Received: 11 December 2018 | Revised: 5 April 2019 | Accepted: 8 May 2019

DOI: 10.1111/2041-210X.13214

APPLICATION

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Anacapa Toolkit: An environmental DNA toolkit for processing multilocus metabarcode datasets

Emily E. Curd[1] | Zack Gold[1] | Gaurav S. Kandlikar[1] | Jesse Gomer[1] | Max Ogden[2] | Taylor O'Connell[1] | Lenore Pipes[3] | Teia M. Schweizer[4] | Laura Rabichow[1] | Meixi Lin[1] | Baochen Shi[5] | Paul H. Barber[1] | Nathan Kraft[1] | Robert Wayne[1] | Rachel S. Meyer[1]

[1]Department of Ecology and Evolutionary Biology, University of California, Los Angeles

[2]Code for Science and Society, Portland, Oregon

[3]Department of Ecology and Evolutionary Biology, University of California, Berkeley

[4]Department of Biology, Colorado State University, Fort Collins

[5]Department of Molecular and Medical Pharmacology, University of California, Los Angeles

Correspondence
Emily Curd
Email: eecurd@g.ucla.edu

Funding information
OISE, Grant/Award Number: 1243541 ; NSF-DEB, Grant/Award Number: 1644641; University of California President's Research Catalyst Award, Grant/Award Number: CA-16-376437; NSF-DGE, Grant/Award Number: 1650604; NSF-GRFP, Grant/Award Number: 2015204395; University of California; Gordon and Betty Moore Foundation, Grant/Award Number: 6864

Handling Editor: Douglas Yu

Abstract

1. Environmental DNA (eDNA) metabarcoding is a promising method to monitor species and community diversity that is rapid, affordable and non-invasive. The longstanding needs of the eDNA community are modular informatics tools, comprehensive and customizable reference databases, flexibility across high-throughput sequencing platforms, fast multilocus metabarcode processing and accurate taxonomic assignment. Improvements in bioinformatics tools make addressing each of these demands within a single toolkit a reality.
2. The new modular metabarcode sequence toolkit *Anacapa* (https://github.com/limey-bean/Anacapa/) addresses the above needs, allowing users to build comprehensive reference databases and assign taxonomy to raw multilocus metabarcode sequence data. A novel aspect of *Anacapa* is its database building module, "Creating Reference libraries Using eXisting tools" (*CRUX*), which generates comprehensive reference databases for specific user-defined metabarcoding loci. The *Quality Control and ASV Parsing* module sorts and processes multiple metabarcoding loci and processes merged, unmerged and unpaired reads maximizing recovered diversity. *DADA2* then detects amplicon sequence variants (ASVs) and the *Anacapa Classifier* module aligns these ASVs to *CRUX*-generated reference databases using *Bowtie2*. Lastly, taxonomy is assigned to ASVs with confidence scores using a Bayesian Lowest Common Ancestor (*BLCA*) method. The *Anacapa Toolkit* also includes an R package, *ranacapa*, for automated results exploration through standard biodiversity statistical analysis.
3. Benchmarking tests verify that the *Anacapa Toolkit* effectively and efficiently generates comprehensive reference databases that capture taxonomic diversity, and can assign taxonomy to both MiSeq and HiSeq-length sequence data. We demonstrate the value of the *Anacapa Toolkit* in assigning taxonomy to seawater eDNA samples collected in southern California.

Zack Gold, Gaurav S. Kandlikar and Jesse Gomer are co-equal second authors.

Robert Wayne and Rachel S. Meyer are co-equal senior authors.

4. The *Anacapa Toolkit* improves the functionality of eDNA and streamlines biodiversity assessment and management by generating metabarcode specific databases, processing multilocus data, retaining a larger proportion of sequencing reads and expanding non-traditional eDNA targets. All the components of the *Anacapa Toolkit* are open and available in a virtual container to ease installation.

KEYWORDS

Bayesian methods, biodiversity, community ecology, environmental DNA, metabarcoding, molecular methods , multilocus metabarcoding processing, sequence data

1 | INTRODUCTION

Rapid and inexpensive biodiversity monitoring tools are critical for maintaining healthy ecosystems and for effective species conservation (Deiner et al., 2017). Environmental DNA (eDNA) is a promising non-invasive approach for biodiversity monitoring that is increasingly used in ecology and conservation research. Although eDNA metabarcoding is a powerful, rapid and cost-effective approach to survey taxa in terrestrial and aquatic ecosystems (Bohmann et al., 2014; Deiner et al., 2017; Kelly, Port, Yamahara, & Crowder, 2014; Taberlet, Coissac, Hajibabaei, & Rieseberg, 2012), three key bioinformatics challenges in sequence processing and taxonomic assignment limit the accuracy and reliability of eDNA approaches.

First, to capture a broad representation of taxonomic diversity, many eDNA studies simultaneously sequence multiple loci per sample (e.g. Stat et al., 2017). However, few metabarcode pipelines are explicitly designed to process multilocus high-throughput sequencing data (but see Arulandhu et al., 2017). As such, researchers must sort and process multiple eDNA metabarcodes independently, substantially increasing computation time with each additional metabarcode.

A second challenge for eDNA metabarcode processing is the lack of robust, locus-specific reference databases (Deiner et al., 2017). Curated databases for select metabarcoding loci offer validated solutions for certain commonly used universal metabarcodes (e.g. UNITE for Fungal *ITS* sequences, Kõljalg et al., 2013), but such curated databases are unlikely to exist for all loci used in metabarcoding studies, especially as the number of target metabarcodes grows. Custom, user-generated databases are a promising solution, but current approaches can be problematic. For example, generating reference databases through in silico PCR will miss reference sequences that do not contain primer recognition sites, a feature of many Sanger-based sequences (Boyer et al., 2016; Ficetola et al., 2010). Furthermore, methods that rely on keyword searches to generate reference databases are sensitive to inaccurate metadata (Machida, Leray, Ho, & Knowlton, 2017) and are susceptible to retrieving sequences that lack the target metabarcode locus. Together, these issues highlight a need for a more comprehensive reference databases to enhance eDNA metabarcoding taxonomic assignment.

A third challenge of existing metabarcode pipelines is that they frequently discard large portions of sequence data, including reads that can be valuable for assigning taxonomy. For example, some pipelines discard unmerged sequences entirely, or only use partial sequence data where full-length alignment with reference metabarcodes is not possible (Port et al., 2015) potentially causing selection bias against certain taxa (Deagle, Jarman, Coissac, Pompanon, & Taberlet, 2014). To attempt to solve this issue, some pipelines employ nested least common ancestor assignments to non-contiguous sequences (See Huson & Weber, 2013), but the lack of joint assignment limits the achievable taxonomic resolution. Few pipelines are specifically designed to handle unmerged paired data (e.g. Bengtsson-Palme et al., 2015), relying heavily on BLAST to assign taxonomy. However, both these approaches usually limit the number of BLAST hits returned, which presents an additional problem because BLAST will prioritize the sequence order within the reference database over the best alignment for taxonomic assignment (Shah, Pop, Nute, & Warnow, 2019). Furthermore, improved handling of unmerged paired sequences would enable researchers to more readily leverage new high-throughput sequencing platforms (e.g. Illumina NovaSeq and 10X) and barcoding loci of longer length (Deiner et al., 2017).

To help resolve these challenges, we developed the *Anacapa Toolkit*, a bioinformatic pipeline with modules for: (1) creating custom reference databases; (2) executing quality control and multilocus read parsing; (3) generating taxonomic assignments for all quality reads produced by HiSeq and MiSeq Illumina platforms; and (4) interactively visualizing taxonomy tables from the *Anacapa Toolkit* using the R package *ranacapa* as described in (Kandlikar et al., 2018).

2 | THE ANACAPA TOOLKIT

The *Anacapa Toolkit* combines components of leading bioinformatics software with custom methods (Figure 1). The first module, Creating Reference libraries Using eXisting tools (*CRUX*), generates custom reference databases. The second module performs raw sequence quality control and employs *DADA2* (Callahan et al., 2016) to infer Amplicon Sequence Variants (ASVs). The third module assigns taxonomy using *Bowtie2* and the Bayesian Lowest Common Ancestor algorithm (*BLCA*; Gao, Lin, Revanna, & Dong, 2017).

All components of the *Anacapa Toolkit* are openly available (https://github.com/limey-bean/Anacapa) and archived in DRYAD

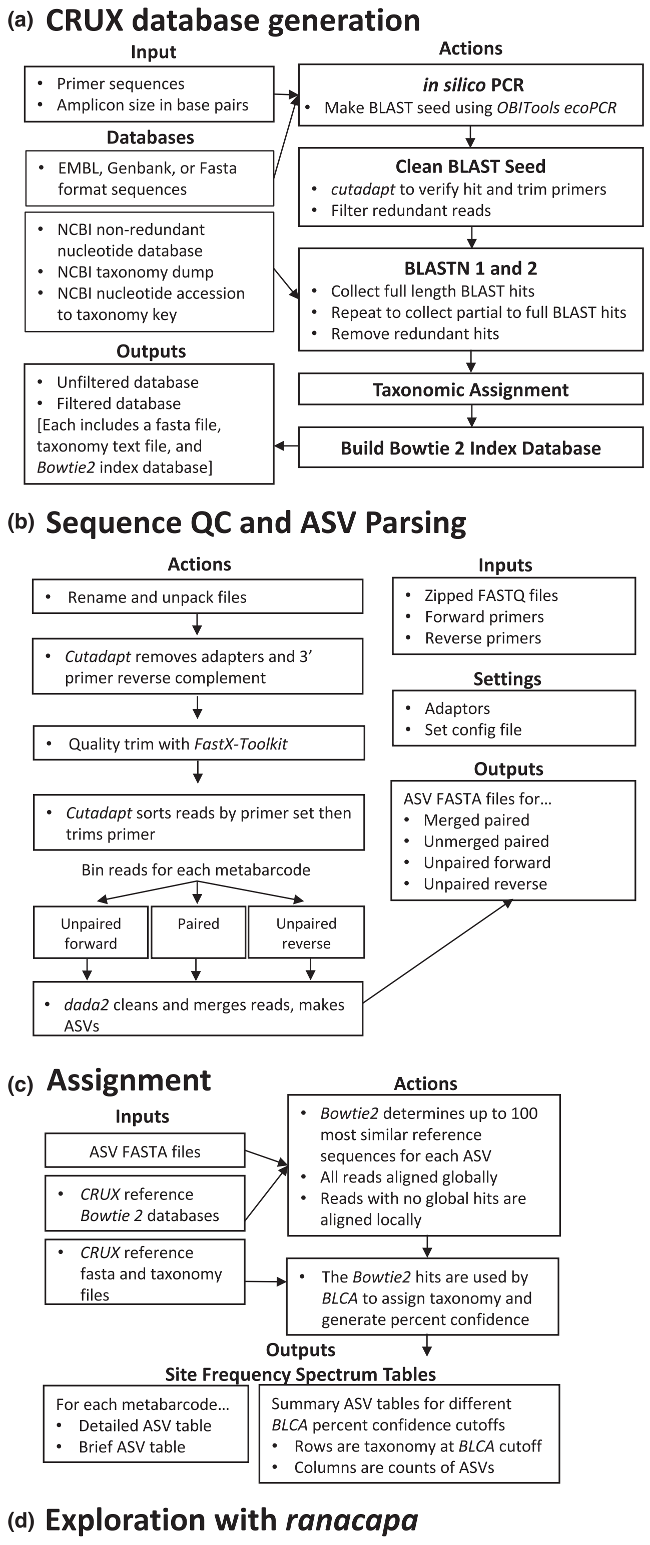

(a) CRUX database generation
Input
• Primer sequences
• Amplicon size in base pairs
Databases
• EMBL, Genbank, or Fasta format sequences
• NCBI non-redundant nucleotide database
• NCBI taxonomy dump
• NCBI nucleotide accession to taxonomy key
Outputs
• Unfiltered database
• Filtered database
[Each includes a fasta file, taxonomy text file, and Bowtie2 index database]
Actions
in silico PCR
• Make BLAST seed using OBITools ecoPCR
Clean BLAST Seed
• cutadapt to verify hit and trim primers
• Filter redundant reads
BLASTN 1 and 2
• Collect full length BLAST hits
• Repeat to collect partial to full BLAST hits
• Remove redundant hits
Taxonomic Assignment
Build Bowtie 2 Index Database
(b) Sequence QC and ASV Parsing
Actions
• Rename and unpack files
• Cutadapt removes adapters and 3' primer reverse complement
• Quality trim with FastX-Toolkit
• Cutadapt sorts reads by primer set then trims primer
Bin reads for each metabarcode
Unpaired forward
Paired
Unpaired reverse
• dada2 cleans and merges reads, makes ASVs
Inputs
• Zipped FASTQ files
• Forward primers
• Reverse primers
Settings
• Adaptors
• Set config file
Outputs
ASV FASTA files for...
• Merged paired
• Unmerged paired
• Unpaired forward
• Unpaired reverse
(c) Assignment
Inputs
ASV FASTA files
• CRUX reference Bowtie 2 databases
• CRUX reference fasta and taxonomy files
Actions
• Bowtie2 determines up to 100 most similar reference sequences for each ASV
• All reads aligned globally
• Reads with no global hits are aligned locally
• The Bowtie2 hits are used by BLCA to assign taxonomy and generate percent confidence
Outputs
Site Frequency Spectrum Tables
For each metabarcode...
• Detailed ASV table
• Brief ASV table
Summary ASV tables for different BLCA percent confidence cutoffs
• Rows are taxonomy at BLCA cutoff
• Columns are counts of ASVs
(d) Exploration with ranacapa

FIGURE 1 Flowchart of the *Anacapa Toolkit*

(https://doi.org/10.5061/dryad.mf0126f). The *Anacapa Toolkit* and several *CRUX*-generated reference databases are available in virtual containers developed with Code for Science and Society (Ogden, 2018). A detailed list of all parameters and their functions is presented in Appendix S1.

2.1 | *CRUX*: Creating reference libraries using existing tools

The *Anacapa Toolkit*'s first module, *CRUX* (Figure 1a; Appendix S1.1), constructs custom reference databases for user-defined primers by querying public databases. *CRUX* first generates metabarcode-specific seed databases by running in silico PCR (Ficetola et al., 2010) on the EMBL standard nucleotide database (Stoesser et al., 2002). To increase the breadth of reference sequences and capture sequences without barcode primers, *CRUX* then uses *blastn* (Camacho et al., 2009) to query the seed databases against the NCBI non-redundant nucleotide database (Pruitt, Tatusova, & Maglott, 2005). *CRUX* de-replicates the *blastn* hits by retaining only the longest version of each sequence and retrieves taxonomy using *Entrez-qiime* (Baker, 2016). For each primer set, *CRUX* generates an "unfiltered" database that contains all accessions and taxonomic paths, a paired "filtered" database that excludes accessions with ambiguities in the taxonomic paths, and a *Bowtie2*-formatted index library (Langmead & Salzberg, 2012).

2.2 | Sequence quality control and ASV parsing

The *Anacapa Toolkit*'s *Quality Control and ASV Parsing* module (Figure 1b; Appendix S1.2) conducts standard DNA sequence quality control and generates ASVs. It uses *cutadapt* (Martin, 2011) and *FastX-toolkit* (Gordon & Hannon, 2010) to trim user-defined primers and adapters and low-quality bases from raw FASTQ files from Illumina sequencing platforms. Next, this module uses *cutadapt* to separate reads from multiple loci within each sample. A custom Python script sorts locus-specific reads into three categories: paired-end reads, forward-only reads and reverse-only reads. These reads are then processed separately through *DADA2* (Callahan et al., 2016) to denoise, dereplicate, merge paired reads and remove chimeric sequences. This step returns ASV FASTA files and ASV count summary tables for four read types: merged paired-end reads, unmerged paired-end reads (filtered based on length and overlap criteria; Appendix S1.2), forward-only reads and reverse-only reads. These files are the inputs for the *Anacapa Classifier* module for assigning taxonomy.

2.3 | Taxonomic assignment: *ANACAPA classifier* assigns taxonomy with *BOWTIe2* and *BLCA*

The *Anacapa Classifier* module (Figure 1c; Appendix S1.3) assigns taxonomy to ASVs using *Bowtie2* and a modified version of *BLCA* (Gao et al., 2017). We verified that our modification to *BLCA* (namely, accepting *Bowtie2*-formatted SAM files rather than BLAST output files) does not influence taxonomic assignment (Appendix S4). In the first step of the *Anacapa Classifier*, *Bowtie2* queries ASVs against metabarcode-specific *CRUX* generated reference databases returning up to 100 alignments per ASV. The module uses *Bowtie2*'s "very-sensitive" preset to ensure high-quality alignments. The *Bowtie2* outputs are then processed with *Bowtie2-BLCA*, using multiple sequence alignment to probabilistically determine taxonomic identity by selecting the lowest common ancestor from the multiple weighted *Bowtie2* hits for each ASV. This module returns both detailed and brief reports of taxonomic assignment, and eight sets of taxonomy tables based on varying bootstrap confidence cutoffs (40–100).

3 | BENCHMARKING THE *ANACAPA TOOLKIT*

To benchmark the performance of the first three modules of the *Anacapa Toolkit*, we performed a series of quantitative tests of these modules on various metabarcodes and sequencing read types. Detailed methods and results from these comparisons are presented in Appendices S3–S5. Briefly, to compare *CRUX*-generated databases to previously published reference databases, we conducted pairwise comparisons examining the total number of metabarcode specific sequences in the reference databases and the phylogenetic breadth of these databases for specific metabarcoding markers (Appendix S3). We found that *CRUX*-generated databases capture more metagenomic sequences and greater taxonomic diversity than published reference databases for three common metabarcoding loci: *CO1*, *12S* and Fungal *ITS* metabarcode loci.

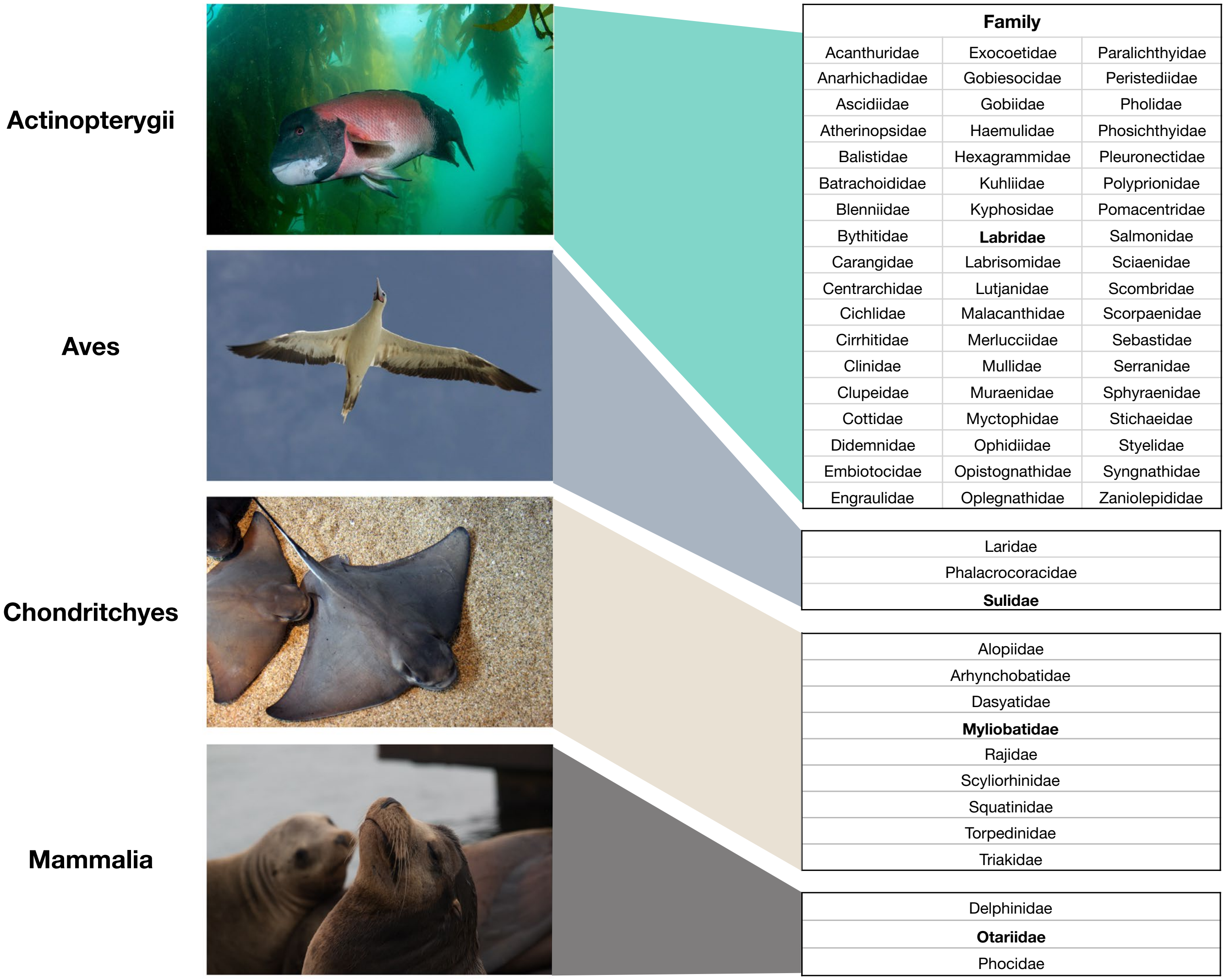
To compare the performance of the *Anacapa Classifier*, we conducted Cross-Validation by Identity using published reference datasets and K-fold (10-fold) *CO1* database comparisons following the methods of (Edgar, 2018; see Appendix S4 for detailed methods and results). Results showed that the *Anacapa Classifier* consistently generated high-accuracy taxonomic assignments comparable to published classifiers. We caution that all classifiers tested had high species-level misclassification and over classification rates (Appendices S3 and S6). We also explored the consequences of varying bootstrap confidence cutoff on assigned taxonomy, and found that the optimal value for the bootstrap confidence cutoff score varied across metabarcoding loci (Appendix S4). Finally, we verified that the *Anacapa Toolkit Quality Control and ASV Parsing* module and *Anacapa Classifier* module can both process longer (e.g. MiSeq) and shorter (e.g. HiSeq) DNA metabarcoding sequences (Appendix S5), expanding the utility of the *Anacapa* pipeline in comparison to existing methods.

4 | CASE STUDY: USING THE *ANACAPA TOOLKIT* TO ASSIGN TAXONOMY TO FIELD COLLECTED EDNA SAMPLES

To test the *Anacapa Toolkit* on field-collected eDNA metabarcoding datasets, we processed 30 seawater samples from kelp forests across the Southern California Channel Islands, including Anacapa Island. Seawater samples were amplified using *12S* (Miya et al., 2015) and *CO1* (Leray et al., 2013) metabarcodes (see Appendix S6 for laboratory preparation and data analysis; Table S2.2). For the *12S* metabarcode, the *CRUX* module was critical for assigning taxonomy because there are no published reference databases for this locus that include the full breadth of amplifiable clades beyond fish taxa (Sato, Miya, Fukunaga, Sado, & Iwasaki, 2018). Sequence data from these samples are available in NCBI (SRA accession SRP140860). Totalled across seawater samples, we generated 15,745,317 paired-end sequencing reads. Of these, 11,866,904 sequences reads were *12S* and 3,878,413 sequences reads were *COI* resulting in 6,876 ASVs and 6,287 ASVs, respectively, after filtering out singletons (Appendix S6). For both loci, we found that 99.5% were merged read pairs and <1% were unmerged paired, forward or reverse only reads. The *Anacapa Toolkit*'s taxonomic assignments indicate that these ASVs matched to 21 eukaryotic phyla that could be further delimited within 49 classes, 295 families, 414 genera and 533 species. Taxa identified included many of interest for natural resource managers including species of special concern (e.g. Basking shark, *Cetorhinus maximus*, and Giant black sea bass, *Stereolepis gigas*) and species that are the subject of focused monitoring efforts (California sheephead, *Semicossyphus pulcher*, and Ochre star, *Pisaster giganteus*) (Figure 2; Table S6.2 andS6.3). These results highlight the ability of eDNA to detect a wide breadth of marine life and its utility for biodiversity monitoring. A detailed and interactive summary of these seawater samples is available as the demo dataset of the *ranacapa* module (https://gauravsk.shinyapps.io/ranacapa/).

5 | CONCLUSION

Biodiversity monitoring initiatives are increasingly using eDNA to inventory communities using multilocus metabarcoding. However, the lack of accurate and easily customizable bioinformatic pipelines is limiting the broader application of eDNA approaches. The *Anacapa Toolkit* provides enhanced functionality for eDNA projects and can be used for other common applications such as gut content analysis (Leray et al., 2013), autonomous reef monitoring structures (Ransome et al., 2017), and microbiomes (Bokulich et al., 2018). Importantly, the *Anacapa Toolkit* is modular and its parameters are easily modifiable, making it easily adapted to user specific needs in several important ways. First, *CRUX* reference databases are compatible with alternative classifiers (Bokulich et al., 2018), and users can append their own reference sequences to *CRUX* databases as needed. Second, the *Quality Control and ASV Parsing* module is designed to process pooled metabarcoding libraries and automatically sort them by barcode and sample. The resulting output files (with the exception of unmerged paired reads) can be analysed by most classifiers. Third, the *Anacapa Bowtie2-BLCA Classifier* can be applied to any high-throughput sequencing data, and process paired and unpaired reads. The robustness of *CRUX*-generated reference databases and the flexibility of the *Anacapa Toolkit* enables studies with a variety of metabarcoding loci to efficiently and transparently assign taxonomy, facilitating a diversity of eDNA approaches, ranging from basic ecology to biodiversity management and conservation.



Family		
Acanthuridae	Exocoetidae	Paralichthyidae
Anarhichadidae	Gobiesocidae	Peristediidae
Ascidiidae	Gobiidae	Pholidae
Atherinopsidae	Haemulidae	Phosichthyidae
Balistidae	Hexagrammidae	Pleuronectidae
Batrachoididae	Kuhliidae	Polyprionidae
Blenniidae	Kyphosidae	Pomacentridae
Bythitidae	**Labridae**	Salmonidae
Carangidae	Labrisomidae	Sciaenidae
Centrarchidae	Lutjanidae	Scombridae
Cichlidae	Malacanthidae	Scorpaenidae
Cirrhitidae	Merlucciidae	Sebastidae
Clinidae	Mullidae	Serranidae
Clupeidae	Muraenidae	Sphyraenidae
Cottidae	Myctophidae	Stichaeidae
Didemnidae	Ophidiidae	Styelidae
Embiotocidae	Opistognathidae	Syngnathidae
Engraulidae	Oplegnathidae	Zaniolepididae

Laridae
Phalacrocoracidae
Sulidae

Alopiidae
Arhynchobatidae
Dasyatidae
Myliobatidae
Rajidae
Scyliorhinidae
Squatinidae
Torpedinidae
Triakidae

Delphinidae
Otariidae
Phocidae

FIGURE 2 Taxonomic assignments from California environmental samples. Highlights the Anacapa Island kelp forest vertebrate families identified from the 12S metabarcodes. Families in bold are featured in the photographs

ACKNOWLEDGEMENTS

The *Anacapa Toolkit* is a product of the CALeDNA community science program (www.ucedna.com). We thank R. Turba, D.R. Ramos, E.W. Rankin, S. Stinson and S. Shirazi for beta-testing, B. Shapiro and R. Nielsen, M. Dawson for guidance, and J.S. Guswa for editing the manuscript. We thank R. D'Auria for helping us implement *Anacapa* on the UCLA Hoffman2 cluster, and S. Chadalapaka for assistance with the MERCED cluster. We thank D. Kushner and J. Sprague for sample collection. We thank Shaun Wilkinson and an anonymous reviewer for providing insightful feedback on the manuscript. This work was funded by the University of California President's Research Catalyst Award (CA-16-376437), NSF-DEB 1644641 (NK), NSF-DGE 1650604 (GK), NSF-GRFP 2015204395 (ZG) and OISE 1243541 (ZG). The Gordon and Betty Moore Foundation (6864) supported the Dat in the Lab project that containerized *Anacapa*.

AUTHORS' CONTRIBUTION

E.E.C. coordinated toolkit development and benchmarking. B.S., E.E.C., G.S.K., J.G., R.S.M., R.W. and Z.G. designed *Anacapa*. B.S., E.E.C., G.K., J.G., M.O. and Z.G. wrote *Anacapa*. E.E.C. and M.O. built the Singularity container. E.E.C., G.K., J.G., L.P., M.L., R.S.M., R.W., T.O. and Z.G. designed or conducted benchmarking studies. E.E.C., L.R., T.M.S. and Z.G. generated the California eDNA libraries. E.E.C., G.K. and Z.G. wrote the manuscript with help from L.P., L.R., M.L., P.B., R.S.M., R.W., and T.O., T.M.S., N.K., P.B. and R.W. provided resources to support the work.

DATA ACCESSIBILITY STATEMENT

CRUX databases are available in the Dryad Digital Repository (https://doi.org/10.5061/dryad.mf0126f). All Software is available in Zenodo (https://doi.org/10.5281/zenodo.1464285, https://

doi.org/10.5281/zenodo.3064152, and https://doi.org/10.5281/zenodo.3064612, https://doi.org/10.5281/zenodo.2602180). The data container is also available at https://doi.org/10.6071/M31H29. Sequence data are available in the NCBI Sequence Read Archive (accession SRP140860). All components of the *Anacapa Toolkit* are openly available (https://github.com/limey-bean/Anacapa, https://github.com/limey-bean/CRUX_Creating-Reference-libraries-Using-eXisting-tools, https://github.com/datproject/anacapa-container, and https://github.com/gauravsk/ranacapa).

ORCID

Emily E. Curd https://orcid.org/0000-0003-0336-6852

Gaurav S. Kandlikar https://orcid.org/0000-0003-3043-6780

REFERENCES

Arulandhu, A. J., Staats, M., Hagelaar, R., Voorhuijzen, M. M., Prins, T. W., Scholtens, I., … Kok, E. (2017). Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. *GigaScience*, *6*(10), 1–18. https://doi.org/10.1093/gigascience/gix080

Baker, C. (2016). bakerccm/entrez_qiime: entrez_qiime v2.0. https://doi.org/10.5281/zenodo.159607

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. https://doi.org/10.1111/1755-0998.12399

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., … de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, *29*(6), 358–367. https://doi.org/10.1016/j.tree.2014.04.003

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., … Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 90. https://doi.org/10.1186/s40168-018-0470-z

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.

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. https://doi.org/10.1038/nmeth.3869

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. *Biology Letters*, *10*(9), https://doi.org/10.1098/rsbl.2014.0562

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

Edgar, R. C. (2018). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ*, *6*, e4652. https://doi.org/10.7717/peerj.4652

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

Gao, X., Lin, H., Revanna, K., & Dong, Q. (2017). A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. *BMC Bioinformatics*, *18*(1), 247. https://doi.org/10.1186/s12859-017-1670-4

Gordon, A., & Hannon, G. J. (2010). Fastx-toolkit. *FASTQ/A Short-Reads Preprocessing Tools (Unpublished)*. Http://Hannonlab. Cshl., Edu/Fastx_toolkit, 5.

Huson, D. H., & Weber, N. (2013). Microbial community analysis using MEGAN. *Methods in Enzymology*, *531*, 465–485. https://doi.org/10.1016/B978-0-12-407863-5.00021-6

Kandlikar, G. S., Gold, Z. J., Cowen, M. C., Meyer, R. S., Freise, A. C., Kraft, N. J. B., & Curd, E. E. (2018). ranacapa: An r package and Shiny web app to explore environmental DNA data with exploratory statistics and interactive visualizations. *F1000Research*, *7*, 1734. https://doi.org/10.12688/f1000research.16680.1

Kelly, R. P., Port, J. A., Yamahara, K. M., & Crowder, L. B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *PLoS ONE*, *9*(1), e86175. https://doi.org/10.1371/journal.pone.0086175.

Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F. S., Bahram, M., … Larsson, K.-H. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, *22*(21), 5271–5277. https://doi.org/10.1111/mec.12481.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., … 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), 34. https://doi.org/10.1186/1742-9994-10-34

Machida, R. J., Leray, M., Ho, S.-L., & Knowlton, N. (2017). Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Scientific Data*, *4*, 170027. https://doi.org/10.1038/sdata.2017.27

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet. Journal*, *17*(1), 10–https://doi.org/10.14806/ej.17.1.200

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., … 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

Ogden, M. (2018). CALeDNA Anacapa/CRUX Dat Container (Linux/HPC) (Version v8). UC Merced Dash. Retrieved fromhttps://doi.org/10.6071/M31H29.

Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., … Kelly, R. P. (2015). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*,*25*(2), 527–541.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *33*(suppl_1), D501–D504.

Ransome, E., Geller, J. B., Timmers, M., Leray, M., Mahardini, A., Sembiring, A., … Meyer, C. P. (2017). The importance of standardization for biodiversity comparisons: A case study using autonomous reef monitoring structures (ARMS) and metabarcoding to measure cryptic diversity on Mo'orea coral reefs, French Polynesia. *PLoS ONE*, *12*(4), e0175066. https://doi.org/10.1371/journal.pone.0175066

Sato, Y., Miya, M., Fukunaga, T., Sado, T., & Iwasaki, W. (2018). MitoFish and MiFish pipeline: a mitochondrial genome database of fish with an analysis pipeline for environmental DNA metabarcoding. *Molecular Biology and Evolution*, *35*(6), 1553–1555. https://doi.org/10.1093/molbev/msy074

Shah, N., Pop, M., Nute, M. G., & Warnow, T. (2019). Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. *Bioinformatics*, *35*(9), 1613–1614. https://doi.org/10.1093/bioinformatics/bty833

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., … 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

Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., … Lombard, V. (2002). The EMBL nucleotide sequence database. *Nucleic Acids Research*, *30*(1), 21–26. https://doi.org/10.1093/nar/30.1.21

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. *Molecular Ecology*, *21*(8), 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x.

SUPPORTING INFORMATION

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

How to cite this article: Curd EE, Gold Z, Kandlikar GS, et al. *Anacapa Toolkit*: An environmental DNA toolkit for processing multilocus metabarcode datasets. *Methods Ecol Evol*. 2019;10:1469–1475. https://doi.org/10.1111/2041-210X.13214