Received: 31 May 2023 | Revised: 12 October 2023 | Accepted: 19 October 2023

DOI: 10.1002/edn3.489

Environmental DNA
Open Access
Dedicated to the study and use of environmental DNA for basic and applied sciences

WILEY

SPECIAL ISSUE ARTICLE

rCRUX: A rapid and versatile tool for generating metabarcoding reference libraries in R

Emily E. Curd[1] | Luna Gal[2,3] | Ramon Gallego[4] | Katherine Silliman[5,6] | Shaun Nielsen[7] | Zachary Gold[3,8]

[1]Vermont Biomedical Research Network, University of Vermont, Burlington, Vermont, USA

[2]Landmark College, Putney, Vermont, USA

[3]California Cooperative Oceanic Fisheries Investigations (CalCOFI), Scripps Institution of Oceanography, University of California San Diego (UCSD), La Jolla, California, USA

[4]Departamento de Biología, Universidad Autónoma de Madrid, Madrid, Spain

[5]Northern Gulf Institute, Mississippi State University, Starkville, Mississippi, USA

[6]NOAA Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida, USA

[7]Independent Researcher, St Leonards, New South Wales, Australia

[8]NOAA Pacific Marine Environmental Laboratory, Seattle, Washington, USA

Correspondence
Emily E. Curd, Vermont Biomedical Research Network, University of Vermont, Burlington, VT, USA.
Email: eguswa@uvm.edu

Zachary Gold, California Cooperative Oceanic Fisheries Investigations (CalCOFI), Scripps Institution of Oceanography, University of California San Diego (UCSD), La Jolla, CA, USA.
Email: zachary.gold@noaa.gov and zjgold@g.ucla.edu

Funding information
California Cooperative Oceanic Fisheries Investigations; National Institute of General Medical Sciences, Grant/Award Number: P20GM103449

Abstract

The sequencing revolution requires accurate taxonomic classification of DNA sequences. The key to making accurate taxonomic assignments is curated, comprehensive reference barcode databases. However, the generation and curation of such databases has remained challenging given the large and continuously growing volumes of both DNA sequence data and novel reference barcode targets. Monitoring and research applications require a greater diversity of specialized gene regions and targeted taxa than are currently curated by professional staff. Thus, there is a growing need for an easy-to-implement computational tool that can generate comprehensive metabarcoding reference libraries for any bespoke locus. We address this need by reimagining CRUX from the Anacapa Toolkit and present the rCRUX package in R which, like its predecessor, relies on sequence homology and PCR primer compatibility instead of keyword searches to avoid limitations of user-defined metadata. The typical workflow involves searching for plausible seed amplicons (*get_seeds_local*() or *get_seeds_remote*()) by simulating in silico PCR to acquire a set of sequences analogous to PCR products containing a user-defined set of primer sequences. Next, these seeds are used to iteratively blast search seed sequences against a local copy of the National Center for Biotechnology Information (NCBI)-formatted *nt* database using a taxonomic rank-based stratified random sampling approach (*blast_seeds*()). This results in a comprehensive set of sequence matches. This database is dereplicated and cleaned (*derep_and_clean_db*()) by identifying identical reference sequences and collapsing the taxonomic path to the lowest taxonomic agreement across all matching reads. This results in a curated, comprehensive database of primer-specific reference barcode sequences from NCBI. Databases can then be compared (*compare_db*()) to determine read and taxonomic overlap. We demonstrate that rCRUX provides more comprehensive reference databases for the MiFish Universal Teleost 12S, Taberlet trnl, fungal ITS, and Leray CO1 loci than CRABS, MetaCurator, RESCRIPt, and ecoPCR reference databases. We then further demonstrate the utility of rCRUX by generating 24 reference databases for 20 metabarcoding loci, many of which lack dedicated reference database curation efforts. The rCRUX package provides a simple-to-use tool

for the generation of curated, comprehensive reference databases for user-defined loci, facilitating accurate and effective taxonomic classification of metabarcoding and DNA sequence efforts broadly.

1 | INTRODUCTION

The fields of freshwater, estuarine, and marine ecology are rapidly embracing high-throughput sequencing to detect, monitor, or assess change in biological communities (Deiner et al., 2017; Takahashi et al., 2023). Fundamental to the efficacy of these molecular biomonitoring efforts, particularly metabarcoding (amplicon sequencing), is the taxonomic assignment of the sequences generated (Bik, 2021; Edgar, 2018). Taxonomic assignment is a complicated bioinformatic process that involves many challenges including the uncertainty around the generated sequencing data, the comparison between those data and a reference database of sequences of known origin, and the bioinformatic decisions that land on a taxonomic identification of the generated sequences (Edgar, 2018; Hleap et al., 2021; Jeunen et al., 2023; Mathon et al., 2021; O'Rourke et al., 2020). Ensuring accurate taxonomic assignment is critical for the adoption of biomolecular monitoring tools including environmental DNA (eDNA) metabarcoding, microbiome, bulk metabarcoding, and gut and diet studies among other applications (Deiner et al., 2017).

Paramount to the success of taxonomic assignment is the comprehensiveness and accuracy of the reference database used to classify query DNA sequences (Banchi et al., 2020; Bucklin et al., 2016; Gold et al., 2021). Large-scale barcoding of life efforts over the past three decades has provided the raw material for such reference databases (Costa & Carvalho, 2007; Darwin Tree of Life Project Consortium, 2022; Hebert et al., 2003; Stoeckle & Hebert, 2008) generating millions of reference barcode sequences publicly available through the National Center for Biotechnology Information (NCBI), the Barcode of Life Data System (BOLD), the European Nucleotide Archive (ENA), and others (Cummins et al., 2022; Ratnasingham & Hebert, 2007; Sayers et al., 2022). Despite the incompleteness of current global reference databases across all domains of life, these large sequence repositories are constantly improving and expanding to allow for accurate identification of DNA sequences needed for a suite of ecological and public health efforts (Beng & Corlett, 2020; Manor et al., 2020; Soon et al., 2013; Taberlet et al., 2018; Thompson et al., 2017). Generating a high-quality reference database from these enormous sequence repositories requires a full accounting of all orthologous sequences, the detection and removal of mislabeled sequences, and the identification of identical sequences across taxa (Curd et al., 2019; Jeunen et al., 2023; Richardson et al., 2020). Parsing and refining these large sequence repositories into curated databases that are comprehensive for specific marker sets remains a significant challenge (Jeunen et al., 2023).

Efforts to address this challenge either rely on the dedicated maintenance and curation of reference databases for specific loci of interest or force the end user to curate their own database, with limited efficacy because they rely on keyword searches, are too computationally demanding, or their installation process is too complicated, requiring a suite of software dependencies. By far, the most successful and widely used reference databases (e.g. Silva, PR2, UNITE, and MitoFish) rely on dedicated staff and resources to maintain and update such repositories (Guillou et al., 2012; Kõljalg et al., 2005; Quast et al., 2012; Zhu et al., 2023). Given the extensive resources needed to curate and maintain such repositories, there are only a handful of such efforts representing only commonly used loci. We cannot expect to have similar dedicated efforts for all metabarcoding loci. This is especially true as novel sequencing technologies allow for longer targets and more immediate in situ sequencing (Zorz et al., 2023). Thus, alternative reference database-generating tools are needed to alleviate taxonomic assignment restrictions at the database level and fill this operational gap in the field.

A commonly utilized approach to generating reference databases relies on keyword searches (Keck & Altermatt, 2023). Such efforts are dependent on the accuracy of associated sequence metadata submitted by users. However, a lack of controlled vocabulary and metadata standards often leads to poor annotations (e.g. CO1, COI, and COX1, all describing the same cytochrome oxidase gene) which frequently limits the comprehensiveness of such reference databases (Curd et al., 2019; Jeunen et al., 2023; Hemsley et al., 2020; Porter & Hajibabaei, 2018). Many tools address these specific limitations in generating comprehensive reference barcode databases for key loci like CO1 (e.g. MIDORI2, CO-ARBitrator, MARES, and COInr; Arranz et al., 2020; Heller et al., 2018; Leray et al., 2022; Meglécz, 2023). However, keyword search-based database generation is particularly susceptible to inadequate capture of orthologous sequences as this requires a priori knowledge of sequence similarities and associated metadata (e.g. MiFish 12S and microbial 16S; Gold et al., 2021; Siddall et al., 2009). Such keyword search-based approaches are useful for a handful of widely used loci (e.g. CO1), but are not flexible enough to be applied to any metabarcoding and sequencing locus (Ahmed et al., 2019; Keck et al., 2022).

To address these limitations, a suite of reference barcode-generating tools were designed based on sequence similarity instead of associated metadata to build comprehensive, curated reference databases (Jeunen et al., 2023; Richardson et al., 2020). CRUX and its counterparts, MetaCurator, Metaxa2, and CRABS, all similarly rely on a two-step database-generating process (Bengtsson-Palme et al., 2015; Curd et al., 2019; Jeunen et al., 2023; Richardson et al., 2020). First, these tools conduct an in silico PCR or analogous seed acquisition step to generate a set of "seed" sequences containing the primer regions. And since not all sequences are submitted with the primer sequences intact, these tools implement a second step which aligns these seed sequences across the large sequence repositories (e.g. GenBank, ENA, and BOLD) to acquire a

comprehensive set of similar sequences. Inherently, these software tools take a brute force approach to generating reference databases that acquire all orthologous sequences and thus, unsurprisingly, require significant computational resources (Jeunen et al., 2023). In addition, these tools often rely on a large number of software dependencies which are difficult to install and maintain on high-performance computing clusters (e.g. CRUX and Metaxa2). Furthermore, these tools are often written in software languages that are not as widely adopted as R and require more advanced command-line computational skills to implement. Together, these limit the adoption and utilization of such reference database-generating tools.

Here, we present rCRUX, a reference database-generating R package (R Core Team, 2022) that relies on efficient iterative BLAST searches to sample all orthologue sequence space (Altschul et al., 1990; Ye et al., 2012), utilizing a smaller set of readily available dependencies. rCRUX provides a simple, easy-to-use reference database-generating tool that facilitates the generation of curated, comprehensive bespoke reference libraries across a diversity of users and platforms including cloud-hosted services.

2 | METHODS

Here, we build on the rationale behind the generation of locus-specific databases outlined first in Curd et al., 2019, which demonstrates that the most comprehensive databases are obtained by way of sequence similarity instead of intended taxonomic identity or sequence description (Curd et al., 2019, Jeunen et al., 2023, Richardson et al., 2020). rCRUX produces reference sequence databases in a three-step process (Figure 1): (1) identification of seed sequences that match the primers of interest, (2) finding homologous and orthologous sequences to those seed sequences via BLAST, and (3) dereplication of the resulting database to reduce redundancy and detect poorly annotated sequences. This can be followed by (4) database comparison tools provided in rCRUX.

2.1 | Installation

In order to install rCRUX onto a computer or cluster, users must first download the rCRUX package and NCBI's BLAST+ toolkit. In addition, users need a blast-formatted nucleotide database which can be downloaded directly from NCBI as well as NCBI taxonomy IDs which can be acquired using taxonomizr's *prepareDatabase*() function (Sherrill-Mix, 2019). We note that these combined databases required over 340 GB of storage as of April 2023.

2.2 | get_seeds: In silico PCR

The first step of rCRUX is an in silico PCR step which takes a set or sets of forward and reverse primer sequences (single or multiple forward and single or multiple reverse primers, which can include degenerate bases) and returns possible full-length barcode sequences containing forward and reverse primer matches along with taxonomic information. This step can be implemented locally through *get_seeds_local*() which uses a modified adaptation of NCBI's primer blast or remotely through *get_seeds_remote*() which submits a web form directly to NCBI's primer blast tool. *get_seeds_local*() avoids querying NCBI's primer BLAST tool and thus is not subject to arbitrary throttling of remote jobs that require significant memory on the NCBI server (Camacho et al., 2009; Ye et al., 2012). Given this limitation, *get_seeds_local*() is the preferred implementation in the rCRUX environment. However, *get_seeds_remote*() provides an alternative tool to test taxonomic breadth of primers without downloading the large (>340 GB) nt database and can allow users to quickly test the taxonomic breadth of a given primer set in R on a personal computer through in silico PCR.

Specifically, *get_seeds_local*() passes the forward and reverse primer sequences for a given PCR product to *run_primer_blastn*(). In the case of a non-degenerate primer set, only two primers will be passed to *run_primer_blast*(). In the case of a degenerate primer set, *get_seeds_local*() will obtain all possible versions of the degenerate primer(s) (using primerTree's *enumerate_primers*() function), randomly sample a user-defined number of forward and reverse primers, and generate a fasta file of selected primers (Cannon et al., 2016). The selected primers are then passed to *run_primer_blastn*() which queries each primer against a blast-formatted database using the task *blastn_short*. This process continues until all of the selected primers are blasted. The result is an output table containing the query subject id, subject NCBI GenInfo Identifier (gi), subject accession version, number of mismatches between the subject and query, subject

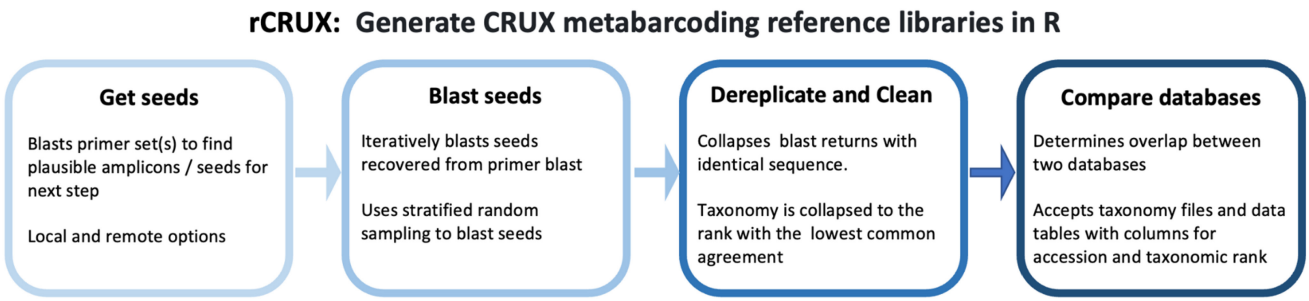

FIGURE 1 Overview of rCRUX workflow.

start base pair location, subject end base pair location, and subject NCBI taxonomic identification (taxID). These returned BLAST hits are then quality controlled to see if they generate plausible amplicons (e.g. amplify the same accession and are in the correct orientation to produce a PCR product). These hits are further filtered for user-specified length and number of mismatches. Lastly, taxonomy is appended to these filtered hits using *get_taxonomy_from_accession*() (Sherrill-Mix, 2019; Zeileis & Grothendieck, 2005; Zhang et al., 2017 Alternatively, *get_seeds_remote*() passes the forward and reverse primer sequences along with user-specified taxID(s) of target organism(s) and databases through *iterative_primer_search* to NCBI's primer blast tool (Ye et al., 2012). Degenerate primers are converted into all possible non-degenerate sets, and a user-defined maximum number of primer combinations is passed to the API. Multiple taxIDs are searched independently, as are multiple databases (e.g. *c*("*nt*," "*refseq_representative_genomes*")). A primer search is then conducted on each resultant combination using *modifiedPrimerTree_Functions* which is a modified version of primerTree's *primer_search*() and primerTree's *parse_primer* to query NCBI's primer BLAST tool, filter the results, and aggregate them into a single dataframe (Cannon et al., 2016). These hits are further filtered for user-specified length and number of mismatches. Lastly, taxonomy is appended to these filtered hits.

The result of the two *get_seeds_** functions is a fasta and taxonomy file containing results with the specified primer sequences along with summary statistics of accessions, taxonomic ranks, and hits returned.

2.3 | blast_seeds: Building comprehensive databases of similar sequences

blast_seeds() takes the output from *get_seeds_local*() or *get_seeds_remote*() and iteratively blasts the seed sequences using a stratified random sampling of a given taxonomic rank, making sure that at least one representative of each rank enters the iterative blast process (default rank is genus). The randomly sampled subset of seeds is then formatted into a multi-line fasta and run through *blastn* to recover all similar sequences based on user-defined sequence similarity parameters (e.g. percent identity, evalue, and query length; Camacho et al., 2009). The resulting blast hits are then dereplicated by accession with only the longest read per accession retained in the output table. All of the subset seeds and seeds recovered through the blast process are then removed from the seeds dataframe, reducing the number of sequences to be blasted. This stratified random sampling process is repeated until there are fewer seed sequences remaining than the max_to_blast parameter, at which point all remaining seeds are blasted. The final aggregated results are cleaned for multiple blast taxIDs, hyphens, and wildcards and are then appended with taxonomy.

Importantly, we note that the blast databases downloaded from NCBI's FTP site utilize representative accessions where identical sequences have been collapsed across multiple accessions even if they have different taxIDs. Here, we identify representative accessions with multiple taxIDs (Katz et al., 2021) and unpack all of the collapsed accessions to allow for the identification of lowest common agreed taxonomy for each representative accession. We do not identify or unpack representative accessions that report a single taxID.

2.4 | derep_and_clean_db: Quality control and curation of reference database

The final step of rCRUX, *derep_and_clean_db*(), takes the output from *blast_seeds*() and conducts quality control and dereplicates the dataset to identify representative sequences. First, all sequences with NA taxonomy for phylum, class, order, family, and genus are removed from the dataset because they typically represent environmental samples with low value for taxonomic classification and are stored separately. Next, all sequences with the same length and composition are collapsed to a single database entry, where the accessions and taxIDs (if there are more than one) are concatenated. The sequences with a clean taxonomic path (e.g. no ranks with multiple entries) are saved. In contrast, sequences with multiple entries for a given taxonomic rank are processed further by removing NAs from rank instances with more than one entry (e.g. "Chordata, NA" will mutate to "Chordata"). Any remaining instances of taxonomic ranks with more than one taxID are reduced to NA (e.g. species rank "*Badis assamensis*, *Badis badis*" will mutate to "NA", but genus rank will remain "*Badis*"). Finally, the resulting taxonomic paths are synonymized to the lowest taxonomic agreement. Lastly, the above cleaned and dereplicated sequences are used to generate a fasta file and taxonomy file of representative NCBI accessions for each sequence.

We note that these dereplicated sequences are analogous to Barcode Index Numbers (BINs; Ratnasingham & Hebert, 2007) which cluster reference sequences and provide a synonymized taxonomy (Fontes et al., 2021). However, our dereplication approach results in unique sequences as opposed to clustered BINs, much like amplicon sequence variants represent unique sequences as opposed to clustered operational taxonomic units.

2.5 | compare_ref_db: Exploring overlap and mismatches between two reference databases

We provide an additional function to compare the overlap and mismatches between any two reference databases. This function provides a summary table, generates a Venn diagram of overlapping accessions and species, and creates a Krona plot of unique taxa to each reference database (Gao et al., 2021; McMurdie & Holmes, 2013; Pauvert, 2020).

2.6 | Benchmarking

We first benchmark the efficacy of rCRUX compared to the original implementation of CRUX. Here, the seed acquisition step was

benchmarked by comparing the rCRUX *get_seeds_local*() in silico PCR function against the CRUX-implemented ecoPCR function (Ficetola et al., 2010) for the MiFish Universal Teleost locus (MiFish; Miya et al., 2015). We built a MiFish Universal Teleost rCRUX database and a corresponding CRUX ecoPCR database using the same underlying EMBL database (Kanz et al., 2005; r143: fun, inv, mam, pln, pro, vrt). We then benchmarked the final databases and compared the results of the *blast_seeds*() steps for both rCRUX and CRUX using the NCBI *nt* blast database (December 2022 download, see Data S1 for details) supplemented with the NCBI *mito* (mitochondrial) blast database. The rCRUX reference database was generated using the optimized parameters outlined in the Data S1 (Figures S1–S8) with the exception of using 4 mismatches between primers and priming sites for rCRUX *get_seeds_local*() to exactly match the parameters employed in the CRUX seed acquisition step. The CRUX reference database was generated using default parameters presented in Curd et al. (2019). All rCRUX reference database statistics are presented in Table 1 and Table S3. For the seed blasting steps, each blast allowed a maximum of 1000 samples at a time, align = "10,000,000," minimum length of 170, and a maximum length of 250. In order to compare compatible taxonomic paths between CRUX and rCRUX, the accessions returned by CRUX were assigned taxonomy using taxonomizr (Sherrill-Mix, 2019).

We next benchmark the efficacy of the rCRUX reference databases against CRABS, RESCRIPt, MetaCurator, and ecoPCR as implemented and presented in Jeunen et al. (2023). We note that we did not remake these reference databases from scratch, instead comparisons were made using the previously generated reference databases made publicly available by Jeunen et al. (2023). To ensure compatible taxonomic assignments, we generated new taxonomic paths for all databases by taxID using taxonomizr. We specifically compare the efficacy of the rCRUX, CRABS, MetaCurator, ecoPCR, and RESCRIPt reference database-generating tools for the MiFish Universal Teleost (MiFish; Miya et al., 2015), Taberlet trnl (trnl; Taberlet et al., 1991; Taberlet et al., 2007), fungal ITS (FITS; Ihrmark et al., 2012; White et al., 1990), and Leray CO1 (CO1; Leray et al., 2019) loci. The trnl and FITS rCRUX databases were made using the NCBI *nt* blast database (December 2022 download, see Data S1 for details). Two rCRUX MiFish databases were made: one using the same NCBI *nt* blast database and an expanded database using the same NCBI *nt* blast database supplemented with an additional custom blast database comprised of all Actinopterygii mitogenomes from the NCBI *mito* database (see Data S1 for details). The expanded database is used for all comparisons below. The CO1 rCRUX database was generated by combining the final results of three unique rCRUX strategies: (1) NCBI *mito* blast database (Data S1 for details), (2) EMBL database to generate seeds as described above, and (3) using a keyword search of "cytochrome c oxidase 1" to generate seeds (see Data S1 for details).

We benchmarked the reference databases by comparing the overlapping NCBI accessions and taxonomic content of the rCRUX databases against the CRABS (All Markers), MetaCurator (All Markers), ecoPCR (All Markers), and RESCRIPt (MiFish, trnl) databases presented in Jeunen et al. (2023) (Richardson et al., 2020; Robeson et al., 2021). We also compared the rCRUX in silico PCR function to ecoPCR implemented by Jeunen et al. (2023) (see Data S1).

Lastly, to demonstrate the value of more comprehensive reference databases for taxonomic assignment we used a taxonomic cross-validation and novel taxa classification framework implemented through the python tool TAXonomic ClassifieR Evaluation Tool (here in tax credit) (Bokulich et al., 2018). Novel taxa analysis can test the performance of a reference database when assigning taxonomy to an undocumented species. Both cross-validation and novel classification analyses were performed for rCRUX, CRABS, ecoPCR, RESCRIPt, and MetaCurator 12S MiFish reference databases. To prepare for cross-validation analyses, we generated 10-fold randomized cross-validation datasets containing test sets and training sets for each reference database. Test set sequences were removed from the corresponding training set. If a taxonomy in any test set was not present in at least 10 sequences in the corresponding training set, the expected taxonomy label was truncated to the nearest common taxonomic rank observed in the training set. To prepare for novel taxa classification analysis, 10 test and training datasets were made for each database at each of three taxonomic levels (L), from species to family. Training sets had all sequences that matched the taxonomy of the corresponding test sequences at taxonomic level L removed.

Taxonomy was assigned to the test datasets using the scikit-learn naïve-bayes classifier implemented through the qiime2 feature-classifier plugin (v2023.5) with default parameters (Bokulich et al., 2018; Pedregosa et al., 2011). We then used tax credit to calculate the total number of exact matches and overclassifications (lineage is correct but longer than expected; e.g. Species instead of Genus), as well as precision, recall, and F-measure (the harmonic mean of precision and recall), as described in Bokulich et al. (2018). We then plot the F-measure for taxonomic assignments either at the species level (cross-validation) or at the genus level when presented with novel species (novel taxa). F-measure scores were compared pairwise between databases using t-tests with Bonferroni's correction for multiple comparisons, as implemented in the Python statsmodels and SciPy libraries.

We then generated 24 rCRUX reference databases for 20 metabarcoding primer sets (Table 1) and made version-controlled, DOI accessions available on the GitHub to provide comprehensive curated reference databases for a suite of bespoke metabarcoding loci that lack dedicated reference databases.

3 | RESULTS

3.1 | Benchmarking rCRUX against CRUX

We compared the newly implemented rCRUX *get_seeds_local*() to CRUX ecoPCR to benchmark the efficacy of this in silico PCR step. The *get_seeds_local*() in silico *PCR* consistently captured a greater

TABLE 1 Databases generated using rCRUX for this publication.

Primer set	Gene	Target	Accessions	Species	Species after dereplication	Citation
MiFish Universal	12S	Teleosts	125,011	21,082	17,990	Miya et al. (2015)
MiFish Universal – expanded blast db	12S	Teleosts	128,649	21,182	18,039	Miya et al. (2015)
Ford 16S	16S	Teleosts	365,827	95,916	79,352	Ford et al. (2016)
MarVer3	16S	Vertebrates	362,131	92,250	73,344	Valsecchi et al. (2020)
MiDeca	16S	Decapods	221,806	79,195	61,819	Komai et al. (2019)
Taberlet trnl	trnl	Plants	139,885	63,720	21,274	Taberlet et al. (1991); Taberlet et al. (2007)
Fungal ITS	ITS	Fungi	1,371,297	228,874	138,089	White et al. (1990); Ihrmark et al. (2012)
Baker Dlp	D loop	Marine mammals	254,404	3693	3382	Baker et al. (2018)
Ceph18S	18S	Cephalopods	576	206	160	de Jonge et al. (2021)
UCR Plant rbcl	rbcl	Plants	202,127	69,414	10,995	McFrederick and Rehan (2016); Spence et al. (2022)
MiSebastes	Cyt b	Rockfish	45,494	6163	4808	Min et al. (2021)
Gu ITS2 Plants	ITS2	Plants	211,113	64,833	54,713	Gu et al. (2013)
teleo	12S	Ffish	108,038	12,534	8178	Valentini et al. (2016)
Coissac trnl	trnl	Plants	162,309	69,505	10,659	Taberlet et al. (2007)
Kelly 16S	16S	Metazoans	540,356	146,979	63,769	Kelly et al. (2016)
18 s SSU3/SSU4	18S	Eukaryotes	188,294	63,757	16,724	McInnes et al. (2017)
16S V4	16S	Prokaryotes	2,015,938	206,944	73,908	Parada et al. (2016)
18S V4	18S	Eukaryotes	427,829	84,651	41,137	Stoeckle and Hebert (2008)
18S V9	18S	Eukaryotes	952,981	165,349	56,855	Amaral-Zettler et al. (2009)
CO1 embl	CO1	Metazoans	5,073,958	940,096	369,488	Leray et al. (2022)
CO1 ncbi mito	CO1	Metazoans	3,269,089	742,818	323,094	Leray et al. (2022)
CO1 searchterm	CO1	Metazoans	2,969,493	702,128	291,583	Leray et al. (2022)
CO1 combined	CO1	Metazoans	5,413,965	990,286	390,508	Leray et al. (2022)
16S V4 phytoplankton	16S	Prokaryotes	4,939,099	325,816	63,808	Walters et al. (2016)

number of species than CRUX ecoPCR for the MiFish 12S Universal Teleost locus (Figure 2a). *get_seeds_local*() captured 92.6% of species that CRUX ecoPCR captured while also capturing an additional 5796 species (44.7% of total species). Likewise, the rCRUX blast_seeds() step captured a greater number of species than CRUX for the MiFish 12S Universal Teleost locus (Figure 2b). rCRUX captured 99.2% of species that CRUX captured while also capturing an additional 7899 species (37.4% of total species). Together, these results demonstrate the improved user time and database size of rCRUX compared to the original CRUX implementation.

We further compared computational resources utilized by rCRUX and CRUX. The in silico PCR step of rCRUX had a shorter system and user time than CRUX, but required more memory (Table S8). The seed blasting step of rCRUX had a longer system and user time than CRUX, but required less memory (Table S8). Overall, the user time for rCRUX was shorter than CRUX, but rCRUX had a longer system time and required more memory (Table S8).

3.2 | Benchmarking rCRUX against CRABS, MetaCurator, RESCRIPt, and ecoPCR

Benchmarking of rCRUX against previously published reference databases demonstrates that rCRUX outperforms CRABS, MetaCurator, RESCRIPt, and ecoPCR by capturing more species for MiFish (Figure 3), trnl (Figure 4), FITS (Figure 5), and CO1 (Figure 6) loci.

For the MiFish reference comparison, only 38.7% of all species ($n = 9089$) were shared across the five reference databases. Each reference database had unique species that were not shared with any other database (range: 12–4376). rCRUX captured 90.9% ($n = 21{,}148$) of all species observed across the MiFish reference databases. rCRUX uniquely had 18.64% ($n = 4376$) of all species observed (Figure 3). The difference in performance also translated into a higher number of unique haplotypes captured for each species (Figure S14). These results demonstrate rCRUX had higher across- and within-species diversity than the other tested databases.

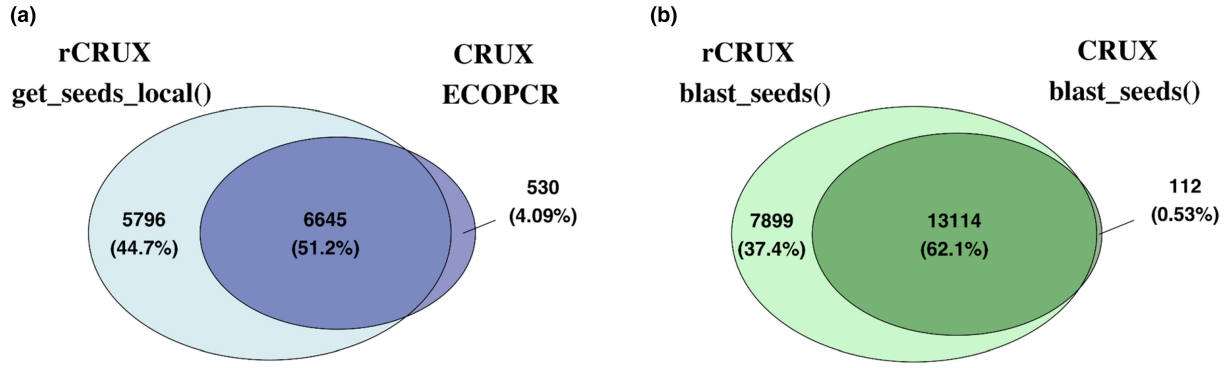

FIGURE 2 Comparison of rCRUX to the original implementation of CRUX. Comparison of number of species captured by (a) rCRUX-implemented *get_seeds_local*() and CRUX-implemented ecoPCR in silico PCR tools and (b) rCRUX- and CRUX-implemented *blast_seeds*() for the MiFish 12S Universal Teleost locus. rCRUX captures the vast majority of species captured by CRUX while also incorporating thousands of additional taxa.

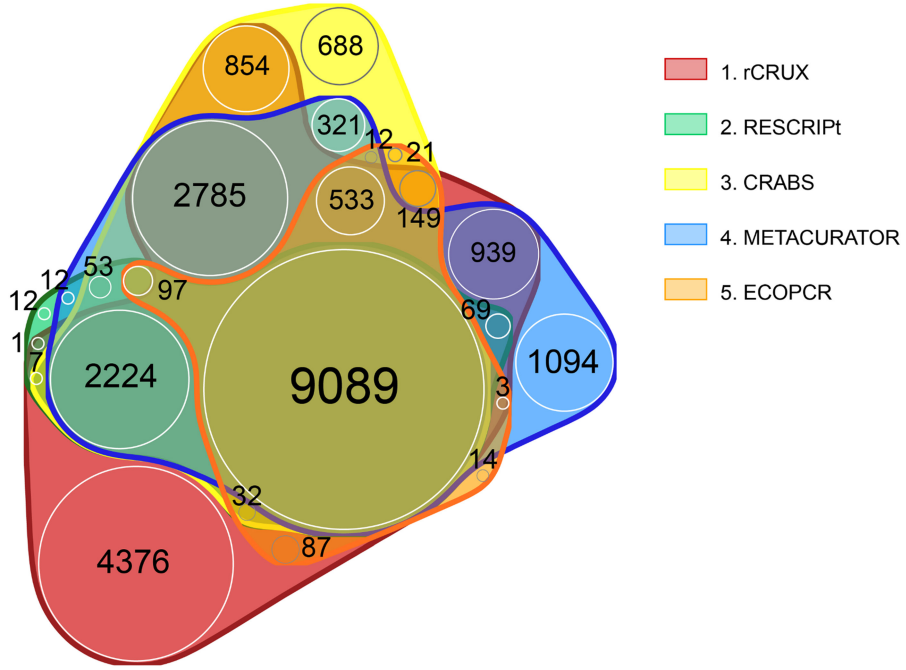

FIGURE 3 MiFISH rCRUX *blast_seeds*() database comparison with CRABS, ecoPCR, MetaCurator, and RESCRIPt databases created by Jeunen et al. (2023).

For the trnl reference database, only 2308 species out of 69,705 species were shared across the five trnl reference databases. Each reference database had unique sequences that were not shared with any other database (range: 1–15,190). rCRUX captured 91.4% ($n = 63,719$) of all species observed across the trnl reference databases. rCRUX uniquely had 21.8% ($n = 15,190$) of all species observed (Figure 4).

For the FITS reference database, only 5.2% of all species ($n = 12,218$) were shared across the 4 reference databases. Each reference database had unique sequences that were not shared with any other database (range: 610–171,358). rCRUX captured 97.2% ($n = 228,873$) of all species observed across the FITS reference databases. rCRUX uniquely had 72.8% ($n = 171,358$) of all species observed (Figure 5).

For the CO1 reference database, only 2.8% of all species ($n = 27,990$) were shared across the 4 reference databases. Each reference database had unique species that were not shared with any other database (range: 4–823,363). rCRUX-combined CO1 database captured 99.6% ($n = 990,286$) of all species observed across the CO1 reference databases. rCRUX-combined CO1 database uniquely had 82.8% ($n = 823,363$) species of all species observed (Figure 6). The three distinct strategies used to generate the rCRUX-combined CO1 database had complementary species (see Data S1).

Limiting the seeds and database generation output comparisons to only Eukaryotic reads had minimal effect on the results (Figures S15–S18). We also note that the rCRUX databases were generated after the other databases; however, they include the majority of species captured by compared methods. Together, these results benchmark

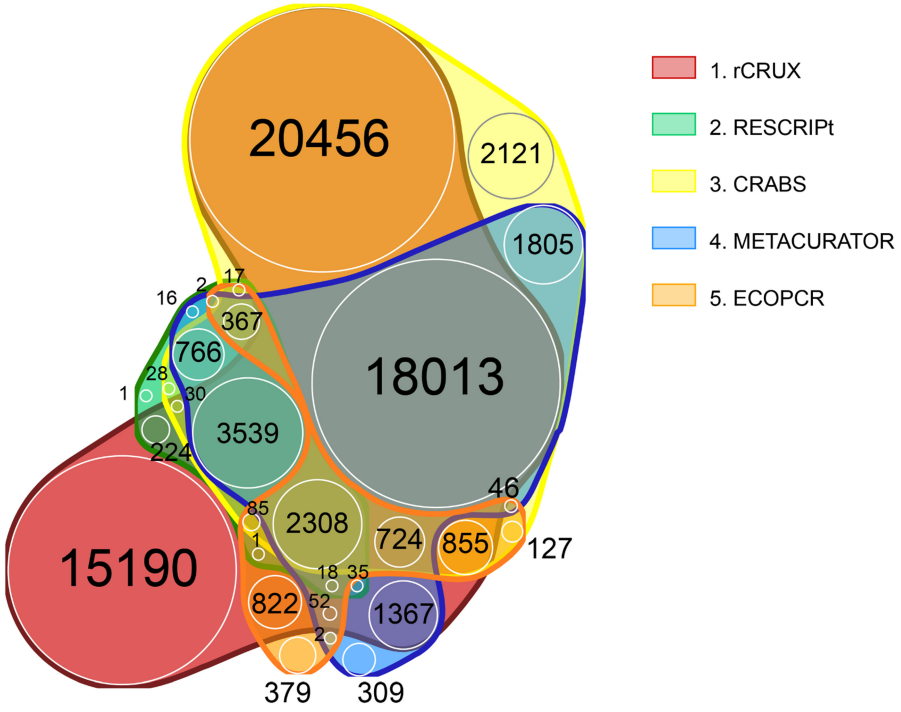

FIGURE 4 trnl rCRUX *blast_seeds*() database comparison with CRABS, ecoPCR, MetaCurator, and RESCRIPt databases created by Jeunen et al. (2023).

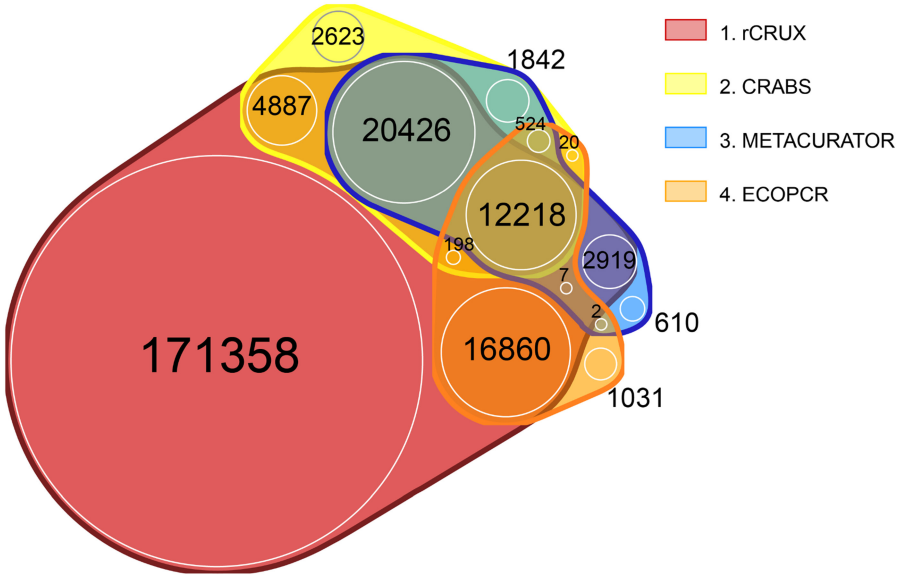

FIGURE 5 FITS rCRUX *blast_seeds*() database comparison with CRABS, ecoPCR, MetaCurator, and RESCRIPt databases created by Jeunen et al. (2023).

rCRUX favorably against CRABS, MetaCurator, ecoPCR, RESCRIPt, and CRUX across a diversity of metabarcoding loci.

3.3 | Cross-validation and novel taxa classification

Cross-validation comparisons of MiFish 12S databases demonstrate that rCRUX had significantly higher average F-measures for species-level assignments (0.605 ± 0.008) than RESCRIPt (0.5 ± 0.008), ecoPCR (0.517 ± 0.009), CRABS (0.0.535 ± 0.011), and MetaCurator (0.515 ± 0.01) reference databases. Similarly, novel classification results demonstrate that rCRUX had significantly higher F-measure for genus-level assignments to species missing from the reference database (0.283 ± 0.006) than RESCRIPt (0.206 ± 0.007), ecoPCR (0.19 ± 0.009), CRABS (0.221 ± 0.007), and MetaCurator (0.222 ± 0.008) reference databases. The rCRUX database was not significantly different in F-measure from the other databases at higher taxonomic levels. Together, these results highlight the

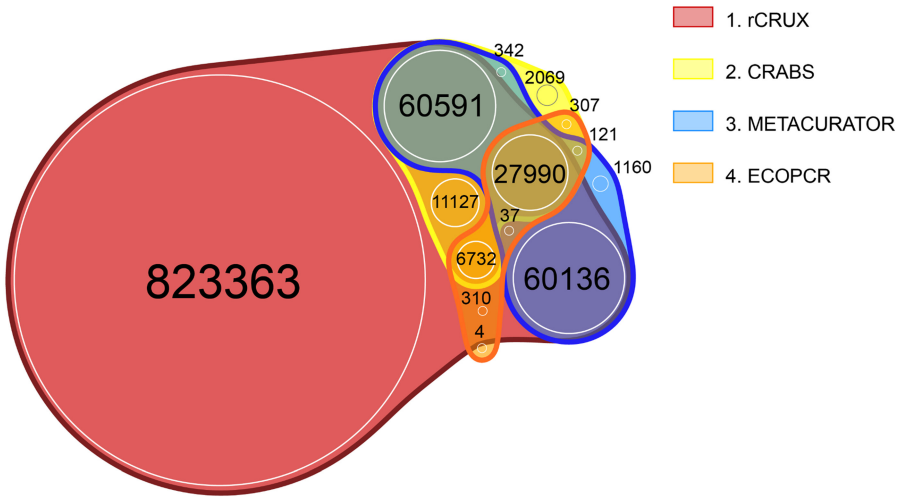

FIGURE 6 CO1 rCRUX *blast_seeds*() database comparison with CRABS, ecoPCR, MetaCurator, and RESCRIPt databases created by Jeunen et al. (2023).

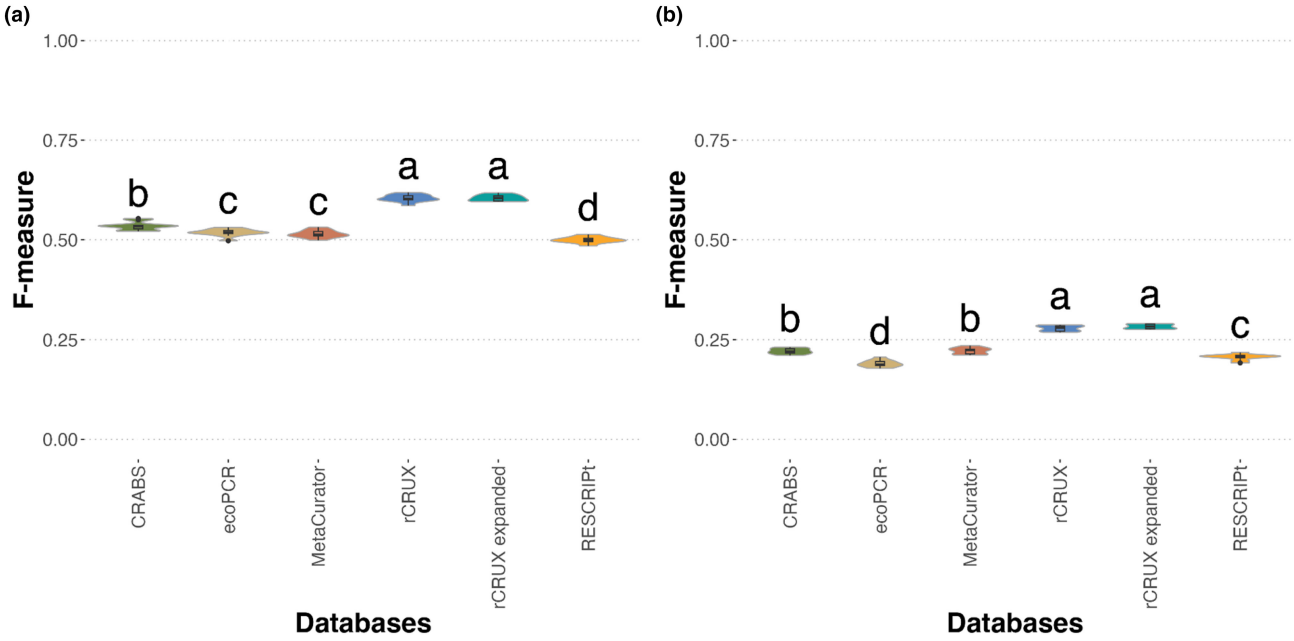

FIGURE 7 Cross-validation and novel taxonomy performance evaluations. rCRUX had significantly higher average F-measure for cross-validation at the species level than RESCRIPt, CRABS, ecoPCR, and MetaCurator 12S reference databases (a). Likewise, rCRUX had significantly higher F-measure for novel species taxonomic assignments at the species level than RESCRIPt, CRABS, ecoPCR, and MetaCurator (b). Violins with different lower-case letters have significantly different means (paired *t*-test, false detection rate-corrected $p < 0.05$).

improved performance of rCRUX for taxonomic assignment compared to other taxonomic reference databases (Figure 7).

3.4 | rCRUX databases

We successfully generated a total of 16 reference databases (Table 1) for a suite of bespoke metabarcoding primer sets. Sizes of these reference databases ranged from 576 to 5,413,965 accessions and 206 (160 unique sequences) to 990,286 (390,508 unique sequences) species.

4 | DISCUSSION

We successfully demonstrate that rCRUX generates comprehensive, curated reference databases for user-defined metabarcoding loci of interest. We benchmarked rCRUX against similar reference database-generating tools, consistently capturing the majority of accessions and species present in those databases and thousands of additional species and accessions not found in CRABS, MetaCurator, ecoPCR, RESCRIPt, and the original implementation of CRUX (Curd et al., 2019; Ficetola et al., 2010; Jeunen et al., 2023; Richardson

et al., 2020; Robeson et al., 2021). We generated 24 reference databases for bespoke metabarcoding loci of interest, including 7 databases for larger universal markers, providing important bioinformatic resources for the broader metabarcoding community. The rCRUX R package presented here provides a valuable tool for the generation and curation of reference databases, enhancing the accuracy, utility, and reproducibility of taxonomic assignment of DNA sequences broadly.

4.1 | Benchmarking rCRUX reference databases

rCRUX-generated reference databases were consistently more comprehensive than other leading databases (Figures 2–6; Figures S15–18). This is partially driven by the efficacy of *get_seeds_local*(), which captured more species and accessions than ecoPCR, serving as a more efficient in silico PCR simulator (Curd et al., 2019; Ficetola et al., 2010) (Figure 1; Figure S12). At the core of *get_seeds_local*() is the NCBI primer blast tool which is a widely used, well-benchmarked, and reproducible tool for the testing of primer sets (Cannon et al., 2016; Hleap et al., 2021; Ye et al., 2012). In addition, we demonstrate that our iterative blasting approach did not impair the efficacy of the *blast_seeds*() step as originally implemented in CRUX, resulting in faster run times and comparably comprehensive reference databases (Figures 2–6; Data S1).

Previous research has highlighted the value of more comprehensive reference databases for improved taxonomic assignment (Curd et al., 2019; Dziedzic et al., 2023; Gold et al., 2021; Jeunen et al., 2023; Keck & Altermatt, 2023; Richardson et al., 2020). Here, we demonstrate through cross-validation and novel classification analysis that the more comprehensive rCRUX databases consistently outperformed the partial RESCIPt, ecoPCR, CRABS, and MetaCurator databases. Thus, the greater diversity and breadth of species and accessions captured in rCRUX-generated reference databases provide an important tool for improving taxonomic classification.

Interestingly, we found that no reference database was completely comprehensive as each generated reference database tested had unique accessions and species (Figures 2–6). These patterns are consistent with previous reference database comparisons and highlight the inherent difficulty of capturing all relevant accessions for a given metabarcoding locus of interest from global public DNA sequence repositories (Curd et al., 2019; Jeunen et al., 2023; Richardson et al., 2020). Although each of the tools compared here shares an underlying strategy for creating comprehensive reference databases for loci of interest, their implementation is different, resulting in distinct subsets of captured reference sequences. These results highlight the challenge of making reproducible reference databases. However, despite these differences across reference database generators, rCRUX captured the vast majority of all species and accessions captured across the tested reference databases, failing to capture at most ~2% of sequences captured by another database. Future reference database-generating efforts may seek to employ multiple distinct generating strategies and combine results to obtain the most comprehensive reference database possible.

Importantly, we demonstrate that rCRUX is nearly perfectly reproducible across subsequent runs of MiFish Teleost rCRUX databases ($n = 10$ runs, coefficient in variation of returned accessions <0.01%, see Data S1 and Table S3), providing high confidence in rCRUX reference databases.

4.2 | FAIR reference database generation

Access to reliable, reproducible, comprehensive, and curated reference databases is critical for improving taxonomic assignment of DNA sequences (Keck & Altermatt, 2023). However, to date, curated, metabarcode-specific reference database-generating tools have not fully adhered to the Findable, Accessible, Interoperable, and Reproducible data management principles (Wilkinson et al., 2016). The generation and curation of reference barcode databases is time- and labor-intensive and requires substantial computational resources and bioinformatic expertise which often limits interoperability and reproducibility across users. Together, this severely limits our ability to generate reference databases quickly and efficiently and limits the number of researchers and scientists who can build the repositories needed to assign taxonomy to DNA sequences (Curd et al., 2019; Shea et al., 2023).

However, recent advances in reference database-generating tools open the door to a broader community of practice in generating reference databases (Jeunen et al., 2023). Given the ubiquity of R users in the molecular biology and ecology fields, rCRUX provides a powerful tool that is straightforward and relatively easy to implement on any computing environment. By providing researchers with an accessible reference database-generating tool, we hope to alleviate the difficulties of building and updating reference databases. Thus, the ability to generate user-specific barcode reference databases will enhance metabarcoding, eDNA, microbiome, and DNA classification research efforts broadly.

One of the motivations for making simple and easy to install, update, and maintain reference database-generating tools was to increase access to these resources across the molecular biology and ecology fields. However, limitations in the utility of reference database-generating software still remain, particularly the scale of computational resources needed. Although the iterative blast implementation of rCRUX reduces computational needs compared to the previous iterations of CRUX, the rCRUX databases presented here still relied on high-performance computing (each run was given a maximum allotment of 250GB of RAM, 40 cores, and 1 week of run time on the University of Vermont – Vermont Advanced Computing Core, RRID:SCR_017762, and 16 of the 24 databases used a fraction of those resources). However, efforts to generate larger reference databases with greater number of available reference barcodes (e.g. CO1, microbial 16S V4, phytoplankton 16S V4, 18S V4, and 18S V9) were challenging, requiring a distinct implementation strategy for

get_seeds_local() (see Data S1) because of a lack of available computational resources to meet the scale of available sequences. Researchers often lack access to computational resources, particularly in developing nations where biodiversity is often the highest and the need for DNA-based taxonomic classification is the greatest (Asase et al., 2022; Barber et al., 2014; Johnson et al., 2022). As cloud computing and high-performance computing resources continue to become increasingly cost-effective, we hope rCRUX and similar reference database-generating tools will become more accessible (Thompson & Thielen, 2023). We note that rCRUX can be successfully implemented on a personal laptop with a 1 TB hard drive, 16 GB of RAM, and 8 cores, given parameters and markers that require fewer computational resources. Importantly, we designed rCRUX to be highly scalable and easy to install through R in any compute environment, allowing for adoption in future cloud computing efforts in which rCRUX could be served to a wide audience like NCBI primerTools or BLAST.

However, to specifically help address issues of access to comprehensive reference databases, we provided 16 reference databases for commonly used or emerging metabarcoding loci and 8 for larger universal loci. These databases will be updated and curated at least annually with a unique DOI, providing important genetic resources to the broader DNA sequencing community including those that lack access to such computational infrastructure. Future efforts will be made to grow the list of available databases as future loci become available and widely adopted (Version Controlled Reference Databases available at https://github.com/CalCOFI/rCRUX).

Lastly, we demonstrate the reproducibility of rCRUX, allowing for users to make identical databases from the same starting parameters and sequence repositories (Table S1). Providing a reproducible and stable tool for the generation of barcode reference databases ensures high-quality genetic resources that adhere to FAIR principles.

4.3 | Broader applications of rCRUX

The most immediate application of rCRUX is the generation of reference databases to support taxonomic assignment of metabarcoding from high-throughput sequencing. However, the utility of rCRUX allows for reference databases to be generated on any blast-formatted database, directly supporting improved taxonomic assignment of a broad range of DNA sequence applications. For example, this allows for the curation of reference barcodes from full- or partial-length mitogenomes (see Data S1), supporting long-read sequencing taxonomic assignment applications (Johri et al., 2019; Ramon-Laca et al. 2022). In addition, rCRUX can be used for building nuclear DNA-based reference databases from whole-genome and transcriptome sequences. Such efforts could be used to develop population-genetic and eRNA-specific reference databases for a diversity of biomonitoring applications (Adams et al., 2019; Greco et al., 2022; McKinney et al., 2022; Sigsgaard et al., 2020; Simon et al., 2019).

Curation of reference databases is important to ensure accurate taxonomic assignment (Bourret et al., 2023; Fontes et al., 2021). rCRUX performs curation of reference sequences by selecting specific marker genes, filtering sequences to relevant lengths, and collapsing accessions with identical sequences to lowest common ancestor analogous to generating BINs (Fontes et al., 2021). Here, we applied a unique sequence dereplication approach to retain as much informative information as possible for taxonomic classification. For example, most salmonids in the NE Pacific have a single base pair difference across the MiFish 12S Teleost locus (Gold et al., 2021; Shelton et al., 2023). Applying a BIN approach, even at 99% clustering, would collapse all of these sequences to genus level (1 base pair/186 bp fragment length, 99.46% sequence similarity). However, the location of these base pair differences is consistent across dozens to hundreds of salmonid sequences, strongly suggesting that these alleles are fixed across each species within this locus and thus can be used for species identification (Gold et al., 2021; Shelton et al., 2023). However, removing duplicate sequences from reference databases can improve the accuracy of taxonomic assignments, particularly when using Bayesian approaches that are biased by the number of reference sequences for a given species (Curd et al., 2019). Thus, the inclusion of this dereplication step provides a valuable curation tool for reference databases.

However, the curation efforts implemented within rCRUX are limited, and future efforts could implement additional quality assurance and quality control measures. For example, recent efforts have demonstrated that local reference databases often perform better than global ones, thus developing tools to subset reference databases to specific geographic regions of interest would provide a valuable curation step (Blackman et al., 2023; Bourret et al., 2023; Gold et al., 2021). However, implementing such an approach would require accurate global repositories of species distributions and detailed inventories of regional biodiversity which are often not readily available (Beck et al., 2014; Costello & Berghe, 2006). In addition, an important curation step that would greatly benefit rCRUX reference databases is the removal of incorrect reference sequences in global sequence repositories (Cheng et al. 2023; Keck et al., 2022). The development and application of automated and efficient tools to remove erroneous sequences in global public DNA sequence repositories would greatly benefit reference database curation efforts broadly. This is particularly an issue for tools like rCRUX which, unlike manually curated databases, rely solely on the accuracy of such global sequence repositories (Leray et al., 2019; Meiklejohn et al., 2019). Specifically, there is a need for the development and application of phylogenetic approaches to readily identify problematic reference databases from public repositories for species assignments (see Jeunen et al., 2023). Although the curation tools presented here provide an important advancement for reference databases, future efforts to enhance reference database curation will greatly improve trust and reliability in taxonomic assignments broadly (Fontes et al., 2021; Keck et al., 2022).

Importantly, we demonstrate that reference database-generating tools like rCRUX paired with taxonomic classification evaluation tools like tax credit provide a valuable resource for designing, validating, and comparing potential metabarcoding targets for a specific research question (Bokulich et al., 2018; Edgar, 2018; Hleap et al., 2021; Mathon et al., 2021). For example, if researchers are deciding which fish metabarcoding loci to use for a given project and have a known target species list (Jerde et al., 2021), rCRUX can be used to conduct an in silico comparison of primer set efficacy. This can be accomplished by first generating rCRUX reference databases for each potential locus, then performing cross-validation of each reference database with tax credit, and then simply cross-referencing the taxonomic resolution of each database against the target taxa list (Gold et al., 2021). The comparison and curation tools provided here allow for the direct comparison of multiple reference databases, serving as a resource for evaluating the relative performance of reference databases on taxonomic assignment. Previous research has demonstrated the value of these kinds of in silico validation and benchmarking approaches for improved taxonomic classification of DNA sequences (Curd et al., 2019; Edgar, 2018; Gold et al., 2021; Jeunen et al., 2023). Thus, rCRUX provides a simple, cost-effective tool for informing scientists and resource managers on the efficacy of taxonomic assignment during the design and development of biomolecular monitoring efforts.

4.4 | Complimentary packages to rCRUX

The rCRUX package provides important novel utility to the wide suite of reference database-managing packages available. We note that such packages can be used in concert to achieve improved reference database management and efficacy. For example, the refdb R package provides a suite of complementary tools that can be used to merge BOLD and GenBank databases which could provide improved blast-formatted nucleotide databases (Keck et al., 2022). In addition, refdb provides a suite of tools to visualize and summarize output reference databases (Keck et al., 2022). Similar utilities to merge GenBank, EMBL, and BOLD databases are available through CRABS, MARES, RESCRIPt, and BAGS and can be used to generate a more comprehensive starting blastDB database, particularly for CO1 genes (Arranz et al., 2020; Fontes et al., 2021; Jeunen et al., 2023; Robeson et al., 2021). In addition, CRABS and MARES also provide tools to output datasets in a greater diversity of formats for use in additional taxonomic classifiers beyond Anacapa and Qiime2 (Bolyen et al., 2019; Curd et al., 2019). The comprehensiveness of rCRUX databases can also be leveraged and used as input into GAPeDNA to better conduct gap analysis for a given locus and target taxa in a specific study region (Marques et al., 2021). Similarly, researchers submitting Omics data to Ocean Biogeographic Information System and complying with Darwin Core standards can use the World Register of Marine Species taxonomy and can readily convert rCRUX taxonomy using the worrms R package (Berry et al., 2021; Chamberlain, 2019; Costello et al., 2013; Grassle & Stocks, 1999; Meyer et al., 2023). Thus, rCRUX provides an important complementary tool to the suite of available reference database management software.

5 | CONCLUSION

Ultimately, rCRUX provides a powerful, reproducible, and reliable tool for the generation of comprehensive and curated reference databases for any genetic loci of interest. By providing users with a simple and accessible reference database-generating R package, rCRUX will ease taxonomic classification as well as validation and benchmarking for bespoke and novel primer sets. Improved ease of implementation over previous iterations of CRUX and a suite of 24 publicly available version-controlled reference databases provide important genetic resources without the need for significant computational resources, facilitating access and adoption of high-quality reference databases and database-generating tools to a broad range of users.

AUTHOR CONTRIBUTIONS

(i) The conception or design of the study: Emily E. Curd, Luna Gal, Ramon Gallego, Shaun Nielsen, and Zachary Gold; (ii) the acquisition, analysis, or interpretation of the data: Emily E. Curd, Luna Gal, Ramon Gallego, Katherine Silliman, Shaun Nielsen, and Zachary Gold; and (iii) the writing of the manuscript: Emily E. Curd, Luna Gal, Ramon Gallego, Katherine Silliman, Shaun Nielsen, and Zachary Gold.

ACKNOWLEDGMENTS

The research reported in this publication was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103449. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIGMS or NIH. Support for the development of rCRUX was provided by the CalCOFI program. This study is a PMEL contribution 5512. This work was supported by the NOAA 'Omics program. This study is a contribution to NOAA Research and Development Database (NRDD) project ID 24755. This work benefited from the amazing input of many including Lenore Pipes, Sarah Stinson, Gaurav Kandlikar, Maura Palacios Mejia, Ryan Kelly, and Kim Parsons. We want to especially acknowledge the late, great Jesse Gomer, coding extraordinaire, rCRUX co-conspirator, and dear friend who tragically passed away before rCRUX was completed. None of this would be possible without Jesse's endless inspiration, creativity, ingenuity, and generosity.

FUNDING INFORMATION

Funding was provided by NOAA Omics program and the CalCOFI program. In addition, research reported in this publication was supported by an Institutional Development Award from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103449.

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interests to report.

DATA AVAILABILITY STATEMENT

The rCRUX package and 24 generated reference databases are available at https://github.com/CalCOFI/rCRUX. Data and code for analysis and figures will be uploaded upon acceptance of the manuscript.

ORCID

Zachary Gold https://orcid.org/0000-0003-0490-7630

REFERENCES

Adams, C. I., Knapp, M., Gemmell, N. J., Jeunen, G. J., Bunce, M., Lamare, M. D., & Taylor, H. R. (2019). Beyond biodiversity: Can environmental DNA (eDNA) cut it as a population genetics tool? *Genes*, *10*(3), 192.

Ahmed, M., Back, M. A., Prior, T., Karssen, G., Lawson, R., Adams, I., & Sapp, M. (2019). Metabarcoding of soil nematodes: The importance of taxonomic coverage and availability of reference sequences in choosing suitable marker (s). *Metabarcoding and Metagenomics*, *3*, e36408.

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.

Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W., & Huse, S. M. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PloS One*, *4*(7), e6372.

Arranz, V., Pearman, W. S., Aguirre, J. D., & Liggins, L. (2020). MARES, a replicable pipeline and curated reference database for marine eukaryote metabarcoding. *Scientific Data*, *7*(1), 209.

Asase, A., Mzumara-Gawa, T. I., Owino, J. O., Peterson, A. T., & Saupe, E. (2022). Replacing "parachute science" with "global science" in ecology and conservation biology. *Conservation Science and Practice*, *4*(5), e517.

Baker, C. S., Steel, D., Nieukirk, S., & Klinck, H. (2018). Environmental DNA (eDNA) from the wake of the whales: Droplet digital PCR for detection and species identification. *Frontiers in Marine Science*, *5*, 133.

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: the Journal of Biological Databases and Curation*, *2020*, baz155. https://doi.org/10.1093/database/baz155

Barber, P. H., Ablan-Lagman, M. C. A., Berlinck, R. G., Cahyani, D., Crandall, E. D., Ravago-Gotanco, R., Juinio-Meñez, M. A., Mahardika, I. G., Shanker, K., Starger, C. J., Toha, A. H. A., Anggoro, A. W., & Willette, D. A. (2014). Advancing biodiversity research in developing countries: The need for changing paradigms. *Bulletin of Marine Science*, *90*(1), 187–210.

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15.

Beng, K. C., & Corlett, R. T. (2020). Applications of environmental DNA (eDNA) in ecology and conservation: Opportunities, challenges and prospects. *Biodiversity and Conservation*, *29*, 2089–2121.

Bengtsson-Palme, J., Hartmann, M., Eriksson, K. M., Pal, C., Thorell, K., Larsson, D. G. J., & Nilsson, R. H. (2015). METAXA2: Improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Molecular Ecology Resources*, *15*(6), 1403–1414.

Berry, O., Jarman, S., Bissett, A., Hope, M., Paeper, C., Bessey, C., Schwartz, M. K., Hale, J., & Bunce, M. (2021). Making environmental DNA (eDNA) biodiversity records globally accessible. *Environmental DNA*, *3*(4), 699–705.

Bik, H. M. (2021). Just keep it simple? Benchmarking the accuracy of taxonomy assignment software in metabarcoding studies. *Molecular Ecology Resources*, *21*, 2187–2189.

Blackman, R. C., Walser, J. C., Rüber, L., Brantschen, J., Villalba, S., Brodersen, J., Seehausen, O., & Altermatt, F. (2023). General principles for assignments of communities from eDNA: Open versus closed taxonomic databases. *Environmental DNA*, *5*(2), 326–342.

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., & Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 1–17.

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.

Bourret, A., Nozères, C., Parent, E., & Parent, G. J. (2023). Maximizing the reliability and the number of species assignments in metabarcoding studies using a curated regional library and a public repository. *Metabarcoding and Metagenomics*, *7*, e98539.

Bucklin, A., Lindeque, P. K., Rodriguez-Ezpeleta, N., Albaina, A., & Lehtiniemi, M. (2016). Metabarcoding of marine zooplankton: Prospects, progress and pitfalls. *Journal of Plankton Research*, *38*(3), 393–400.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 1–9.

Cannon, M. V., Hester, J., Shalkhauser, A., Chan, E. R., Logue, K., Small, S. T., & Serre, D. (2016). In silico assessment of primers for eDNA studies using PrimerTree and application to characterize the biodiversity surrounding the Cuyahoga River. *Scientific Reports*, *6*(1), 1–11.

Chamberlain, S. (2019). Worrms: World register of marine species (WoRMS) client. R Package Version 0.4.0.

Cheng, Z., Li, Q., Deng, J., Liu, Q., & Huang, X. (2023). The devil is in the details: Problems in DNA barcoding practices indicated by systematic evaluation of insect barcodes. *Frontiers in Ecology and Evolution*, *11*, 1149839.

Costa, F. O., & Carvalho, G. R. (2007). The barcode of life initiative: Synopsis and prospective societal impacts of DNA barcoding of fish. *Genomics, Society and Policy*, *3*(2), 29.

Costello, M. J., & Berghe, E. V. (2006). 'Ocean biodiversity informatics': A new era in marine biology research and management. *Marine Ecology Progress Series*, *316*, 203–214.

Costello, M. J., Bouchet, P., Boxshall, G., Fauchald, K., Gordon, D., Hoeksema, B. W., Poore, G. C., van Soest, R. W., Stöhr, S., Walter, T. C., & Vanhoorne, B. (2013). Global coordination and standardisation in marine biodiversity through the world register of marine species (WoRMS) and related databases. *PLoS One*, *8*(1), e51629.

Cummins, C., Ahamed, A., Aslam, R., Burgin, J., Devraj, R., Edbali, O., Gupta, D., Harrison, P. W., Haseeb, M., & Holt, S. (2022). The european nucleotide archive in 2021. *Nucleic Acids Research*, *50*(D1), D106–D110.

Curd, E. E., Gold, Z., Kandlikar, G. S., Gomer, J., Ogden, M., O'Connell, T., Pipes, L., Schweizer, T. M., Rabichow, L., & Lin, M. (2019). Anacapa toolkit: An environmental DNA toolkit for processing multilocus metabarcode datasets. *Methods in Ecology and Evolution*, *10*(9), 1469–1475.

Darwin Tree of Life Project Consortium. (2022). Sequence locally, think globally: The Darwin tree of life project. *Proceedings of the*

National Academy of Sciences of the United States of America, *119*(4), e2115642118.

de Jonge, D. S., Merten, V., Bayer, T., Puebla, O., Reusch, T. B., & Hoving, H. J. T. (2021). A novel metabarcoding primer pair for environmental DNA analysis of Cephalopoda (Mollusca) targeting the nuclear 18S rRNA region. *Royal Society Open Science*, *8*(2), 201388.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., & De Vere, N. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895.

Dziedzic, E., Sidlauskas, B., Cronn, R., Anthony, J., Cornwell, T., Friesen, T. A., Konstantinidis, P., Penaluna, B. E., Stein, S., & Levi, T. (2023). Creating, curating and evaluating a mitogenomic reference database to improve regional species identification using environmental DNA. *Molecular Ecology Resources*, *23*, 1880–1904.

Edgar, R. C. (2018). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ*, *6*, e4652.

Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., Taberlet, P., & Pompanon, F. (2010). An in silico approach for the evaluation of DNA barcodes. *BMC Genomics*, *11*(1), 1–10.

Fontes, J. T., Vieira, P. E., Ekrem, T., Soares, P., & Costa, F. O. (2021). BAGS: An automated Barcode, Audit & Grade System for DNA barcode reference libraries.

Ford, M. J., Hempelmann, J., Hanson, M. B., Ayres, K. L., Baird, R. W., Emmons, C. K., Lundin, J. I., Schorr, G. S., Wasser, S. K., & Park, L. K. (2016). Estimation of a killer whale (*Orcinus orca*) population's diet using sequencing analysis of DNA from feces. *PloS One*, *11*(1), e0144956.

Gao, C. H., Yu, G., & Cai, P. (2021). ggVennDiagram: An intuitive, easy-to-use, and highly customizable R Package to generate venn diagram. *Frontiers in Genetics*, *12*, 706907. https://doi.org/10.3389/fgene.2021.706907

Gold, Z., Curd, E. E., Goodwin, K. D., Choi, E. S., Frable, B. W., Thompson, A. R., Walker, H. J., Jr., Burton, R. S., Kacev, D., & Martz, L. D. (2021). Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem. *Molecular Ecology Resources*, *21*(7), 2546–2564.

Grassle, J. F., & Stocks, K. I. (1999). A global ocean biogeographic information system (OBIS) for the census of marine life. *Oceanography*, *12*(3), 12–14.

Greco, M., Lejzerowicz, F., Reo, E., Caruso, A., Maccotta, A., Coccioni, R., Pawlowski, J., & Frontalini, F. (2022). Environmental RNA outperforms eDNA metabarcoding in assessing impact of marine pollution: A chromium-spiked mesocosm test. *Chemosphere*, *298*, 134239.

Gu, W., Song, J., Cao, Y., Sun, Q., Yao, H., Wu, Q., Chao, J., Zhou, J., Xue, W., & Duan, J. (2013). Application of the ITS2 region for barcoding medicinal plants of Selaginellaceae in Pteridophyta. *PLoS One*, *8*(6), e67818.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., & Decelle, J. (2012). The protist ribosomal reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), D597–D604.

Hebert, P. D., Ratnasingham, S., & De Waard, J. R. (2003). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(suppl_1), S96–S99.

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

Hemsley, J., Qin, J., & Bratt, S. E. (2020). Data to knowledge in action: A longitudinal analysis of GenBank metadata. *Proceedings of the Association for Information Science and Technology*, *57*(1), e253.

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

Ihrmark, K., Bödeker, I. T., Cruz-Martinez, K., Friberg, H., Kubartova, A., Schenck, J., Strid, Y., Stenlid, J., Brandström-Durling, M., Clemmensen, K. E., & Lindahl, B. D. (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.

Jerde, C. L., Mahon, A. R., Campbell, T., McElroy, M. E., Pin, K., Childress, J. N., Armstrong, M. N., Zehnpfennig, J. R., Kelson, S. J., Koning, A. A., Ngor, P. B., Nuon, V., So, N., Chandra, S., & Hogan, Z. S. (2021). Are genetic reference libraries sufficient for environmental DNA metabarcoding of Mekong River basin fish? *Water*, *13*(13), 1767.

Jeunen, G. J., Dowle, E., Edgecombe, J., von Ammon, U., Gemmell, N. J., & Cross, H. (2023). Crabs—A software program to generate curated reference databases for metabarcoding sequencing data. *Molecular Ecology Resources*, *23*(3), 725–738.

Johnson, A., Saypanya, S., Hansel, T., & Rao, M. (2022). More than an academic exercise: Structuring international partnerships to build research and professional capacity for conservation impact. *Conservation Science and Practice*, *4*(5), e539.

Johri, S., Solanki, J., Cantu, V. A., Fellows, S. R., Edwards, R. A., Moreno, I., Vyas, A., & Dinsdale, E. A. (2019). Genome skimming' with the MinION hand-held sequencer identifies CITES-listed shark species in India's exports market. *Scientific Reports*, *9*(1), 1–13.

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

Katz, K. S., Shutov, O., Lapoint, R., Kimelman, M., Brister, J. R., & O'Sullivan, C. (2021). STAT: A fast, scalable, MinHash-based k-mer tool to assess sequence read archive next generation sequence submissions. *Genome Biology*, *22*, 270.

Keck, F., Couton, M., & Altermatt, F. (2022). Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. *Molecular Ecology Resources*, *23*(4), 742–755.

Keck, F., & Altermatt, F. (2023). Management of DNA reference libraries for barcoding and metabarcoding studies with the R package refdb. *Molecular Ecology Resources*, *23*(2), 511–518.

Kelly, R. P., O'Donnell, J. L., Lowell, N. C., Shelton, A. O., Samhouri, J. F., Hennessey, S. M., Feist, B. E., & Williams, G. D. (2016). Genetic signatures of ecological diversity along an urbanization gradient. *PeerJ*, *4*, e2444.

Kõljalg, U., Larsson, K. H., Abarenkov, K., Nilsson, R. H., Alexander, I. J., Eberhardt, U., Erland, S., Høiland, K., Kjøller, R., Larsson, E., Pennanen, T., Sen, R., Taylor, A. F. S., Tedersoo, L., Vrålstad, T., & Ursing, B. M. (2005). UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *The New Phytologist*, *166*, 1063–1068. https://doi.org/10.1111/j.1469-8137.2005.01376.x

Komai, T., Gotoh, R. O., Sado, T., & Miya, M. (2019). Development of a new set of PCR primers for eDNA metabarcoding decapod crustaceans. *Metabarcoding and Metagenomics*, *3*, e33835.

Leray, M., Knowlton, N., Ho, S. L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences*, *116*(45), 22651–22656.

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

Manor, O., Dai, C. L., Kornilov, S. A., Smith, B., Price, N. D., Lovejoy, J. C., Gibbons, S. M., & Magis, A. T. (2020). Health and disease markers correlate with gut microbiome composition across thousands of people. *Nature Communications*, *11*(1), 5206.

Marques, V., Milhau, T., Albouy, C., Dejean, T., Manel, S., Mouillot, D., & Juhel, J. B. (2021). GAPeDNA: Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. *Diversity and Distributions*, *27*(10), 1880–1892.

Mathon, L., Valentini, A., Guérin, P. E., Normandeau, E., Noel, C., Lionnet, C., Boulanger, E., Thuiller, W., Bernatchez, L., & Mouillot, D. (2021). Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. *Molecular Ecology Resources*, *21*(7), 2565–2579.

McFrederick, Q. S., & Rehan, S. M. (2016). Characterization of pollen and bacterial community composition in brood provisions of a small carpenter bee. *Molecular Ecology*, *25*(10), 2302–2311.

McInnes, J. C., Alderman, R., Deagle, B. E., Lea, M. A., Raymond, B., & Jarman, S. N. (2017). Optimised scat collection protocols for dietary DNA metabarcoding in vertebrates. *Methods in Ecology and Evolution*, *8*(2), 192–202.

McKinney, G. J., Barry, P. D., Pascal, C., Seeb, J. E., Seeb, L. W., & McPhee, M. V. (2022). A new genotyping-in-thousands-by-sequencing single nucleotide polymorphism panel for mixed-stock analysis of chum Salmon from coastal Western Alaska. *North American Journal of Fisheries Management*, *42*(5), 1134–1143.

McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, *8*(4), e61217.

Meglécz, E. (2023). COInr and mkCOInr: Building and customizing a non-redundant barcoding reference database from BOLD and NCBI using a semi-automated pipeline. *Molecular Ecology Resources*, *23*(4), 933–945.

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.

Meyer, R., Appeltans, W., Duncan, W. D., Dimitrova, M., Gan, Y. M., Stjernegaard Jeppesen, T., Mungall, C., Paul, D. L., Provoost, P., Robertson, T., Schriml, L., Suominen, S., Walls, R., Sweetlove, M., Ung, V., Van de Putte, A., Wallis, E., Wieczorek, J., & Buttigieg, P. L. (2023). Aligning standards communities for omics biodiversity data: Sustainable darwin core-MIxS interoperability. *Biodiversity Data Journal*, *11*, e112420. https://doi.org/10.3897/BDJ.11.e112420

Min, M. A., Barber, P. H., & Gold, Z. (2021). MiSebastes: An eDNA metabarcoding primer set for rockfishes (genus Sebastes). *Conservation Genetics Resources*, *13*, 447–456.

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H., & Kondoh, M. (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.

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

Parada, A. E., Needham, D. M., & Fuhrman, J. A. (2016). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology*, *18*(5), 1403–1414.

Pauvert, C. (2020). Psadd: Additions to phyloseq package for microbiome analysis. R Packag, Version 0.1, 2.

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

Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology*, *27*(2), 313–338.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ramon-Laca, A., Gallego, R., & Nichols, K. (2023). Affordable de novo generation of fish mitogenomes using amplification-free enrichment of mitochondrial DNA and deep sequencing of long fragments. *Molecular Ecology Resources*, *23*(4), 818–832.

Ratnasingham, S., & Hebert, P. D. (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. (2021). RESCRIPt: Reproducible sequence taxonomy reference database management. *PLoS Computational Biology*, *17*(11), e1009581.

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

Shea, M. M., Kuppermann, J., Rogers, M. P., Smith, D. S., Edwards, P., & Boehm, A. B. (2023). Systematic review of marine environmental DNA metabarcoding studies: Toward best practices for data usability and accessibility. *PeerJ*, *11*, e14993.

Shelton, A. O., Gold, Z. J., Jensen, A. J., D′ Agnese, E., Andruszkiewicz Allan, E., Van Cise, A., Gallego, R., Ramón-Laca, A., Garber-Yonts, M., Parsons, K., & Kelly, R. P. (2023). Toward quantitative metabarcoding. *Ecology*, *104*(2), e3906.

Sherrill-Mix, S. (2019). Taxonomizr: Functions to work with NCBI accessions and taxonomy. *See https://CRAN. R-project. org/package= taxonomizr*.

Siddall, M. E., Fontanella, F. M., Watson, S. C., Kvist, S., & Erséus, C. (2009). Barcoding bamboozled by bacteria: Convergence to metazoan mitochondrial primer targets by marine microbes. *Systematic Biology*, *58*(4), 445–451.

Sigsgaard, E. E., Jensen, M. R., Winkelmann, I. E., Møller, P. R., Hansen, M. M., & Thomsen, P. F. (2020). Population-level inferences from environmental DNA—Current status and future perspectives. *Evolutionary Applications*, *13*(2), 245–262.

Simon, H. Y., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell*, *178*(4), 779–794.

Soon, W. W., Hariharan, M., & Snyder, M. P. (2013). High-throughput sequencing for biology and medicine. *Molecular Systems Biology*, *9*(1), 640.

Spence, A. R., Wilson Rankin, E. E., & Tingley, M. W. (2022). DNA metabarcoding reveals broadly overlapping diets in three sympatric north American hummingbirds. *The Auk*, *139*(1), ukab074.

Stoeckle, M. Y., & Hebert, P. D. (2008). Barcode of life. *Scientific American*, *299*(4), 82–89.

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). DNA sequencing. In *Environmental DNA: For biodiversity research and monitoring*. Oxford University Press.

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

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. https://doi.org/10.1093/nar/gkl938

Takahashi, M., Saccò, M., Kestel, J. H., Nester, G., Campbell, M. A., Van Der Heyde, M., Heydenrych, M. J., Juszkiewicz, D. J., Nevill, P., & Dawkins, K. L. (2023). Aquatic environmental DNA: A review of

the macro-organismal biomonitoring revolution. *Science of the Total Environment*, *873*, 162322.

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., Prill, R. J., Tripathi, A., Gibbons, S. M., & Ackermann, G. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, *551*(7681), 457–463.

Thompson, L. R., & Thielen, P. (2023). Decoding dissolved information: Environmental DNA sequencing at global scale to monitor a changing ocean. *Current Opinion in Biotechnology*, *81*, 102936.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., & Gaboriaud, C. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, *25*(4), 929–942.

Valsecchi, E., Bylemans, J., Goodman, S. J., Lombardi, R., Carr, I., Castellano, L., Galimberti, A., & Galli, P. (2020). Novel universal primers for metabarcoding environmental DNA surveys of marine mammals and other marine vertebrates. *Environmental DNA*, *2*(4), 460–476.

Walters, W., Hyde, E. R., Berg-Lyons, D., Ackermann, G., Humphrey, G., Parada, A., Gilbert, J. A., Jansson, J. K., Caporaso, J. G., Fuhrman, J. A., & Apprill, A. (2016). Improved bacterial 16S rRNA gene (V4 and V4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. *Msystems*, *1*(1), e00009-15.

White, T. J., Bruns, T., Lee, S. J. W. T., & 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.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., & Bouwman, J. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9.

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*, 1–11.

Zeileis, A., & Grothendieck, G. (2005). Zoo: An s3 class and methods for indexed totally ordered observations.

Zhang, J., Pei, N., Mi, X., & Zhang, M. J. (2017). Package 'phylotools'. dimension 12.

Zhu, T., Sato, Y., Sado, T., Miya, M., & Iwasaki, W. (2023). MitoFish, MitoAnnotator, and MiFish pipeline: Updates in 10 years. *Molecular Biology and Evolution*, *40*(3), msad035.

Zorz, J., Li, C., Chakraborty, A., Gittins, D. A., Surcon, T., Morrison, N., Bennett, R., MacDonald, A., & Hubert, C. R. J. (2023). SituSeq: An offline protocol for rapid and remote nanopore 16S rRNA amplicon sequence analysis. *ISME Communications*, *3*(1), 33.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Curd, E. E., Gal, L., Gallego, R., Silliman, K., Nielsen, S., & Gold, Z. (2024). rCRUX: A rapid and versatile tool for generating metabarcoding reference libraries in R. *Environmental DNA*, *6*, e489. https://doi.org/10.1002/edn3.489