# CRABS – A software program to generate curated reference databases for metabarcoding sequencing data

Gert-Jan Jeunen[1], Eddy Dowle[2], Jonika Edgecombe[1], Ulla von Ammon[3], Neil Gemmell[1], and Hugh Cross[1]

[1]University of Otago
[2]University of Colorado
[3]Cawthron Institute

June 1, 2022

**Abstract**

The measurement of biodiversity is an integral aspect of life science research. With the establishment of second- and third-generation sequencing technologies, an increasing amount of metabarcoding data is being generated as we seek to describe the extent and patterns of biodiversity in multiple contexts. The reliability and accuracy of taxonomically assigning metabarcoding sequencing data has been shown to be critically influenced by the quality and completeness of reference databases. Custom, curated, eukaryotic reference databases, however, are scarce, as are the software programs for generating them. Here, we present CRABS (Creating Reference databases for Amplicon-Based Sequencing), a software package to create custom reference databases for metabarcoding studies. CRABS includes tools to download sequences from multiple online repositories (i.e., NCBI, BOLD, EMBL, MitoFish), retrieve amplicon regions through in silico PCR analysis and pairwise global alignments, curate the database through multiple filtering parameters (e.g., dereplication, sequence length, sequence quality, unresolved taxonomy), export the reference database in multiple formats for the immediate use in taxonomy assignment software, and investigate the reference database through implemented visualizations for diversity, primer efficiency, reference sequence length, and taxonomic resolution. CRABS is a versatile tool for generating curated reference databases of user-specified genetic markers to aid taxonomy assignment from metabarcoding sequencing data. CRABS is available for download as a conda package and via GitHub (https://github.com/gjeunen/reference_database_creator).

Title: CRABS – A software program to generate curated reference databases for metabarcoding sequencing data

Running title: Creating reference databases for amplicon-based sequencing

Gert-Jan Jeunen[1], Eddy Dowle[1], Jonika Edgecombe[1], Ulla von Ammon[2], Neil J. Gemmell[1], Hugh Cross[1,3]

[1] Department of Anatomy, School of Biomedical Sciences, University of Otago, Dunedin 9016, New Zealand.

[2] Coastal and Freshwater Group, Cawthron Institute, 98 Halifax Street East, Nelson 7010, New Zealand.

[3]National Ecological Observatory Network, 1685 38$^{th}$ Street, Suite 100, Boulder, CO 80301, USA.

Corresponding author: Dr. Gert-Jan Jeunen

Department of Anatomy

University of Otago

Dunedin 9016

New Zealand

*gjeunen@gmail.com*



## ABSTRACT

The measurement of biodiversity is an integral aspect of life science research. With the establishment of second- and third-generation sequencing technologies, an increasing amount of metabarcoding data is being generated as we seek to describe the extent and patterns of biodiversity in multiple contexts. The reliability and accuracy of taxonomically assigning metabarcoding sequencing data has been shown to be critically influenced by the quality and completeness of reference databases. Custom, curated, eukaryotic reference databases, however, are scarce, as are the software programs for generating them. Here, we present CRABS (Creating Reference databases for Amplicon-Based Sequencing), a software package to create custom reference databases for metabarcoding studies. CRABS includes tools to download sequences from multiple online repositories (i.e., NCBI, BOLD, EMBL, MitoFish), retrieve amplicon regions through *in silico* PCR analysis and pairwise global alignments, curate the database through multiple filtering parameters (e.g., dereplication, sequence length, sequence quality, unresolved taxonomy), export the reference database in multiple formats for the immediate use in taxonomy assignment software, and investigate the reference database through implemented visualizations for diversity, primer efficiency, reference sequence length, and taxonomic resolution. CRABS is a versatile tool for generating curated reference databases of user-specified genetic markers to aid taxonomy assignment from metabarcoding sequencing data. CRABS is available for download as a conda package and via GitHub (https://github.com/gjeunen/reference_database_creator).

**Keywords:** *Reference database curation, environmental DNA, eDNA, ancient DNA, aDNA, taxonomy assignment, python*

## 1 | INTRODUCTION

Investigating, classifying, and understanding Earth's biodiversity is a fundamental component of many disciplines, including ecology, evolution, taxonomy, and paleobiology (Soulé, 1985). While essential, a thorough understanding of diversity patterns has traditionally been hard to achieve, due to the complexity of biological systems (Chave, 2013; Kovalenko et al., 2012).

In the last two decades, the advent of high-throughput DNA sequencing has enabled researchers to understand biodiversity at an unprecedented scope (Hagelberg et al., 2015; Hugenholtz & Tyson, 2008). Untargeted (i.e., metagenomic) sequencing approaches, such as Illumina shotgun sequencing, take full advantage of the gigabytes of data generated in a single sequencing run to uncover the complete complexity of biodiversity present within a sample (Bovo et al., 2020; Cowart et al., 2018; Key et al., 2017). However, with the lack of reference genomes (Scott et al., 2021) and the immense barcoding effort already undertaken (Hebert & Gregory, 2005), targeted sequencing strategies (i.e., metabarcoding), targeting one or several gene regions through PCR amplification (Jeunen et al., 2019; Seersholm et al., 2018) or capture-enrichment (Ávila-Arcos et al., 2011; Seeber et al., 2019), are frequently used to increase the percentage of reads to be taxonomically assigned by enriching the sequencing library for barcoding gene regions (Stat et al., 2017). Additionally, by enriching the library for a select few gene regions, metabarcoding is a more cost-friendly alternative compared to metagenomic sequencing, by reducing the required sequencing depth per sample, as well as the computational time and effort (Stat et al., 2017; Taberlet et al., 2012).

The popularity of metabarcoding analyses has led to the development of multiple tools aiming to improve the taxonomic assignment accuracy of the biological community present in samples. Most of the taxonomy assignment programs can be split into four distinct methods (Hleap et al., 2021), including sequence similarity methods (BLAST [Altschul et al., 1990] and Kraken2 [Wood & Salzberg, 2014]), sequence composition methods (RDP [Qiong et al., 2007] and IDTaxa [Murali et al., 2018]), phylogenetics methods (EPA [Barbera et al., 2019] and pplacer [Matsen et al., 2010]), and probabilistic methods (Protax [Somervuo et al., 2016]).

Comparative studies have revealed that, irrespective of the taxonomic assignment method used, comprehensive, curated, and well-annotated reference databases are critical for accurate taxonomy assignment (Gold et

al., 2021; Hleap et al., 2021; Leray et al., 2022). The early adoption of metabarcoding in microbial research, as well as a focus on the 16S rRNA gene for bacterial species identification (Johnson et al., 2019), have led to the creation of curated reference databases used to assign taxonomy in the majority of microbiome studies, e.g., RDP (Qiong et al., 2007). Metabarcoding research exploring eukaryotic diversity, on the other hand, employs a wide variety of primer sets targeting a broad range of gene regions (Zhang et al., 2020), including cytochrome*c* oxidase subunit I (COI), 16S ribosomal RNA (16S rRNA), 18S ribosomal RNA (18S rRNA), and nuclear ribosomal internal transcribed spacer regions (nrITS). Hence, the majority of eukaryotic metabarcoding research utilizes global databases to assign taxonomy, such as NCBI (Johnson et al., 2008) and EMBL (Kanz et al., 2005). However, the lack of curation with global databases allow for the entry of sequences with missing species ID (environmental studies), erroneous identification, and duplication, factors which contribute to a reduced accuracy of taxonomic assignment algorithms (Gold et al., 2021; Hleap et al., 2021; Leray et al., 2018).

Multiple curated eukaryotic reference databases, as well as pipelines to generate them, have, therefore, been published in recent years, including BOLD (Ratnasingham & Hebert, 2007), UNITE (Kõljalg et al., 2005), PLANiTS (Banchi et al., 2020), MitoFish (Iwasaki et al., 2013), MARES (Arranz et al., 2020), Meta-Fish-Lib (Collins et al., 2021), and MIDORI2 (Leray et al., 2022). The missing built-in flexibility and the large number of reference databases to cover all target gene regions and taxonomic groups, however, favours the development and use of software programs able to generate curated reference databases that are customized through user-specified parameters. Existing software packages are able to extract amplicon regions through *in silico* PCR (ecoPCR/OBITools; hereafter named "ecoPCR"; Boyer et al., 2016; Ficetola et al., 2010), local alignments (RESCRIPt; Robeson et al., 2020), and profile hidden Markov models (MetaCurator; Richardson et al., 2020). An easy-to-use software program able to complete the full reference database creation workflow from start to finish on a personal computer with true flexibility, limited storage requirements, and fast results would further aid in increasing taxonomic assignment accuracy for user-specific experimental designs.

Here, we introduce CRABS (Creating Reference databases for Amplicon-Based Sequencing), a software package to generate curated reference databases and assess the incorporated diversity and taxonomic resolution. To determine the flexibility and efficiency of CRABS, we compare reference databases generated by CRABS to ecoPCR, MetaCurator, and RESCRIPt for four widely-used primer sets in metabarcoding research: MiFish-E/U (Chondrichthyes/Actinopterygii; Miya et al., 2015), mlCOIintF/jgHC02198 (Eukaryota; Leray et al., 2013), Taberlet c/h (Plantae; Taberlet et al., 1991, 2007), and gITS7/ITS4 (Fungi; Ihrmark et al., 2012; White et al., 1990). Additionally, we assess the quality of the generated reference databases through taxonomic assignment of published sequencing data. We show that the reference databases generated by CRABS are equivalent to or outperform available tools based on incorporated diversity. Additionally, CRABS is feature-rich, highly versatile in its implementation, and requires relatively limited computational resources.

## 2 | METHODS

### *2.1 | CRABS workflow*

We present CRABS as a versatile software program to generate curated reference databases for metagenomic analyses. The CRABS workflow consists of five modules (Figure 1), including: (i)*sequence retrieval* : download and import data from online repositories into CRABS; (ii) *amplicon extraction* : extract amplicon regions through *in silico* PCR analysis and pairwise global alignments; (iii) *database curation* : clean up the database through dereplication, sequence parameters, and metadata information; (iv) *database format* : export the database into various formats to use CRABS-generated reference databases into most taxonomic assignment software packages; and (v) *database exploration* : generate a summary overview of the final reference database through multiple post-processing functions and visualizations. A brief explanation for each of the five modules is provided below. Additional information for each function with example code can be found on GitHub (https://github.com/gjeunen/reference_database_creator).

### *2.2 | Module 1: sequence retrieval*

CRABS supports the download of sequencing data from four major online repositories using the '*db_download* ' function, including Barcode of Life Data System (BOLD), National Center for Biotechnology Information (NCBI), European Molecular Biology Laboratory (EMBL), and Mitochondrial Genome Database of Fish (MitoFish). Upon downloading, sequencing data will be processed to CRABS format, a simple two-line fasta format with NCBI accession numbers as header information. When accession numbers are not available, CRABS will generate unique sequence IDs using the following format: '*CRABS_* [num]*:species_name* '. Sequence data can be merged using the '*db_merge* ' function when data from multiple sources are downloaded. Additionally, in-house generated data, in fasta format, can be imported into CRABS format using the '*db_import* ' function.

*2.3 | Module 2: amplicon extraction*

Once sequences have been downloaded and imported into CRABS, the amplicon region can be extracted. For amplicon-based sequencing, we recommend a two-step approach, whereby (i) amplicons are found by locating the forward and reverse primer-binding regions through *in silico* PCR analysis using the '*insilico_pcr* ' function and (ii) using the amplicons retrieved by *in silico* PCR as seed sequences for a pairwise global alignment analysis (function: '*pga* ') to retrieve amplicon regions with missing info on the primer-binding regions.

CRABS utilizes cutadapt v3.5 (Martin, 2011) to accomplish the *in silico* PCR analysis. Forward and reverse primer sequences that were used to generate the sequencing library can be specified through the '*–fwd* ' and '*–rev* ' parameters, respectively and a maximum number of errors in the primer-binding sites can be specified using the '*–error* ' parameter. Sequences for which primer-binding regions could not be located are reverse complemented using VSEARCH v2.21.1 (Rognes et al., 2016). Reverse complemented sequences are subjected once more to the *in silico* PCR analysis, thereby targeting sequences that were deposited on the negative strand in online repositories.

The *in silico* PCR analysis will fail to retrieve amplicons when one or both primer-binding regions are missing from the deposited sequence. This can occur when a partial gene sequence is deposited or when the deposited sequence has been generated using one or both primers used in the metabarcoding analysis, as these primer-binding regions are removed prior to online archiving. To recover these amplicon regions, CRABS implements a pairwise global alignment (function: '*pga* ') using the '*–usearch_* global' algorithm in VSEARCH, whereby amplicon regions retrieved through *in silico* PCR analysis are used as seed sequences (parameter: '*–database* '). The '*pga* ' function can be restricted to only include alignments starting or ending within the length of the forward or reverse primer-binding region (parameter: '*–filter_method strict* '), to minimize the risk of including erroneous sequences. Additionally, the similarity and query coverage thresholds within the '*pga* ' function can be specified using the '*–percid* ' and '*–coverage* ' parameters, respectively.

*2.4 | Module 3: database curation*

Taxonomic lineages are based on the NCBI taxonomy information, which must be downloaded first using the '*db_download –source taxonomy* ' function. Afterwards, taxonomic lineages can be created using the '*assign_tax* ' function. The taxonomic lineage consists of seven levels by default, including: (i) kingdom; (ii) phylum; (iii) class; (iv) order; (v) family; (vi) genus; and (vii) species. However, users can specify any taxonomic lineage through the parameter '*–ranks* '. From this point forward, CRABS will output a tab-delimited file, whereby all information for each sequence record (including the sequence) is contained on a single line.

The reference database can be dereplicated in three ways using the '*–method* ' parameter within the '*dereplicate* ' function, including: (i) *strict* : unique sequences will be kept, irrespective of taxonomy; (ii) *single_species* : a single sequence is retained for each species in the database; and (iii)*uniq_species* : all unique sequences are retained for each species in the database.

The reference database can be further curated, using the '*seq_cleanup* ' function, by filtering sequences on (i) minimum and maximum sequence length (consistent with amplicon length range), (ii) number of ambiguous

bases, (iii) environmental sequences, (iv) unspecified species names, and (v) missing taxonomic information.

Lastly, the reference database can be curated using the '*geo_cleanup* ' function, by excluding species not occurring inside the area of interest. Recent research has provided evidence for increased taxonomic assignment accuracy when geographical species locations are considered (Gold et al., 2021; Murali et al., 2018). Users can, therefore, provide a text file containing species names using the '*–species* ' parameter and CRABS will make a subset from the reference database by only including sequences assigned to these species and providing information on which species are present and absent in the database.

*2.5 | Module 4: database format*

Once the database is curated and cleaned, CRABS can output the database in six different formats (function: '*tax_format* ') to allow the reference database to be implemented in most taxonomic assignment software programs. Implemented formats are: (i) SINTAX, incorporated in VSEARCH (Rognes et al., 2016) and USEARCH (Edgar, 2010, 2016); (ii) RDP, incorporated in the RDP classifier (Qiong et al., 2007); (iii) QIIf (flat-file format), incorporated in QIIME and QIIME2 (Bolyen et al., 2019; Caporaso et al., 2010); (iv) DAD, incorporated in DADA2 (Callahan et al., 2016); (v) DADS, incorporated in DADA2; and (vi) IDT, incorporated in IDTaxa (Murali et al., 2018).

*2.6 | Module 5: database exploration*

Once the reference database has been curated to the user's specifications, CRABS can provide five outputs to provide information about the generated reference database using the '*visualization* ' function. The '*diversity* ' method will generate a horizontal bar plot displaying number of species and number of sequences for each taxonomic group at a user-specified taxonomic level. The '*amplicon_length* ' method will generate a line graph displaying the amplicon length distribution for each taxonomic group at a user-specified taxonomic level. Included in the legend are the total number of sequences assigned to each taxonomic group. The '*db_completeness* ' method generates a table containing the number of closely related species that are present in the reference database for a user-provided list of species. The '*phylo* ' method will generate phylogenetic trees for a user-provided list of species to determine the taxonomic resolution of the amplicon region for each species at a user-specified taxonomic level. Finally, the '*primer_efficiency* ' method will generate a graph for the forward and reverse primers, displaying the proportion of base pair occurrence in the primer-binding region for a user-specified taxonomic group.

*2.7 | Evaluation of CRABS for multiple common metabarcoding primer sets*

To test the performance of CRABS, reference databases were generated for four widely-used primer sets in metabarcoding research (Supplement 1), including MiFish-E/U (Miya et al., 2015), mlCOIintF/jgHC02198 (Leray et al., 2013), Taberlet c/h (Taberlet et al., 1991, 2007), and gITS7/ITS4 (Ihrmark et al., 2012; White et al., 1990). CRABS-generated reference databases were benchmarked for included diversity against databases created by ecoPCR (Ficetola et al., 2010), MetaCurator (Richardson et al., 2020), and the QIIME RESCRIPt plugin (Robeson et al., 2020). Tutorial workflows for each of the four software packages were followed to create the reference databases, except for the final database curation step, which was conducted by CRABS, to enable a true comparison between software programs. For RESCRIPt, due to alignment difficulties in the extraction step, the standard QIIME '*extract_* reads' tool was used instead. Missing functionality within the reference database creation workflow in MetaCurator, ecoPCR, and RESCRIPt was also covered by CRABS (Supplement 2). Additionally, the quality of all reference databases was determined by comparing taxonomy assignments of OTUs (Operational Taxonomic Units) created from publicly available sequencing data: MiFish-E/U sequencing data was used from Jeunen et al. (2020); mlCOIintF/jgHC02198 sequencing data was used from Jeunen et al. (2019); Taberlet c/h sequencing data was used from (Cross et al., *unpubl.* ); and gITS7/ITS4 sequencing data was used from Cross et al. (2017). A detailed comparison of the features implemented in CRABS, ecoPCR (Ficetola et al., 2010), MetaCurator (Richardson et al., 2020), and QIIME RESCRIPt (Robeson et al., 2020) is presented in Table 1, including database download options, amplicon extraction methodology, database clean-up parameters, and database output options. Bioinformatic scripts and reference databases for each assay and software program are provided in Supplement 2 and Supplement

3, respectively.

## 3 | RESULTS

### *3.1 Exploring CRABS generated reference databases through incorporated visualizations*

By downloading sequencing data from EMBL, MitoFish, and NCBI online repositories, the CRABS generated MiFish-E/U reference database incorporated 28,350 sequences, covering 16,906 species. The '*–method diversity* ' visualization (Figure 2.a) shows that the majority of sequences belong to the class Actinopteri (36.2%), followed by Mammalia (19.7%), Amphibia (16.9%), and Lepidosauria (10.6%). The '*–method amplicon_ length* ' visualization displays an average amplicon length of ˜180 bp, with a slightly larger average amplicon size for birds and amphibians compared to the target taxonomic group of fish (Supplement 4.a). Additionally, the '*–method phylo* ' visualization identified species-level taxonomic resolution might not always be obtainable for this amplicon region (Supplement 4.b).

The Taberlet c/h primer set is designed to target land plants and the CRABS reference database was built using the NCBI and EMBL online repositories. The curated CRABS reference database consisted of 71,031 sequences, covering 51,366 species. Based on the '*–method diversity* ' visualization output, the majority of sequences belong to the classes Magnoliopsida (90.0%), Bryopsida (5.0%), and Jungermanniopsida (3.8%) within the phylum Streptophyta (99.9%; Supplement 4.c). Average amplicon length showed large variations within the phylum Streptophyta, with amplicon size ranging from <100 bp to ˜180 bp (visualization method: '*–method amplicon_ length* '; Supplement 4.d). Despite the large variation in amplicon sizes, the '*–method primer_ efficiency* ' visualization revealed only two places in the primer-binding regions with a significant proportion of mismatch occurrence for species within the phylum Streptophyta (Figure 2.b).

For the mlCOIintF/jgHC02198 primer set that is designed to target eukaryotes, the CRABS reference database was built using the BOLD, EMBL, MitoFish, and NCBI online repositories. The reference database included 590,228 sequences covering 109,545 species, with the phyla Arthropoda (72.1%), Chordata (17.4%), and Mollusca (4.8%) most abundantly present (visualization method: '*–method diversity* '; Supplement 4.e). The '*–method amplicon_ length* ' visualization displays an average amplicon length of ˜313 bp, with high consistency between taxonomic groups (Supplement 4.f). Furthermore, the '*–method phylo* ' revealed intraspecific variation is present in the amplicon region for a majority of taxa (Figure 2.c; example: genus Apteryx [Kiwi]).

CRABS generated a reference database containing 339,286 sequences and covering 42,961 species for the gITS7/ITS4 primer set that is designed to target fungi, by incorporating sequencing data from NCBI and EMBL. According to the '*–method diversity* ' visualization, the majority of sequences belong to the phyla Ascomycota (67.9%), Basidiomycota (28.9%), and Mucoromycota (3.0%; Supplement 4.g). Amplicon size was restricted between 100 bp to 500 bp during database curation (Supplement 2). However, the '*–method amplicon_ length* ' visualization indicates this range could be too restrictive for the maximum length size (Figure 2.d). The '*–method primer_ efficiency* ' visualization determined >80% of taxa within the phylum Ascomycota showed no diversity at the degenerate base locations in the primer-binding regions, while also showing a larger variation in base pair composition at the 3' end of the reverse primer (Supplement 4.h).

### *3.2 | Comparing incorporated diversity between reference databases*

For each of the four primer sets tested in this study, reference databases generated by CRABS contain the largest number of sequences and species compared to ecoPCR, MetaCurator, and RESCRIPt (Figure 3), except for the increased number of species contained within the mlCOIintF/jgHC02198 reference database generated by MetaCurator (Figure 3c). Reference databases generated by MetaCurator contain the second largest number of sequences and species, followed by RESCRIPt and ecoPCR (Figure 3). RESCRIPt was unable to generate reference databases for the mlCOIintF/jgHC02198 and gITS7/ITS4 primer sets. Both primer sets amplify a wide variety of taxonomic groups and contain degenerate bases in the primer-binding region, which resulted in difficulties to create local alignments and implement the *in silico* PCR step in the standard QIIME toolkit (QIIME extract_reads).

A significant overlap in incorporated species was observed between reference databases, with 80.6% of species incorporated in more than one reference database for the MiFish-E/U primer set, 52.9% for Taberlet c/h, 56.3% for mlCOIintF/jgHC02198, and 54.8% for the gITS7/ITS4 primer set (Figure 4). Similarly, a significant overlap in sequence ID's was observed between reference databases, with 55.3% of sequence ID's incorporated in more than one reference database for the MiFish-E/U primer set, 37.1% for Taberlet c/h, 26.8% for mlCOIintF/jgHC02198, and 27.6% for the gITS7/ITS4 primer set (Supplement 5). Interestingly, the amplicon region retrieved by the software packages was only identical between CRABS, ecoPCR, and RESCRIPt, as MetaCurator failed to recover the full amplicon region, except for the gITS7/ITS4 amplicon region (Figure 4).

### *3.3 | Taxonomic assignment differences between reference databases*

Reference database choice did not significantly impact the number of OTUs assigned to a specific taxonomic rank (Figure 5). On average, 10.6% ± 0.9% of OTUs failed to be assigned a taxonomy for the MiFish-E/U sequencing data, 5.1% ± 0.9% of OTUs for the Taberlet c/h sequencing data, 73.8% ± 3.6% of OTUs for the mlCOIintF/jgHC02198 sequencing data, and 16.8% ± 0.6% of OTUs for the gITS7/ITS4 sequencing data. Additionally, high similarity was achieved between reference databases as to which OTUs were able to be assigned a taxonomy (Supplement 6).

The achieved taxonomic assignment of OTUs between reference databases, on the other hand, showed limited overlap, with 39.5% identical taxonomy assignments for the MiFish-E/U sequencing data, 25.0% for the Taberlet c/h sequencing data, 28.3% for the mlCOIintF/jgHC02198 sequencing data, and 30.0% for the gITS7/ITS4 sequencing data (Figure 6). The limited overlap in taxonomy assignment resulted from differences in taxonomic resolution for a specific OTU, rather than the assignment of OTUs to different taxonomic lineages. No consistency was observed for which reference database achieved higher taxonomic resolution across OTUs within each sequencing data set (Supplement 6).

## 4 | DISCUSSION

The necessity to generate and curate reference databases to increase taxonomic assignment accuracy and resolution in eukaryotic metabarcoding research has recently come to light (Gold et al., 2021; Hleap et al., 2021). Since metabarcoding research targets a broad range of taxonomic groups and gene regions through a vast number of primer sets, flexibility is required from software packages to suit user-specific needs. Here, we present CRABS, an easy-to-use software program with a full suite of features to generate, curate, and explore reference databases.

### *4.1 | Sequence retrieval*

The increased diversity contained within CRABS-generated reference databases compared to the other software programs tested can partially be explained by CRABS' ability to access multiple online sequencing repositories, including BOLD, EMBL, NCBI, and MitoFish. As sequence data only partially overlaps between online repositories (Arranz et al., 2020; Meiklejohn et al., 2019; Porter & Hajibabaei, 2018), CRABS facilitates the generation of reference databases using the largest proportion of available sequences, thereby increasing the diversity included in the final curated reference database. MetaCurator was the second-best performing software package in our comparison, but does not incorporate a function to download sequencing data. By using CRABS to download sequencing data from multiple online repositories in the MetaCurator pipeline, we could have influenced the output diversity achieved from that program.

While CRABS can access multiple online repositories, downloaded file sizes and time requirements are kept to a minimum by solely downloading the gene region or taxonomic group of interest using the '*db_download –query* ' parameter. Additionally, sequence length restrictions can be specified to exclude genome sequences, further reducing file sizes and speeding up the process. EcoPCR, on the other hand, recommends the download of the entire EMBL database, taking up >2 TB of storage and a significant amount of time. While time- and size-inefficient, ecoPCR-generated reference databases included several species that were missed by CRABS due to the initial sequence length exceeding the restriction parameter. Another benefit of utilizing the full

online repository is the identification of issues around co-amplification from unintended taxonomic groups (Banos et al., 2018). For example, ecoPCR identified the co-amplification of plants for the gITS7/ITS4 (fungal) primer set, taxa not incorporated in the CRABS reference database for which initial sequencing data was restricted to fungal ITS sequences. To avoid the need to download complete online repositories, we recommend using primer-specificity testing software, such as Primer-BLAST (Ye et al., 2012), to determine which taxonomic groups need to be included in the initial sequence download by CRABS.

*4.2 | Amplicon extraction*

The extraction of the amplicon region from sequences deposited in online repositories is a crucial part in the creation of curated reference databases. While the different methodologies implemented in software packages can be effective, CRABS' combined implementation of *in silico* PCR analysis and pairwise global alignments resulted in the most complete reference databases for each of the four primer sets. In particular, using amplicon regions extracted from the *in silico*PCR analysis as seed sequences for pairwise global alignments substantially increased the diversity included in the final reference database, thereby outperforming an "*in silico* PCR-only" approach. The proportion of additional barcodes retrieved by the pairwise global alignment step will be heavily influenced by the chosen primer set, with lower success for metabarcoding primers located within the traditional barcoding region. Caution is warranted in the relaxed parameter settings in the '*pga* ' function, as it may increase the inclusion of false-positive hits in the reference database. We, therefore, recommend the use of '*pga –filter_ method strict* ' to reduce the chance of including erroneous sequences.

MetaCurator, employing profile hidden Markov models to extract amplicon regions, is the only software package in our test not taking into account any information about the primer-binding regions. Instead, MetaCurator imports up to 10 user-provided seed sequences that are trimmed to the exact marker of interest. Without information about primer-binding regions, MetaCurator often failed to recover several base pairs at the beginning and end of the target amplicon region. While no effect on taxonomy assignment was observed in our comparison, further studies are required to determine the impact of partial reference sequences on taxonomy assignment. Additionally, the amplicon extraction step implemented in MetaCurator came at a great computational cost, taking hundreds of more CPU hours than any other method. Furthermore, MetaCurator was unable to handle the high number of sequences included in the mlCOIintF/jgHC02198 and gITS7/ITS4 database trials. Once the number of query sequences reached approximately half a million, MetaCurator could not run to completion (the program did not finish in our trial of a two week runtime). To circumvent the issue, larger files were split into subsets, which were run separately, after which the results were combined, furthermore adding to the computational cost and potentially introducing other issues with running subsets separately.

Even greater computational problems were encountered for the QIIME RESCRIPt plug-in. When following the author's guidelines to create a database using RESCRIPt, we were unable to proceed past the alignment step. Due to extreme sequence divergence across taxonomic groups and/or excessive indels, alignments around a seed group of taxa contained an average gap percentage of >70%. From these poor alignments – even at the genus level – no amplicons were retained during the amplicon extraction step in the pipeline. In light of this, we opted to use the *in silico* PCR step in the standard QIIME toolkit (QIIME extract_reads), rather than relying on the local alignment method implemented in RESCRIPt. While successful for the MiFish-E/U and Taberlet c/h primer sets, no results could be obtained for the mlCOIintF/jgHC02198 and gITS7/ITS4 primer sets due to the presence of degenerate bases in the primer sequences.

*4.3 | Database curation*

Comprehensive and well-annotated reference databases are of crucial importance to increase taxonomy assignment accuracy (Hleap et al., 2021). While database curation parameters between software packages were not compared in our experiment, CRABS boasts the most complete set of features, with the only software program able to take into account geographical species locations, a parameter shown to increase taxonomic assignment accuracy (Gold et al., 2021; Murali et al., 2018). Additionally, the efficiency of the taxonomy-dependent dereplication function differed greatly between software packages, with CRABS and RESCRIPt

handling large datasets within seconds, while MetaCurator failed to handle datasets greater than ~500,000 sequences.

### *4.4 | Database exploration*

Alongside the functionality to generate curated reference databases, CRABS facilitates the exploration of reference databases using multiple visualizations, thereby providing essential information on how to interpret taxonomy assignment results. For example, machine learning classifiers, such as SINTAX (Edgar, 2016) and RDP (Qiong et al., 2007), are known to overclassify sequences in situations where the correct label lies outside the scope of the reference database (Dave, 1991; Murali et al., 2018). Hence, it is crucial to determine the completeness of the reference database for missing barcodes of closely related species, implemented in the CRABS '*–methoddb_ completeness* ' visualization. It should be noted, however, that the '*–method db_ completeness* ' visualization should be interpreted as a guide only, as the function is built around the NCBI taxonomy database (Federhen, 2012), known to exhibit errors (Schoch et al., 2020). Therefore, consulting the primary literature remains essential.

Metabarcoding analyses attempt to classify sequences to species-level resolution, based on the variations present in the amplicon region. However, these partial gene sequences do not always permit species-level resolution and the variation might not be consistent between taxonomic groups, as has repeatedly been observed (Porter & Hajibabaei, 2020). The generation of amplicon-based phylogenetic trees, implemented in the CRABS '*–method phylo* ' visualization, can provide guidance about the resolution of the amplicon region for specific taxonomic groups, thereby aiding in the assignment of taxonomy at the correct resolution.

With multiple primer sets available for specific taxonomic groups targeting various gene regions (Zhang et al., 2020), CRABS visualizations could aid in determining the optimal primer set for a specific experimental design. For example, '*–method db_ completeness* ' might provide information about which gene regions contains the largest number of barcodes for taxa of interest, while the '*–method diversity* ' visualization gives insight into issues surrounding unintended co-amplification. The '*–method amplicon_ length* ', on the other hand, could determine which primer sets can be multiplexed on an Illumina sequencing run and the '*–method phylo* ' can visualize which amplicon obtains the highest taxonomic resolution. Finally, the '*–method primer_ efficiency* ' shows which primer set contains the least amount of mismatches for target taxa or provides information on how to optimise the primer sequences for the taxa of interest.

### *4.5 | Implemented features within CRABS*

As shown in Table 1, CRABS is equally feature-rich or richer in certain feature categories, compared with the other three software packages. In particular, CRABS' support for downloading sequencing data of interest from multiple online repositories is distinctive. Additionally, the ability to use extracted amplicons from the *in silico* PCR analysis as a database for pairwise global alignment analysis enables the retrieval of a larger portion of amplicon regions. Furthermore, CRABS incorporates the most comprehensive set of database curation parameters, as well database export formats, thereby facilitating reference databases to be immediately used in taxonomy assignment software packages. The implemented CRABS visualizations, also, allow for a thorough investigation of the reference database to aid taxonomy assignment of sequencing data. Finally, easy installation of the conda package, simple parameter settings, and a fully documented step-by-step workflow for reference database curation renders CRABS user and analysis friendly.

## 5 | CONCLUSION

We present a reliable, flexible, and user-friendly software package to create curated reference databases. CRABS successfully generated reference databases for four widely used primer sets in metabarcoding research, incorporating higher diversity than ecoPCR, RESCRIPt, and MetaCurator, while also reducing computational requirements. CRABS provides a full suite of features that allows for the generation of curated reference databases in a limited timeframe, while facilitating the flexibility needed to cover user's needs. Furthermore, CRABS offers detailed visualizations to explore the reference database for completeness, included diversity, amplicon length variation, taxonomic resolution of the amplicon region, and primer-binding

efficiency, enabling a high level of quality control during taxonomy assignment of metabarcoding datasets. Such a feature set makes CRABS a powerful and valuable tool for curated reference database creation in the rapidly expanding field of metabarcoding.

## 6 | ACKNOWLEDGEMENTS

This work was funded through the Marsden Fast-Start (MFP-21-UOO-087; Molecular time-capsules of oceans past – reconstructing Antarctica's marine ecosystems using historical environmental DNA from marine invertebrate collections) and by the New Zealand Ministry of Business, Innovation and Employment funding (CAWX1904 – A toolbox to underpin and enable tomorrow's marine biosecurity system).

## 7 | AUTHOR CONTRIBUTIONS

GJJ wrote the CRABS code, with significant input from HC. HC developed the conda package. ED, JE, UvA, and NJG conducted the debugging of the program. GJJ conducted the comparative experiment. GJJ wrote the manuscript, with substantial help from HC. All authors contributed to the manuscript writing.

## 8 | DATA AVAILABILITY STATEMENT

The CRABS code is available on the public GitHub repo (https://github.com/gjeunen/reference_database_creator), as well as via conda packaging. The reference databases that were created to compare software programs are available as supplemental files (Supplement 3), as well as the code used to generate them (Supplement 2). The sequencing data used in this manuscript to compare the impact on taxonomy assignment between reference databases was obtained from published literature. The generated OTUs, plus their sequences and taxonomy assignments are available as supplemental files (Supplement 6)

## 9 | REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* , *215* (3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

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), 1–8.

Ávila-Arcos, M. C., Cappellini, E., Romero-Navarro, J. A., Wales, N., Moreno-Mayar, J. V., Rasmussen, M., Fordyce, S. L., Montiel, R., Vielle-Calzada, J.-P., Willerslev, E., & Gilbert, M. T. P. (2011). Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Scientific Reports* ,*1* (1), 74. https://doi.org/10.1038/srep00074

Banchi, E., Ametrano, C. G., Greco, S., Stanković, D., Muggia, L., & Pallavicini, A. (2020). PLANiTS: a curated sequence reference dataset for plant ITS DNA metabarcoding. *Database* , *2020* .

Banos, S., Lentendu, G., Kopf, A., Wubet, T., Glöckner, F. O., & Reich, M. (2018). A comprehensive fungi-specific 18S rRNA gene sequence primer toolkit suited for diverse research issues and sequencing platforms.*BMC Microbiology* , *18* (1), 190. https://doi.org/10.1186/s12866-018-1331-4

Barbera, P., Kozlov, A. M., Czech, L., Morel, B., Darriba, D., Flouri, T., & Stamatakis, A. (2019). EPA-ng: massively parallel evolutionary placement of genetic sequences. *Systematic Biology* , *68* (2), 365–369.

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

Bovo, S., Utzeri, V. J., Ribani, A., Cabbri, R., & Fontanesi, L. (2020). Shotgun sequencing of honey DNA can describe honey bee derived environmental signatures and the honey bee hologenome complexity.*Scientific Reports* , *10* (1), 9279. https://doi.org/10.1038/s41598-020-66127-1

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: a unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources* , *16* (1), 176–182. https://doi.org/doi:10.1111/1755-0998.12428

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* , *13* (7), 581–583.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., ... Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* , *7* (5), 335–336. https://doi.org/10.1038/nmeth.f.303

Chave, J. (2013). The problem of pattern and scale in ecology: what have we learned in 20 years? *Ecology Letters* , *16* , 4–16.

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

Cowart, D. A., Murphy, K. R., & Cheng, C.-H. C. H. C. (2018). Metagenomic sequencing of environmental DNA reveals marine faunal assemblages from the West Antarctic Peninsula. *Marine Genomics* ,*37* (September 2017), 148–160. https://doi.org/https://doi.org/10.1016/j.margen.2017.11.003

Cross, H., Sonstebo, J. H., Nagy, N. E., Timmermann, V., Solheim, H., Borja, I., Kauserud, H., Carlsen, T., Rzepka, B., Wasak, K., Vivian-Smith, A., & Hietala, A. M. (2017). Fungal diversity and seasonal succession in ash leaves infected by the invasive ascomycete Hymenoscyphus fraxineus. *New Phytologist* , *213* (3), 1405–1417. https://doi.org/https://doi.org/10.1111/nph.14204

Dave, R. N. (1991). Characterization and detection of noise in clustering. *Pattern Recognition Letters* , *12* (11), 657–664.

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

Edgar, R. C. (2016). SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *BioRxiv* , 74161. https://doi.org/10.1101/074161

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research* , *40* (D1), D136–D143. https://doi.org/10.1093/nar/gkr1178

Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessiere, J., Taberlet, P., & Pompanon, F. (2010). An In silico approach for the evaluation of DNA barcodes. *BMC Genomics* ,*11* (1), 434. https://doi.org/10.1186/1471-2164-11-434

Gold, Z., Curd, E., Goodwin, K., Choi, E., Frable, B., Thompson, A., Jr, H. J. W., Burton, R., Kacev, D., & Barber, P. (2021). *Improving Metabarcoding Taxonomic Assignment: A Case Study of Fishes in a Large Marine Ecosystem* . https://doi.org/10.22541/au.161407483.33882798/v1

Hagelberg, E., Hofreiter, M., & Keyser, C. (2015). Ancient DNA: the first three decades. *Philosophical Transactions of the Royal Society B: Biological Sciences* , *370* (1660), 20130371. https://doi.org/10.1098/rstb.2013.0371

Hebert, P. D. N., & Gregory, T. R. (2005). The Promise of DNA Barcoding for Taxonomy. *Systematic Biology* , *54* (5), 852–859. https://doi.org/10.1080/10635150500354886

Hleap, J. S., Littlefair, J. E., Steinke, D., Hebert, P. D. N., & Cristescu, M. E. (2021). Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. *Molecular Ecology Resources* , *21* (7), 2190–2203.

Hugenholtz, P., & Tyson, G. W. (2008). Metagenomics. *Nature* ,*455* (7212), 481–483. https://doi.org/10.1038/455481a

Ihrmark, K., Bodeker, I., Cruz-Martinez, K., Friberg, H., Kubartova, A., Schenck, J., Strid, Y., Stenlid, J., Brandstrom-Durling, M., & Clemmensen, K. E. (2012). New primers to amplify the fungal ITS2 region–evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiology Ecology* , *82* (3), 666–677.

Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T. P., Sado, T., Mabuchi, K., Takeshima, H., & Miya, M. (2013). MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Molecular Biology and Evolution* , *30* (11), 2531–2540.

Jeunen, G.-J., Urban, L., Lewis, R., Knapp, M., Lamare, M., Rayment, W., Dawson, S., & Gemmell, N. (2020). *Marine environmental DNA (eDNA) for biodiversity assessments: a one-to-one comparison between eDNA and baited remote underwater video (BRUV) surveys*.https://doi.org/10.22541/au.160278512.26241559/v1

Jeunen, G. J., Knapp, M., Spencer, H. G., Lamare, M. D., Taylor, H. R., Stat, M., Bunce, M., & Gemmell, N. J. (2019). Environmental DNA (eDNA) metabarcoding reveals strong discrimination among diverse marine habitats connected by water movement. *Molecular Ecology Resources* , *19* (2), 426–438. https://doi.org/10.1111/1755-0998.12982

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications* , *10* (1), 5029. https://doi.org/10.1038/s41467-019-13036-1

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface.*Nucleic Acids Research* , *36* (suppl_2), W5–W9.

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., & Cochrane, G. (2005). The EMBL nucleotide sequence database. *Nucleic Acids Research* ,*33* (suppl_1), D29–D33.

Key, F. M., Posth, C., Krause, J., Herbig, A., & Bos, K. I. (2017). Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication. *Trends in Genetics* , *33* (8), 508–520. https://doi.org/https://doi.org/10.1016/j.tig.2017.05.005

Koljalg, U., Larsson, K., Abarenkov, K., Nilsson, R. H., Alexander, I. J., Eberhardt, U., Erland, S., Hoiland, K., Kjoller, R., & Larsson, E. (2005). UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist* ,*166* (3), 1063–1068.

Kovalenko, K. E., Thomaz, S. M., & Warfe, D. M. (2012). Habitat complexity: approaches and future directions. *Hydrobiologia* ,*685* (1), 1–17. https://doi.org/10.1007/s10750-011-0974-z

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.

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* , *n/a* (n/a). https://doi.org/https://doi.org/10.1002/edn3.303

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., & Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology* , *10* (1), 1–14. https://doi.org/10.1186/1742-9994-10-34

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal* , *17* (1), 10–12.

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* , *11* (1), 1–16.

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

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H., Kondoh, M., & Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Royal Society Open Science* ,*2* (7), 150088. https://doi.org/10.1098/rsos.150088

Murali, A., Bhargava, A., & Wright, E. S. (2018). IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences.*Microbiome* , *6* (1), 1–14.

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

Porter, T. M., & Hajibabaei, M. (2020). Putting COI Metabarcoding in Context: The Utility of Exact Sequence Variants (ESVs) in Biodiversity Analysis . In *Frontiers in Ecology and Evolution* (Vol. 8). https://www.frontiersin.org/article/10.3389/fevo.2020.00248

Qiong, W., M., G. G., M., T. J., & R., C. J. (2007). Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology* , *73* (16), 5261–5267. https://doi.org/10.1128/AEM.00062-07

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (http://www.barcodinglife. org). *Molecular Ecology Notes* , *7* (3), 355–364.

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

Robeson, M. S., O'Rourke, D. R., Kaehler, B. D., Ziemski, M., Dillon, M. R., Foster, J. T., & Bokulich, N. A. (2020). RESCRIPt: Reproducible sequence taxonomy reference database management for the masses.*Biorxiv* .

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahe, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* ,*4* , e2584.

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* , *2020* , baaa062. https://doi.org/10.1093/database/baaa062

Scott, H., L., K. J., & B., F. P. (2021). Toward a genome sequence for every animal: Where are we now? *Proceedings of the National Academy of Sciences* , *118* (52), e2109019118. https://doi.org/10.1073/pnas.2109019118

Seeber, P. A., McEwen, G. K., Lober, U., Forster, D. W., East, M. L., Melzheimer, J., & Greenwood, A. D. (2019). Terrestrial mammal surveillance using hybridization capture of environmental DNA from African waterholes. *Molecular Ecology Resources* , *19* (6), 1486–1496. https://doi.org/https://doi.org/10.1111/1755-0998.13069

Seersholm, F. V, Cole, T. L., Grealy, A., Rawlence, N. J., Greig, K., Knapp, M., Stat, M., Hansen, A. J., Easton, L. J., Shepherd, L., Tennyson, A. J. D., Scofield, R. P., Walter, R., & Bunce, M. (2018). Subsistence practices, past biodiversity, and anthropogenic impacts revealed by New Zealand-wide ancient DNA survey. *Proceedings of the National Academy of Sciences* , *115* (30), 7771 LP – 7776. https://doi.org/10.1073/pnas.1803573115

Somervuo, P., Koskela, S., Pennanen, J., Henrik Nilsson, R., & Ovaskainen, O. (2016). Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics* , *32* (19), 2920–2927. https://doi.org/10.1093/bioinformatics/btw346

Soule, M. E. (1985). What is conservation biology? *BioScience* ,*35* (11), 727–734.

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., Harvey, E. S., & Bunce, M. (2017). Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports* , *7* (1), 12240. https://doi.org/10.1038/s41598-017-12501-5

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

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G., Brochmann, C., & Willerslev, E. (2007). Power and limitations of the chloroplast trn L (UAA) intron for plant DNA barcoding . *Nucleic Acids Research* , *35* (3), e14–e14. https://doi.org/10.1093/nar/gkl938

Taberlet, P., Gielly, L., Pautou, G., & Bouvet, J. (1991). Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* , *17* (5), 1105–1109.

White, T. J., Bruns, T., Lee, S., & Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics.*PCR Protocols: A Guide to Methods and Applications* , *18* (1), 315–322.

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* ,*15* (3), 1–12.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., & Madden, T. L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* ,*13* (1), 134. https://doi.org/10.1186/1471-2105-13-134

Zhang, S., Zhao, J., & Yao, M. (2020). A comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. *Methods in Ecology and Evolution* , *11* (12), 1609–1625. https://doi.org/https://doi.org/10.1111/2041-210X.13485

## 10 | FIGURE HEADINGS

Figure 1: Schematic depiction of CRABS workflow and incorporated features. The five CRABS modules are indicated in blue. CRABS implemented functions within each module are underlined.

Figure 2: CRABS generated reference database exploration through CRABS incorporated visualizations. (a) horizontal bar graph displaying number of species (blue) and number of sequences (orange) contained within each class from the MiFish-E/U reference database. (b) Bar graph displaying the proportion of base occurrence for the primer-binding regions of the Taberlet c/h primer set, with base 'A' indicated by orange, base 'C' indicated in green, base 'G' indicated in yellow, and base 'T' indicated in red. Degenerate bases (label 'other') are indicated in grey. Sequences of the Taberlet c/h primer set are presented on the bottom of the figure. (c) A phylogenetic tree depicting the variation observed within the mlCOIintF/jgHC02198 amplicon region for the genus Apteryx (kiwi). (d) Amplicon length distribution of the gITS7/ITS4 primer set, with the overall amplicon length distribution shaded grey and most abundant phyla represented by coloured lines (Ascomycota: blue; Basidiomycota: orange; Mucoromycota: green). Figures are taken straight from CRABS output without further editing.

Figure 3: Reference database comparison between CRABS (blue), ecoPCR (purple), MetaCurator (green), RESCRIPt (yellow) for the (a) MiFish-E/U, (b) Taberlet c/h, (c) mlCOIintF/jgHC02198, and (d) gITS7/ITS4 primer sets. Number of sequences are depicted on the primary y-axis and number of species are depicted on the secondary y-axis.

Figure 4: Venn diagrams displaying the proportional overlap of species and identical sequences between CRABS (blue), ecoPCR (purple), MetaCurator (green), and RESCRIPt (yellow) for the (a) MiFish-E/U, (b) Taberlet c/h, (c) mlCOIintF/jgHC02198, and (d) gITS7/ITS4 primer sets. Percentage values between brackets indicate the proportion of species and identical sequences incorporated into the specific reference database for each of the four software programs.

Figure 5: Observed differences in proportion of OTUs classified between reference databases created by CRABS, ecoPCR, MetaCurator, and RESCRIPt for the (a) MiFish-E/U, (b) Taberlet c/h, (c) mlCOIintF/jgHC02198, and (d) gITS7/ITS4 primer sets. Taxonomy assignment of OTUs was performed through the SINTAX algorithm implemented in VSEARCH. Proportion of unassigned OTUs is indicated in black, species in dark blue, genus in light blue, family in dark red, order in light red, class in dark purple, phylum in light purple, and kingdom in green.

Figure 6: Venn diagrams displaying the proportional overlap in taxonomy assignment of OTUs between CRABS (blue), ecoPCR (purple), MetaCurator (green), RESCRIPt (yellow) for the (a) MiFish-E/U, (b) Taberlet c/h, (c) mlCOIintF/jgHC02198, and (d) gITS7/ITS4 primer sets.

## 11 | TABLE HEADINGS

Table 1: Comparison of the implemented features in each of the four tested software packages, including CRABS, ecoPCR, MetaCurator, and RESCRIPt. Amplicon extraction methodologies implemented in current software: (i) *in* silico PCR; (ii) LA – Multiple Sequence Alignment; (iii) PGA – Pairwise Global Alignment; and (iv) PHMM – Profile Hidden Markov Models. * indicates the code is not implemented in the software package, but provided as shell scripts in the tutorial workflow.

| Module | Feature | CRABS | ecoPCR | MetaCurator | RESCRIPt |
|---|---|---|---|---|---|
| Sequence retrieval | NCBI | P | | P* | P |
| | EMBL | P | P* | | |
| | BOLD | P | | | P* |
| | MitoFish | P | | | |
| | In-house | P | | | P* |
| Amplicon extraction | *in silico PCR* | P | P | | |
| | MSA | | | | P |
| | PGA | P | | | |
| | PHMM | | | P | |
| Database curation | Dereplication: taxonomy | P | P | P | P |
| | Dereplication: no taxonomy | P | | | P |
| | Curation: sequence | P | P | | P |
| | Curation: header | P | P | | |
| | Curation: taxonomy | P | P | P | P |
| Database formatting | Sintax | P | | | |
| | RDP | P | | | |
| | QIIME | P | | P | P |
| | DADA | P | | | |
| | IDT | P | | | |
| | OBITools | | P | | |
| Database exploration | Taxonomic resolution | P | | | |
| | Included diversity | P | | | |
| | Amplicon length | P | | | |

| Module | Feature | CRABS | ecoPCR | MetaCurator | RESCRIPt |
|---|---|---|---|---|---|
| | Primer efficiency | P | | | |
| | Database completeness | P | | | |