taxalogue: a toolkit to create comprehensive CO1 reference databases (#82743)

First submission

Guidance from your Editor

Please submit by **18 Mar 2023** for the benefit of the authors (and your token reward) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files

Download and review all files from the materials page.

4 Figure file(s)
2 Table file(s)

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

taxalogue: a toolkit to create comprehensive CO1 reference databases

Niklas W Noll [Corresp., 1], **Christoph Scherber** [1], **Livia Schäffler** [1]

[1] Centre for Biodiversity Monitoring and Conservation Science, Leibniz Institute for the Analysis of Biodiversity Change, Bonn, North Rhine-Westphalia, Germany

Corresponding Author: Niklas W Noll
Email address: N.Noll@leibniz-lib.de

Background. Taxonomic identification through DNA Barcodes gained considerable traction through the invention of next-generation sequencing and DNA metabarcoding. Metabarcoding allows for the simultaneous identification of thousands of organisms from bulk samples with high taxonomic resolution. However, reliable identifications can only be achieved with comprehensive and curated reference databases. Therefore, custom reference databases are often created to meet the needs of specific research questions. Due to taxonomic inconsistencies, formatting issues, and technical difficulties, building a custom reference database requires tremendous effort. Here, we present *taxalogue*, an easy-to-use software for creating comprehensive and customized reference databases.

Methods. *taxalogue* collects DNA sequences from several online sources (BOLD, GenBank, and GBOL) and combines them into a reference database. Taxonomic incongruencies between the different data sources can be harmonized according to available taxonomies (NCBI taxonomy or GBIF backbone). Dereplication and various filtering options are available regarding sequence quality or metadata information. *taxalogue* is implemented in the open-source ruby programming language, and the source code is available at https://github.com/nwnoll/taxalogue. We benchmark four reference databases by sequence identity against eight queries from different localities and trapping devices. Subsamples from each reference database were used to compare how well another one is covered.

Results. *taxalogue* produces reference databases that have, for most tested queries, the best coverage at high identities and therefore enables more accurate, reliable predictions with higher certainty than the other benchmarked reference databases. Additionally, the performance of *taxalogue* is more consistent while providing good coverage for a variety of habitats, regions, and sampling methods. *taxalogue* simplifies the creation of reference databases and makes the process reproducible and transparent. Multiple available output formats for commonly used downstream applications facilitate the easy adoption of *taxalogue* in many different software pipelines. The resulting reference databases improve the taxonomic classification accuracy through high coverage of the query sequences at high identities.

taxalogue: a toolkit to create comprehensive CO1 reference databases

Niklas W Noll[1], Christoph Scherber[1], Livia Schäffler[1]

[1]Centre for Biodiversity Monitoring and Conservation Science, Leibniz Institute for the Analysis of Biodiversity Change, Bonn, North Rhine-Westphalia, Germany

Corresponding Author:
Niklas Noll[1]
Adenauerallee 127, Bonn, 53113, Germany
Email address: nwnoll88@gmail.com

Abstract

Background. Taxonomic identification through DNA Barcodes gained considerable traction through the invention of next-generation sequencing and DNA metabarcoding. Metabarcoding allows for the simultaneous identification of thousands of organisms from bulk samples with high taxonomic resolution. However, reliable identifications can only be achieved with comprehensive and curated reference databases. Therefore, custom reference databases are often created to meet the needs of specific research questions. Due to taxonomic inconsistencies, formatting issues, and technical difficulties, building a custom reference database requires tremendous effort. Here, we present *taxalogue*, an easy-to-use software for creating comprehensive and customized reference databases.

Methods. *taxalogue* collects DNA sequences from several online sources (BOLD, GenBank, and GBOL) and combines them into a reference database. Taxonomic incongruencies between the different data sources can be harmonized according to available taxonomies (NCBI taxonomy or GBIF backbone). Dereplication and various filtering options are available regarding sequence quality or metadata information. *taxalogue* is implemented in the open-source ruby programming language, and the source code is available at https://github.com/nwnoll/taxalogue. We benchmark four reference databases by sequence identity against eight queries from different localities and trapping devices. Subsamples from each reference database were used to compare how well another one is covered.

Results. *taxalogue* produces reference databases that have, for most tested queries, the best coverage at high identities and therefore enables more accurate, reliable predictions with higher certainty than the other benchmarked reference databases. Additionally, the performance of *taxalogue* is more consistent while providing good coverage for a variety of habitats, regions, and sampling methods. *taxalogue* simplifies the creation of reference databases and makes the

process reproducible and transparent. Multiple available output formats for commonly used downstream applications facilitate the easy adoption of *taxalogue* in many different software pipelines. The resulting reference databases improve the taxonomic classification accuracy through high coverage of the query sequences at high identities.

Keywords: CO1, DNA Barcoding, metabarcoding, reference database, taxon, taxonomic harmonization, taxonomic identification

Introduction

Great effort is currently taken to arrive at a comprehensive DNA barcode reference database for all life on Earth (Hobern & Hebert, 2019), which has also been fundamental for the mission of the international Barcode of Life Consortium (International Barcode of Life, 2022): saving the living planet and cataloging all multicellular species before the first half of the century. DNA Barcodes are short marker-gene sequences that are ideally conserved at species level with sufficient genetic differentiation to distinguish even closely related sister taxa (Hebert et al., 2003; Hebert, Ratnasingham & DeWaard, 2003). Many different barcode markers are used for different taxa, but the most often used animal barcode is the folmer region (Folmer et al., 1994) of the mitochondrial CO1 gene, which is part of the respiratory complex and is known to have, in general, a high resolution until species level (e.g., Hebert et al., 2003; Hebert, Ratnasingham & DeWaard, 2003; Fišer & Buzan, 2014; Huemer et al., 2014). To identify specimens even without taxonomic expertise, the same barcode region from unknown organisms is sequenced and compared to barcode sequences of already identified specimens stored in a reference database. New sequences can be compared directly with an online source database using identification services such as those provided by GenBank (Sayers et al., 2022), BOLD (Ratnasingham & Hebert, 2007), or GBOL (Geiger et al., 2016a). Since large online databases are subject to constant changes, self-created reference databases are often used instead; they require more work and expertise but provide full control over the sequences and make taxonomic identification reproducible. Given the large number of sequences generated by metabarcoding, where the DNA from many organisms is simultaneously sequenced (Taberlet et al., 2012), a self-created reference database can also speed up the identification process (Macher, Macher & Leese, 2017).

The primary goal of a DNA barcode reference database is to provide taxon names for sequences. Taxon names are like other carefully circumscribed abstractions: good names subsume ecological observations and evolutionary theories (Franz, 2005). Therefore, scientific species names are a link to the accumulated knowledge of a species in time (Grimaldi & Engel, 2005) and much of biology relies on them (Agnarsson and Kuntner, 2007). However, synonyms, taxonomic disagreements and revisions have received little attention in using DNA barcode reference databases (Leray et al.; 2019; Pappalardo et al., 2021; Piper et al., 2021). Their effects on the interpretation of metabarcoding results remain unexplored, even though proper taxonomic name usage is a prerequisite for any reliable conclusion (e.g., Bortolus, 2008). Taxa lists derived from metabarcoding results depend on the composition of the used reference database: taxon names in the reference database might be based on a particular taxonomic opinion, used identification

literature, prior taxonomic harmonization (Ratnasingham & Hebert, 2007; Schoch et al., 2020), reverse taxonomy (identification by its sequence and not morphology) (Weigand et al., 2019), and more. Even correct names in source databases could convey distinct taxonomic concepts (Berendsohn & Geoffroy, 2007). The meaning of a name is unclear without mentioning the taxonomic circumscription on which an identifier based the specimen identification (Berendsohn, 1995). ~~Harmonization of~~ taxon names is an often-used step to ensure an up-to-date taxonomy and successful data integration from multiple sources (Grenié et al., 2022). Since manual harmonization might not be actionable for studies investigating a broad range of taxa, or a diverse taxon such as Arthropoda, an automated approach might be the most obvious. Data aggregators such as NCBI Taxonomy (Shoch et al., 2020) or GBIF (GBIF Secretariat, 2021) provide a resolved taxonomy by acting as a decisive authority in the case of taxonomic disagreements and can be used to automatically harmonize data from different sources.

Besides the influences of nomenclature and taxonomy on the source databases, data quality and coverage are also essential for the condition of the used reference database. Comprehensive taxonomic coverage of a reference database is necessary for reliable identifications (Meyer & Paulay, 2005; Vences et al., 2005; Ekrem et al., 2007). A sufficient sampling of each taxon has been stressed as an initial requirement for DNA Barcoding (Sperling, 2003) ~~and~~ its importance continues to be emphasized (Phillips, Gillis & Hanner, 2019). For taxa with high intraspecific variation, sampling the whole geographic range might be necessary for appropriate identification (Lou & Golding, 2012; Geiger et al., 2016b). However, the observed genetic differentiation between closely related taxa might also decrease with an increase in the geographic scale of the reference database, impairing the identification process. Therefore, regional reference databases have been suggested (Bergsten et al., 2012). Despite significant efforts to complete these reference databases, commonly used sources such as GenBank (Sayers et al., 2022) and BOLD (Ratnasingham & Hebert, 2007) still have exclusive CO1 records (Porter et al., 2014; Macher, Macher & Leese, 2017; Curry et al., 2018; Porter & Hajibabaei, 2018a; Pentinsaari et al., 2020; O`Rourke et al. 2020; Porter & Hajibabaei, 2020; Robeson et al. 2021, Nakazato & Jinbo, 2022) and coverage is reduced when using just one source. ~~Filtering~~ may become necessary when data quality in reference databases is insufficient (Meyer & Paulay, 2005; Nilsson et al., 2006; Collins & Cruickshank, 2013).

The ~~above~~mentioned problems and circumstances ~~make it clear~~ that care is required when creating a reference database. Several software solutions have been developed to create custom reference databases (Macher, Macher & Leese, 2017; Bengtsson-Palme et al., 2018; Palmer et al., 2018; Richardson et al., 2018; Heller et al., 2018; Keller et al., 2020; Arranz et al., 2020; Robeson et al., 2021; Piper et al., 2021; Meglécz, 2023; Keck & Altermatt, 2022) or to provide ready-to-use reference databases (Leray et al., 2018; Porter & Hajibabaei, 2018b; O`Rourke et al., 2020; Leray, Knowlton & Machida, 2022; Magoga et al., 2022). However, only some can integrate multiple CO1 database sources (Macher, Macher & Leese, 2017; Bengtsson-Palme et al., 2018; Porter & Hajibabaei, 2018a; Arranz et al., 2020; Piper et al., 2021; Meglécz, 2023; Keck & Altermatt, 2022). To the best of our knowledge, no software currently available allows

the exploration of distinct taxonomic harmonization strategies while also including data from GBOL, having extensive sequence filtering options, creating reference databases with different geographical scales (countries, continents, biogeographic realms, or user-defined ArcGIS shape files), dereplication, and providing multiple ready-to-use outputs for common downstream analysis applications. To close this gap, we developed *taxalogue* (https://github.com/nwnoll/taxalogue). In this paper, we demonstrate the suitability of this toolkit to create comprehensive and customized reference databases and compare them with already available CO1 reference databases for arthropods.

Materials & Methods

The current version of *taxalogue* can create reference databases of the CO1 Folmer region (Folmer et al., 1994) for animals. CO1 sequences from animal specimens are referred to as "sequences" or "records" in the following. We envisage the implementation of additional markers and a broader range of taxa for upcoming versions. See Fig. 1 for an overview of *taxalogue* main functions and consider using *taxalogue* with the "--help" command, or visit the GitHub webpage (https://github.com/nwnoll/taxalogue).

Backbone taxonomy

taxalogue automatically downloads backbone taxonomy files and imports them into an SQLite (Hipp, 2022) database. *taxalogue* relies on a backbone taxonomy database to check and format taxonomic information from multiple sources. Users can use the "setup" subcommand to reset the taxonomies or import them separately. We optimized the database model for query speed through indexing, which decreases program runtime after the database has been built. However, importing millions of taxonomic records into the database will take hours, depending on the machine used. *taxalogue* provides the option to use the GBIF backbone taxonomy (GBIF Secretariat, 2021), NCBI taxonomy (Schoch et al., 2020) or none. *taxalogue* resolves and imports homonyms based on a list provided by GBIF (GBIF, 2022).

Download

taxalogue collects data from up to three different online sources to generate various outputs that users could use as a reference database for taxonomic assignment of DNA sequences. The online sources currently available are BOLD (http://www.boldsystems.org/), NCBI GenBank (www.ncbi.nlm.nih.gov/genbank/), and GBOL (https://bolgermany.de/gbol1/ergebnisse/results). The retrieval of sequences and specimen information, such as taxonomic name and locality, varies between the three sources, as explained below. To prevent unnecessary downloads, *taxalogue* checks if the user has already downloaded records for a taxon.

NCBI GenBank: Many attempts to download records via web queries (e.g., https://eutils.ncbi.nlm.nih.gov/entrez/eutils/) yielded incomplete downloads, even if we implemented the recommended waiting times. Therefore, the primary download strategy used in *taxalogue* is downloading the whole GenBank release for the user-specified taxon. If, for

example, the user wants records of the taxon Arthropoda, *taxalogue* will download all invertebrate records (gbinv*.seq.gz) from the latest GenBank release (https://ftp.ncbi.nlm.nih.gov/genbank/). We implemented waiting times to avoid server overload. If a download fails, *taxalogue* restarts it after an extended waiting period. The download of the current GenBank release ensures the complete retrieval of all records for a particular taxon but has the disadvantage of needing more disk space.

BOLD: The user-specified taxon is queried against the public data API (http://www.boldsystems.org/index.php/api_home) for combined data. In general, queries for taxa with many records available, as in Arthropoda, will fail. The taxon for which the download failed will be subdivided into the next lower taxa to circumvent this problem. Supported taxonomic ranks are kingdom, phylum, class, order, family, genus, and species. As with the NCBI GenBank download, we implemented waiting times and retries to avoid overloading the server. We parallelized the download to speed up data retrieval, and the user can specify the number of threads used to retrieve the records. The default value is five threads, and users should make changes with caution. Too many threads could overload the BOLD server and ultimately result in a complete shutdown for the user. Because of this, we recommend not to increase the number of threads used simultaneously to more than the default value of five.

A downside of the taxonomic subdivision into lower ranks is that records only determined to the taxon rank for which the download has failed ~~won't~~ be included. If, for example, the user specified to download all Arthropoda records, the downloaded results will not include those records that have only Arthropoda as name information. However, it is a benign problem since higher taxonomic ranks (e.g., Arthropoda) would still be covered by lower ranks (e.g., Coleoptera) of that taxon in the subsequent taxonomic assignment step. This is because taxonomic assignment to higher ranks requires less sequence similarity than lower taxonomic ranks. These are rare cases, and records with a greater taxonomic resolution are preferred.

GBOL: The latest GBOL dataset release (bolgermany.de/gbol1/release/GBOL_Dataset_Release-20210128.zip) is provided as a zip file. *taxalogue* will download the file and extract the CSV file. Since the GBOL release has some rank inconsistencies, meaning that not all ranks are used at the same position in the higher classification, *taxalogue* will add those missing ranks. Depending on the user-specified options, this might be necessary to enable merging of all three source databases. The GBOL database is intended as a reference barcode source for Germany. Therefore, it consists mainly of specimens collected in Germany. Since these specimens might also occur in neighboring countries or could be invasive in, for example, North America, it might still be of value to include these records in reference databases for studies from other countries.

Filtering

The user can filter records by properties such as the number of ambiguous bases (Ns), length, minimal available taxonomic rank, and others. More information is available with the "filter --help" command. It is also possible to only retain records collected in one or multiple countries,

continents, or biogeographic realms ("region --help" will provide more information). Since some records have the same sequence, a dereplication step is applied by default. Dereplication removes redundant data and decreases the size of the reference database, which could speed up further downstream analysis. During dereplication, multiple comparisons occur if records have the same sequence but differing taxonomic information. If everything except the taxonomic resolution remained unaltered, the dereplication procedure will favor records with greater taxonomic resolution. The lowest common ancestor is chosen for records with differing taxonomic information at the same rank, given they also have the same number of records. *taxalogue* will choose a record as the correct one if it has more records. Even though we are aware that this is subject to taxonomic bias, it is a pragmatic way to conserve taxonomic resolution; for a reference, see Leray et al., 2019, who investigated clusters with multiple taxon names, and in 95% of cases the most abundant taxon name was labeled as the correct one. *taxalogue* processes the GenBank format and amino acid translation with functions from the ruby gem "bio" version 2.0.1 (Goto et al. 2010).

Harmonization

Harmonization means that the taxonomy of a record is mapped onto a backbone taxonomy. The taxonomy from the downloaded record is mapped against, for example, the NCBI taxonomy, and only the standard ranks (kingdom, phylum, class, order, family, genus, and species) will be displayed in the reference database. This action is optional and does not need to be used, although it is the current default setting (to disable harmonization, use the "taxonomy --unmapped" option). It also checks if the taxon of the record is the currently accepted taxon, according to the backbone taxonomy. If the downloaded record has a taxon name considered a synonym, it will replace the name with the accepted name unless the user allows synonyms. This action will be noted and is available in the comparison file. If *taxalogue* could find neither the accepted name nor a synonym, the next higher taxon from the downloaded record is checked against the backbone taxonomy until it finds a match. If it finds a match, it will display the matched higher rank as the actual determination. This action is not without drawbacks and is, therefore, optional. Since some taxonomic classifiers compare the taxon information of each rank, synonyms would be regarded as different taxa and result in a lower bootstrap value, which could lead to the exclusion of some ranks for some sequences.

The already mentioned "taxonomy --unmapped" option does not do any harmonization. It merges the downloads without mapping them onto a backbone taxonomy. This has some consequences, for example: the records from the GBOL Database provide the kingdom name Animalia, whereas the NCBI GenBank records use the name Metazoa, and the BOLD records do not have any kingdom information available. The same taxa with differing taxonomic information on some ranks might affect downstream analysis. If the user runs *taxalogue* with the "--unmapped" option, users should be aware that different taxonomic classifications within your dataset might occur. A ruby script "scripts/replace_taxon_name_for_rank.rb" can change taxon names for each rank.

Name cleaning

Since many names from online sources include digits or terms specifying accuracy and are not part of a valid taxonomic name, some name cleaning will be performed. Digits are not allowed and will be erased from the name. Terms belonging to open nomenclature, like aff., cff. and others were taken from Matthews, 1973 and will be erased, leaving only the name parts that could be considered valid (e.g., "*Apis* cf. *mellifera*" would result in "*Apis*"). Or in other words: *taxalogue* only uses name parts, where the identifier of that particular specimen has been sure about the correctness of the identification. Also, other name parts as sp. or spp. will be erased. If harmonization is enabled and no representative of this name could be found for this name, the ruby library biodiversity (~> 5.1, >= 5.1.2) is used to check if the name stem could be found. This is helpful for rare cases where some record names have used suffixes that are not present in the name of the backbone taxonomy.

Output formats

taxalogue provides multiple output formats for the reference database. Differing output formats provide distinct information depth. The table format is a tab-separated text file that contains location information. *taxalogue* creates it by default and is required for some optional processing (e.g., "scripts/replace_taxon_name_for_rank.rb" relies on the table file). A fasta file and a comparison file are also created by default. The comparison file shows the accepted names according to a chosen backbone taxonomy and their synonyms. Additionally, output files in the format for dada2, kraken2, qiime2, SINTAX can be generated.

Case Study

To test a reference database created by *taxalogue* against three published CO1 reference databases, we searched metabarcoding publications for OTU sequences or mock communities to use them as queries. The tested reference databases consist of records from different sources and filtering procedures (see Table 1). The used query datasets are shown in Table 2 and were selected to cover different regions of the world and different sampling methods. Any preprocessing and filtering of the databases is described in "ref_db_taxalogue/worklow_ref_db_taxalogue.txt" and in "benchmark/workflow_benchmark.txt".

The main method used to compare the reference databases was a top-hit identity distribution (THID) (Edgar, 2018). A THID shows the distances between a query dataset, e.g., OTU sequences, and a reference database. The number of best hits between a query sequence and a reference database is used herewith as a function of sequence identity. We generated the THIDs with VSEARCH version 2.14.1 (Rognes et al., 2016), with the "--usearch_global" (Edgar, 2010) command and the essential options "--id 0.7 --maxaccepts 8 --maxrejects 128 --top_hits_only --maxhits 1 --userfields query+target+id". Computed identities were subsequently rounded to integers and summarized with a custom script. We created Figs. 2-4 with R version 4.1.3 (R Core

Team, 2022). See the folder "benchmark" in the supplements for complete commands, scripts, and the whole workflow.

Based on the aforementioned THID data, we calculated ranks for all reference database/query combinations at 100% identity. Ranks ranged from 1 to 4, whereas rank 1 means the fewest best hits at 100% identity and rank 4 the most. Violin and box plots are shown in Fig. 3 to illustrate the performance and consistency of the individual databases.

We further investigated the midori, *taxalogue*, and tidybug reference databases: 10 x 5,000 sequences were randomly subsampled for each reference database with the "--fastx_subsample" option of VSEARCH version 2.14.1 (Rognes et al., 2016). Each subsample was subsequently used as a query against all reference databases, itself excluded, with the same commands as for the THID generation. We excluded the porter reference database for this benchmark since it only included a subset of BOLD records until the end of 2015 and was primarily composed of GenBank records. This biases the representation of subsampled sequences. By chance alone, fewer BOLD than GenBank sequences would be sampled as queries; therefore, tidybug, with only BOLD sequences, gets fewer best hits at high identities.

Results

The top-hit identity distribution (THID) for four reference databases and eight distinct query datasets is shown in Fig. 2. The THIDs show how well a reference database represents a query. Most THIDs show a skew to the left to higher identities, which means that the highest proportion of queries has their best matches to very similar reference database sequences. Therefore, reference databases with more hits at high percent similarities have better coverage of the queries. The reference database created by *taxalogue* shows for most queries the best coverage at high identities and fewer hits with low identity. This is also true for the tidybug and midori reference databases, but here we see more variation with different query datasets (see Fig. 3). The porter reference database reflects, in all cases, the query datasets the least good. However, kick samples from Canada (see Fig. 2C) had a peak at 98% sequence identity, with only a small number of best hits at higher identities. Additionally, Malaise trap samples from China (see Fig. 2E) had a peak around 84% sequence identity and a smaller, second peak at 100% identity for most reference databases.

In Fig. 4, the THIDs of three reference databases are shown with subsampled queries taken from other reference databases than themselves. *taxalogue* had the most hits at 100% identity with sequences from midori or tidybug. Accordingly, the reference databases midori and tidybug had more hits with lower identities, like 99% and 98%. This shows that *taxalogue* provides more exclusive sequences and generally offers better coverage than the other reference databases.

Discussion

We presented *taxalogue*, a new toolkit to create reproducible reference databases. Using a case study, we showed that *taxalogue* creates reference databases that generally best represent the test cases from multiple areas and trapping devices. *taxalogue* addresses mentioned issues of the current source and reference databases. However, some problems require a major structural change in the source databases, and our approach represents only the most appropriate solution under the given circumstances.

Comprehensive reference databases

Reference databases are the foundation of taxonomic identification via metabarcoding, and resulting taxa lists depend on the quality of the underlying reference database. Therefore, creating a reference database should have a high priority. We showed in our case study that combining records from multiple source databases generally leads to a better representation of the test cases. A better representation with higher identities between query and reference database is crucial for correct taxonomic predictions (Edgar, 2018). *taxalogue* produces a reference database with the best coverage at high identities for most tested queries, enabling more accurate and reliable predictions with higher certainty than the other reference databases tested . Yet, we cannot conclude that a better representation by sequence identity would result in a more reliable reference database per se. More extensive reference databases have higher coverage, but this may be due to records with incorrect annotations (Edgar, 2018). The reference database from the case study created with *taxalogue* consists of records from three source databases. This potentially increases the total amount of erroneous records since several cases of misidentifications have been found in the source databases GenBank and BOLD (e.g., Meier & Dikow, 2004; Becker, Hanner & Steinke, 2011; Lis & Lis, 2011; Lis, Lis & Ziaja, 2016; Jin et al., 2020; Radulovici et al., 2021; Kjærandsen, 2022). But since Leray et al., 2019 found a surprisingly low error rate in GenBank for animal CO1 sequences at the genus level and similar results were found at the species level in a study that investigated both GenBank and BOLD (Jin et al., 2020), the baseline of expected errors should be low. Furthermore, the also included GBOL source database has more strict quality standards, and only records from species experts are accepted (Coleman & Radulovici, 2020); even though this has not been empirically tested, we would expect a similar error rate for records from GBOL. Although the general trend of source database quality points in a positive direction, the methodology of the aforementioned studies prevents a conclusion in this regard. Leray et al., 2019 did not investigate incongruities of species names. Due to increased difficulty in assigning species names (e.g., Sweeney et al., 2011, Ko et al., 2013), we expect a higher error proportion at the species level. Since Jin et al., 2020 do not mention any measures to account for synonyms, the true error proportion might also be different. Furthermore, they identified a sequence as erroneous if the second-best hit (best hit would be itself) had a different taxonomic name, potentially leaving out other matches at 100% identity that could tag a record as erroneous.

Mock communities (samples with known compositions) could potentially be used to test the taxonomic assignment from differing reference databases. The results of the reference databases could be compared with the names of the mock community and subsequently summarized as a confusion matrix. However, this approach poses some problems. Since records from the reference databases are usually determined morphologically in the same way as records from the mock communities, the preference for one of these identifications would be arbitrary. Furthermore, taxonomic mismatches between results are potentially due to synonyms or distinct taxonomic opinions of the taxon concept used. In our opinion, it is impossible to make an objective decision to accept the names of a mock community or the names of the reference databases as the truth. Therefore, we refrained from comparing the taxonomic assignments from the reference databases with those from a mock community.

As expected, merging records from commonly used source databases increases the coverage of a reference database (Porter et al., 2014; Macher, Macher & Leese, 2017; Curry et al., 2018; Porter & Hajibabaei, 2018a; O'Rourke et al. 2020; Porter & Hajibabaei, 2020; Robeson et al. 2021; Nakazato & Jinbo, 2022), at least for the reference database created by *taxalogue.* The porter reference database (Porter & Hajibabaei, 2018b), which also consists of records from GenBank and BOLD (see Table 1), on the other hand, has the lowest coverage of all tested queries. This is explainable because the last retrieval of GenBank records is from 2019, and it only uses the BOLD data releases, which are no longer updated since the end of 2015. The porter reference database also only uses records identified at species level, thereby discarding many records. This point illustrates that the source usage and the filtering of reference databases directly impact taxonomic coverage. However, an unexpected result is that the *taxalogue* reference database has a lower number of best hits at 100% identity with the Honduras query (see Fig. 2D). Since *taxalogue* downloaded records for all Arthropoda from BOLD, just as was done for tidybug. Still, tidybug has better coverage of the Honduras query, so either some records have been deleted from BOLD in the meantime, or *taxalogue* failed to download the respective records. After examining the missing sequences, we found that the missing sequences belonged to taxa with too many records in most cases. As mentioned in the Methods section, downloads from taxa with numerous records are rarely successful due to read timeouts. *taxalogue* circumvents this problem by subdividing the failed taxon into lower taxa. This leads to the problem that records only identified until the failed taxon level won't be included. However, it is a relatively benign problem since the taxonomic resolution of those missing sequences is low (mostly family level and above). As this only occurs with record-rich taxa, we would expect a sufficient number of records with a high enough sequence identity to assign queries to, at least, the missing level. This problem emphasizes the need for reproducible and transparent creation of reference databases. Since *taxalogue* logs the essential steps of the reference database generation, such issues are quickly resolved. Furthermore, it makes the creation of a reference database reproducible, which is indispensable for future replication or comparison.

Reproducibility

Many BOLD records are private and not downloadable. Therefore, none of the reference databases tested do include private records. Solely consisting of downloadable records their coverage is of course reduced. For aquatic biota, up to 50% of sequences in some taxa were only available as private records (Weigand et al., 2019). The BOLD identification system allows comparing user-provided queries with private records, and included 9,458,738 records that could be used if private and public sequences were considered, but only 2,429,025 records were available if choosing only public sequences (accessed on the 4th of March 2022). Identification including private records increased the success rate from 43.3% to 78.6% for invasive pests, when using records from BOLD only (Madden et al., 2019). However, the usage of private records is flagged with a warning since the underlying database consists of unvalidated information. Furthermore, if the user compares the queries against all barcode records, no probability of placement is available. Another issue with this approach is that the records cannot be investigated and filtered based on meta-information or sequence quality. If a query has a hit with a private record, the user cannot investigate the sequence, which did cause problems in diagnosing pests (Hodgetts et al., 2016). And since the BOLD source database constantly changes, the taxonomic identification is not reproducible (Federhen, 2011).

For some private data, thorough reprocessing and curating misidentifications within the BOLD workbench might be the most important reason to delay a release (Becker et al., 2011). Additionally, the BOLD identification engine is largely a "black box" where the exact classification method is unknown. Several studies showed that classification methods varied in suitability on distinct reference databases compositions (e.g., Meier et al, 2006; Wilson et al., 2011; Virgilio, 2012; Bergsten, 2012; Lou & Golding, 2012), so adjusting the classification method to the used reference database is crucial. In response to Federhen, 2011, BOLD added the option to identify a query against an annually created, time-stamped and archived reference database version (Ratnasingham and Hebert, 2011). These archived versions are a snapshot in time. They do not consider information deleted or changed over a year and therefore do not provide a reproducible identification if a user chooses the current version. The current version can change just within one day. Identification with a current version could consequently result in different outcomes within a single day. A reproducible identification with private data could only be achieved if the identification was based on one of the archived versions of the reference database, and only if BOLD does not change the classification method. However, the usage of an archived version seems very unlikely since the latest version is from July 2019. BOLD did not add any versions for 2020, 2021, or 2022. Archived versions are also not available for fungal or plant records. To preserve reproducibility and good scientific practice we refrained from adding any functionality that would incorporate private data. Nonetheless, software solutions providing this service have been developed (e.g., https://github.com/VascoElbrecht/JAMP; Yang et al., 2020; Buchner et al., 2021).

Geographic scale of reference databases

A reference database should be tailored to the needs of a particular research question. Our case study compared global reference databases for Arthropoda, whereas a global scope might not be necessary for other research questions and could even hamper taxonomic identification (Bergsten et al., 2012). Using larger reference databases will certainly increase the number of erroneous sequences, although it is unclear if it increases the proportion of false positives. But using less comprehensive databases comes with the cost of potential false negatives. The main incentive to have an extensive database is to identify organisms at a higher taxonomic resolution with a more reliable identification (Meyer & Paulay, 2005; Vences et al., 2005; Ekrem, Willassen & Stur, 2007). Since a comprehensive database is also needed to distinguish closely related taxa with a great range (Lou & Golding, 2012; Geiger et al., 2016b), it is unknown at what point a local database might be the right choice to avoid the effect of decreased interspecific divergence of allopatrically distributed sister taxa in a geographically expanded dataset (Bergsten et al., 2012). Furthermore, the effects of geographical scale will differ between taxa and areas, and local reference databases could exclude invasive species or populations that have been recently shifting their ranges (Bergsten et al., 2012).

Which form of error is more acceptable has to be decided individually for each research question and could guide the reference database creation. To our knowledge, no current software is available with more extensive geographical filtering options than *taxalogue*. Reference databases could be filtered by multiple countries, continents, biogeographic realms, ecoregions, and even custom shapefiles. Geographic filtering reduces the effect of lower identification success due to a decreased genetic differentiation between closely related taxa in geographically broader reference databases (Bergsten et al., 2012). However, since online source databases hold records with missing location information (Nilsson et al., 2006; Porter & Hajibabaei, 2018a), or the available records are not evenly distributed across countries and continents (Porter & Hajibabaei, 2018a), geographical filtering has its limitations. Additionally, records rarely possess information about the coordinate reference system used – although most GPS trackers use the WGS84 (EPSG:4326) by default.

Taxonomic harmonization

Some data aggregators approximate a long-envisioned unitary taxonomy: a consensus classification and an entry point for additional taxonomic and nomenclatural information (Thompson, 1993; Godfray, 2002). *taxalogue* uses such unitary taxonomies to harmonize taxon names automatically. Harmonized taxon names are helpful due to the increasing usage of hierarchical classifiers in the taxonomic assignment step of a metabarcoding pipeline (Piper et al. 2021). Hierarchical classifiers depend on the taxonomic congruency between records since incongruent taxonomic information would introduce an artificial bias, leading to decreased identification success with lower taxonomic resolution. Other classification methods also benefit from harmonized reference databases because otherwise, a reference database could simultaneously consist of synonyms and the currently accepted name for one taxon, resulting in arbitrary assignments to the accepted or synonymized name. Taxonomic harmonization is already applied directly in NCBI and BOLD (Schoch et al., 2020) and indirectly through the automated

identification of specimens without prior taxonomic assignment (Ratnasingham and Hebert, 2007). To what extent users have harmonized identifications before uploading data to the source databases and on what basis is unknown. This indicates that taxonomic harmonizations occur to different and partly unknown degrees, even within a single source database. Data integration across multiple source databases, as in our test case, amplifies this problem since the records from different sources might also be harmonized to varying degrees. Piper et al., 2021 recommend taxonomic harmonization as a default step, just as other filtering procedures.

Even though taxonomic harmonization provides a clear advantage for further downstream analysis, criticism exists against synchronizing data to a particular unitary taxonomy. Such a taxonomy is algorithmically or socially resolved, even if no consensus has yet been reached in the taxonomist community (Senderov, 2018). A synthesized conclusion without clear consensus is suspected to decrease taxonomic stability (Pauly, Hillis & Cannatella, 2009) and trust in data aggregators (Franz, 2018). Although macroecologists, conservationists, administrators and others depend on stable species lists for reliable predictions (Hey et al., 2003; Isaac, Mallet & Mace, 2004; Padial & De la Riva, 2006), the independence of taxonomy as a scientific endeavor has been stressed to be of utmost importance (e.g., Dubois, 1998). A top-down administration is in stark contrast to taxonomic tradition (Godfray, 2002), where a taxon could be seen as a falsifiable scientific hypothesis that has to withstand time (Haszprunar, 2011). Scientists expressed concerns that such an administration would lead to authoritarianism (Thiele and Yeates, 2002) and about the data quality of biodiversity data aggregators (e.g., Franz 2018). Even though we should preserve taxonomic independence, a non-taxonomist still has difficulties deciding which taxon name is most appropriate (Grenié et al., 2022). This problem is aggravated when very diverse taxa are studied or when different data sources are used (Sterner and Franz., 2017). Users should weigh the advantages of taxonomic harmonization against the disadvantages and decide accordingly.

taxalogue harmonizes with global backbone taxonomies, but regional or taxon-specific taxonomies may better represent the scientific consensus for that particular group. Since integrating many specialized taxonomies, with distinct scales and taxonomic breadth, is an enormous challenge and selecting appropriate taxonomies would still be opinion-based, we provide the commonly used NCBI Taxonomy, GBIF Taxonomy or no harmonization at all as options in *taxalogue*. However, we would like to point out that, for example, a specialized taxonomic harmonization, as found in Arranz et al., 2020, might be a more appropriate choice for marine samples.

Taxon concepts

Several studies showed that using taxon concepts *sensu* Berendsohn, 1995 is necessary to unambiguously determine the meaning of a taxon name (e.g., Berendsohn, 1995; Kennedy, Kukla & Paterson, 2005; Franz, Peet & Weakley, 2006). However, current source databases for sequence data do not provide this information. Of the major source databases, only BOLD provides a separate field for the used identification literature, which could help to derive the used

taxon concept. Unfortunately, providing information for this field is not an obligatory upload prerequisite. Furthermore, BOLD did not define this field's semantics. Therefore, it is unclear how to use this information. Consequently, reference database creation tools cannot provide taxon names in combination with the used taxon concepts. Instead, *taxalogue* approximates the idea of a reconciliation group (Patterson et al., 2010) with the option to generate a "comparison" file. This file aggregates previously used names for a taxon and aims to ease the information retrieval for all taxon names in the reference database. The taxonomic database Avibase is an example of how taxon concepts have already been implemented successfully (Lepage, Vaidya & Guralnick, 2014) and could guide further improvement of the source databases.

Outlook

Since *taxalogue* combines sequences from up to three source databases, a user can achieve comprehensive coverage without relying on private and unreliable data, which posed problems in the past (e.g., Federhen, 2011; Hodgetts et al., 2016). As a result, the reference database is reproducible and can be tailored to the particular research question. With comprehensive options to define the scale of the reference database, the user can exploit the advantages of a comprehensive (Meyer & Paulay, 2005; Vences et al., 2005; Ekrem et al., 2007) and a local database (Bergsten et al., 2012) simultaneously. Furthermore, the options for taxonomic harmonization unlock the possibility of investigating their effects on the interpretation of taxa lists. The latter points to potential questions that future research still needs to address: To what extent do taxonomic harmonizations influence the significance of metabarcoding results? Or whether the absence of the taxon concept *sensu* Berendsohn, 1995 in the source databases impedes the application of metabarcoding for ecological or macroevolutionary questions?

Acknowledgments

We would like to thank Leighton Thomas for proofreading an earlier draft of the manuscript.

References

Agnarsson, I., & Kuntner, M. (2007). Taxonomy in a Changing World: Seeking Solutions for a Science in Crisis. *Systematic Biology*, *56*(3), 531–539. https://doi.org/10.1080/10635150701424546

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). https://doi.org/10.1038/s41597-020-0549-9

Becker, S., Hanner, R., & Steinke, D. (2011). Five years of FISH-BOL: Brief status report. *Mitochondrial DNA*, *22*(SUPPL. 1), 3–9. https://doi.org/10.3109/19401736.2010.535528

Bengtsson-Palme, J., Richardson, R. T., Meola, M., Wurzbacher, C., Tremblay, É. D., Thorell, K., … 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

Berendsohn, W. G. (1995). The Concept of "Potential Taxa" in Databases. *Taxon*, *44*(2), 207–212.

Berendsohn, W. G., & Geoffroy, M. (2016). Networking Taxonomic Concepts - Uniting without 'Unitary-ism.' In *Biodiversity Databases: Techniques, Politics, and Applications* (pp. 13–22). https://doi.org/10.1201/9781439832547-3

Bergsten, J., Bilton, D. T., Fujisawa, T., Elliott, M., Monaghan, M. T., Balke, M., … Vogler, A. P. (2012). The Effect of Geographical Scale of Sampling on DNA Barcoding. *Systematic Biology*, *61*(5), 851–869. https://doi.org/10.1093/sysbio/sys037

Bortolus, A. (2008). Error Cascades in the Biological Sciences: The Unwanted Consequences of Using Bad Taxonomy in Ecology. *AMBIO: A Journal of the Human Environment*, *37*(2), 114–118. https://doi.org/10.1579/0044-7447(2008)37[114:ECITBS]2.0.CO;2

Coleman, C. O., & Radulovici, A. E. (2020). Challenges for the future of taxonomy: talents, databases and knowledge growth. *Megataxa*, *1*(1), 28–34–28–34. https://doi.org/10.11646/megataxa.1.1.5

Collins, R. A., & Cruickshank, R. H. (2013). The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*, *13*(6), 969–975. https://doi.org/10.1111/1755-0998.12046

Curry, C. J., Gibson, J. F., Shokralla, S., Hajibabaei, M., & Baird, D. J. (2018). Identifying north American freshwater invertebrates using DNA barcodes: Are existing COI sequence libraries fit for purpose? *Freshwater Science*, *37*(1), 178–189. https://doi.org/10.1086/696613

Dubois, A. (1998). Lists of European species of amphibians and reptiles: Will we soon be reaching "stability"? *Amphibia Reptilia*, *19*(1), 1–28. https://doi.org/10.1163/156853898X00304

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Edgar, R. C. (2018). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ*, *2018*(4), e4652. https://doi.org/10.7717/peerj.4652

Ekrem, T., Willassen, E., & Stur, E. (2007). A comprehensive DNA sequence library is essential for identification with DNA barcodes. *Molecular Phylogenetics and Evolution*, *43*(2), 530–542. https://doi.org/10.1016/j.ympev.2006.11.021

Federhen, S. (2011). Comment on 'Birdstrikes and barcoding: can DNA methods help make the airways safer?' *Molecular Ecology Resources*, *11*(6), 937–938. https://doi.org/10.1111/j.1755-0998.2011.03054.x

Fišer Pečnikar, Ž., & Buzan, E. V. (2014). 20 years since the introduction of DNA barcoding: From theory to application. *Journal of Applied Genetics*, Vol. 55, pp. 43–52. https://doi.org/10.1007/s13353-013-0180-y

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

Franz, N. M. (2005). On the lack of good scientific reasons for the growing phylogeny/classification gap. *Cladistics*, *21*(5), 495–500. https://doi.org/10.1111/j.1096-0031.2005.00080.x

Franz, N. M., & Sterner, B. W. (2018). To increase trust, change the social design behind aggregated biodiversity data. *Database*, *2018*(2018). https://doi.org/10.1093/database/bax100

Franz, N. M., Peet, R. K., & Weakley, A. S. (2006). On the Use of Taxonomic Concepts in Support of Biodiversity Research and Taxonomy. *Systematics Association Special Volume*, *76*.

Geiger, M., Astrin, J., Borsch, T., Burkhardt, U., Grobe, P., Hand, R., Hausmann, A., Hohberg, K., Krogmann, L., Lutz, M., Monje, C., Misof, B., Morinière, J., Müller, K., Pietsch, S., Quandt, D., Rulik, B., Scholler, M., Traunspurger, W., … Wägele, W. (2016a). How to tackle the molecular species inventory for an industrialized nation-lessons from the first phase of the German Barcode of Life initiative GBOL (2012-2015) 1. *Genome*, *59*(9), 661–670. https://doi.org/10.1139/gen-2015-0185

Geiger, M. F., Moriniere, J., Hausmann, A., Haszprunar, G., Wägele, W., Hebert, P. D. N., & Rulik, B. (2016b). Testing the Global Malaise Trap Program - How well does the current barcode reference library identify flying insects in Germany? *Biodiversity Data Journal*, *4*(1). https://doi.org/10.3897/BDJ.4.e10671

Godfray, H. C. J. (2002). Challenges for taxonomy. *Nature*, Vol. 417, pp. 17–19. https://doi.org/10.1038/417017a

Goto, N., Prins, P., Nakao, M., Bonnal, R., Aerts, J., & Katayama, T. (2010). BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics*, *26*(20), 2617–2619. https://doi.org/10.1093/bioinformatics/btq475

Grenié, M., Berti, E., Carvajal Quintero, J., Dädlow, G. M. L., Sagouis, A., & Winter, M. (2022). Harmonizing taxon names in biodiversity data: a review of tools, databases, and best practices. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210x.13802

Grimaldi, D., & Engel, M. (2005). *Evolution of the Insects*. Cambridge University Press.

Haszprunar, G. (2011). Species delimitations-not “only descriptive.” *Organisms Diversity and Evolution*, Vol. 11, pp. 249–252. https://doi.org/10.1007/s13127-011-0047-1

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 of the Royal Society B: Biological Sciences*, *270*(SUPPL. 1). https://doi.org/10.1098/rsbl.2003.0025

Heller, P., Casaletto, J., Ruiz, G., & Geller, J. (2018). Data Descriptor: A database of metazoan cytochrome c oxidase subunit I gene sequences derived from GenBank with CO-ARBitrator. *Scientific Data*, *5*(1), 1–7. https://doi.org/10.1038/sdata.2018.156

Hey, J., Waples, R. S., Arnold, M. L., Butlin, R. K., & Harrison, R. G. (2003). Understanding and confronting species uncertainty in biology and conservation. *Trends in Ecology and Evolution*, Vol. 18, pp. 597–603. https://doi.org/10.1016/j.tree.2003.08.014

Hipp, D. R. (2022). SQLite.

Hobern, D., & Hebert, P. (2019). BIOSCAN - Revealing Eukaryote Diversity, Dynamics, and Interactions. *Biodiversity Information Science and Standards*, *3*. https://doi.org/10.3897/biss.3.37333

Hodgetts, J., Ostojá-Starzewski, J. C., Prior, T., Lawson, R., Hall, J., Boonham, N., & Wilson, J. J. (2016). DNA barcoding for biosecurity: Case studies from the UK plant protection program1. *Genome*, *59*(11), 1033–1048. https://doi.org/10.1139/gen-2016-0010

Huemer, P., Mutanen, M., Sefc, K. M., & Hebert, P. D. N. (2014). Testing DNA Barcode Performance in 1000 Species of European Lepidoptera: Large Geographic Distances Have Small Genetic Impacts. *PLoS ONE*, *9*(12), e115774. https://doi.org/10.1371/journal.pone.0115774

Isaac, N. J. B., Mallet, J., & Mace, G. M. (2004). Taxonomic inflation: Its influence on macroecology and conservation. *Trends in Ecology and Evolution*, *19*(9), 464–469. https://doi.org/10.1016/j.tree.2004.06.004

Jin, S., Kim, K. Y., Kim, M. S., & Park, C. (2020). An assessment of the taxonomic reliability of dna barcode sequences in publicly available databases. *Algae*, *35*(3), 293–301. https://doi.org/10.4490/algae.2020.35.9.4

Keck, F., & Altermatt, F. (2022). Management of DNA reference libraries for barcoding and metabarcoding studies with the R package refdb. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13723

Kennedy, J. B., Kukla, R., & Paterson, T. (2005). Scientific names are ambiguous as identifiers for biological taxa: Their context and definition are required for accurate data integration. *Lecture Notes in Bioinformatics (Subseries of Lecture Notes in Computer Science)*, *3615*, 80–95. https://doi.org/10.1007/11530084_8

Kjærandsen, J. (2022). Current State of DNA Barcoding of Sciaroidea (Diptera) - Highlighting the Need to Build the Reference Library. *Insects*, *13*(2), 147. https://doi.org/10.3390/insects13020147

Ko, H.-L., Wang, Y.-T., Chiu, T.-S., Lee, M.-A., Leu, M.-Y., Chang, K.-Z., … Shao, K.-T. (2013). Evaluating the Accuracy of Morphological Identification of Larval Fishes by Applying DNA Barcoding. *PLoS ONE*, *8*(1), e53451. https://doi.org/10.1371/journal.pone.0053451

Lepage, D., Vaidya, G., & Guralnick, R. (2014). Avibase - A database system for managing and organizing taxonomic concepts. *ZooKeys*, *420*(420), 117–135. https://doi.org/10.3897/zookeys.420.7089

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

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

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

International Barcode of Life (2022). International Barcode of Life Program Overview. Retrieved from https://ibol.org/programs/program-overview/

Lis, J. A., & Lis, B. (2011). Is accurate taxon identification important for molecular studies? Several cases of faux pas in pentatomoid bugs (Hemiptera: Heteroptera: Pentatomoidea). *Zootaxa*, Vol. 2932, pp. 47–50. https://doi.org/10.11646/zootaxa.2932.1.5

Lis, J. A., Lis, B., & Ziaja, D. J. (2016). In BOLD we trust? A commentary on the reliability of specimen identification for DNA barcoding: A case study on burrower bugs (Hemiptera: Heteroptera: Cydnidae). *Zootaxa*, *4114*(1), 83–86. https://doi.org/10.11646/zootaxa.4114.1.6

Lou, M., & Golding, G. B. (2012). The effect of sampling from subdivided populations on species identification with DNA barcodes using a Bayesian statistical approach. *Molecular Phylogenetics and Evolution*, *65*(2), 765–773. https://doi.org/10.1016/j.ympev.2012.07.033

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

Madden, M. J. L., Young, R. G., Brown, J. W., Miller, S. E., Frewin, A. J., & Hanner, R. H. (2019). Using DNA barcoding to improve invasive pest identification at U.S. ports-of-entry. *PLoS ONE*, *14*(9), e0222291. https://doi.org/10.1371/journal.pone.0222291

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*. https://doi.org/10.1093/database/baac055

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*, *0*(0), 1–13. https://doi.org/https://doi.org/10.1111/1755-0998.13756

Meier, R., & Dikow, T. (2004). Significance of Specimen Databases from Taxonomic Revisions for Estimating and Mapping the Global Species Diversity of Invertebrates and Repatriating Reliable Specimen Data. *Conservation Biology*, *18*(2), 478–488. https://doi.org/10.1111/j.1523-1739.2004.00233.x

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

Meyer, C. P., & Paulay, G. (2005). DNA Barcoding: Error Rates Based on Comprehensive Sampling. *PLoS Biology*, *3*(12), e422. https://doi.org/10.1371/journal.pbio.0030422

Nakazato, T., & Jinbo, U. (2022). Cross-sectional use of barcode of life data system and GenBank as DNA barcoding databases for the advancement of museomics. *Frontiers in Ecology and Evolution*, *10*, 1015. https://doi.org/10.3389/fevo.2022.966605

Nilsson, R. H., Ryberg, M., Kristiansson, E., Abarenkov, K., Larsson, K.-H., & Kõljalg, U. (2006). Taxonomic Reliability of DNA Sequences in Public Sequence Databases: A Fungal Perspective. *PLoS ONE*, *1*(1), e59. https://doi.org/10.1371/journal.pone.0000059

O'Rourke, D. R., Bokulich, N. A., Jusino, M. A., MacManes, M. D., & Foster, J. T. (2020). A total crapshoot? Evaluating bioinformatic decisions in animal diet metabarcoding analyses. *Ecology and Evolution*, *10*(18), 9721–9739. https://doi.org/10.1002/ece3.6594

Padial, J. M., & De la Riva, I. (2006). Taxonomic Inflation and the Stability of Species Lists: The Perils of Ostrich's Behavior. *Systematic Biology*, *55*(5), 859–867. https://doi.org/10.1080/1063515060081588

Palmer, J. M., Jusino, M. A., Banik, M. T., & Lindner, D. L. (2018). Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data. *PeerJ*, *2018*(5), e4925. https://doi.org/10.7717/peerj.4925

Pappalardo, P., Collins, A. G., Pagenkopp Lohan, K. M., Hanson, K. M., Truskey, S. B., Jaeckle, W., … Osborn, K. J. (2021). The role of taxonomic expertise in interpretation

of metabarcoding studies. *ICES Journal of Marine Science*, *78*(9), 3397–3410. https://doi.org/10.1093/icesjms/fsab082

Patterson, D. J., Cooper, J., Kirk, P. M., Pyle, R. L., & Remsen, D. P. (2010). Names are key to the big new biology. *Trends in Ecology and Evolution*, Vol. 25, pp. 686–691. https://doi.org/10.1016/j.tree.2010.09.004

Pauly, G. B., Hillis, D. M., & Cannatella, D. C. (2009). Taxonomic Freedom and the Role of Official Lists of Species Names. *Herpetologica*, *65*(2), 115–128. https://doi.org/10.1655/08-031R1.1

Pentinsaari, M., Ratnasingham, S., Miller, S. E., & Hebert, P. D. N. (2020). BOLD and GenBank revisited – Do identification errors arise in the lab or in the sequence libraries? *PLoS ONE*, *15*(4), e0231814. https://doi.org/10.1371/journal.pone.0231814

Phillips, J. D., Gillis, D. J., & Hanner, R. H. (2019). Incomplete estimates of genetic diversity within species: Implications for DNA barcoding. *Ecology and Evolution*, Vol. 9, pp. 2996–3010. https://doi.org/10.1002/ece3.4757

Piper, A. M., Cogan, N. O. I., Cunningham, J. P., & Blacket, M. J. (2021). Computational Evaluation of DNA Metabarcoding for Universal Diagnostics of Invasive Insect Pests. *BioRxiv*, 2021.03.16.435710. https://doi.org/10.1101/2021.03.16.435710

Porter, T. M., Gibson, J. F., Shokralla, S., Baird, D. J., Golding, G. B., & Hajibabaei, M. (2014). Rapid and accurate taxonomic classification of insect (class Insecta) cytochrome *c* oxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier. *Molecular Ecology Resources*, *14*(5), n/a-n/a. https://doi.org/10.1111/1755-0998.12240

Porter, T. M., & Hajibabaei, M. (2018a). 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. (2018b). Automated high throughput animal CO1 metabarcode classification. *Scientific Reports*, *8*(1), 4226. https://doi.org/10.1038/s41598-018-22505-4

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. https://doi.org/10.3389/fevo.2020.00248

Radulovici, A. E., Vieira, P. E., Duarte, S., Teixeira, M. A. L., Borges, L. M. S., Deagle, B. E., … Costa, F. O. (2021). Revision and annotation of DNA barcode records for marine invertebrates: Report of the 8th iBOL conference hackathon. *Metabarcoding and Metagenomics*, *5*, 207–217. https://doi.org/10.3897/mbmg.5.67862

Ratnasingham, S., & Hebert, P. D. N. (2011). BOLD's role in barcode data management and analysis: A response. *Molecular Ecology Resources*, Vol. 11, pp. 941–942. https://doi.org/10.1111/j.1755-0998.2011.03067.x

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

Richardson, R. T., Bengtsson-Palme, J., Gardiner, M. M., & Johnson, R. M. (2018). A reference cytochrome c oxidase subunit I database curated for hierarchical classification of arthropod metabarcoding data. *PeerJ*, *2018*(6), e5126. https://doi.org/10.7717/peerj.5126

Robeson, M. S., O'Rourke, D. R., Kaehler, B. D., Ziemski, M., Dillon, M. R., Foster, J. T., & Bokulich, N. A. (2021). RESCRIPt: Reproducible sequence taxonomy reference database management. *PLOS Computational Biology*, *17*(11), e1009581. https://doi.org/10.1371/journal.pcbi.1009581

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *2016*(10), e2584. https://doi.org/10.7717/peerj.2584

Matthews, S. C. . (1973). Notes on open nomenclature and on synonymy lists. *Paleontology*, *16*(Part 4), 713–719.

Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Schoch, C. L., Sherry, S. T., & Karsch-Mizrachi, I. (2022). GenBank. *Nucleic Acids Research*, *50*(D1), D161–D164. https://doi.org/10.1093/nar/gkab1135

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., … Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, *2020*. https://doi.org/10.1093/database/baaa062

Senderov, V., Simov, K., Franz, N., Stoev, P., Catapano, T., Agosti, D., … Penev, L. (2018). OpenBiodiv-O: Ontology of the OpenBiodiv knowledge management system. *Journal of Biomedical Semantics*, *9*(1). https://doi.org/10.1186/s13326-017-0174-5

Sperling, F. (2003). Opinion DNA Barcoding: Deus ex Machina. In *Newsletter of the Biological Survey*.

Sterner, B., & Franz, N. M. (2017). Taxonomy for Humans or Computers? Cognitive Pragmatics for Big Data. *Biological Theory*, *12*(2), 99–111. https://doi.org/10.1007/s13752-017-0259-5

Sweeney, B. W., Battle, J. M., Jackson, J. K., & Dapkey, T. (2011). Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? *Journal of the North American Benthological Society*, *30*(1), 195–216. https://doi.org/10.1899/10-016.1

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Thiele, K., & Yeates, D. (2002). Tension arises from duality at the heart of taxonomy. *Nature*, Vol. 419, p. 337. https://doi.org/10.1038/419337a

Thompson, C. F. (1993). Names: the Keys to Biodiversity | The Diptera Site. Retrieved September 5, 2022, from Talk - Biodiversity from 1986 to the 21st Century website: https://diptera.myspecies.info/content/names-keys-biodiversity

Vences, M., Thomas, M., Bonett, R. M., & Vieites, D. R. (2005). Deciphering amphibian diversity through DNA barcoding: Chances and challenges. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1462), 1859–1868. https://doi.org/10.1098/rstb.2005.1717

Virgilio, M., Jordaens, K., Breman, F. C., Backeljau, T., & De Meyer, M. (2012). Identifying Insects with Incomplete DNA Barcode Libraries, African Fruit Flies (Diptera: Tephritidae) as a Test Case. *PLoS ONE*, *7*(2), e31581. https://doi.org/10.1371/journal.pone.0031581

Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., … Ekrem, T. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Science of the Total Environment*, Vol. 678, pp. 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

Wilson, J. J., Rougerie, R., Schonfeld, J., Janzen, D. H., Hallwachs, W., Hajibabaei, M., … Hebert, P. D. N. (2011). When species matches are unavailable are DNA barcodes correctly assigned to higher taxa? An assessment using sphingid moths. *BMC Ecology*, *11*(1), 18. https://doi.org/10.1186/1472-6785-11-18

Yang, C., Zheng, Y., Tan, S., Meng, G., Rao, W., Yang, C., … Liu, S. (2020). Efficient COI barcoding using high throughput single-end 400 bp sequencing. *BMC Genomics*, *21*(1), 862. https://doi.org/10.1186/s12864-020-07255-w

Table 1 (on next page)

Summary of reference databases used in the benchmark.

database=Arthropoda CO1 reference database name, #sequences=total number of sequences, min sequence length=smallest sequence length in reference database, BOLD=download date, GBOL=download date, GenBank=download date, reference=publication reference. *BOLD data releases from December 31, 2010 till December 31, 2015

database	#sequences	min sequence length	BOLD	GBOL	GenBank	reference
midori	2,086,807	100 bp	none	none	2022-02-15	Leray et al., 2018
porter	888,696	500 bp	2015-12-31*	none	2019-04	Porter & Hajibabaei, 2018b
taxalogue	2,921,104	400 bp	2022-02-02	2021-01-28	2021-12-15	this publication
tidybug	1,841,946	100 bp	2019-02-24	none	none	O'Rourke et al., 2020

Table 2 (on next page)

Summary of the query datasets used in the benchmark.

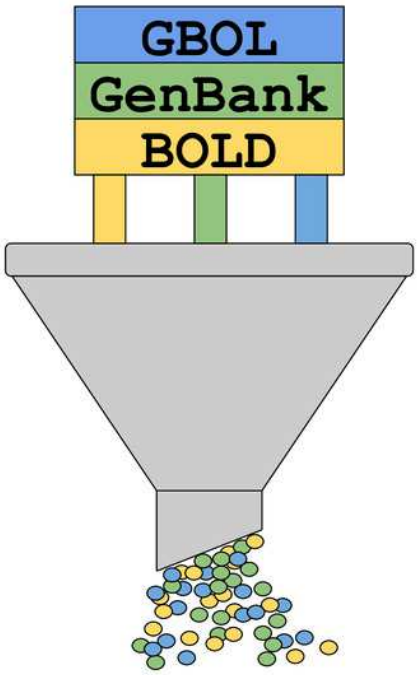
country=country of sample, sampling method=device or method for sampling of specimens, habitat=natural habitat where sampling did take place, *mock=sampled from multiple locations and potentially different habitats, taxon=expected organism group, reference=publication reference.

country	sampling method	habitat	taxon	reference
Canada	kick net	benthic zone	Macrozoobenthos	Gibson et al., 2014
Canada	Malaise trap	grassland, forested pond	Arthropoda	Steinke et al., 2021
China	Malaise trap	mock*	Arthropoda	Yu et al., 2012
China	Malaise trap	mock*	Arthropoda	Yang et al. 2021
Costa Rica	Malaise trap	rainforest	Arthropoda	Gibson et al., 2014
Germany	Malaise trap	meadow	Arthropoda	Elbrecht et al., 2021
Honduras	canopy fogging	canopy	Arthropoda	Creedy et al., 2019
Portugal	automatic light traps	cork oak woodlands	Arthropoda	Mata et al., 2021

Figure 1

Overview of main taxalogue functions from the download of records to output generation.

For more information use taxalogue with “--help”. K=kingdom, P=phylum, C=class, O=order, F=family, G=genus, S=species

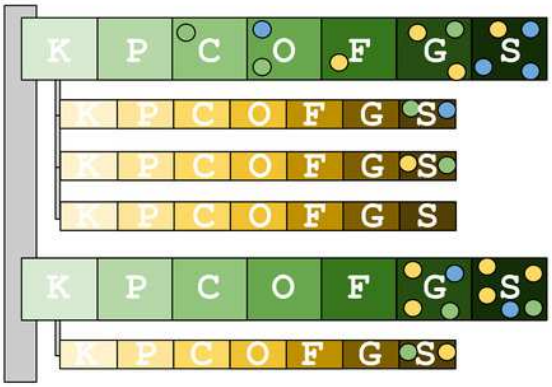


Download

from **all databases**, or pick the ones you want

Filter

by **sequence properties** (e.g. Ns, length)
by **taxonomic lineage** (e.g. name, rank)
by **other metadata** (e.g. location, realm)



Harmonize

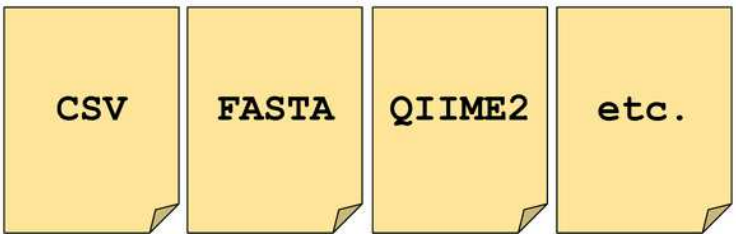
taxon name to a **reference taxonomy** (e.g. NCBI, GBIF) or

allow synonyms according to chosen reference taxonomy

rank	taxon	sequence	#
species	*A. cerana*	ACCTAG	1
species	***A. florea***	**ACCTAG**	**9**
family	Apidae	ACCTAG	5

Dereplicate

and choose taxon if the same sequence has **differing taxon assignments** (e.g. LCA, random)



Output

generate outputs in **several formats** (e.g. QIIME2, FASTA)

Figure 2

Top-hit identity distribution for 4 reference databases and 8 queries.

The number of best hits as a function of sequence identity for a selection of CO1 reference databases and query datasets (see Table 1 for reference database descriptions and Table 2 for the query datasets), depicted as a stair step diagram. identity=percent similarity between a query sequence and its best hit in a reference database. hits=number of best matches between query and reference database sequences at a certain identity percentage.

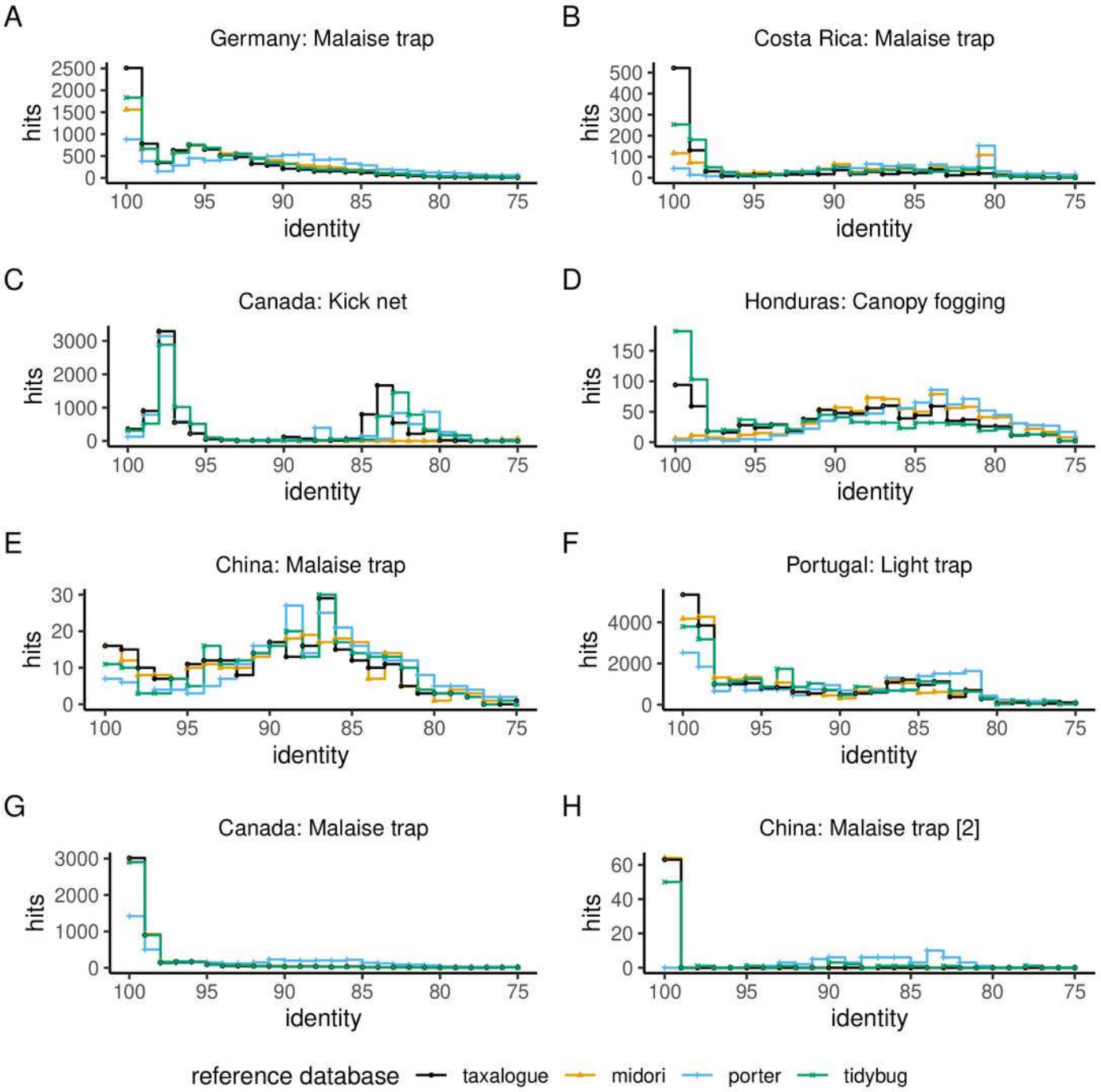

A
Germany: Malaise trap
B
Costa Rica: Malaise trap
C
Canada: Kick net
D
Honduras: Canopy fogging
E
China: Malaise trap
F
Portugal: Light trap
G
Canada: Malaise trap
H
China: Malaise trap [2]
hits
identity
reference database
taxalogue
midori
porter
tidybug

Figure 3

Violin plots showing the top-hit identity distributions at 100% identity from all reference database/query combinations.

The width of a violin indicates how often a reference database achieved a rank; the height shows the variation in achieved ranks. Each reference database was tested with 8 queries. rank=Ranks range from 1 to 4, where rank 1 corresponds to the fewest best hits at 100% identity and rank 4 to the highest number of best hits (if reference databases had the same number of best hits, they share the next lower rank), reference database=Arthropoda CO1 reference database name.

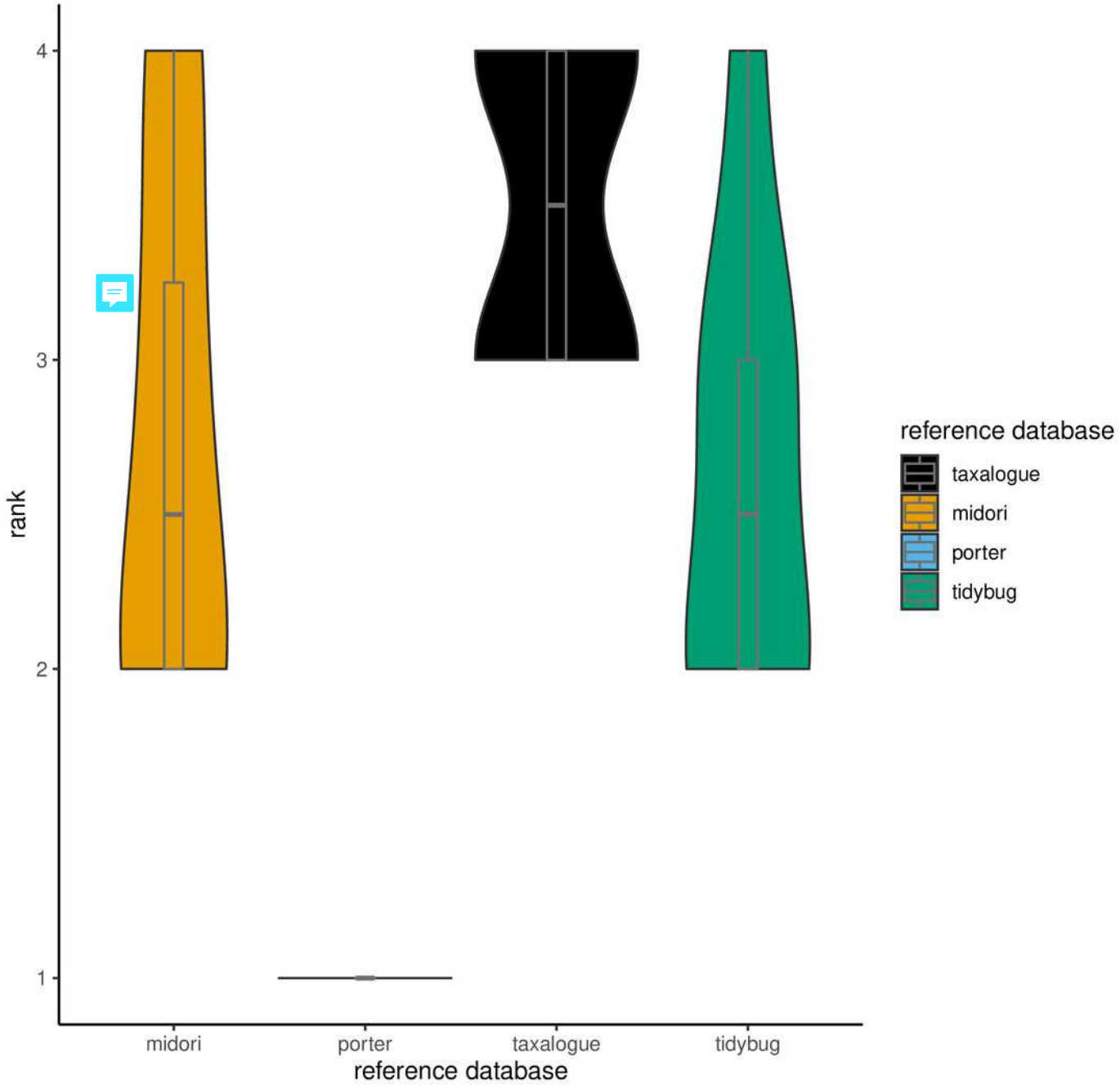
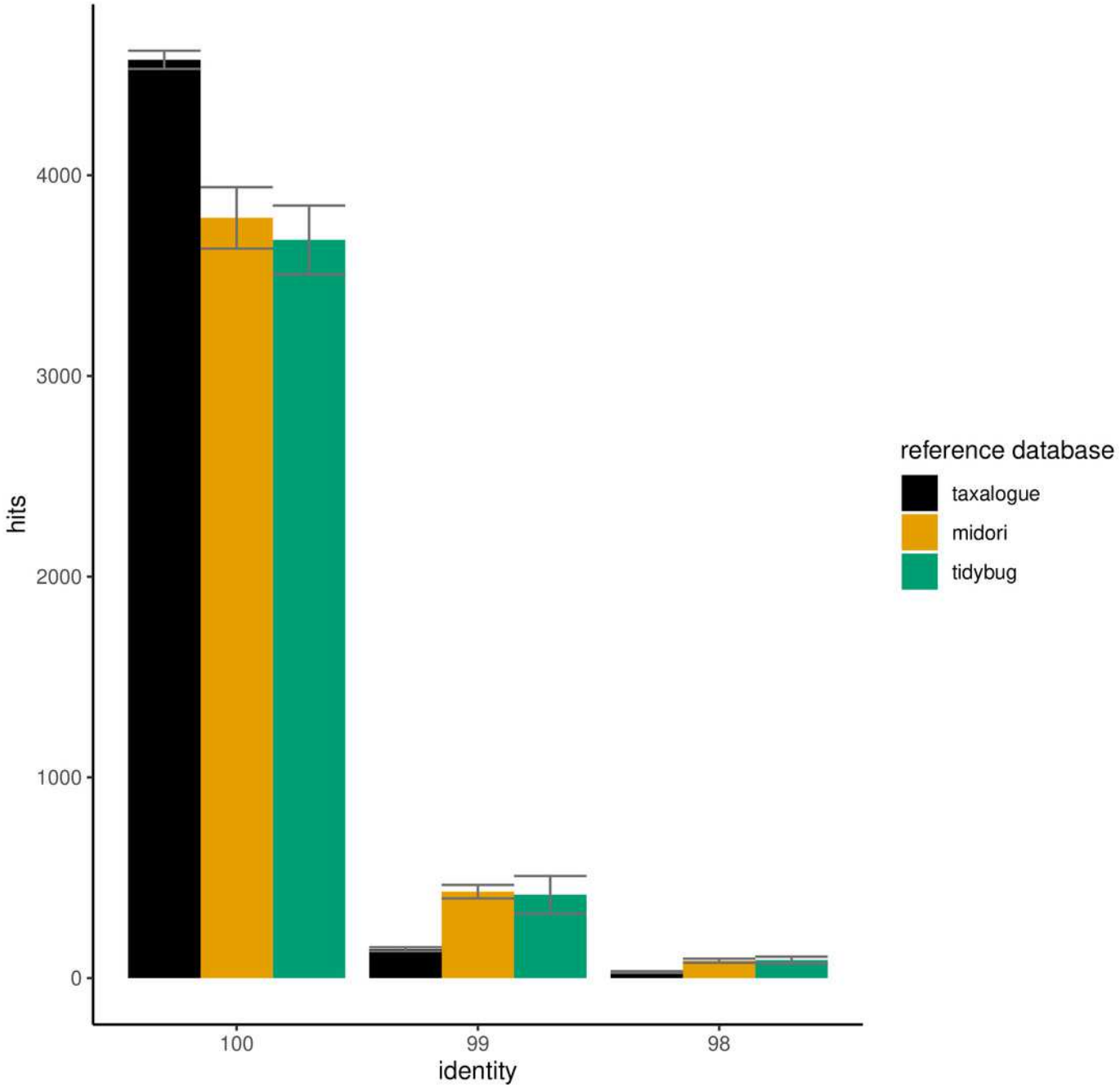

4
3
2
1
rank
midori
porter
taxalogue
tidybug
reference database
reference database
taxalogue
midori
porter
tidybug

Figure 4

Top-hit identity distributions for the taxalogue, midori and tidybug reference databases queried against each other.

Each reference database was queried with 20 * 5,000 randomly selected sequences from all the other reference databases (e.g., taxalogue was queried against sequences from midori and tidybug; midori was queried against sequences from taxalogue and tidybug, etc.).Only hits at 100, 99, and 98 percent identity were considered (see Table 1 for reference database descriptions). Each query consists of 5,000 randomly selected sequences. The whiskers show the standard deviation per reference database at a certain identity. identity=percent similarity between a query sequence and its best hit in a reference database. hits=number of best matches between query and reference database sequences at a certain identity percentage.


hits
4000
3000
2000
1000
0
100
99
98
identity
reference database
taxalogue
midori
tidybug