Database, 2022, 1–7
DOI: https://doi.org/10.1093/database/baac055
Original article

Curation of a reference database of COI sequences for insect identification through DNA metabarcoding: COins

Giulia Magoga[1,*], Giobbe Forni[1], Matteo Brunetti[1], Aycan Meral[1], Alberto Spada[1], Alessio De Biase[2] and Matteo Montagna[3,4,*]

[1]Department of Agricultural and Environmental Sciences, University of Milan, Via Celoria 2, Milano 20133, Italy
[2]Department of Biology and Biotechnology 'Charles Darwin', Sapienza University of Rome, Viale dell'Università 32, Rome 00185, Italy
[3]Department of Agricultural Sciences, University of Naples Federico II, Via Università 100, Portici 80055, Italy
[4]Interuniversity Center for Studies on Bioinspired Agro-Environmental Technology (BAT Center), University of Naples Federico II, Via Università 100, Naples 80055, Italy

[*]Corresponding author: Giulia Magoga Tel: +39 02 50316782; E-mail: giulia.magoga@unimi.it;
Correspondence may also be addressed to Matteo Montagna. Tel: +39 0812539014; E-mail: matteo.montagna@unina.it

Citation details: 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) Vol. 2022: article ID baac055; DOI: https://doi.org/10.1093/database/baac055

Abstract

DNA metabarcoding is a widespread approach for the molecular identification of organisms. While the associated wet-lab and data processing procedures are well established and highly efficient, the reference databases for taxonomic assignment can be implemented to improve the accuracy of identifications. Insects are among the organisms for which DNA-based identification is most commonly used; yet, a DNA-metabarcoding reference database specifically curated for their species identification using software requiring local databases is lacking. Here, we present COins, a database of 5′ region *cytochrome c oxidase subunit I* sequences (*COI*-5P) of insects that includes over 532 000 representative sequences of >106 000 species specifically formatted for the QIIME2 software platform. Through a combination of automated and manually curated steps, we developed this database starting from all *COI* sequences available in the Barcode of Life Data System for insects, focusing on sequences that comply with several standards, including a species-level identification. COins was validated on previously published DNA-metabarcoding sequences data (bulk samples from Malaise traps) and its efficiency compared with other publicly available reference databases (not specific for insects). COins can allow an increase of up to 30% of species-level identifications and thus can represent a valuable resource for the taxonomic assignment of insects' DNA-metabarcoding data, especially when species-level identification is needed https://doi.org/10.6084/m9.figshare.19130465.v1.

Introduction

DNA metabarcoding is a popular method in molecular taxonomy widely used for organisms' identification starting from short DNA sequences of one or a few genes (1, 2). This method has wide applicability in many different fields in which the identification of living organisms is required (3–6). DNA-based identification methods are more useful on organisms for which identification using other approaches is problematic, requires vast expertise or takes a long time. Due to their species richness and ubiquity and to the high level of specialization required for their morphological identification, insects represent one of the groups for which DNA-based identification is commonly used (7–10).

Currently, DNA-metabarcoding wet-lab protocols and raw data analysis pipelines are well established, allowing researchers to obtain high-quality results both from insect environmental DNA and insect community DNA samples (11). Nevertheless, the choices of the DNA marker(s) and of the reference database for sequences' taxonomic assignment are two key aspects that can affect the accuracy of identifications.

Depending on the aim of the study, different DNA markers are deemed as appropriate for insect molecular identification (12, 13). Some of them have a wide taxa coverage but a low taxonomic resolution, e.g. 16S ribosomal RNA (rRNA) and 18S rRNA (13, 14), while others permit more specific identifications, e.g. *cytochrome oxidase subunit I* (*COI* (14)). The choice of the marker is usually driven by the amount of data available as reference, in particular when prior knowledge on the sampled insect taxa is lacking (as in the case of biodiversity surveys; insectivorous animals' diets characterization (15, 16)). In this case, *COI* is the best choice thanks to the high number of publicly available sequence data stored within online repositories (major ones being Barcode of Life Data (BOLD) System and GenBank (17, 18)), especially for the 5′ end (*COI*-5P), the region that can be amplified using universal DNA-barcoding primers, such as LCO1490/HCO2198 (19). In recent years, a consistent number of DNA-metabarcoding primers targeting this region have been developed (20, 21) and demonstrated to work effectively on different insect taxa (21).

Regarding reference databases of *COI* sequences, some types of software directly connecting to BOLD system

Received 1 March 2022; Revised 19 May 2022; Accepted 17 June 2022

databases for taxonomic assignments of operational taxonomic units (OTUs)/amplicon sequence variants (ASVs) are available (22, 23), but other tools (e.g. QIIME (24), Ribosomal Database Project (RDP) classifier (25) BLAST+ (26)) need local reference databases. For the latter, some publicly available ready-to-use databases have been developed (e.g. MIDORI (27)). Nevertheless, in some cases, the use of self-developed reference databases can increase taxa identification accuracy. Generating a customized reference database can be a challenging process, and until recently, it has been possible only using self-developed pipelines. Recently, Robeson and colleagues (28) released RESCRIPT, a largely automated tool for creating metabarcoding reference databases starting from online repositories of public sequence data. Exploiting the large amount of publicly available data for developing DNA-metabarcoding reference databases for insects' identification can be convenient, considering that the identification success and accuracy using DNA metabarcoding is strongly dependent on the completeness of the reference database in terms of taxa representative sequences (29). Indeed, such kinds of references for insects can be developed de novo only by combining the forces of multiple researchers having different taxonomic competence, due to the high taxa richness and intra-taxon diversity (in terms of *COI* variability) characterizing this group. Although the large amount of *COI* sequence data stored in online repositories may sound like a fundamental asset for creating a good reference database, dealing with data developed by other people and generated in the context of different studies can be a double-edged sword. In fact, sequences of undesired origin [nuclear mitochondrial pseudogenes (numts) and contaminants] or related to the wrong taxonomy are commonly released (11), and identifying them among a huge amount of data can be challenging. The presence of erroneous sequences in a reference database can lead to the misidentification of taxa. However, fully automatizing their filtering within a bioinformatic pipeline is unlikely to succeed; as a result, a manual curation step is always fundamental (28).

Here, we present COins, a curated reference database for insect molecular identification that can be used with software requiring a local reference database. This novel database leverages the *COI* sequence of the 5′ region published on BOLD and has been developed using a combination of automated and manual curation steps. The goal was to develop a tool allowing more accurate and specific identification to be obtained than by using the resources currently available for the molecular identification of insect taxa.

Materials and methods

Data mining and database curation

All DNA sequences of insects publicly available on the BOLD COI database were downloaded (search query 'Insecta') on 18 September 2020 along with the information on the specimens from which they were generated. All subsequent steps were performed through *ad hoc* bash and R software (R Core Team, 2019) scripts unless otherwise specified. From the mined dataset, COI-5P and full gene sequences were selected, and any sequences that lacked the relevant taxonomic information on the specimens they were developed from (i.e. order, family, genus and species of belonging) were removed. Then, multi-FASTA format files were generated for each insect order included in the dataset, and the related taxonomy was reported as sequences ids (insect orders sub-datasets, Step 1; Figure 1). In order to avoid the presence within the datasets of sequences annotated as insect but actually belonging to other organisms, the sequences were compared with selected *Homo sapiens, Wolbachia and Rickettsia* sequences using BLAST+ (30), and the entries matching non-insect sequences with an *e*-value < $1e^{-20}$ were removed (Step 2; Figure 1). Additionally, sequences >150 amino acids were identified using Transeq from the EMBOSS software (31) and removed (Step 2; Figure 1). As a further step, the sequences associated with an invalid species-level taxonomy were identified and removed using two methods: (i) an *ad hoc* script looking for key terms included in species names, i.e. 'sp.', 'cf.', 'cfr.', 'group', 'nr.' and numbers and (ii) a manual filtering for detecting further invalid names non-identifiable through key terms (e.g. collection localities, collectors or species author names instead of species name, alphabetic codes replacing species name and many others) (Step 2; Figure 1). To further verify sequences homology, each dataset was aligned at the codon level using MAFFT software (32), and the identity of all sequences introducing gaps in the alignments was verified using the BLASTn tool (33) and, in case of incongruence, removed (Step 3; Figure 1). The resulting datasets were then trimmed to keep only the *COI* gene 'Folmer region' (19); all sequences having a length of ≤420 bp were then discarded using the R library spider (34) (Step 4; Figure 1). In the subsequent step, one representative sequence for each haplotype of the species included in the datasets was selected using the R package haplotypes (https://biolsystematics.wordpress.com/r/) (Step 5; Figure 1). All insect orders' subdatasets were then combined into a single multi-FASTA file in which BOLD process ids were used as sequences' identifiers, and sequences' associated taxonomy was stored in a separate file (Step 6; Figure 1). Within this file, five Rickettsiales sequences amplified from insects using the Folmer primers pair (19) were also included. Due to these sequences' similarity to insects' *COI*-5P, including them in the database could allow the avoidance of misassignment. Finally, the database was formatted for QIIME2 (.qza files available at https://doi.org/10.6084/m9.figshare.19130465.v1). Any further updated version of the database will be published at the same link. COins will be updated whenever a meaningful number of *COI* insect sequences will be published on the BOLD system.

Reference database efficiency test

DNA-metabarcoding raw data (obtained from 54 bulk samples collected with Malaise traps) developed in Kirse *et al.* (35), using mlCOIintF (5′–ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT GGW ACW GGW TGA ACW GTW TAY CCY CC–3′) and dgHCO2198 (5′–GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC TTA AAC TTC AGG GTG ACC AAA RAA YCA–3′) primers pair, were obtained from the Sequence Read Archive (SRA) archive (project accession number PRJNA68109) and used to test the efficiency of the developed database. The bioinformatic analyses were performed using the QIIME2 platform (24). Raw sequences were denoised with the DADA2 algorithm (36) to remove errors and obtain the actual biological sequences (ASVs).

The ASV taxonomic assignment was then performed using two approaches: (i) BLAST+ local alignment between query and reference reads (sequence identity = 97%, minimum

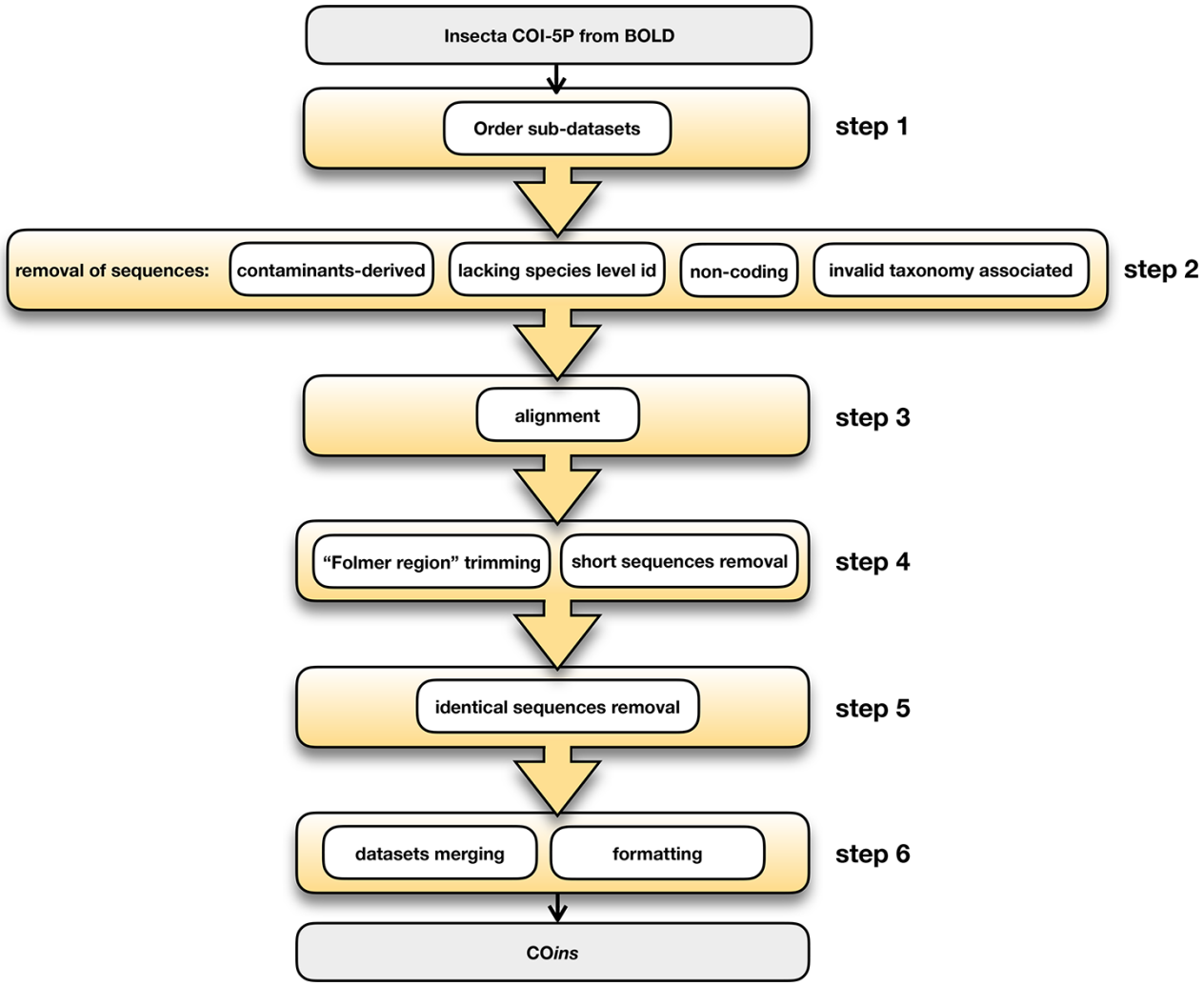

Figure 1. COins database development steps.

consensus among top hits = 80% (26)) and (ii) the naïve Bayes taxonomic classifier trained on the reference database using the fit-classifier sklearn method (confidence = 0.97 (37, 38)). Three different databases were used as reference: (i) the database developed in this study, hereafter named COins; (ii) MIDORI CO1 unique version 245 (Leray *et al.*, in preparation; http://www.reference-midori.info/download.php#) and (iii) a reference database of *COI* sequences created using RESCRIPT software starting from animals' *COI* sequences registered in BOLD (retrieving date July–August 2020), hereafter named ResBO (database available at https://osf.io/d4jra/).

Results

The database

A total of 5 065 234 insect *COI* sequences were mined from BOLD. After filtering (up to Step 4; Figure 1), 3 745 421 sequences were lost (mainly due to the removal of sequences lacking species-level identification). At the end of Step 6 (Figure 1), the database was composed of 532 617 unique sequences, belonging to >106 000 species of 27 different insect orders. The most represented order within COins is Lepidoptera, followed by Diptera and Coleoptera (Table 1). Only a few sequences of Zoraptera and Notoptera are present (Table 1).

Two metadata files associated with COins are available. The first one comprises the information on the identification procedure of the voucher specimens from which *COI* sequences included in the database were generated. The same information is reported also for all identical sequences within haplotypes that were removed in Step 5 of the database curation (Figure 1). The second file reports the information on identical sequences belonging to different species present within the database. These files can be consulted when any specific molecular identification obtained using COins is doubtful (available at https://doi.org/10.6084/m9.figshare.19130465.v1).

Database efficiency test

The 54 DNA-metabarcoding samples (32) used to test the database efficiency, included a total of 27 348 365 raw reads (mean per sample = 506,451.2 reads), after denoising and filtering 8312 ASVs were obtained. The two algorithms adopted in this study (BLAST+-based and fit-classifier sklearn) demonstrated a good congruence in the taxonomic assignments of

Table 1. Number of unique sequences for each insect order included in the database

Order	Number of sequences
Archaeognatha	79
Blattodea	1558
Coleoptera	65 684
Dermaptera	140
Diptera	122 306
Embioptera	69
Ephemeroptera	7150
Hemiptera	28 494
Hymenoptera	58 124
Lepidoptera	209 290
Mantodea	378
Mecoptera	304
Megaloptera	281
Neuroptera	1821
Notoptera	3
Odonata	5142
Orthoptera	7369
Phasmatodea	172
Plecoptera	4733
Psocodea	1800
Raphidioptera	41
Siphonaptera	473
Strepsiptera	56
Thysanoptera	1778
Trichoptera	15 321
Zoraptera	2
Zygentoma	49

the ASVs detected, with COins sharing the highest number of ASVs' unique identifications between algorithms than the other databases, i.e. 80.6% in comparison to 73.6% for MIDORI and 67.8% for ResBO (Figure 2).

The taxonomic assignments of these ASVs using as reference ResBO resulted in 2381 (using BLAST+ algorithm) and 2870 (fit-classifier sklearn algorithm) ASVs assigned to the Insecta class. COins identified 2368 (BLAST+) and 8026 (fit-classifier sklearn) Insecta ASVs, while MIDORI identified 1876 (BLAST+) and 3273 (fit-classifier sklearn) ASVs. Among them, order-level assignments were obtained for 2374 (BLAST+) and 2008 (fit-classifier sklearn) ASVs adopting ResBO as reference; 2367 (BLAST+) and 2611 (fit-classifier sklearn) ASVs using COins and 1864 (BLAST+) and 2219 (fit-classifier sklearn) ASVs using MIDORI (Figure 3). Regarding species-level assignments, the following results were obtained: ResBO identified 1530 (BLAST+) and 1608 (fit-classifier sklearn) ASVs to species; COins 2117 (BLAST+) and 2243 (fit-classifier sklearn) ASVs to species and MIDORI 1594 (BLAST+) and 1584 ASVs (fit-classifier sklearn) to species (Figure 3).

Among the species-level identified ASVs using BLAST+-based algorithm, 825 different species were recognized by MIDORI: 27 of them were shared with ResBO, which identified 887 species (Figure 4a). The highest number of species was found using COins, i.e. 1051, 184 of them in common with ResBO (Figure 4a). Using the BLAST+-based algorithm, 41.4% of the species were identically identified by the three reference databases (Figure 4a). A similar situation was observed when fit-classifier sklearn algorithm was applied, in fact 836 different species were identified by MIDORI (Figure 4b), 29 of them were shared with ResBO, which identified 866 species, and COins detected 1108, 202 in common with the last database (Figure 4b). Using this algorithm, the percentage of common species recognized by the three databases was 40.1% (Figure 4b).

COins identified some ASVs as belonging to Rickettsiales (<20), these ASVs were assigned to Insecta, Arthropoda or remained unassigned when using the other reference databases.

Discussion

In this study, a reference database of *COI* sequences (5′ region) for insects' taxonomic assignment using DNA metabarcoding was developed, starting from the data available on BOLD. These data were filtered according to several criteria in order to remove sequences, which might be potential sources of error during taxonomic assignments of the ASVs. Different motivations for sequence removal—along with their implications—are discussed below.

Sequences associated with incorrect or invalid taxonomy. The most common situation was the presence of sequences annotated as insect but instead derived from other organisms, in particular *Homo sapiens* and also the most common bacterial endosymbionts of insects (e.g. *Wolbachia* and *Rickettsia*). The latter is an already well-known problem related to online reference databases (39). Filtering *COI* sequences separately as sub-datasets for each insect order allowed us to detect further inconsistencies between sequences' variability and their associated taxonomy. In particular, during the

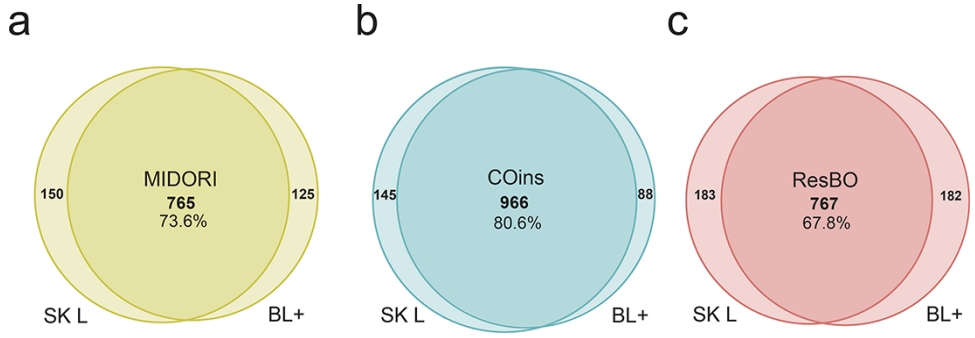

Figure 2. Number of ASVs identified by the two taxonomic assignment algorithms adopted in this study, i.e. the machine learning-based algorithm fit-classifier sklearn (SK L) and the BLAST+ (BL+) algorithm, using each database: (a) MIDORI database, (b) COins database and (c) ResBO database. Numbers of common identifications between the two algorithms are also expressed in percentages.

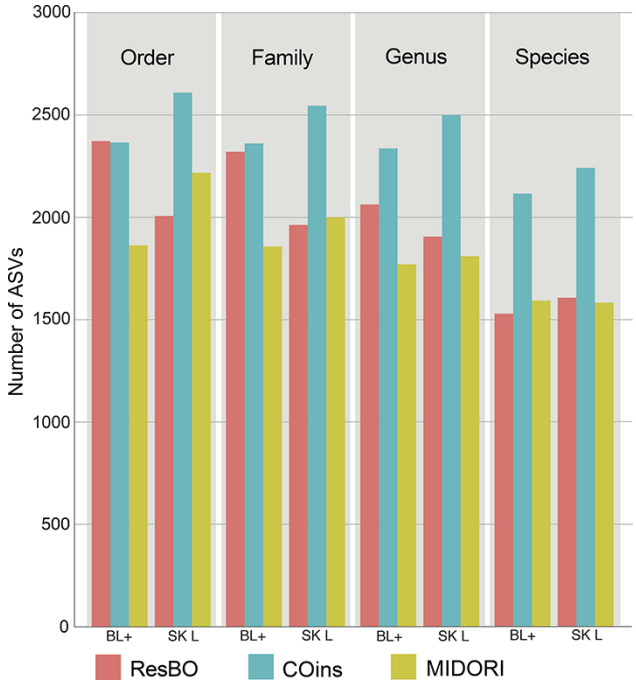

Figure 3. Number of ASVs assigned to the different taxonomic levels (from order to species) when using ResBO, COins and MIDORI as reference. Numbers of assignments obtained using the BLAST+ (BL+) and fit-classifier sklearn (SK L) algorithms are specified too.

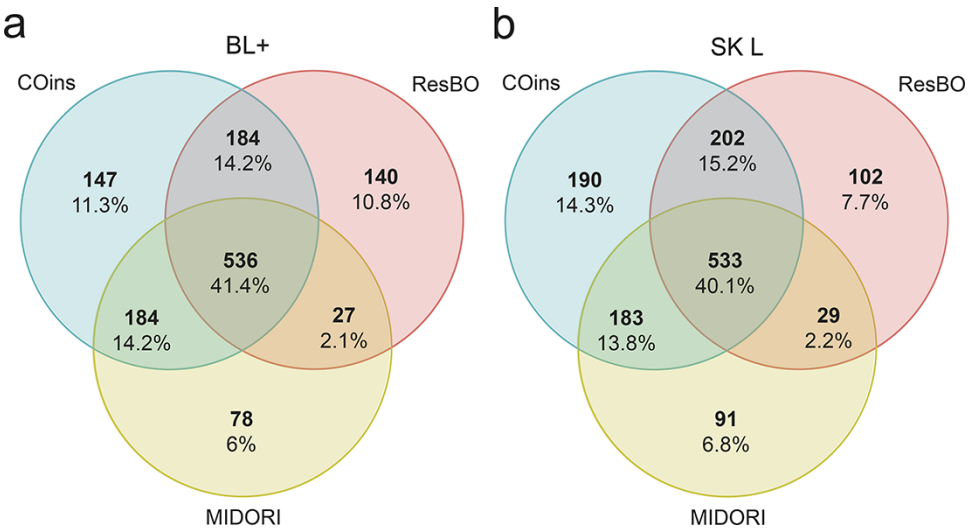

Figure 4. Number of species identified using each database MIDORI, COins and ResBO. (a) Number of species identified adopting the BLAST+ algorithm (BL+). (b) Number of species identified adopting fit-classifier sklearn algorithm (SK L). All values are also reported as percentages.

alignment step, some sequences showing low overall homology with the others in the same sub-dataset were found to be related to misidentifications at the order level. Within this study framework, the official validity of all sequences-associated taxonomic names was intentionally not investigated, because of ongoing debates on the taxonomic status of some insect taxa. As a matter of fact, the increasingly common use of molecular taxonomy has introduced a bias in insect taxonomy: frequently, new species are recognized based on molecular information (e.g. through molecular species delimitation or in the context of DNA-barcoding studies) and named, but never, or only much later, formally described. These species names are not considered valid according with the International Code of Zoological Nomenclature (40) until the formal description of the species is published, but online databases include the reference sequences which allow their identification under the new species name. Nonetheless, the filters applied to the sequences, the manual filter in particular, allowed the detection and discarding of many invalid species names unrelated to the above-mentioned situations and possibly linked with the absence of species-level morphological identification (e.g. genera names followed by numeric or alphabetic codes, but also geographical names or person names replacing specific epithets). In case of doubt, the

scientific works within which the sequences were developed were consulted.

Non-coding sequences were possibly derived from the amplification of numts (41), from sequencing errors, or from the lack of proper editing of electropherograms before data publication. This issue was particularly evident in the database alignment step, where many sequences were discarded since they introduced one or two bases' gaps in the alignment.

Sequences not associated with species-level taxonomy within a reference database, especially if identified at the highest taxonomic ranks, appear to reduce the accuracy of the molecular identification, hindering the reaching of identifications at lower taxonomic levels. This scenario is also a likely explanation for some of the results achieved in the present study, i.e. the cases in which COins assigned the ASV at the species level, while ResBO assigned the same ASVs to a higher taxonomic level, despite the two databases include the same species-level identified reference sequences. At the same time, excluding from a reference database, the sequences not identified at the species level could potentially increase the number of missing identifications, especially when those sequences belong to the only representative of a specific taxon within the database.

Some of the sequences discarded from the database are clearly related to errors, and they could be the results of the lack of care of some BOLD users, as indeed is also a common situation in the case of other databases. The BOLD team routinely perform data curation, in particular checking discordant Barcode Index Number and suppressing potential erroneous sequences from the online database (42). As in the case of this study, the curation is performed manually. It is a time-consuming process done periodically, thus leaving some erroneous sequences in place for a while. This is why using publicly available data for developing DNA-metabarcoding reference databases for local use should always require a manual curation step (28).

The efficiency test on COins showed how this database has an identification efficiency comparable to that of the other databases (MIDORI and ResBO) at the highest taxonomic ranks (e.g. order and family), but it allows the assignment of a considerably higher number of ASVs to the species and genus levels, with a notable increase between 25% and 30% of species-level identifications.

The performed analyses also allowed observation to be made on the effect of using different assignment algorithms. The machine learning-based algorithm (fit-classifier sklearn) was found to assign a higher number of ASVs at any taxonomic level, compared with the BLAST+ algorithm (Figure 3). An evident bias of the use of the fit-classifier sklearn algorithm in association with COins is that almost all the ASVs detected in the samples analysed were assigned to Insecta (8026 ASVs out of 8312) even if some of them likely belong to other classes (e.g. sequences that MIDORI and ResBO assigned to Collembola or Arachnida). This is related to the underlying principle of machine learning-based algorithms, which assumes that all existing taxa are included in the reference used for the assignment (37, 38). Yet, this drawback is only associated with higher level taxonomic assignments and does not affect the accuracy of low-level ones. As a matter of fact, COins was the database for which the highest congruence between identification achieved through the two algorithms used in this study was achieved (Figure 2).

The results obtained using COins highlight the importance of manual curation during the development of reference databases for local use. The effort required is however undeniable. Unfortunately, fully automated filters that make sequences downloaded from public resources readily usable for metabarcoding taxonomic assignment are not yet available. In the meantime, it is necessary, albeit expensive and time-consuming, especially in terms of updating, to make high-quality data available for those metabarcoding software platforms that use local reference databases. Moreover, the direct interaction between software such as QIIME2 with the online BOLD COI database for metazoan ASV/OTU taxonomic assignment is also advisable.

Funding

The authors acknowledge the support of the Article Processing Charge (APC) central fund of the University of Milan (Italy) and the Department of Agricultural and Environmental Sciences of the University of Milan (Italy) which provided the postdoc fellowship of the first author (years 2020–2022).

Conflict of interest

None declared.

References

1. Hajibabaei,M., Shokralla,S., Zhou,X. *et al.* (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One*, **6**, e17497. 10.1371/journal.pone.0017497.
2. Taberlet,P., Coissac,E., Pompanon,F. *et al.* (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.*, **21**, 2045–2050. 10.1111/j.1365-294X.2012.05470.x.
3. Staats,M., Arulandhu,A., Gravendeel,B. *et al.* (2016) Advances in DNA metabarcoding for food and wildlife forensic species identification. *Anal. Bioanal. Chem.*, **408**, 4615–4630. 10.1007/s00216-016-9595-8.
4. Montagna,M., Berruti,A., Bianciotto,V. *et al.* (2018) Differential biodiversity responses between kingdoms (plants, fungi, bacteria and metazoa) along an Alpine succession gradient. *Mol. Ecol.*, **27**, 3671–3685. 10.1111/mec.14817.
5. Zhang,G., Liu,J., Gao,M. *et al.* (2020) Tracing the edible and medicinal plant Pueraria montana and its products in the marketplace yields subspecies level distinction using DNA barcoding and DNA metabarcoding. *Front. Pharmacol.*, **11**, 336. 10.3389/fphar.2020.00336.
6. Brunetti,M., Magoga,G., Gionechetti,F. *et al.* (2021) Does diet breadth affect the complexity of the phytophagous insect microbiota? The case study of Chrysomelidae. *Environ. Microbiol.* 10.1111/1462-2920.15847.
7. deWaard,J., Mitchell,R., Keena,A. *et al.* (2010) Towards a global barcode library for Lymantria (Lepidoptera: Lymantriinae) tussock moths of biosecurity concern. *PLoS One*, **5**, e14280. 10.1371/journal.pone.0014280.
8. Marullo,R., Mercati,F. and Vono,G. (2020) DNA barcoding: a reliable method for the identification of thrips species (Thysanoptera, Thripidae) collected on sticky traps in onion fields. *Insects*, **11**, 489. 10.3390/insects11080489.
9. Magoga,G., Fontaneto,D. and Montagna,M. (2021) Factors affecting the efficiency of molecular species delimitation in a species-rich insect family. *Mol. Ecol. Resour.*, **21**, 1475–1489. 10.1111/1755-0998.13352.

10. Gadawski,P., Montagna,M., Rossaro,B. *et al.* (2022) DNA barcoding of chironomidae from the Lake Skadar region: reference library and a comparative analysis of the European fauna. *Divers Distrib.*, **00**, 1–20. 10.1111/ddi.13504.
11. Deiner,K., Bik,H., Mächler,E. *et al.* (2017) Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol. Ecol.*, **26**, 5872–5895. 10.1111/mec.14350.
12. Batovska,J., Piper,A., Valenzuela,I. *et al.* (2021) Developing a non-destructive metabarcoding protocol for detection of pest insects in bulk trap catches. *Sci. Rep.*, **11**, 7946. 10.1038/s41598-021-85855-6.
13. Ficetola,G., Boyer,F., Valentini,A. *et al.* (2021) Comparison of markers for the monitoring of freshwater benthic biodiversity through DNA metabarcoding. *Mol. Ecol.*, **30**, 3189–3202. 10.1111/mec.15632.
14. Marquina,D., Andersson,A.F. and Ronquist,F. (2018) New mitochondrial primers for metabarcoding of insects, designed and evaluated using in silico methods. *Mol. Ecol. Resour.*, **19**, 90–104. 10.1111/1755-0998.12942.
15. Alberdi,A., Razgour,O., Aizpurua,O. *et al.* (2020) DNA metabarcoding and spatial modelling link diet diversification with distribution homogeneity in European bats. *Nat. Commun.*, **11**, 1154. 10.1038/s41467-020-14961-2.
16. Hardulak,L.A., Morinière,J., Hausmann,A. *et al.* (2020) DNA metabarcoding for biodiversity monitoring in a national park: screening for invasive and pest species. *Mol. Ecol. Resour.*, **20**, 1542–1557. 10.1111/1755-0998.13212.
17. Ratnasingham,S. and Hebert,P.D.N. (2007) BOLD: the Barcode of Life Data System (www.barcodinglife.org). *Mol. Ecol. Notes*, **7**, 355–364. 10.1111/j.1471-8286.2007.01678.x.
18. Clark,K., Karsch-Mizrachi,I., Lipman,D.J. *et al.* (2016) GenBank. *Nucleic Acids Res.*, **44**, D67–D72. 10.1093/nar/gkv1276.
19. Folmer,O., Black,M., Hoeh,W. *et al.* (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotech.*, **3**, 7881515.
20. Brandon-Mong,G., Gan,H., Sing,K. *et al.* (2015) DNA metabarcoding of insects and allies: an evaluation of primers and pipelines. *Bull. Entomol. Res.*, **105**, 717–727. 10.1017/S0007485315000681.
21. Elbrecht,V., Braukmann,T., Ivanova,N.V. *et al.* (2019) Validation of COI metabarcoding primers for terrestrial arthropods. *Peer J.*, **7**, e7745. 10.7717/peerj.7745.
22. Ratnasingham,S. (2019) mBRAVE: the multiplex barcode research and visualization environment. *BISS*, **3**, e37986. 10.3897/biss.3.37986.
23. Buchner,D. and Leese,F. (2020) BOLDigger – a Python package to identify and organise sequences with the Barcode of Life Data systems. *MBMG*, **4**, e53535. 10.3897/mbmg.4.53535.
24. Bolyen,E., Rideout,J.R., Dillon,M.R. *et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.*, **37**, 852–857. 10.1038/s41587-019-0209-9.
25. Wang,Q., Garrity,G., Tiedje,J. *et al.* (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267. 10.1128/aem.00062–07.
26. Camacho,C., Coulouris,G., Avagyan,V. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinform.*, **10**, 421. 10.1186/1471-2105-10-421.
27. Machida,R., Leray,M., Ho,S.L. *et al.* (2017) Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Sci. Data*, **4**, 170027. 10.1038/sdata.2017.27.
28. Robeson,M., O'Rourke,D., Kaehler,B. *et al.* (2021) RESCRIPt: reproducible sequence taxonomy reference database management. *PLoS Comput. Biol.*, **17**, e1009581. 10.1371/journal.pcbi.1009581.
29. Beentjes,K., Speksnijder,A., Schilthuizen,M. *et al.* (2019) Increased performance of DNA metabarcoding of macroinvertebrates by taxonomic sorting. *PLoS One*, **14**, e0226527. 10.1371/journal.pone.0226527.
30. Altschul,S.F., Madden,T.L., Schäffer,A.A. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402. 10.1093/nar/25.17.3389.
31. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 276–277. 10.1016/s0168-9525(00)02024-2.
32. Katoh,K. and Standley,D. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780. 10.1093/molbev/mst010.
33. Altschul,S., Gish,W., Miller,W. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410. 10.1016/s0022-2836(05)80360-2.
34. Brown,S., Collins,R., Boyer,S. *et al.* (2012) SPIDER: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Mol. Ecol. Resour.*, **12**, 562–565. 10.1111/j.1755-0998.2011.03108.x.
35. Kirse,A., Bourlat,S., Langen,K. *et al.* (2021) Metabarcoding Malaise traps and soil eDNA reveals seasonal and local arthropod diversity shifts. *Sci. Rep.*, **11**, 10498. doi: 10.1038/s41598-021-89950-6.
36. Callahan,B., McMurdie,P., Rosen,M. *et al.* (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods*, **13**, 581–583. 10.1038/nmeth.3869.
37. Pedregosa,F., Varoquaux,G., Gramfort,A. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
38. Bokulich,N., Kaehler,B., Rideout,J. *et al.* (2018) Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's aq2-feature-classifier plugin. *Microbiome*, **6**. 10.1186/s40168-018-0470-z.
39. Smith,M., Bertrand,C., Crosby,K. *et al.* (2012) Wolbachia and DNA barcoding insects: patterns, potential, and problems. *PLoS One*, **7**, e36514. 10.1371/journal.pone.0036514.
40. International Commission on Zoological Nomenclature (1999) International Code of Zoological Nomenclature. 4th edition. International Trust for Zoological Nomenclature, London, UK.
41. Song,H., Buhay,J.E., Whiting,M.F. *et al.* (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proc. Natl. Acad. Sci.*, **105**, 13486–13491. 10.1073/pnas.0803076105.
42. Coleman,C.O. and Radulovici,A. (2020) Challenges for the future of taxonomy: talents, databases and knowledge growth. *Megataxa*, **1**, 28–34. 10.11646/megataxa.1.1.5.