic lineages from different databases (e.g., the genus *Vexillata* is a nematode belonging to the family Ornithostrongylidae according to BOLD, but to the family Trichostrongylidae according to NCBI taxonomy), which further complicates pooling of taxonomic information into a single coherent system.

Merging of COI sequences from the NCBI-nt and BOLD has been attempted in different programs. BOLD_NCBI_MERGER (Macher et al., 2017) uses a simple method based on identical taxon names. METACOXI (Balech et al., 2022) obtains NCBI taxIDs and taxonomic lineages based on ENA flat files, when available. However, when this information is not available (the sequence is present only in BOLD), NCBI taxIDs are determined by simply matching taxon names to the NCBI taxonomy, without checking for homonymy. Furthermore, taxon names not present in the NCBI taxonomy do not receive a taxID, and therefore a taxID system is incomplete.

A further difficulty of creating custom (local) databases is the download of sequences from the original sources. NCBI provides different means of accessing data: a whole database can be downloaded via ftp sites, and filtered subsequently, or Application Programming Interfaces (APIs) are provided for targeted downloads (Kans, 2021). On the other hand, BOLD systems do not provide an easy way to download the whole public data set, and the use of BOLD APIs needs considerable optimization to be able to access large data sets. Although the BOLD R package (https://docs.ropensci.org/bold/) is available to download data from BOLD, it is subject to failure for large taxa and takes several hours or days, according to requested data size.

The MKCOINR tool was designed to create the COInr database, which includes all COI sequences from NCBI-nt and BOLD sequences, irrespective of the region of the gene covered and the taxonomic group.

All sequences have a taxID, and all taxIDs form a coherent system compatible with, but not limited to, the NCBI taxIDs, allowing the user to unambiguously obtain taxonomic lineages even for taxon names with homonyms. Sequence redundancy within taxa is eliminated to reduce database size, without losing information. This database is freely available and can be easily and quickly downloaded from Zenodo (https://zenodo.org/record/6555985; Meglécz, 2022a, 2022b), thus saving the most complicated and time-consuming steps of custom database creation. Users can customize the downloaded database using MKCOINR scripts and format them to be able to use it with their preferred taxonomic assignment tool. It is a semi-automated pipeline using command lines in a Linux environment. It is possible to add local sequences, select or eliminate sequences of a list of taxa, filter sequences for minimum taxonomic resolution, and choose a gene region. The COInr database is planned to be updated annually, but all scripts are available with detailed documentation to re-create it at any time or produce a different database by modifying some of the filtering options.

2 | MATERIAL AND METHODS

MKCOINR is a series of Perl scripts that can be executed in command line, thus being easily integrated into other pipelines. The scripts were written for Linux OS and can run on MacOS or other Unix environments. The Windows Subsystem Linux (https://docs.microsoft.com/en-us/windows/wsl/) allows Windows users to run MKCOINR scripts. Particular care was taken to reduce dependencies to easy-to-instal, third-party programs without the use of special packages. BLAST (Altschul et al., 1997), VSEARCH (Rognes et al., 2016), CUTADAPT (Martin, 2011) and NSDPY (Hebert & Meglécz, 2022) can all be installed either through the Python Package Index (PyPI) or standard program repositories.

Figure 1 represents a complete flowchart of the pipeline. A tutorial and detailed documentation is available at https://github.com/meglecz/mkCOInr (Meglécz, 2022a, 2022b).

2.1 | Construction of the COInr database

2.1.1 | NCBI

NCBI sequences were downloaded with the NSDPY (Hebert & Meglécz, 2022) python package using the following request:

```
nsdpy -r "COI OR COX1 OR CO1 OR COXI OR
(complete[Title] AND genome[Title] AND
Mitochondrion[Filter])" -T -v --cds
```

This allowed the download of all coding DNA sequences (CDS) returned with the keyword search for COI, CO1, COXI or COX1, and CDS from complete mitochondrial genomes. The scope of this search was intentionally very wide, and the downloaded sequences were further filtered by the *format_ncbi.pl* script to (i) only retain CDS with gene and protein names corresponding to COI, and (ii) eliminate genes with introns and sequences from environmental or metagenomic samples. Sequences with more than five consecutive internal Ns, and outside of the length range of 100–2000 nucleotides were also eliminated. Open nomenclature was not accepted in taxon names. If the taxID did not correspond to a correct Latin name format, the smallest taxon with a correct Latin name in the lineage was chosen for the sequence (e.g., *Acentrella* sp. AMI 1, taxID: 888165, rank: species was replaced by *Acentrella*, taxID: 248176, rank: genus). Sequences were then subjected to taxonomically aware dereplication by the *dereplicate.pl* script. Within each taxID, all sequences that were a substring of another sequence were eliminated. This allows the size of the database to be reduced without losing information and keeping intraspecific variability.

2.1.2 | BOLD

A list of taxa was established from the taxonomy page of BOLD Systems (https://www.boldsystems.org/index.php/TaxBrowser_Home), where each taxon had fewer than 500,000 specimen records. All public sequences of the above list and associated information were downloaded from BOLD, using the *download_bold.pl* script that uses the BOLD APIs. For each taxon, the integrity of the downloaded files and the number of records were checked, and the download was repeated automatically in case of failure. From the raw downloaded files, COI sequences (COI-5P, COI-3P) were selected if they did not contain more than five consecutive internal Ns and were in the length range of 100–2000 nucleotides. As for NCBI sequences, the smallest taxon in the BOLD lineage with a correct Latin name was chosen for the sequence to avoid open nomenclature. All unique lineages were then listed with the corresponding sequence identifiers (sequenceID) and for each lineage a taxID was determined using the *add_taxids.pl* script: the smallest taxon is identified in each BOLD lineage, where the name matches a taxon name in the NCBI taxonomy database (including synonyms), and at least 60% of the taxon names in the BOLD lineage match the NCBI lineage. For example, for the BOLD lineage "Chordata, Actinopterygii, Trachiniformes, Pinguipedidae, *Parapercis*, *Parapercis somaliensis*," the genus *Parapercis* matches the 215,380 NCBI taxID, even if the orders are different in BOLD and NCBI (Trachiniformes and Uranoscopiformes, respectively). In the next step, a taxon under the smallest taxon with NCBI taxID was attributed to an arbitrary, negative taxID, and the new taxID was integrated into the taxID system, with the NCBI taxID as a parent. The newly created taxID was then added to the taxID system and it was characterized by a taxon name, a taxonomic rank and the taxID of its direct parent, forming a hierarchical system. This hierarchical taxID system allows the creation of the lineage of any taxID unambiguously, even in the case of homonymy and synonymy. As for NCBI sequences, the filtered BOLD data set was dereplicated by the *dereplicate.pl* script.

To compare the effect of using only correct Latin names (as in COInr) or accepting all taxon names present in the input databases, the above pipeline was run a second time using systematically the

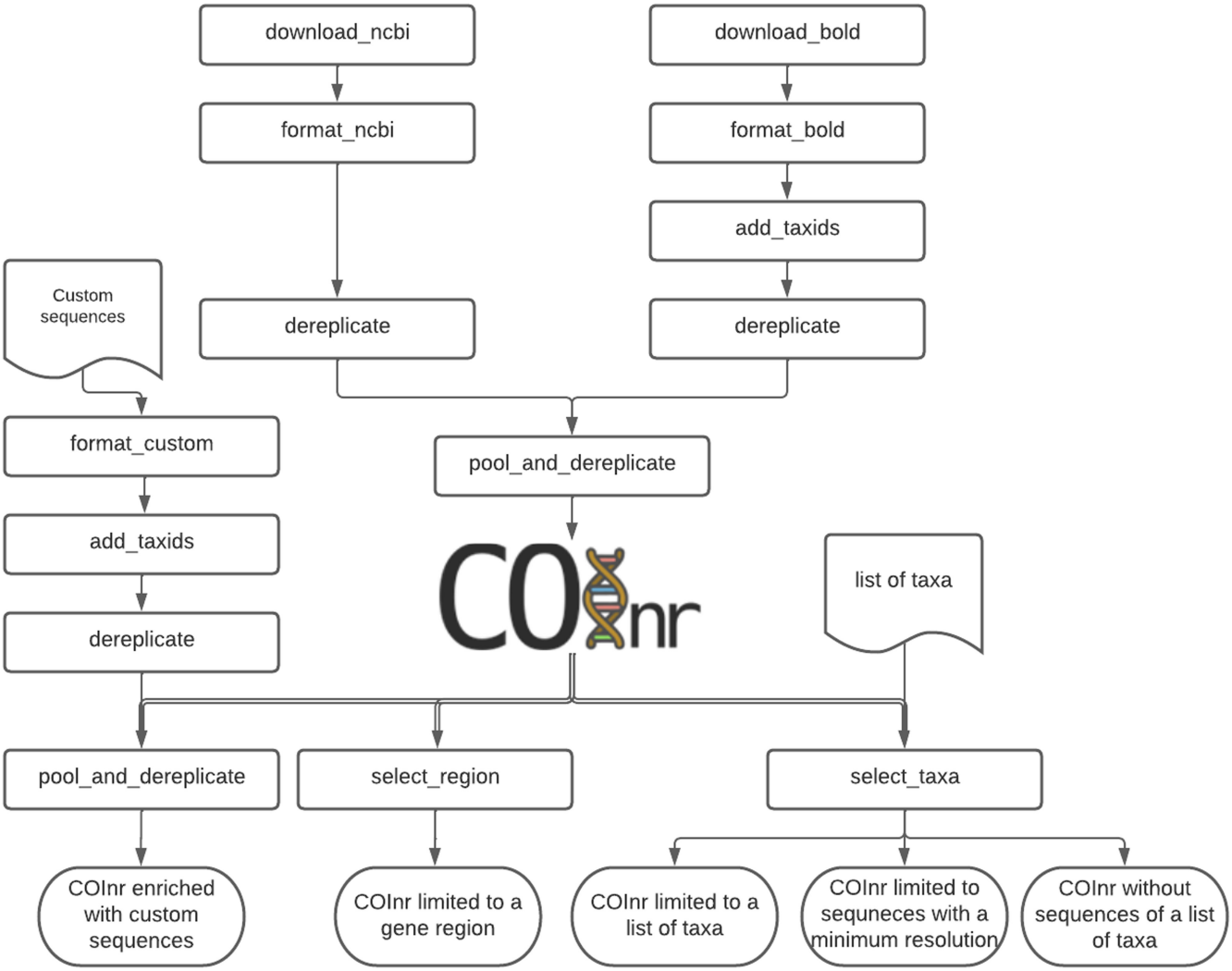

FIGURE 1 Flowchart of MKCOINR. Double lines represent the different options for customizing the COInr database. These steps can also be consecutive

smallest taxon in each lineage, even if it did not correspond to a correct Latin name.

2.2 | The COInr database

The BOLD and NCBI data sets were pooled into one single data set by the *pool_and_dereplicate.pl* script, where sequences for the taxIDs shared by the two source databases were dereplicated, while sequences from taxIDs unique to one of the sources were simply added to the combined database. This database is a starting point to create more specific custom databases according to the users' needs.

The core database consists of two simple-to-parse tsv files (tab-separated values). The sequence file has three columns (sequenceIDs, taxIDs and sequences), and contains sequences of all taxonomic groups that can cover any COI region, with variable taxonomic resolution from species to phylum level. The taxonomy file contains taxIDs, scientific names, parent taxIDs, taxonomic rank and taxonomic level index. The taxonomic level index contains integers from 0 to 8, each corresponding to a major taxonomic level (rank) similar to those used in RDP CLASSIFIER (Wang et al., 2007): root, superkingdom, kingdom, phylum, class, order, family, genus, species. Intermediate taxonomic levels have 0.5 added to the next major taxon level index (e.g., 7.5 for subgenus). This file allows the reconstruction of the complete lineages of all taxa or the ranked lineages containing only the major taxonomic ranks.

2.3 | Customizing the COInr database

The COInr database can be modified according to users' needs. Sequences can be selected for a list of taxa or, by contrast, removed from the database through the *select_taxa.pl* script. The script will also produce a lineage and a taxID for each taxon in the taxon list, allowing users to check for potential errors due to homonyms. In case of incoherence, the taxon list enriched by the correct taxIDs can be used to rerun the script with more precise selection. The same script

also allows sequences to be selected with a minimum taxonomic resolution.

The *select_region.pl* script trims the sequences to a specific region of the COI gene. Using the usearch_global command of VSEARCH (Rognes et al., 2016), sequences of the database are aligned to a small bait file, which contains a taxonomically diverse pool of sequences already trimmed to target region. The USEARCH_GLOBAL program is similar to BLAST since it aligns each query sequence (sequences of the COInr) to the sequences of its database (the bait file in this case). Contrary to BLAST, USEARCH_GLOBAL produces global alignments. The best alignment of each query sequence is used to trim the query sequence according to the alignment positions. The bait file can be provided by the users or can be produced by the same script by making an E-PCR on the core database. The E-PCR (electronic PCR) uses CUTADAPT (Martin, 2011) to select for a particular subregion of COI using a pair of primer sequences.

The COInr database can also be completed by custom sequences. Users will need a taxon name and sequenceID for each custom sequence. The *format_custom.pl* script will produce a lineage file with all input taxa, which should be checked, and eventually corrected and completed by the users. The *add_taxids.pl* script will add taxIDs to each lineage and complete the input taxonomy file (part of the COInr database). Sequences should then be dereplicated by the *dereplicate.pl* script and added to the COInr database using the *pool_and_derelplicate.pl*.

Figure 1 represents the customizing options on MKCOINR, each of them starting from the COInr database. However, the different steps can also be successive to produce a final database. For example, it is possible to start by selecting sequences for a list of taxa, then adding custom sequences to the newly created database, which in turn can be trimmed to the target region.

2.4 | Format database

The very simple format of the database (sequence file and taxonomy file both in tsv format) allows users to easily obtain a database in their desired format. The *format_db.pl* script can produce databases ready to use for BLAST (Altschul et al., 1997), VTAM (González et al., 2020), RDP CLASSIFIER (Wang et al., 2007) and QIIME (Bolyen et al., 2019). The "full" option will produce a single tsv file with sequence IDs, ranked lineages, taxIDs and the sequences, allowing user to parse, and produce basic statistics on the database content (e.g., number of sequences of each taxon).

2.5 | Benchmarking the *select_region* script

Identifying and trimming sequences to a target region is one of the most difficult steps in customizing the COInr database. It can produce false positives and false negatives if the search parameters are not set properly. The detailed protocol of the benchmarking and the associated scripts are found in the github repository of MKCOINR (https://github.com/meglecz/mkCOInr). Briefly, a positive test data set was produced by downloading CDS of all complete mitochondrial genomes from NCBI-nt and the whole COI gene sequences were identified by the *format_ncbi.pl* script. Sequences shorter than 1100 bp were filtered out to avoid using erroneously annotated incomplete sequences. Two negative data sets were also produced. The negative-mito data set was derived from the above downloaded mitogenomes. After filtering out COI genes and selecting genes with length between 700 and 2000 bp, sequences were randomly selected to match the size of the positive data set. The negative-chloroplast data set was produced similarly, but from the complete chloroplast genomes downloaded from NCBI-nt.

First, the E-PCR option of the *select_region.pl* script was tested on the three test data sets. Using this option, the bait file is produced by E-PCR, then sequences not recognized by E-PCR are aligned to this bait file to trim them. I varied the min_overlap (the minimum overlap between the primer and the sequence; 10 or 20) and the trim_error (maximum proportion of mismatch between primer and sequence; 0.2 or 0.3) for the E-PCR, and the identity parameter (minimum similarity threshold to accept a usearch_global hit; 0.6 or 0.7) for the usearch_global step.

To test how variable a bait data set should be if it is provided by the user without using the E-PCR option, bait files were produced of varying diversity. The trimmed sequences of the most reliable E-PCR-based trimming of the positive data set was used to produce the baits. I randomly sampled one or five sequences per phylum, class or order and each random sampling was repeated 10 times. Each of the resulting 60 bait files was used to trim the positive and the two negative test data sets by using two identity thresholds (0.6 and 0.7).

3 | RESULTS

Table 1 summarizes the number of taxa and sequences in the initial databases before and after taxonomically aware dereplication, and after pooling and dereplicating sequences from BOLD and NCBI-nt to the COInr database. After initial quality control, the NCBI and BOLD databases contained 3.9 million and 7.6 million COI sequences respectively, belonging to approximately 200,000 taxa with correct Latin names in both databases. Taxonomically aware dereplication within each of the source databases resulted in 1.7 million and 2.8 M nonredundant sequences, corresponding to 58% and 63% reduction in the NCBI and BOLD databases, respectively. The total number of taxa was 268,438 after pooling NCBI and BOLD, 69% of which was shared between the input databases, 14% and 17% unique to NCBI and BOLD, respectively. After pooling the databases and dereplication, 90% of the sequences were from taxa present in both databases, while 4% and 6% were specific to NCBI and BOLD, respectively. Overall, the 11.5 million input sequences were reduced to 3.3 million by eliminating redundancy between the two input databases, and within each taxon.

TABLE 1 The number of taxa and COI sequences of the input databases (NCBI-nt, BOLD), and in the COInr database (May 2022). COInr is the result of pooling and taxonomically aware dereplication of sequences in the input databases

	No. of taxIDs	No. of sequences
After initial quality control		
NCBI	221,565	3,920,624
BOLD	231,425	7,590,488
After dereplication within input database		
NCBI	221,565	1,657,602
BOLD	231,425	2,843,248
After pooling and dereplication (COInr)		
Shared by BOLD and NCBI	184,552	2,944,524
Unique to NCBI	37,013	124,811
Unique to BOLD	46,873	190,319
Total	**268,438**	**3,259,654**

Apart from sequences of animals, which comprised 99% of the database and corresponded 97% of the species, other Eukaryotes (plants, Fungi) and even some Bacteria and Archaea sequences were also present in the database (Table 2). Within Metazoa, 83% of the sequences were from Arthropoda, which corresponds to 74% of the animal species of the database.

Downloading the COI sequences from NCBI and all sequences from BOLD took approximately 24 and 40 h, respectively. It is clearly the most time-consuming step. The rest of the pipeline until obtaining the final COInr database took approximately 1 h on a desktop computer (Intel Core i7-7700, CPU @ 3.6 GHz, 32 Gb RAM). This compares to the few seconds of downloading the COInr database from Zenodo that clearly saves a tremendous amount of time for users. Creating a custom database from COInr varied in length as a function of the desired characteristics of the database. As an example, the creation of a custom database explained in the tutorial (Eukaryotes, noninsect sequences with known genera, enriched by custom sequences, limited to a 313-bp region, formatted to the RDP CLASSIFIER) takes approximately 30 min.

To evaluate the effect of using nonstandard taxon names, corresponding to open nomenclature (e.g., *Allograpta* aff. *argentipila*, *Alona guttata group*, *Macrobiotus* cf. *hufelandi*) or correct Latin names completed by arbitrary identifiers (e.g., *Macrobathra* sp. ACL2485, *Ablemma* BioLep730, *Abacarus* sp. GD111), two databases were created: COInr, where only correct Latin names were used, and the all-names database created by the same pipeline, with the exception that all taxon names were accepted regardless of their format (e.g., *Lepidoptera* sp. 096 PS-2011 was used instead of the taxID of *Lepidoptera* order). The total number of taxa in NCBI was more than three times higher when using all names (769,956 vs. 221,565). This difference was smaller, yet considerable for the number of BOLD taxa (322,927 vs. 231,425) for the all-names and Latin names databases (Table 3).

The proportion of identical sequences shared by different taxa was also higher, when accepting all taxon names compared to using only Latin names, especially for NCBI: 4.0% vs. 1.4% for NCBI, 1.1% vs. 0.9% for BOLD. Similarly, the proportion of taxIDs sharing identical sequences was higher using all names: 28.8% vs. 9.8% for NCBI, 13.2% vs. 11.0% for BOLD. The same tendency was observed for the proportion of the taxIDs that had only sequences identical to other taxa: 25.5% vs. 1.8% for NCBI, 5.6% vs. 1.6% for BOLD (Table 3).

The results of benchmarking the *select_region.pl* command using the E-PCR option is summarized in Table 4. In the positive data set, the increase of the trim_error parameter from 0.2 to 0.3 decreased the false negative rates from 2–2.6% to 0.5%. The effect of the other parameters was negligible. In the negative (non-COI) data sets, increasing the trim_error rate and decreasing the identity threshold increased the number of false positives (FP), while the increase of the minimum overlap between primer and sequence decreased the number of false positives. Among the tested parameter combinations, 0.3 as trim_error, 20 as min_overlap and 0.7 as identity threshold provided the best compromise to minimize both false positive and false negative rates.

When using the bait_file option of the *select_region.pl* script on the negative test data sets, for all bait files and for both identity thresholds (0.6 and 0.7) the false positive rates were 0.007% in the negative-mito data set and 0 in the negative-chloro data set. Figure 2 summarizes the effect of diversity of the sequences in the bait files and the identity threshold when running the *select_region.pl* script on the positive data set. Sensitivity (percentage of true positives among all COI sequences) was high in all cases and increased with decreasing identity thresholds. When selecting one random sequence for each phylum, class or order, sensitivity increased steadily, reaching 98% and 99% for identity thresholds of 70% and 60%, respectively. Increasing the number of sequences per taxon also increased the sensitivity.

4 | DISCUSSION

The need for high-quality databases can be measured by the number of published databases and methods of their construction. Several existing tools are based on sequences in NCBI databases. Some of them such as the CRUX database Builder integrated to Anacapa (Curd et al., 2019), Metataxa2 Database Builder (Bengtsson-Palme et al., 2018), MetaCurator (Richardson et al., 2020), BCDATABASER (Keller et al., 2020) or the database construction module of BAGPIPE (Papadopoulou et al., 2015) are not specific to a marker, while others are more specialized. For example, the MIDORI database (Leray et al., 2018; Machida et al., 2017) contains mitochondrial sequences of 13 protein-coding genes and the CO-ARBITRATOR (Heller et al., 2018) is designed to build an auto-curated database of Metazoan COI sequences. All the above-mentioned databases and tools are based exclusively on NCBI databases or on a data set already containing a coherent system of lineages. Several COI-specific databases containing sequences from NCBI databases and BOLD have also been

TABLE 2 The number of taxa and sequences by phylum

	Class	Order	Family	Genus	Species	Seqences
Eukaryota						
Metazoa	126	679	5793	60,175	251,755	3,227,851
Arthropoda	20	135	2486	41,975	185,721	2,692,056
Chordata	14	178	1202	8646	35,960	272,027
Mollusca	9	69	649	4213	14,860	134,996
Annelida	3	27	152	1035	3603	39,322
Platyhelminthes	7	45	231	915	2275	21,776
Echinodermata	6	47	185	709	1854	19,590
Nematoda	3	20	169	608	1873	14,117
Cnidaria	7	29	268	896	2474	11,212
Rotifera	3	9	29	78	270	6452
Porifera	5	33	130	412	1147	3707
Nemertea	4	10	40	120	347	3032
Acanthocephala	5	10	21	62	149	1811
Tardigrada	3	7	24	68	234	1615
Bryozoa	4	7	69	132	286	1296
Chaetognatha	2	5	10	23	47	1051
Onychophora	2	2	3	38	111	989
Sipuncula	1	5	9	24	74	526
Other	28	41	116	221	470	2276
Viridiplantae	30	115	280	990	1834	2362
Streptophyta	17	90	235	920	1722	2174
Other	13	25	45	70	112	188
Fungi	32	71	147	265	739	1984
Ascomycota	13	38	71	139	433	1108
Basidiomycota	8	20	61	105	261	585
Other	11	13	15	21	45	291
Undef.	55	202	444	1306	4928	26,604
Rhodophyta	4	37	130	628	2228	13,191
Oomycota	1	11	18	57	804	3738
undef	19	69	141	344	834	3685
Apicomplexa	3	5	13	32	351	2951
Ciliophora	6	21	60	103	291	1489
Bacillariophyta	5	24	36	61	206	920
Other	17	35	46	81	214	630
Archaea	1	2	2	2	2	2
Bacteria	7	14	16	33	46	850
Viruses	1	1	1	1	1	1

published and they were designed to fill specific needs such as having a database specific to a taxon, a geographical region or a taxonomic assignment tool. The Eukaryote CO1 Reference Set For The RDP CLASSIFIER (Porter & Hajibabaei, 2018a) is specifically designed for the RDP CLASSIFIER and focuses on Arthropoda and Chordata and also contains outgroups. It contains NCBI and BOLD sequences of at least 500 bp, and sequences were screened to remove typical contaminants (e.g., mislabelled human or bacterial sequences). Meta-Fish-Lib (Collins et al., 2021) and MitoFish (Sato et al., 2018) are reference libraries of fishes, and COins (Magoga et al., 2022) is a manually curated COI database of insects. The COInr database and the MKCOINR tool are most comparable to the MARES (Arranz et al., 2020) and METACOXI (Balech et al., 2022) databases and their underlying pipelines, since they all contain sequences from both NCBI-nt and BOLD databases and the pipelines are also provided to create a new database specific to the users' needs. However, the creation of custom databases based on the MARES pipeline starts from downloading sequences and therefore includes long

TABLE 3 Comparison of the number of sequences and taxIDs when accepting all taxon names or using only formal Latin names

	NCBI Latin names	NCBI All names	BOLD Latin names	BOLD All names
Total number of sequences	1,630,665	1,768,768	2,815,860	2,826,583
Percentage of sequences present in different taxIDs	1.44	3.99	0.87	1.08
Total number of taxIDs	221,565	769,956	231,425	322,927
Percentage of taxIDs sharing sequences with another taxIDs	9.80	28.91	10.97	13.21
Percentage of taxIDs without unique sequences	1.82	25.45	1.57	5.59

TABLE 4 Percentage of true and false positives after running the *select_region* script using the E-PCR option with different parameter settings

Trim_error	Min_overlap	Identity	Percentage trimmed	Percentage untrimmed
Positive test data set			Percentage TP	Percentage FN
0.2	10	0.6	98	2
0.2	10	0.7	97.4	2.6
0.2	20	0.6	98	2
0.2	20	0.7	97.4	2.6
0.3	10	0.6	99.5	0.5
0.3	10	0.7	99.5	0.5
0.3	20	0.6	99.5	0.5
0.3	20	0.7	99.5	0.5
Negative-chloro test data set			Percentage FP	Percentage TN
0.2	10	0.6	0	100
0.2	10	0.7	0	100
0.2	20	0.6	0	100
0.2	20	0.7	0	100
0.3	10	0.6	13.4	86.6
0.3	10	0.7	13.1	86.9
0.3	20	0.6	0	100
0.3	20	0.7	0	100
Negative-mito test data set			Percentage FP	Percentage TN
0.2	10	0.6	0	100
0.2	10	0.7	0	100
0.2	20	0.6	0	100
0.2	20	0.7	0	100
0.3	10	0.6	19.4	80.6
0.3	10	0.7	5.2	94.8
0.3	20	0.6	12.4	87.6
0.3	20	0.7	0.4	99.6

Note: In the positive data set, trimmed sequences are true positives (TP), and untrimmed sequences are false negatives (FN). In the negative data sets, trimmed sequences are false positives (FP), and untrimmed sequences are true negatives (TN).

and difficult steps. METACOXI is a COI database including metazoan sequences in an easy-to-parse format. However, no tools are provided at present for the creation of custom databases, and therefore basic programming skills are necessary to obtain a ready-to-use custom database from it. This is relatively easy for some tasks, such as selection of sequences of a taxon, but needs considerable effort to select a specific gene region, or to format the database for RDP CLASSIFIER (Wang et al., 2007) or QIIME2 (Bolyen et al., 2019). COInr and MKCOINR fill a gap by both providing a comprehensive, easy-to-access database and a versatile tool to customize it.

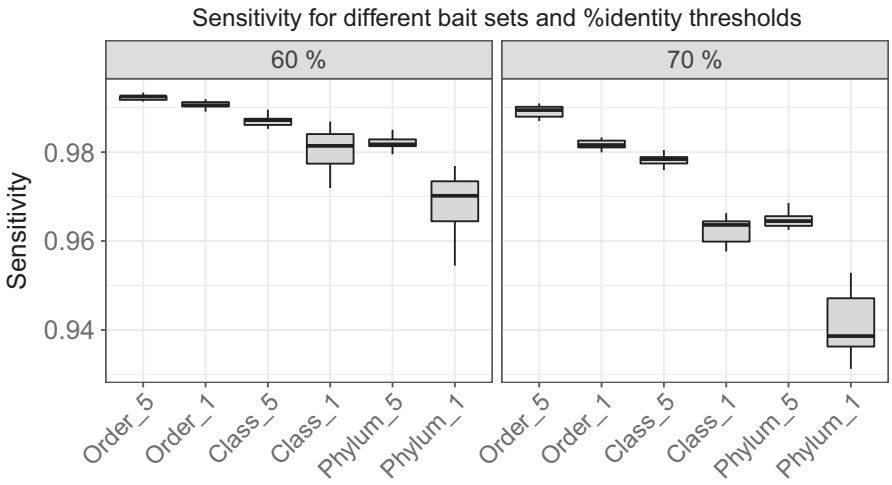

FIGURE 2 Sensitivity of the *select_region.pl* script using the bait_file option in the function of bait complexity and identity threshold. Taxrank_X refers to a bait file containing X randomly selected sequences for each taxon of the taxonomic rank

4.1 | Use of accepted Latin names

Both BOLD and NCBI contain a large number of taxon names at a species level, with unique taxIDs, which do not correspond to the binomial nomenclature. In most cases they correspond to taxon names of a higher level completed by an identifier or simply completing the taxon name by "sp." In principle, they could be proxies of species, but they in fact reflect a lack of information. This phenomenon is particularly pronounced in NCBI, where the total number of taxa including all names is more than three times higher than the number of distinct Latin names. For example, many genus names in NCBI are completed by the sampleID of BOLD and used as species names (e.g., Platynothrus sp. BIOUG14078-H10). The utility of nonstandard taxon names is questionable for most metabarcoding applications. When accepting all names as they appear in the input database, a high proportion of the COI sequences are shared between taxa, and most importantly a high proportion of taxa contain only sequences that are identical to sequences of other taxa. Therefore, keeping artificial identifiers as species names, when they do not necessarily correspond to species, they are uninformative for most users and in many cases they cannot be distinguished from sequences of other taxa, inflates uselessly the number of taxa and thus hinders efficient, taxonomically aware reduction of redundancy. The COInr database uses only taxa with correct Latin name format. To avoid the loss of sequences, sequences with incorrect taxon names are attributed to the lowest taxon in the lineage with a Latin name. Therefore, sequences are kept in the database, with a conservative level of taxonomic information, resulting in a more efficient dereplication, and thus a smaller database without the loss of crucial information. This particularity should be kept in mind when comparing the number of taxa to other databases that do not follow this strategy. However, for users who wish to include nonstandard names, the pipeline can be re-run with deactivating the *check_name* option, thus keeping all taxon names as they appear in the source database.

4.2 | Selecting the target region

The COInr database includes sequences that can cover any region of the COI gene. For taxonomic assignment methods based on sequence similarity (Clemente et al., 2011; Huson et al., 2007; Kahlke & Ralph, 2019; Wood & Salzberg, 2014) the database can be used as it is, because sequences of the nontarget region will not be returned by BLAST or other similarity searches. The only disadvantage would be the database size, which could be eventually reduced by selecting only the region of the sequence that covers the target region. On the other hand, for taxonomic assignment based on sequence composition or phylogeny (Murali et al., 2018; Nguyen et al., 2014; Rosen et al., 2011; Wang et al., 2007), or for the use of the database for phylogenetic, mitogenomic or genomic studies it is preferable to trim sequences to the target region. This can be done using the MKCOINR tool. It is possible to select only full-length sequences covering the whole target region. However, this comes at the price of losing partial sequences, and thus some taxa. Therefore, MKCOINR can also select sequences that cover user-defined portions of the target region to increase taxonomic coverage.

The *select_region.pl* command of MKCOINR depends on several parameters that influence the sensitivity (proportion of true positives in the positive data set) and the specificity (proportion of true negatives in the negative data sets) of this step. According to benchmarking, when using the E-PCR option, relaxed parameter settings increase considerably the false positive rates, since they allow the primers to match to nontarget sites, and this problem is amplified in the next step when sequences produced by E-PCR are used in the bait file to trim the remaining sequences. On the other hand, parameters that are too stringent can reduce sensitivity. A balance between false positives and false negatives was achieved, however, with 0–0.4% of false positives and 0.5% of false negatives for a carefully selected parameter combination (trim_error: 0.3; min_overlap: 20; identity: 0.7). Alternatively, users can choose to provide a bait file to replace the E-PCR step. This option is more robust for avoiding false positives, irrespective of the bait file diversity and the identity threshold. On the other hand, the more diverse the bait file, the higher the sensitivity. As a rule of thumb, five random sequences per taxonomic class or one to five sequences per order lead to a high sensitivity (>98%) and precision (near 100%).

The choice between using the E-PCR or the bait-file option depends on the aim of the database. The use of bait files makes the selection of the target region robust against false positives with a

high sensitivity. However, in some taxa with introns, or considerable length variation of the COI gene, it is important to include bait sequences of taxa with COI of atypical length. In this case, the E-PCR option can help to capture the variability of the data set, but the parameters should be carefully chosen to find the balance between sensitivity and specificity.

4.3 | Selecting the target groups

Using a large database with a wide taxonomic scope is convenient for users analysing different data sets with a varied taxonomic origin, since the same database can be used and can give a good first approximation of taxonomic assignment of sequences. It can also be helpful to detect contaminant sequences that are not expected in the study (e.g., human sequences or model species studied in the same laboratory) or sequences outside of the target group of the study (e.g., bacteria, algae and fungi when focusing on animals). By using a generalist database, these sequences can be identified and eliminated. On the other hand, the presence of reference sequences from taxa not relevant to the study can also have disadvantages: the database size is higher and therefore the speed of taxonomic assignment is lower with generalist databases. Moreover, sequences can be assigned to unexpected taxa if the taxonomic coverage of the target group is incomplete. This can be avoided with databases specific to the target group (Axtner et al., 2019; Mathon et al., 2021; Valentini et al., 2016). For example, many sequences from marine samples can be erroneously assigned to insects when using a generalized database, which is the combined result of the facts that most marine groups are insufficiently covered in the reference databases (Mugnai et al., 2021), and an overwhelming majority of the sequences are from insects (73%). Therefore, the possibility to easily create custom databases specifically tailored to the users' needs is particularly important, and MKCOINR provides the necessary tools to make this selection.

4.4 | Selecting sequences with different taxonomic resolution

Another consideration when creating custom databases is whether to keep reference sequences with incomplete lineages. Most sequences of a reference database assigned to an insect order without further precision are likely to be useless, since most insect reference sequences are determined at least to the genus level, and the taxonomic coverage of this group is wide. By contrast, for less well-covered groups, especially if species or higher-level groups are difficult to identify morphologically (e.g., Nematoda, Rotifera), reference sequences with partial lineages are still informative.

4.5 | Database curation

Erroneously annotated sequences in the reference database can have serious consequences on taxonomic assignments. Unfortunately, both the NCBI and BOLD databases contain mislabelled sequences (Bidartondo et al., 2008; Meiklejohn et al., 2019). This problem should be addressed from the source by a public sequence database that can incorporate a community-curated annotation and allows third parties to improve the annotations of sequences. In the BOLD database, the detection of taxonomic incoherencies is principally based on BINs (Barcoding Index Numbers) (Ratnasingham & Hebert, 2013), and applications such as BAGS (Fontes et al., 2021) allow us to automatically flag some of them. However, the BIN system is based on the existence of a barcoding gap, which does not exist for all taxa (Meier et al., 2006; Roe & Sperling, 2007; Rubinoff et al., 2006). Therefore, human expertise with curation jams is still very much needed for taxonomic revision (Radulovici et al., 2021). On the other hand, errors in primary sequence data in NCBI-nt can only be corrected by the authors, which is inefficient and unsustainable.

In the field of mycology, considerable progress has been done to identify undescribed taxa using a Taxon Hypothesis (Kõljalg et al., 2020), to make concerted effort to identify high-quality sequences and to re-annotate erroneous or insufficiently annotated public ITS sequences (R. Henrik Nilsson et al., 2014) and include the improved annotations to the UNITE database (Rolf Henrik Nilsson et al., 2019). This database can be used as a reference for automated curation of some other error types such as chimeras (R. Henrik Nilsson et al., 2015). A similar approach would also be desirable for Metazoa, especially for taxa that are difficult to distinguish morphologically.

Given the lack of sufficient curation effort of the source databases, ideally, a local database derived from them should be curated to identify incorrectly assigned sequences. Published semi-automatic methods aiming to curate databases are not applicable to large databases (millions of sequences), since either the run time would be prohibitive or include a manual step for curation (Collins et al., 2021; Kozlov et al., 2016; Rulik et al., 2017). The COInr database is too large to be able to run an automatic curation step, which should be kept in mind when using the full database. However, if a small custom database is created from COInr, this curation step becomes feasible and strongly recommended.

5 | CONCLUSIONS

The COInr is a comprehensive database of COI sequences, and its major aim is to serve as a reference database for barcoding and metabarcoding studies. It can be used for taxonomic assignments of COI sequences as it is, since it is not limited in its taxonomic scope, or to a particular region of the gene. It is also a good starting point to create local, custom databases, since it saves the most time-intensive and complicated steps of database creation: (i) downloading a large number of sequences, (ii) creation of a coherent taxID system to avoid ambiguity due to homonymy and synonymy, and (iii) sequence dereplication.

The MKCOINR package provides the necessary tools to both to re-create a whole COInr database, between the planned annual

updates, and produce a custom database starting from COInr. The possibility of refining the taxonomic composition of the database, selection of the gene region and formatting the output to widely used database formats (BLAST, RDP, QIIME) are filling the need for an easy way of creating customized COI databases.

AUTHOR CONTRIBUTIONS

EM designed the research, wrote the scripts, analysed the data and wrote the manuscript.

ACKNOWLEDGEMENTS

I thank Francesco Mugnai for testing mkCOnr and making valuable comments on its use, documentation and the present paper and Gabriel Nève for language editing.

CONFLICT OF INTEREST

The author declare that she has no competing interests.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://doi.org/10.5281/zenodo.6555985

DATA AVAILABILITY STATEMENT

Data Accessibility: The complete COI database can be downloaded from https://doi.org/10.5281/zenodo.6555985 (Meglécz, 2022). All scripts are available in https://github.com/meglecz/mkCOInr including full documentation and they are also archived in Zenodo at https://doi.org/10.5281/zenodo.6961340 (Meglécz, 2022).Benefits Generated: Benefits from this research accrue from the sharing of my data and results on public databases as described above.

ORCID

Emese Meglécz https://orcid.org/0000-0001-7547-4858

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., & Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the metazoa. *Molecular Ecology*, *27*(20), 3968–3975. https://doi.org/10.1111/mec.14844

Arranz, V., Pearman, W. S., Aguirre, J. D., & Liggins, L. (2020). MARES, a replicable pipeline and curated reference database for marine eukaryote metabarcoding. *Scientific Data*, *7*(1), 209. https://doi.org/10.1038/s41597-020-0549-9

Axtner, J., Crampton-Platt, A., Hörig, L. A., Mohamed, A., Xu, C. C. Y., Yu, D. W., & Wilting, A. (2019). An efficient and robust laboratory workflow and tetrapod database for larger scale environmental DNA studies. *GigaScience*, *8*(4), giz029. https://doi.org/10.1093/gigascience/giz029

Balech, B., Sandionigi, A., Marzano, M., Pesole, G., & Santamaria, M. (2022). MetaCOXI: An integrated collection of metazoan mitochondrial cytochrome oxidase subunit-I DNA sequences. *Database*, *2022*, baab084. https://doi.org/10.1093/database/baab084

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

Bengtsson-Palme, J., Richardson, R. T., Meola, M., Wurzbacher, C., Tremblay, É. D., Thorell, K., Kanger, K., Eriksson, K. M., Bilodeau, G. J., Johnson, R. M., Hartmann, M., & Nilsson, R. H. (2018). Metaxa2 database builder: Enabling taxonomic identification from metagenomic or metabarcoding data using any genetic marker. *Bioinformatics*, *34*(23), 4027–4033. https://doi.org/10.1093/bioinformatics/bty482

Bidartondo, M. I., Bruns, T. D., Blackwell, M., Edwards, I., Taylor, A. F. S., Horton, T., Zhang, N., Koljalg, U., May, G., Kuyper, T. W., Bever, J. D., Gilbert, G., Taylor, J. W., DeSantis, T. Z., Pringle, A., Borneman, J., Thorn, G., Berbee, M., Mueller, G. M., … Hughes, K. (2008). Preserving accuracy in GenBank. *Science*, *319*(5870), 1616a. https://doi.org/10.1126/science.319.5870.1616a

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., & Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 90. https://doi.org/10.1186/s40168-018-0470-z

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Bruns, T. D., White, T. J., & Taylor, J. W. (1991). Fungal molecular systematics. *Annual Review of Ecology and Systematics*, *22*, 525–564.

Clemente, J. C., Jansson, J., & Valiente, G. (2011). Flexible taxonomic assignment of ambiguous sequencing reads. *BMC Bioinformatics*, *12*(1), 8. https://doi.org/10.1186/1471-2105-12-8

Collins, R. A., Trauzzi, G., Maltby, K. M., Gibson, T. I., Ratcliffe, F. C., Hallam, J., Rainbird, S., Maclaine, J., Henderson, P. A., Sims, D. W., Mariani, S., & Genner, M. J. (2021). Meta-fish-lib: A generalised, dynamic DNA reference library pipeline for metabarcoding of fishes. *Journal of Fish Biology*, *99*(4), 1446–1454. https://doi.org/10.1111/jfb.14852

Compson, Z. G., McClenaghan, B., Singer, G. A. C., Fahner, N. A., & Hajibabaei, M. (2020). Metabarcoding from microbes to mammals: Comprehensive bioassessment on a global scale. *Frontiers in Ecology and Evolution*, *8*, 581835. Retrieved from. https://www.frontiersin.org/article/10.3389/fevo.2020.581835

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., Potter, C., & Bik, H. M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, *7*(9), 1008–1018. https://doi.org/10.1111/2041-210X.12574

Curd, E. E., Gold, Z., Kandlikar, G. S., Gomer, J., Ogden, M., O'Connell, T., Pipes, L., Schweizer, T. M., Rabichow, L., Lin, M., Shi, B., Barber, P. H., Kraft, N., Wayne, R., & Meyer, R. S. (2019). Anacapa toolkit: An environmental DNA toolkit for processing multilocus metabarcode datasets. *Methods in Ecology and Evolution*, *10*(9), 1469–1475. https://doi.org/10.1111/2041-210X.13214

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072. https://doi.org/10.1128/AEM.03006-05

Fontes, J. T., Vieira, P. E., Ekrem, T., Soares, P., & Costa, F. O. (2021). BAGS: An automated barcode, Audit & Grade System for DNA barcode reference libraries. *Molecular Ecology Resources*, *21*(2), 573–583. https://doi.org/10.1111/1755-0998.13262

González, A., Dubut, V., Corse, E., Mekdad, R., Dechatre, T., & Meglécz, E. (2020). VTAM: A robust pipeline for validating metabarcoding data using internal controls. *BioRxiv*, *2020*, 11.06.371187. https://doi.org/10.1101/2020.11.06.371187

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

Hebert, P. D. N., Ratnasingham, S., & deWaard, J. R. (2003). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings Biological Sciences*, *270*(Suppl 1), S96–S99. https://doi.org/10.1098/rsbl.2003.0025

Hebert, R., & Meglécz, E. (2022). NSDPY: A python package to download DNA sequences from NCBI. *SoftwareX*, *18*, 101038. https://doi.org/10.1016/j.softx.2022.101038

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

Ho, S. Y. W., & Gilbert, M. T. P. (2010). Ancient mitogenomics. *Mitochondrion*, *10*(1), 1–11. https://doi.org/10.1016/j.mito.2009.09.005

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386. https://doi.org/10.1101/gr.5969107

Kahlke, T., & Ralph, P. J. (2019). BASTA – Taxonomic classification of sequences and sequence bins using last common ancestor estimations. *Methods in Ecology and Evolution*, *10*(1), 100–103. https://doi.org/10.1111/2041-210X.13095

Kans, J. (2021). Entrez direct: E-utilities on the Unix command line. In *Entrez programming utilities help [Internet]*. National Center for Biotechnology Information (US). Retrieved from. https://www.ncbi.nlm.nih.gov/books/NBK179288/

Keller, A., Hohlfeld, S., Kolter, A., Schultz, J., Gemeinholzer, B., & Ankenbrand, M. J. (2020). BCdataBaser: On-the-fly reference database creation for (meta-)barcoding. *Bioinformatics*, *36*(8), 2630–2631. https://doi.org/10.1093/bioinformatics/btz960

Khater, E. I. M., Baig, F., Kamal, H. A., Powell, J. R., & Saleh, A. A. (2021). Molecular Phylogenetics and population genetics of the dengue vector Aedes aegypti from the Arabian peninsula. *Journal of Medical Entomology*, *58*(6), 2161–2176. https://doi.org/10.1093/jme/tjab112

Kõljalg, U., Nilsson, H. R., Schigel, D., Tedersoo, L., Larsson, K.-H., May, T. W., Taylor, A. F. S., Jeppesen, T. S., Frøslev, T. G., Lindahl, B. D., Põldmaa, K., Saar, I., Suija, A., Savchenko, A., Yatsiuk, I., Adojaan, K., Ivanov, F., Piirmann, T., Pöhönen, R., … Abarenkov, K. (2020). The taxon hypothesis paradigm—On the unambiguous detection and communication of taxa. *Microorganisms*, *8*(12), 1910. https://doi.org/10.3390/microorganisms8121910

Kozlov, A. M., Zhang, J., Yilmaz, P., Glöckner, F. O., & Stamatakis, A. (2016). Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Research*, *44*(11), 5022–5033. https://doi.org/10.1093/nar/gkw396

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

Macher, J. N., Macher, T. H., & Leese, F. (2017). Combining NCBI and BOLD databases for OTU assignment in metabarcoding and metagenomic datasets: The BOLD_NCBI _Merger. *Metabarcoding and Metagenomics*, *1*, e22262. https://doi.org/10.3897/mbmg.1.22262

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

Magoga, G., Forni, G., Brunetti, M., Meral, A., Spada, A., De Biase, A., & Montagna, M. (2022). Curation of a reference database of COI sequences for insect identification through DNA metabarcoding: COins. *Database*, *2022*, baac055. https://doi.org/10.1093/database/baac055

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

Mathon, L., Valentini, A., Guérin, P.-E., Normandeau, E., Noel, C., Lionnet, C., Boulanger, E., Thuiller, W., Bernatchez, L., Mouillot, D., Dejean, T., & Manel, S. (2021). Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. *Molecular Ecology Resources*, *21*(7), 2565–2579. https://doi.org/10.1111/1755-0998.13430

Meglécz, E. (2022a). *COInr a comprehensive, non-redundant COI database from NCBI-nt and BOLD [data set]*. Zenodo. https://doi.org/10.5281/zenodo.6555985

Meglécz, E. (2022b). *meglecz/mkCOInr: MkCOInr-v.0.2.0*. Zenodo. doi:10.5281/zenodo.6961340

Meier, R., Shiyang, K., Vaidya, G., & Ng, P. K. L. (2006). DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology*, *55*(5), 715–728. https://doi.org/10.1080/10635150600969864

Meiklejohn, K. A., Damaso, N., & Robertson, J. M. (2019). Assessment of BOLD and GenBank – Their accuracy and reliability for the identification of biological materials. *PLoS One*, *14*(6), e0217084. https://doi.org/10.1371/journal.pone.0217084

Mugnai, F., Meglécz, E., Costantini, F., Abbiati, M., Bavestrello, G., Bertasi, F., Bo, M., Capa, M., Chenuil, A., Colangelo, M. A., De Clerck, O., Gutiérrez, J. M., Lattanzi, L., Leduc, M., Martin, D., Matterson, K. O., Mikac, B., Plaisance, L., Ponti, M., … Wangensteen, O. S. (2021). Are well-studied marine biodiversity hotspots still blackspots for animal barcoding? *Global Ecology and Conservation*, *32*, e01909. https://doi.org/10.1016/j.gecco.2021.e01909

Murali, A., Bhargava, A., & Wright, E. S. (2018). IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*, *6*(1), 140. https://doi.org/10.1186/s40168-018-0521-5

Nguyen, N. P., Mirarab, S., Liu, B., Pop, M., & Warnow, T. (2014). TIPP: Taxonomic identification and phylogenetic profiling. *Bioinformatics*, *30*(24), 3548–3555. https://doi.org/10.1093/bioinformatics/btu721

Nilsson, R. H., Hyde, K. D., Pawłowska, J., Ryberg, M., Tedersoo, L., Aas, A. B., Alias, S. A., Alves, A., Anderson, C. L., Antonelli, A., Arnold, A. E., Bahnmann, B., Bahram, M., Bengtsson-Palme, J., Berlin, A., Branco, S., Chomnunti, P., Dissanayake, A., Drenkhan, R., … Abarenkov, K. (2014). Improving ITS sequence data for identification of plant pathogenic fungi. *Fungal Diversity*, *67*(1), 11–19. https://doi.org/10.1007/s13225-014-0291-8

Nilsson, R. H., Tedersoo, L., Ryberg, M., Kristiansson, E., Hartmann, M., Unterseher, M., Porter, T. M., Bengtsson-Palme, J., Walker, D. M., de Sousa, F., Gamper, H. A., Larsson, E., Larsson, K.-H., Kõljalg, U., Edgar, R. C., & Abarenkov, K. (2015). A comprehensive, automatically updated fungal ITS sequence dataset for reference-based chimera control in environmental sequencing efforts. *Microbes and Environments*, *30*(2), 145–150. https://doi.org/10.1264/jsme2.ME14121

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

Papadopoulou, A., Chesters, D., Coronado, I., De la Cadena, G., Cardoso, A., Reyes, J. C., Maes, J.-M., Rueda, R. M., & Gómez-Zurita, J. (2015). Automated DNA-based plant identification for large-scale biodiversity assessment. *Molecular Ecology Resources*, *15*(1), 136–152. https://doi.org/10.1111/1755-0998.12256

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Porter, T. M., & Hajibabaei, M. (2018a). Automated high throughput animal CO1 metabarcode classification. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-22505-4

Porter, T. M., & Hajibabaei, M. (2018b). Over 2.5 million COI sequences in GenBank and growing. *PLoS One*, *13*(9), e0200177. https://doi.org/10.1371/journal.pone.0200177

Porter, T. M., & Hajibabaei, M. (2020). Putting COI Metabarcoding in context: The utility of exact sequence variants (ESVs) in biodiversity analysis. *Frontiers in Ecology and Evolution*, *8*, 248. Retrieved from. https://www.frontiersin.org/article/10.3389/fevo.2020.00248

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, *35*(21), 7188–7196. https://doi.org/10.1093/nar/gkm864

Radulovici, A. E., Vieira, P. E., Duarte, S., Teixeira, M. A. L., Borges, L. M. S., Deagle, B., Majaneva, S., Redmond, N., Schultz, J. A., & Costa, F. O. (2021). Revision and annotation of DNA barcode records for marine invertebrates: Report of the 8th iBOL conference hackathon. *BioRxiv*, 2021.03.07.434272. https://doi.org/10.1101/2021.03.07.434272

Ratnasingham, S., & Hebert, P. D. N. (2007). Bold: The barcode of life data system. *Molecular Ecology Notes*, *7*(3), 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-based registry for all animal species: The barcode index number (BIN) system. *PLoS One*, *8*(7), e66213. https://doi.org/10.1371/journal.pone.0066213

Richardson, R. T., Sponsler, D. B., McMinn-Sauder, H., & Johnson, R. M. (2020). MetaCurator: A hidden Markov model-based toolkit for extracting and curating sequences from taxonomically-informative genetic markers. *Methods in Ecology and Evolution*, *11*(1), 181–186. https://doi.org/10.1111/2041-210X.13314

Roe, A. D., & Sperling, F. A. H. (2007). Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. *Molecular Phylogenetics and Evolution*, *44*(1), 325–345. https://doi.org/10.1016/j.ympev.2006.12.005

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Rosen, G. L., Reichenberger, E. R., & Rosenfeld, A. M. (2011). NBC: The Naïve Bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, *27*(1), 127–129. https://doi.org/10.1093/bioinformatics/btq619

Rubinoff, D., Cameron, S., & Will, K. (2006). A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. *Journal of Heredity*, *97*(6), 581–594. https://doi.org/10.1093/jhered/esl036

Rulik, B., Eberle, J., von der Mark, L., Thormann, J., Jung, M., Köhler, F., Apfel, W., Weigel, A., Kopetz, A., Köhler, J., Fritzlar, F., Hartmann, M., Hadulla, K., Schmidt, J., Hörren, T., Krebs, D., Theves, F., Eulitz, U., Skale, A., … Ahrens, D. (2017). Using taxonomic consistency with semi-automated data pre-processing for high quality DNA barcodes. *Methods in Ecology and Evolution*, *8*(12), 1878–1887. https://doi.org/10.1111/2041-210X.12824

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

Santamaria, M., Fosso, B., Consiglio, A., Caro, G. D., Grillo, G., Licciulli, F., Liuni, S., Marzano, M., Alonso-Alemany, D., Valiente, G., & Pesole, G. (2012). Reference databases for taxonomic assignment in metagenomics. *Briefings in Bioinformatics*, *13*(6), 682–695. https://doi.org/10.1093/bib/bbs036

Sato, Y., Miya, M., Fukunaga, T., Sado, T., & Iwasaki, W. (2018). MitoFish and MiFish pipeline: A mitochondrial genome database of fish with an analysis pipeline for environmental DNA Metabarcoding. *Molecular Biology and Evolution*, *35*(6), 1553–1555. https://doi.org/10.1093/molbev/msy074

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Slater-Baker, M. R., Austin, A. D., Whitfield, J. B., & Fagan-Jeffries, E. P. (2022). First record of miracine parasitoid wasps (hymenoptera: Braconidae) from Australia: Molecular phylogenetics and morphology reveal multiple new species. *Austral Entomology*, *61*(1), 49–67. https://doi.org/10.1111/aen.12582

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of next-generation sequencing technologies. *Current Protocols in Molecular Biology*, *122*(1), e59. https://doi.org/10.1002/cpmb.59

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G. H., Geniez, P., Pont, D., Argillier, C., Baudoin, J.-M., … Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, *25*(4), 929–942. https://doi.org/10.1111/mec.13428

Vijapure, T., Sukumaran, S., & Manohar, C. S. (2019). Molecular characterization and phylogenetics of Indian polychaete fauna: Scope for implementation in ecological monitoring. *Aquatic Ecology*, *53*(4), 665–677. https://doi.org/10.1007/s10452-019-09717-0

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

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. https://doi.org/10.1186/gb-2014-15-3-r46

Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y., Pang, X., Xu, H., Zhu, Y., Xiao, P., & Chen, S. (2010). Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS One*, *5*(10), e13102. https://doi.org/10.1371/journal.pone.0013102

How to cite this article: Meglécz, E. (2023). COInr and mkCOInr: Building and customizing a nonredundant barcoding reference database from BOLD and NCBI using a semi-automated pipeline. *Molecular Ecology Resources*, *00*, 1–13. https://doi.org/10.1111/1755-0998.13756