Reference Sequence Browser: An R application with a User-Friendly GUI to rapidly query sequence databases

Samuel Rapp[1], Jorge Tapias Gomez[2], Sriram Ramesh[1], Benjamin Levine[1], Daniel Tapias-Gomez[3], Dickson Chung[1], and Zia Truong[1]

[1]**University of California Santa Cruz, Santa Cruz, CA, 95064, USA**
[2]**Cornell University, Ithaca, NY, 14850, USA**
[3]**University of Texas Southwestern, TX, 75390, USA**

Abstract

Land managers, researchers, and regulators increasingly utilize environmental DNA (eDNA) techniques to monitor species richness, presence, and absence. In order to properly develop a biological assay for eDNA metabarcoding or quantitative PCR, scientists must be able to find not only reference sequences (previously identified sequences in a genomics database) that match their target taxa but also reference sequences that match non-target taxa. Determining which taxa have publicly available sequences in a time-efficient and accurate manner currently requires computational skills to search, manipulate, and parse multiple unconnected DNA sequence databases.

Our team iteratively designed a Graphic User Interface (GUI) Shiny application called the Reference Sequence Browser (RSB) that provides users efficient and intuitive access to multiple genetic databases regardless of computer programming expertise. The application returns the number of publicly accessible barcode markers per organism in the NCBI Nucleotide, BOLD, or CALeDNA CRUX Metabarcoding Reference Databases. Depending on the database, we offer various search filters such as min and max sequence length or country of origin. Users can then download the FASTA/GenBank files from the RSB web tool, view statistics about the data, and explore results to determine details about the availability or absence of reference sequences.

Keywords

environmental DNA, metabarcoding, shiny, NCBI, BOLD, CRUX, data access, reference sequence

Introduction

Environmental DNA (eDNA) is an emerging field that has helped answer questions in several disciplines, including molecular ecology, environmental sciences, conservation biology, and paleontology (Seymour, 2019; Thomsen and Willerslev, 2015). eDNA techniques are advantageous to traditional monitoring due to their non-invasive nature, ability to monitor aquatic communities, and the relative ease of training field workers.

Environmental DNA workflows begin with acquiring samples from the environment under study so that genetic material can be extracted and then analyzed with various forms of Polymerase Chain Reaction (PCR) assays (Thomsen and Willerslev, 2015). Depending on the specific goals of the study, eDNA primers are designed at varying levels of taxonomic inclusivity, often using either taxa-specific quantitative PCR (qPCR) or DNA metabarcoding PCR. Taxa-specific qPCR uses eDNA to detect a single targeted taxon, while eDNA metabarcoding aims to assess taxa richness.

Both qPCR assay design and taxonomic assignment in metabarcoding depend heavily on the "availability of DNA reference sequences in public data facilities (e.g., National Center for Biotechnical Information (NCBI), The Barcoding Of Life Data System (BOLD))" (Andersen et al., 2016). Reference sequences are pre-labeled sequences of DNA that allow researchers to either identify the unlabeled DNA they collect during their studies or design new species-specific primers for qPCR (Klymus et al., 2022). However, high-quality reference sequences are unavailable for many organisms, which has become a key factor limiting the broad application of eDNA techniques (Nagarajan et al., 2022). As such, screening genomic databases (e.g., NCBI Nucleotide and BOLD) or personalized reference sequence databases such as CRUX (Creating Reference Libraries Using eXisting tools) (Curd et al., 2019) for reference sequence availability and gaps is essential.

Additionally, detecting the absence of reference sequences is not the only challenge that eDNA scientists face. Through our own experience and interviews conducted with over 70 scientists (see Acknowledgements), we found that downloading the potentially thousands of sequences needed to create personalized reference databases is a tedious task for both new and experienced labs. Using the official websites for sequence databases, like NCBI Nucleotide or BOLD, to manually search for and download quality sequences can take days. Some researchers avoid this time consuming process by writing scripts but not all researchers possess those skills.

Completing both tasks in a time-efficient and accurate manner requires computational skills to search, manipulate, and parse multiple DNA sequence databases. Ecologists, however, often need more training in this area. For example, in California, most (74 - 80%) University of California and California State University undergraduates in ecological and environmental sciences had no formal training in programming skills in any language (Hernandez et al., 2012). Bioinformatic assessments of large genetic databases can be challenging and time-consuming for classically trained ecologists and may result in bioinformatic work becoming a study expense and delaying the project.

To address these issues, we built the Reference Sequence Browser (RSB), a Graphical User Interface (GUI) tool that allows researchers to perform customized batch searches and downloads for reference sequences on the following publicly available databases: NCBI nucleotide, BOLD, and CALeDNA CRUX databases for metabarcoding (Winter, 2017; Chamberlain, 2021; Curd et al., 2019). Those interested in a simple but powerful way of viewing and downloading reference sequences from multiple genomics databases for one or multiple organisms and barcoding loci simultaneously will find that this tool saves time and provides insight into reference sequence availability and gaps. With the RSB, researchers can efficiently develop qPCR assays, determine the efficacy of particular barcoding markers to match an eDNA project's objective, and engage in more deliberate and specific sequencing efforts to address gaps in reference sequence availability.

Methods

Operation

The RSB is a Shiny GUI app (Chang et al., 2023), which is available online at `http://shiny.eeb.ucla.edu/BlueWaltzBio/Coverage/` or for download at `https://github.com/SamuelLRapp/BlueWaltzBio`. RSB uses the rentrez and bold CRAN packages to access the live NCBI and BOLD databases respectively and the Taxize package for some ease-of-use features. The app then visualizes the output using the shiny, ggplot, and plotly packages.

Alternatively, the application can be installed locally. This has been tested for Windows, Mac-OSX, and Ubuntu. The tool uses R version 4.1.2. To install the app, download the zip file found here `https://github.com/SamuelLRapp/BlueWaltzBio/releases/tag/v0.1-beta.0` and then execute the script "rsbPackages.R" to install the correct versions of the required libraries. After that, the app can be run by opening an R console in the directory where the zip file was extracted and executing the command "shiny::runApp('Coverage')". A complete list of our app's package dependencies can be found in "rsbPackages.xlsx" in the zip file above and this paper's "References" section.

Input File Structure

All search tabs use the same CSV file format for uploading large inputs. There are two columns: "OrganismNames" and "Barcodes". Every search tab uses the "OrganismNames" column, but "Barcodes" is only used by the NCBI Nucleotide search tab. Every pairing of entries in the "OrganismNames" and "Barcodes" columns will be searched for, so the two columns do not need to have the same number of elements. Figure 1 shows an example.

	A	B
1	OrganismNames	Barcodes
2	Codium fragile subsp. tomentosoides	12S
3	Heterosiphonia	ITS1
4	Prorocentrum minimum	COI
5	Sargassum muticum	
6	Bonnemaisonia hamifera	
7	Ficopomatus enigmaticus	
8	Teredo navalis	
9	Marenzelleria viridis	
10	Membranipora membranacea	
11	Bugula neritina	
12	Cordylophora caspia	

Figure 1. CSV format for uploading large inputs using the sample dataset.

Implementation

0.1 General Features and Workflow:

The GUI is split into different tabs for each reference database, which the user can select by clicking on the respective tab at the top of the application window. Regardless of which database the user will be querying, the general workflow remains the same. Users will first be presented with a screen allowing them to upload a formatted CSV file with their query specifications (see the "Operation" section to learn more about the CSV file template). While uploading a CSV file is optional, we highly recommend using this feature when doing large or expansive searches. Afterward, the RSB will present users with a screen where they may manually adjust their search parameters. After running a search, the app will display a table summarizing the results, and a list of download buttons will be present for the user to selectively download either the sequences or any metadata they will need for use with external bioinformatics tools.

Users can also select whether they want to apply the taxize CRAN package's autocorrect feature to their list of organisms of interest. Instead of overriding the user's original input information, the tool will append any corrections to the user's list of organisms. For more information about how taxize's corrections work, see the taxize CRAN package documentation (Chamberlain et al., 2020).

Additionally, a tabular summary of the search results can be downloaded from any of the three database tabs, allowing users to view statistics about the overall quantity of data in each given reference database that matches their queries. These tables help users to quickly notice any missing information they may need for their study, and how overrepresented or underrepresented certain organisms or barcodes are in the search results. The rows of the summary table represent the various barcoding loci. The four columns of the table are the number of sequences found for this barcode, the percentage of all sequences found from the user's search that this barcode accounts for, the number of organisms that have at least one sequence for this barcode, and the number of organisms that have no sequences for this barcode. Figure 2 shows an example of a summary table

Summary of Search Results

For each barcode, we display the total number of database entries found, the percentage of the total number of database entries that each marker/gene accounts for, and the number of organisms with at least one or no sequences found

Show 10 entries Search:

	Barcodes	Number of Sequences Found	Percent of Total Sequences Found	Number of Organisms with at Least one Sequence	Number of Organisms with no Sequences
1	Total	5671	100	25	1
2	18S	103	1.82	23	3
3	16S	1334	23.52	14	12
4	PITS	0	0	0	26
5	CO1	2946	51.95	23	3
6	FITS	1	0.02	1	25
7	trnL	0	0	0	26
8	Vert12S	1287	22.69	3	23

Showing 1 to 8 of 8 entries Previous 1 Next

Figure 2. Example of the CRUX summary table as shown in the RSB Shiny app. All summary tables across databases have the same behavior and formatting.

While the summary tables are appropriate for seeing aggregated statistics about the search results of a user's query, the RSB also provides a more detailed view of a user's search results via the Coverage Matrix. The Coverage Matrix is a table that allows users to see how much of their needs can be "covered" by the data in the database they are querying. Each row of the table corresponds to an organism name, each column corresponds to a barcoding loci, and each cell displays the number of sequences found for the given organism-barcode pair. This layout allows the user to quickly scan for zeros or other low numbers in the table and understand where there are currently gaps in the sequence information publicly available to the eDNA community. After any search, users can download the Coverage Matrix (in future sections abbreviated as "CM") and the summary table.

0.2 CRUX database specific features

The CRUX database search tab aims to identify the level of representation of a user-defined set of organisms in the public CALeDNA reference databases. This feature may be used by either scientists interested in using the public CALeDNA reference database or parties interested in working with CALeDNA scientists to conduct a study. The site displays the taxonomic resolution and reference sequence abundance available within the CALeDNA databases for different organisms, which informs the user of how well the databases may fulfill the needs of their study. The databases were created by running the CRUX Pipeline, part of the Anacapa Toolkit (Curd et al., 2019), on a snapshot of the NCBI Nucleotide database in 2019.

The CM for the CRUX database search tab is slightly different from the other CMs in the tool. The rows still represent an organism, but the columns represent one of the public CRUX databases, and values in the table may not always be numbers. When the tool does not find direct matches in a database, it will instead search for higher taxonomic ranks of that organism until it finds a match. If the RSB performs lower-resolution searches it will output the rank at which it found a sequence instead of displaying a numerical value. The tool works its way up the following taxonomic ranks: species, genus, family, order, class, phylum, and domain.

The cells of the outputted table will show one of the following:

1. The number of sequences in a database, if direct matches are found.
2. If no direct matches exist, the next most specific taxonomic rank for which sequences exist.
3. "0" if nothing is found at any taxonomic rank.

Instances of an Organism Found in Each CRUX Metabarcoding Database

Show 10 entries Search:

	18S	16S	PITS	CO1	FITS	trnL	Vert12S
Canis lupus	2	1270	phylum	1297	order	0	1283
Codium fragile subsp. tomentosoides	order	genus	class	class	order	phylum	0
Heterosiphonia	4	family	phylum	23	order	0	0
Prorocentrum minimum	14	12	0	11	class	0	0
Sargassum muticum	2	1	0	20	1	class	0
Bonnemaisonia hamifera	2	genus	phylum	6	class	0	0
Ficopomatus enigmaticus	1	family	0	family	class	0	class
Teredo navalis	4	4	0	4	phylum	0	class
Marenzelleria viridis	1	1	0	25	class	0	order
Membranipora membranacea	3	genus	0	64	0	0	0

Showing 1 to 10 of 26 entries Previous 1 2 3 Next

Figure 3. Example of the CM in the CRUX database as shown in the RSB Shiny app

For example, in the CRUX CM shown in Figure 3, there are no sequences for Canis lupus in the PITS database at the genus-species level. Therefore, the tool would search for any sequences belonging to the genus Canis. These sequences are also not present in the database, so it would search for any sequences from the family of Canis lupus, Canidae. It keeps going up in taxonomic rank until it finally finds a sequence belonging to the Canis lupus phylum, Chordata. Thus the word "phylum" is displayed in that cell of the CM.

0.3 NCBI Nucleotide database specific features

The NCBI Nucleotide search tab allows the user to both examine the quantity of existing sequence data for any set of organism names and barcodes and also quickly download those sequences. Unlike the other tabs, this tab allows users to specify which barcode loci they want to search for. As with the organism names, this information can be input either through the CSV file that users can upload or by manually typing any barcode names into a textbox provided in the GUI (See "General Features and Workflow" section for more details). We also provided buttons on the GUI that add groupings of commonly used barcodes and any alternate names they may have to the search (see the "Discussion" section for more information).

By default, the tool will only search for sequences by matching against the "ORGN" and "GENE" Genbank metadata fields. Using this setting may lead to the user finding fewer sequences than what truly exists in NCBI Nucleotide. Therefore, we provide the option to toggle between metadata searches and full-text searches. Due to the inherent inaccuracy of full-text searches, we recommend that users only switch to full-text searches when metadata searches return too few sequences for the needs of the user.

However, the tool can still find entries in NCBI Nucleotide that include sequence data outside the user's requested barcodes. False positives can occur because full mitochondrial genomes or other large sequences often have their metadata labeled with all barcode loci contained within the sequence. To filter out these large sequences, we provide an optional feature that lets users specify a minimum and maximum base pair length for each barcode sequence. Figure 4 shows the UI elements through which the user can specify minimum and maximum sequence lengths for their query.

Once the search is complete, the CM and the summary table will be visible, as described in the "General Features and Workflow" section above. If the user wishes to view the exact search queries sent to NCBI to get these results, a few examples are displayed below the CM (as shown in Figure 6). The user can download a CSV file containing the complete list of search queries by clicking a button below the table. The contents of the CSV file follow the same dimensions as the CM. There is also a button for downloading the CM table as a CSV file for usage in external tools or storage. Figure 5 & Figure 7 show the CM and the CSV file representation of the CM respectively. Figure 6 & Figure 8 respectively show

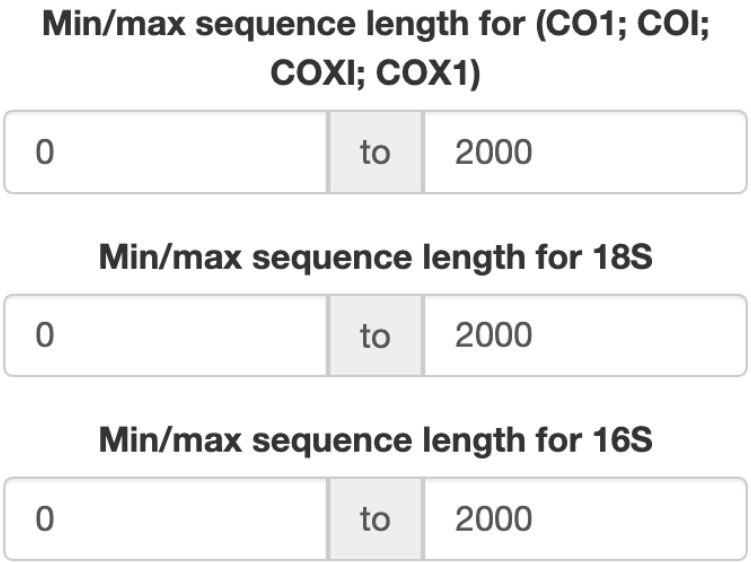

Figure 4. Example of NCBI number range inputs for the user to specify the sequence length (minimum and maximum) for each barcode as shown in the RSB Shiny app

examples of the short list of NCBI search queries and the CSV file containing all of them. These queries can be taken directly into NCBI Nucleotide web portal to get the same results we present.

Two additional download buttons also exist for retrieving the .fasta and .genbank file contents associated with each sequence. Currently, the tool only supports downloading all the fasta sequences into one file. However, in the future, we intend to give users the option of either downloading one file or separating the sequences into different files by barcode.

0.4 BOLD database specific features

The BOLD (Barcode of Life) pipeline of the Reference Sequence Browser screens the BOLD database for genetic barcode coverage. At the beginning of the pipeline users can upload a CSV file containing all the species they wish to search. They may instead manually input this list, as described in the "General Features and Workflow" section above. Once inputted, the application will search through the most up-to-date version of the BOLD database using the BOLD Systems API Package for R (Chamberlain, 2021). Once the search is complete, the users will see a list of all species of interest not found in the database. This list will help quickly check whether the species of interest have reference sequences in BOLD.

Next in the pipeline is the filter tab, which allows users to manipulate the results gathered by filtering the results by country of origin and giving the option to remove any entries also present in NCBI Nucleotide. This allows users to avoid downloading duplicate FASTA files between BOLD and NCBI Nucleotide. All subsequent tables and visualizations depend on this filtering, which can be updated by returning to the filter tab at any time. This way, users can adjust their filters to gather more or fewer sequences as they see fit.

Once the search is complete, the app will first produce a CM and summary table as described in the "General Features and Workflow" section above. This data depends on both the results gathered from BOLD and also the filters the user selected. Additionally, the app provides users with buttons to download the CM, summary table, and the associated FASTA files. Afterward, the RSB presents various tables and graphs in the "Country Data", "Plot Total Sequences per Country", and "Plot Unique Sequences per Country" tabs to help users analyze the data and fine-tune their filters.

In the "Country Data" tab, there are two main tables. The first table has organism names as rows and countries as columns, and each intersection of a row and column shows how many records are in the BOLD database. The second table only displays information when some of the user's species of interest are absent in all selected countries. When this happens, the organism names are rows and the three columns show the top three countries with the most sequences for that species, ordered from highest to lowest. The purpose of the first table is to provide the user with a breakdown of which countries their sequences are coming from, while the second table suggests additional countries to add to the filter when the current list of selected countries has no sequences for some of their target organisms. The user can download either table by using the buttons below them. Figure 9 & Figure 10 display these two country data tables.

Following these two tables, we provide two plots in the "Plot Total Sequences per Country" and "Plot Unique Sequences per Country" tabs to visualize the geographic distribution of the data. The first plot is a treemap that provides a graphical representation of the distribution of the sequence record amongst the selected countries. The second plot is a bar graph that shows the number of unique species per country. Each of these plots also has a download button to save them. Figure 11 shows an example of the treemap, while Figure 12 shows an example of the bar graph.

Finally, there is a table of BOLD entries with unlabeled barcodes. For each organism where this is the case, we display a list of BOLD UIDs so that users may explore these results further by searching in BOLD and looking at the sequences themselves. The rows are the organisms, and the only column displays the list of IDs. The tool also provides a button to download this table at the bottom of the page. Figure 13 displays an example of this table. In our experience, entries in this table have no sequences.

Discussion

Screening of and access to available reference sequences is necessary for the development of new quantitative PCR (qPCR) primers, to understand the efficacy of existing primers, and to understand gaps in what taxa can be identified in eDNA samples for given DNA barcodes. However, access to reference sequence information for the numerous organisms eDNA scientists hope to detect or exclude via primer design requires learning bioinformatic skills or investing large amounts of time into manual searches. With the RSB, users can easily screen NCBI, BOLD, and the public CALeDNA CRUX databases for reference sequence availability and download sequences for numerous organisms at multiple DNA barcodes. Therefore, users can generate deduplicated up-to-date local sequence databases from the combined pools of BOLD and NCBI for primer design by using the NCBI filter in the BOLD tab. Additionally, summary statistics and other data visualization features allow users to glean other insights about reference sequence availability. Therefore, this tool can be used to explore gaps in reference sequence availability and help direct future sequencing efforts and funds. The RSB GUI bridges the gap between ecologists and computer scientists by providing efficient and intuitive access to NCBI, BOLD and CRUX databases without requiring any programming.

While the RSB will be very helpful to eDNA scientists of all backgrounds and levels of experience, the tool does have limitations and space for some improvements. Text searches in NCBI Nucleotide create the possibility of false positives and negatives, which we attempted to address via the [GENE] and [ORGN] metadata search options. Updating the UI to better highlight oddities in the data, such as homonyms in organism names, would give users more control over the quality of their datasets. Additionally, new features, such as the ability to search via higher taxonomic classification and geographic searches, could open up new use cases. Finally, a great quality-of-life feature would be to change the fasta download format such that one fasta file was downloaded for each barcode.

Software availability

- A Shiny app, is available at `http://shiny.eeb.ucla.edu/BlueWaltzBio/Coverage/`
- Source code is available from GitHub: `https://github.com/SamuelLRapp/BlueWaltzBio`
- Software license (GPL-3)

Number of Sequences Found Per Organism-Barcode Pairing

Show 10 entries Search:

	16S	18S	(CO1; COI; COXI; COX1)
Canis Lupus	0	0	2075
Codium fragile subsp. tomentosoides	0	0	0
Heterosiphonia	0	0	34
Prorocentrum minimum	0	0	20
Sargassum muticum	0	0	34
Bonnemaisonia hamifera	0	0	10
Ficopomatus enigmaticus	0	0	22
Teredo navalis	0	0	158
Marenzelleria viridis	0	0	17
Membranipora membranacea	0	0	77

Showing 1 to 10 of 26 entries Previous 1 2 3 Next

Figure 5. Example of the NCBI CM table as shown in the RSB Shiny app.

An Example of the Search Queries Sent to NCBI

Below we show an example of some of the queries we are sending to NCBI Nucleotide. If you wish to check the validity of our results, you can go to the NCBI website and paste these queries into their online search tool and see if you get the same results. If you want to see all of the queries used in this search, you can download them using the button located below.

Show 10 entries Search:

16S	(16S[GENE] AND Canis Lupus[ORGN])
18S	(18S[GENE] AND Canis Lupus[ORGN])
(CO1; COI; COXI; COX1)	(CO1[GENE] AND Canis Lupus[ORGN]) OR (COI[GENE] AND Canis Lupus[ORGN]) OR (COXI[GENE] AND Canis Lupus[ORGN]) OR (COX1[GENE] AND Canis Lupus[ORGN])

Showing 1 to 3 of 3 entries Previous 1 Next

Figure 6. Example of the NCBI search queries table as shown in the RSB Shiny app.

	16S	18S	(CO1; COI; COXI; COX1)
Canis Lupus	0	0	2075
Codium fragile subsp. tom	0	0	0
Heterosiphonia	0	0	34
Prorocentrum minimum	0	0	20
Sargassum muticum	0	0	34
Bonnemaisonia hamifera	0	0	10
Ficopomatus enigmaticus	0	0	22
Teredo navalis	0	0	158
Marenzelleria viridis	0	0	17
Membranipora membrana	0	0	77
Bugula neritina	0	0	384
Cordylophora caspia	0	0	10
Ectopleura crocea	0	0	22
Mysis diluviana	0	0	29
Caprella mutica	0	0	320
Amphibalanus improvisus	0	0	54
Bythotrephes longimanus	0	0	27
Cercopagis pengoi	0	0	7
Rhithropanopeus harrisii	0	0	295
Orconectes	0	0	203
Procambarus clarkii	0	0	436
Homarus americanus	0	0	19
Crepidula fornicata	0	0	66
Mytilopsis leucophaeata	0	0	113
Dreissena polymorpha	0	0	107
Crassostrea gigas	0	0	370

Figure 7. Example of the downloaded CSV file containing the NCBI CM.

	16S	18S	(CO1; COI; COXI; COX1)
Canis Lupus	(16S[GENE] AND Canis Lupus[ORGN])	(18S[GENE] AND Canis Lupus[ORGN	(CO1[GENE] AND Canis Lupus[ORGN]) OR (COI[GENE] A
Codium fragile subsp. tomentosoides	(16S[GENE] AND Codium fragile subsp	(18S[GENE] AND Codium fragile sub	(CO1[GENE] AND Codium fragile subsp. tomentosoides[
Heterosiphonia	(16S[GENE] AND Heterosiphonia[ORGI	(18S[GENE] AND Heterosiphonia[OF	(CO1[GENE] AND Heterosiphonia[ORGN]) OR (COI[GEN
Prorocentrum minimum	(16S[GENE] AND Prorocentrum minim	(18S[GENE] AND Prorocentrum min	(CO1[GENE] AND Prorocentrum minimum[ORGN]) OR (
Sargassum muticum	(16S[GENE] AND Sargassum muticum[	(18S[GENE] AND Sargassum muticu	(CO1[GENE] AND Sargassum muticum[ORGN]) OR (COI
Bonnemaisonia hamifera	(16S[GENE] AND Bonnemaisonia hami	(18S[GENE] AND Bonnemaisonia ha	(CO1[GENE] AND Bonnemaisonia hamifera[ORGN]) OR
Ficopomatus enigmaticus	(16S[GENE] AND Ficopomatus enigma	(18S[GENE] AND Ficopomatus enigr	(CO1[GENE] AND Ficopomatus enigmaticus[ORGN]) OR
Teredo navalis	(16S[GENE] AND Teredo navalis[ORGN	(18S[GENE] AND Teredo navalis[OR	(CO1[GENE] AND Teredo navalis[ORGN]) OR (COI[GENE
Marenzelleria viridis	(16S[GENE] AND Marenzelleria viridis[	(18S[GENE] AND Marenzelleria viric	(CO1[GENE] AND Marenzelleria viridis[ORGN]) OR (COI
Membranipora membranacea	(16S[GENE] AND Membranipora mem	(18S[GENE] AND Membranipora me	(CO1[GENE] AND Membranipora membranacea[ORGN]
Bugula neritina	(16S[GENE] AND Bugula neritina[ORG	(18S[GENE] AND Bugula neritina[OF	(CO1[GENE] AND Bugula neritina[ORGN]) OR (COI[GEN
Cordylophora caspia	(16S[GENE] AND Cordylophora caspia[	(18S[GENE] AND Cordylophora casp	(CO1[GENE] AND Cordylophora caspia[ORGN]) OR (COI[
Ectopleura crocea	(16S[GENE] AND Ectopleura crocea[OF	(18S[GENE] AND Ectopleura crocea[	(CO1[GENE] AND Ectopleura crocea[ORGN]) OR (COI[G
Mysis diluviana	(16S[GENE] AND Mysis diluviana[ORG	(18S[GENE] AND Mysis diluviana[OF	(CO1[GENE] AND Mysis diluviana[ORGN]) OR (COI[GEN
Caprella mutica	(16S[GENE] AND Caprella mutica[ORG	(18S[GENE] AND Caprella mutica[O	(CO1[GENE] AND Caprella mutica[ORGN]) OR (COI[GEN
Amphibalanus improvisus	(16S[GENE] AND Amphibalanus impro	(18S[GENE] AND Amphibalanus imp	(CO1[GENE] AND Amphibalanus improvisus[ORGN]) OR
Bythotrephes longimanus	(16S[GENE] AND Bythotrephes longim	(18S[GENE] AND Bythotrephes long	(CO1[GENE] AND Bythotrephes longimanus[ORGN]) OR
Cercopagis pengoi	(16S[GENE] AND Cercopagis pengoi[OI	(18S[GENE] AND Cercopagis pengoi	(CO1[GENE] AND Cercopagis pengoi[ORGN]) OR (COI[G
Rhithropanopeus harrisii	(16S[GENE] AND Rhithropanopeus har	(18S[GENE] AND Rhithropanopeus h	(CO1[GENE] AND Rhithropanopeus harrisii[ORGN]) OR
Orconectes	(16S[GENE] AND Orconectes[ORGN])	(18S[GENE] AND Orconectes[ORGN	(CO1[GENE] AND Orconectes[ORGN]) OR (COI[GENE] A
Procambarus clarkii	(16S[GENE] AND Procambarus clarkii[(	(18S[GENE] AND Procambarus clark	(CO1[GENE] AND Procambarus clarkii[ORGN]) OR (COI[
Homarus americanus	(16S[GENE] AND Homarus americanus	(18S[GENE] AND Homarus americar	(CO1[GENE] AND Homarus americanus[ORGN]) OR (CO
Crepidula fornicata	(16S[GENE] AND Crepidula fornicata[C	(18S[GENE] AND Crepidula fornicata	(CO1[GENE] AND Crepidula fornicata[ORGN]) OR (COI[(
Mytilopsis leucophaeata	(16S[GENE] AND Mytilopsis leucophae	(18S[GENE] AND Mytilopsis leucoph	(CO1[GENE] AND Mytilopsis leucophaeata[ORGN]) OR (
Dreissena polymorpha	(16S[GENE] AND Dreissena polymorph	(18S[GENE] AND Dreissena polymor	(CO1[GENE] AND Dreissena polymorpha[ORGN]) OR (C
Crassostrea gigas	(16S[GENE] AND Crassostrea gigas[OR	(18S[GENE] AND Crassostrea gigas[	(CO1[GENE] AND Crassostrea gigas[ORGN]) OR (COI[GI

Figure 8. Example of the downloaded CSV file containing the NCBI search queries.

Total Number of Barcodes Found by Country for Each Unique Species

Show 10 entries Search:

	Canada	Ireland	United States
Amphibalanus improvisus	0	0	13
Bonnemaisonia hamifera	3	0	1
Bugula neritina	0	0	227
Bythotrephes longimanus	2	0	1
Canis lupus	17	0	33
Caprella mutica	15	0	1
Cercopagis pengoi	2	0	5
Cordylophora caspia	0	0	115
Crepidula fornicata	17	0	37
Dreissena polymorpha	12	0	6

Showing 1 to 10 of 49 entries Previous 1 2 3 4 5 Next

Figure 9. Example of the BOLD country data table as shown in the RSB Shiny app

Suggested Countries to Add to Your Filter

For those species that have no barcodes found in the country(s) filtered we provide the top 3 unselected countries with the most sequence results.

Show 10 entries Search:

	1st	2nd	3rd
Codium fragile subsp. tomentosoides	No Country Listed	NA	NA
Ficopomatus enigmaticus	No Country Listed	Spain	NA
Heterosiphonia crispella	China	NA	NA
Heterosiphonia gibbesii	United States Virgin Islands	NA	NA
Heterosiphonia japonica	China	France	Netherlands
Heterosiphonia pulchra	South Korea	NA	NA
Orconectes acares	No Country Listed	NA	NA
Orconectes alabamensis	No Country Listed	NA	NA
Orconectes bisectus	No Country Listed	NA	NA
Orconectes burri	No Country Listed	NA	NA

Showing 1 to 10 of 75 entries Previous 1 2 3 4 5 … 8 Next

Figure 10. Example of the BOLD Absent country data table as shown in the RSB Shiny app

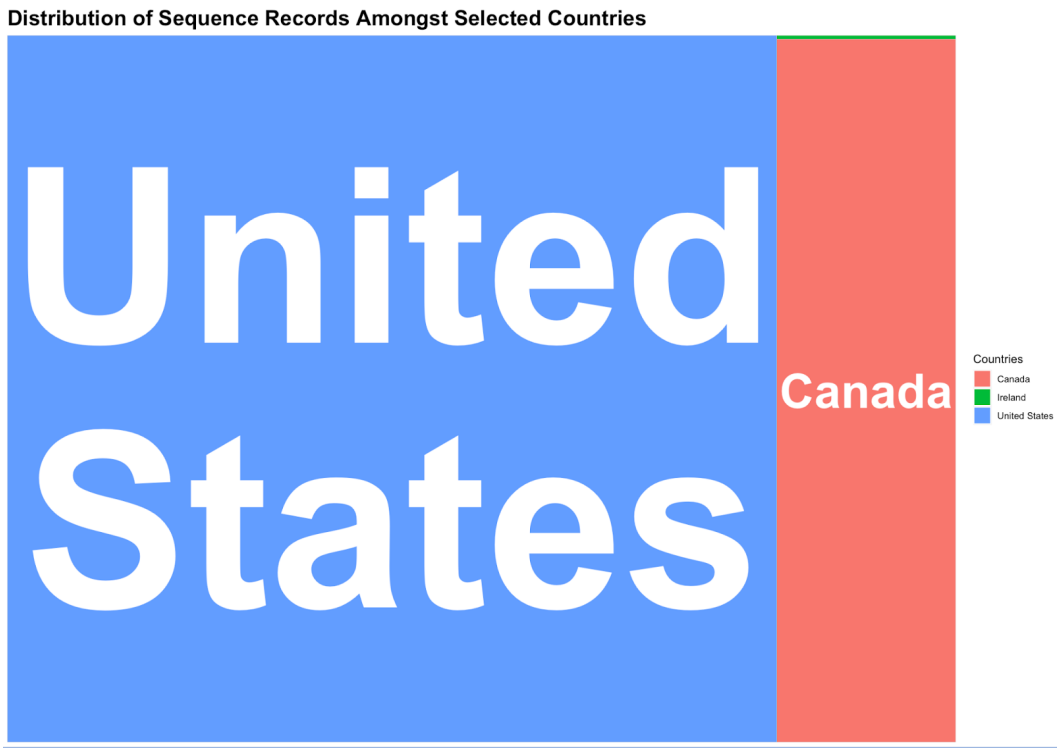

Figure 11. Example of the BOLD treemap output representing the number of sequences per country as shown in the RSB Shiny app

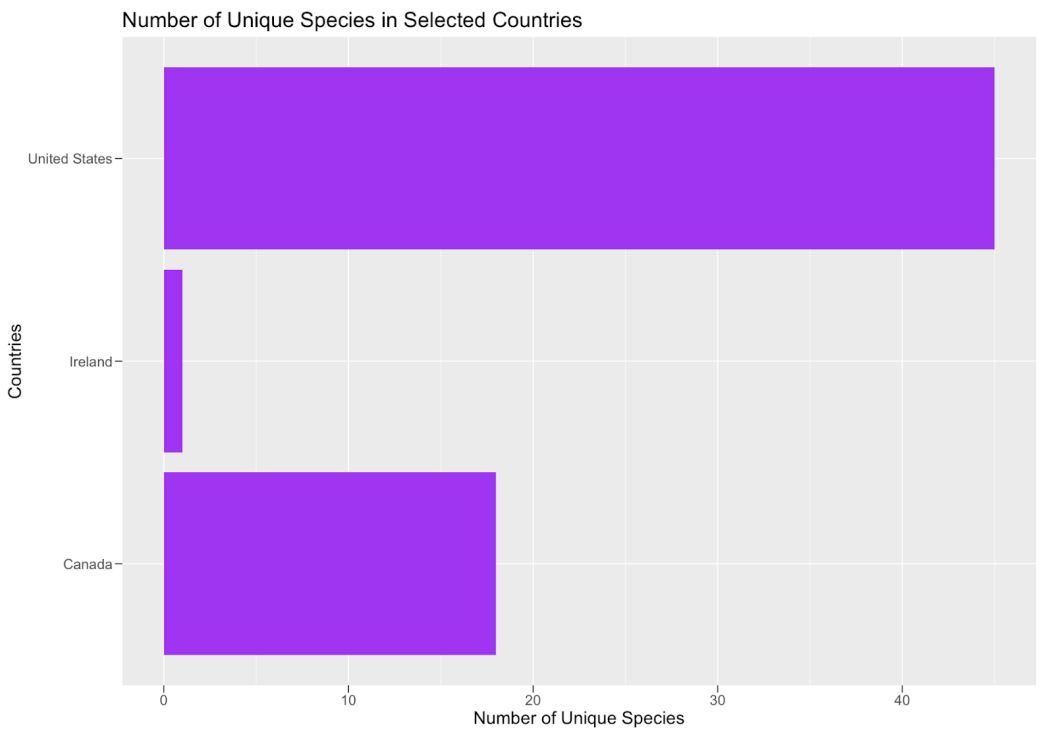

Figure 12. Example of the BOLD bar graph output representing the number of unique sequences per country as shown in the RSB Shiny app

Bold UIDs for Entries With Unlabeled Barcodes

For those entries in BOLD that have unlabeled barcodes, we provide their respective BOLD UIDs so that scientists may explore these results further.

Show 10 entries Search:

	Entries where Barcode was NA
Bythotrephes longimanus	BYTHO114-22, CAISN938-13, NJCGS015-09, NJCGS016-09
Canis lupus	NOMAM074-09, ZSILG432-22, NOMAM145-17
Caprella mutica	BNSB040-21, DUTCH667-19, DUTCH668-19, GBCMA12312-16, QHAK022-20, QHAK026-20, QHAK032-20, QHAK034-20, QHAK037-20, SWEMA466-15, SWEMA864-15, BNSB038-21, BNSB039-21, BNSB035-21, BNSB036-21, BNSB037-21
Cordylophora caspia	ANNMO567-20, SWEMA984-15
Crepidula fornicata	BNAGB664-14, LITOR363-10, LITOR364-10, LITOR365-10, SWEMA033-15
Ectopleura crocea	KHA729-14, KHA730-14
Heterosiphonia japonica	PNMIA1825-14
Heterosiphonia plumosa	PNMIA011-14, PNMIA076-14, PNMIA1024-14, PNMIA105-14, PNMIA1133-14, PNMIA1161-14, PNMIA1177-14, PNMIA1190-14, PNMIA1443-14, PNMIA1583-14, PNMIA1648-14, PNMIA1714-14, PNMIA1762-14, PNMIA183-14, PNMIA1956-14, PNMIA1976-14, PNMIA1989-14, PNMIA200-14, PNMIA217-14, PNMIA225-14, PNMIA240-14, PNMIA266-14, PNMIA301-14, PNMIA404-14, PNMIA436-14, PNMIA446-14, PNMIA454-14, PNMIA596-14, PNMIA644-14, PNMIA731-14, PNMIA846-14, PNMIA900-14, PNMIA934-14

Figure 13. Example of the BOLD table representing unique entries with unlabeled barcodes, which require manual searching to determine what the sequences represent as shown in the RSB Shiny app

Acknowledgements

We thank Hailey Nava for helping during the interview and research phase, Nitya Jain for helping in the BOLD pipeline particularly with the creation of the graphs, Jay Ferreira for helping us refactor and test our code, Dr. Rachel Meyer for guidance, review of the written process and helping us host our tool in the UCLA EEB servers. Lastly, we would like to thank all the scientists we have interviewed in the last 3 years who helped us identify workflow bottlenecks in the field of eDNA and improve this tool including but not limited to: Shane Farrel (University of Maine); Dr. Jonathan Coddington (Smithsonian National Museum of Natural History); Dr. Eva Aylagas Martinez (King Abdullah University of Science and Technology); Dr. Kristin Saltonstall (Smithsonian Tropical Research Institute); Hailey Shchepanik (King Abdullah University of Science and Technology); Dr. Daniel Marquina (York University); Abigail Wells (Northwest Fisheries Science Center); Dr. Ann Holmes (University of California: Davis); Dominik Buchner (University of Duisburg-Essen); Dr. Florian Leese (University of Duisburg-Essen); Dr. Shaili Johri (Stanford University); Dr. Gaurav Kandlikar (University of Missouri, Columbia); Dr. Lynsey Harper (Freshwater Biological Association); Charles Nye (Monterey Bay Aquarium Research Institute); Dr. Patrick Berry (National Oceanographic and Atmospheric Administration); Dr. Vasco Elbrecht (Hamilton Bonaduz AG); Dr. Virginie Marques (ETH Zurich); Dr. Jay P. Graham (University of California: Berkeley School of Public Health); Dr. Sam Chew Chin (Molecular Ecology and Fisheries policy Monmouth University); Dr. Haris Zafeiropoulos (Rega Institute for Medical Research); Wai-Yin Kwan (UC Conservation Genomics Consortium); Terry Vanech (K&S Ranch); Dr. Zach Gold (Pacific Marine Environmental Laboratory); Dr. Scott Miller (National Museum of Natural History); Dr. Nuala Oleary (National Center for Biotechnology Information); Dr. Eric Cox (National Institutes of Health); Dr. Meixi Lin (Carnegie Institution for Science); Dr. Emily Curd (University of Vermont); Dr. Niamph M. Redmond (Universitè Paul Sabatier); Dr. Elif Demir-Hilto (Oceankind); Dr. Meghan M. Shea (Stanford University); Patrick Shea (City College of the City University of New York); Dr. Collin Closek (Stanford University); Kiumars Edalati (MD Candidate at University of Washington School of Medicine); Dr. Ramon Gallego Simon (University of Washington Seattle); Dr. Matt Cannon (University of Maryland, Baltimore); Dr. Wes Larson (National Oceanic and Atmospheric Administration); Dr. Matthew R. Snyder (National Oceanic and Atmospheric Administration); Dr. Caren S. Goldberg (Washington State University); Dr. Kevin Morey (University of Guelph); Bryn Philips (University of California: Davis); Dr. Taylor Wilcox (National Genomics Center for Wildlife and Fish Conservation), Dr. Neil Davies (University of California Gump South Pacific Research Station); Kirsten Harper (National Oceanographic and Atmospheric Administration); Dr. Michael K. Young (United States Forest Service Rocky Mountain Research Station); Dr. Matthew D. McCarthy (University of California: Santa Cruz); Mark Y Stoeckle (The Rockefeller University); Luisa Valiela (Environmental Protection Specialist); Dr. Jason Thompson (Oceankind); Dr. Ryan Kelly (University of Washington); Dr. Josh D. Kapp (University of California: Santa Cruz); Anna Worth (University of California: Santa Cruz); Dr. Rebecca Vega Thurber (Oregon State University); Dr. Bernie Tershy (University of California: Santa Cruz); Dr. Ranjiv Khush (Aquaya); Dr. Jonathan Fong (Lingnan University); Dr. Loren M. Miller (Minnesota Aquatic Invasive Species Research Center); Maria Nguyen (University of California: Berkeley); Dr. Rachel Carlson (University of California: Davis); Dr. Ana Ramon-Laca (Madrid Museo Nacional de Ciencias Naturales); Dr. Corinne Gibble (California Department of Fish and Wildlife); Dr. Josh Kapp (University of California: Santa Cruz); Dr. Susanna Theroux (Southern California Coastal Water Research Project)

References

Mathew Seymour. Rapid progression and future of environmental dna research. *Communications Biology*, 2, 12 2019. doi: 10.1038/s42003-019-0330-9.

Philip Francis Thomsen and Eske Willerslev. Environmental dna – an emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, 183:4–18, 2015. ISSN 0006-3207. doi: https://doi.org/10.1016/j.biocon.2014.11.019. URL https://www.sciencedirect.com/science/article/pii/S0006320714004443. Special Issue: Environmental DNA: A powerful new tool for biological conservation.

Jesper Andersen, Emilie Kallenbach, Martin Hesselsoe, Steen Knudsen, Peter Moller, Dorte Bekkevold, Brian Hansen, and Jens Thaulow. Steps toward nation-wide monitoring of non-indigenous species in danish marine waters under the marine strategy framework directive. 04 2016. doi: 10.13140/RG.2.1.3003.6889.

Katy E Klymus, Robert A Hrabik, Nathan L Thompson, and Robert S Cornman. Genome resequencing clarifies phylogeny and reveals patterns of selection in the toxicogenomics model pimephales promelas. *PeerJ*, 10:e13954, 2022.

Raman Nagarajan, Mallory Bedwell, Ann Holmes, Thiago Sanches, Shawn Acuña, Melinda Baerwald, Matthew Barnes, Scott Blankenship, Richard Connon, Kristy Deiner, Daphne Gille, Caren Goldberg, Margaret Hunter, Christopher Jerde, Gordon Luikart, Rachel Meyer, Alison Watts, and Andrea Schreier. Environmental dna methods for ecological monitoring and biodiversity assessment in estuaries. *Estuaries and Coasts*, 45, 06 2022. doi: 10.1007/s12237-022-01080-y.

Emily E. Curd, Zack Gold, Gaurav S. Kandlikar, Jesse Gomer, Max Ogden, Taylor O'Connell, Lenore Pipes, Teia M. Schweizer, Laura Rabichow, Meixi Lin, Baochen Shi, Paul H. Barber, Nathan Kraft, Robert Wayne, and Rachel S. Meyer. Anacapa toolkit: An environmental dna toolkit for processing multilocus metabarcode datasets. *Methods in Ecology and Evolution*, 10(9):1469–1475, 2019. doi: https://doi.org/10.1111/2041-210X.13214. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13214.

Rebecca R. Hernandez, Matthew S. Mayernik, Michelle L. Murphy-Mariscal, and Michael F. Allen. Advanced Technologies and Data Management Practices in Environmental Science: Lessons from Academia. *BioScience*, 62(12):1067–1076, 12 2012. ISSN 0006-3568. doi: 10.1525/bio.2012.62.12.8. URL https://doi.org/10.1525/bio.2012.62.12.8.

David J. Winter. rentrez: an r package for the ncbi eutils api. *The R Journal*, 9:520–526, 2017.

Scott Chamberlain. *bold: Interface to Bold Systems API*, 2021. URL https://CRAN.R-project.org/package=bold. R package version 1.2.0.

Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. *shiny: Web Application Framework for R*, 2023. URL https://CRAN.R-project.org/package=shiny. R package version 1.7.4.1.

Scott Chamberlain, Eduard Szoecs, Zachary Foster, Zebulun Arendsee, Carl Boettiger, Karthik Ram, Ignasi Bartomeus, John Baumgartner, James O'Donnell, Jari Oksanen, Bastian Greshake Tzovaras, Philippe Marchand, Vinh Tran, Maëlle Salmon, Gaopeng Li, and Matthias Grenié. *taxize: Taxonomic information from around the web*, 2020. URL https://github.com/ropensci/taxize. R package version 0.9.98.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

David Wilkins. *treemapify: Draw Treemaps in 'ggplot2'*, 2021. URL https://CRAN.R-project.org/package=treemapify. R package version 2.5.5.

Hadley Wickham, Maximilian Girlich, and Edgar Ruiz. *dbplyr: A 'dplyr' Back End for Databases*, 2021. URL https://CRAN.R-project.org/package=dbplyr. R package version 2.1.1.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2022. URL https://CRAN.R-project.org/package=dplyr. R package version 1.0.8.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Hannah Weigand, Arne J. Beermann, Fedor Čiampor, Filipe O. Costa, Zoltán Csabai, Sofia Duarte, Matthias F. Geiger, Michał Grabowski, Frédéric Rimet, Björn Rulik, Malin Strand, Nikolaus Szucsich, Alexander M. Weigand, Endre Willassen, Sofia A. Wyler, Agnès Bouchez, Angel Borja, Zuzana Čiamporová Zaťovičová, Sónia Ferreira, Klaas-Douwe B. Dijkstra, Ursula Eisendle, Jörg Freyhof, Piotr Gadawski, Wolfram Graf, Arne Haegerbaeumer, Berry B. van der Hoorn, Bella Japoshvili, Lujza Keresztes,

Emre Keskin, Florian Leese, Jan N. Macher, Tomasz Mamos, Guy Paz, Vladimir Pešić, Daniela Maric Pfannkuchen, Martin Andreas Pfannkuchen, Benjamin W. Price, Buki Rinkevich, Marcos A.L. Teixeira, Gábor Várbíró, and Torbjørn Ekrem. Dna barcode reference libraries for the monitoring of aquatic biota in europe: Gap-analysis and recommendations for future work. *Science of The Total Environment*, 678:499–524, 2019. ISSN 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2019.04.247. URL https://www.sciencedirect.com/science/article/pii/S0048969719317760.

Sebastian Warnholz. *modules: Self Contained Units of Source Code*, 2021. URL https://CRAN.R-project.org/package=modules. R package version 0.10.0.

Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

Carson Sievert. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC, 2020. ISBN 9781138331457. URL https://plotly-r.com.

The pandas development team. pandas-dev/pandas: Pandas, feb 2020. URL https://doi.org/10.5281/zenodo.3509134.

Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

Victor Perrier, Fanny Meyer, and David Granjon. *shinyWidgets: Custom Inputs Widgets for Shiny*, 2022. URL https://CRAN.R-project.org/package=shinyWidgets. R package version 0.6.4.

R Special Interest Group on Databases (R-SIG-DB), Hadley Wickham, and Kirill Müller. *DBI: R Database Interface*, 2021. URL https://CRAN.R-project.org/package=DBI. R package version 1.1.2.

Kirill Müller, Hadley Wickham, David A. James, and Seth Falcon. *RSQLite: SQLite Interface for R*, 2021. URL https://CRAN.R-project.org/package=RSQLite. R package version 2.2.9.

Virginie Marques, Tristan Milhau, Camille Albouy, Tony Dejean, Stéphanie Manel, David Mouillot, and Jean-Baptiste Juhel. Gapedna: Assessing and mapping global species gaps in genetic databases for edna metabarcoding. *Diversity and Distributions*, 27(10):1880–1892, 2021. doi: https://doi.org/10.1111/ddi.13142. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ddi.13142.

Ian Lyttle. *vembedr: Embed Video in HTML*, 2021. URL https://CRAN.R-project.org/package=vembedr. R package version 0.1.5.

David Kahle. mpoly: Multivariate polynomials in R. *The R Journal*, 5(1):162–170, 2013. URL https://journal.r-project.org/archive/2013-1/kahle.pdf.

Ian E. Fellows. *ipc: Tools for Message Passing Between Processes*, 2019. URL https://CRAN.R-project.org/package=ipc. R package version 0.1.3.

Joe Cheng. *promises: Abstractions for Promise-Based Asynchronous Programming*, 2021. URL https://CRAN.R-project.org/package=promises. R package version 1.2.0.1.

Winston Chang and Barbara Borges Ribeiro. *shinydashboard: Create Dashboards with 'Shiny'*, 2021. URL https://CRAN.R-project.org/package=shinydashboard. R package version 0.7.2.

Henrik Bengtsson. A unifying framework for parallel and distributed processing in r using futures, 2021. URL https://journal.r-project.org/archive/2021/RJ-2021-048/index.html. 10.32614/RJ-2021-048.

Eric Bailey. *shinyBS: Twitter Bootstrap Components for Shiny*, 2022. URL https://CRAN.R-project.org/package=shinyBS. R package version 0.61.1.

Dean Attali and Tristan Edwards. *shinyalert: Easily Create Pretty Popup Messages (Modals) in 'Shiny'*, 2021. URL https://CRAN.R-project.org/package=shinyalert. R package version 3.0.0.

Andras Sali and Dean Attali. *shinycssloaders: Add Loading Animations to a 'shiny' Output While It's Recalculating*, 2020. URL https://CRAN.R-project.org/package=shinycssloaders. R package version 1.0.0.

Dean Attali. *shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds*, 2021. URL https://CRAN.R-project.org/package=shinyjs. R package version 2.1.0.