**Title Page:**

**Anacapa: an environmental DNA toolkit for processing multilocus metabarcode datasets**

**Emily Curd (eecurd@g.ucla.edu), Zack Gold (zackgold@ucla.edu)\*, Gaurav Kandlikar (gkandlikar@ucla.edu)\*, Jesse Gomer (jessegomer@gmail.com)\*, Max Ogden (max@maxogden.com), Lenore Pipes (lpipes@berkeley.edu), Teia Schweizer (teiaschweizer@ucla.edu), Laura Rabichow (rabichl@ucla.edu), Meixi Lin (mexilin@ucla.edu), Baochen Shi (biosbc@gmail.com), Paul Barber (paulbarber@ucla.edu), Nathan Kraft, (nkraft@ucla.edu), Robert Wayne (rwayne@ucla.edu), Rachel S. Meyer (rsmeyer@ucla.edu)**

* Corresponding authors = Emily Curd and Rachel Meyer
* \* equal contributions

**Abstract (≤250 words) [currently 247 words]**

Environmental DNA (eDNA) metabarcoding with next generation sequencing (NGS) holds promise as a noninvasive approach to monitor species and communities. Longstanding needs of the eDNA community are for flexible informatic tools, comprehensive and customizable reference databases, flexibility across NGS platforms, fast and parallelized multilocus metabarcode processing, and precise taxonomic assignment for which they can assess accuracy. To address these needs, we developed Anacapa, a flexible metabarcode sequence toolkit, as part of an initiative to broaden the exploration of eDNA from soils, sediments, and water and evaluate value for biodiversity management. With Anacapa, users create their own reference databases that we demonstrate are more comprehensive than major curated databases. We demonstrate that Anacapa can reliably assign taxonomy to both the longer sequences generated through MiSeq sequencing, as well as to the higher-throughput but shorter HiSeq-generated sequences. Anacapa uses Dada2 to dereplicate, remove chimeras, and detect amplicon sequence variants (ASVs). ASVs are then matched to reference datasets using Bowtie2, followed by Bayesian Least Common Ancestor (BLCA) algorithm that assigns confidence for each step in the taxonomic path. Anacapa's performance matches or exceeds that of other taxonomy assignment methods, and we show this both with mock community data and with eDNA from California terrestrial and marine samples. The toolkit also includes an R package, *ranacapa*, for automated results exploration through standard biodiversity statistical analysis. Anacapa software and source code are freely available and it also comes in a Singularity containerized version, can be accessed here: http://www.ucedna.com/software/

**Introduction**

A promising approach for assessing and monitoring terrestrial, aquatic, and marine ecosystems is the application of environmental DNA (eDNA). This technique relies on the fact that organisms leave trace amounts of dissociated DNA that accumulate in the environment (Taberlet *et al.* 2012). In a relatively small water or soil samples the DNA of archaea, bacteria, viruses, algae, fungi, plants, and animals can be detected and selectively amplified through the technique of metabarcoding (Deiner *et al.* 2017)*.* Increasingly, eDNA metabarcoding is being highlighted as an important tool for non-invasively monitoring biodiversity, providing a rapid, reliable, accurate, and cost effective technique to survey a broad array taxa and ecosystems (Bohmann *et al.* 2014; Kelly *et al.* 2014). However, there are key methodological and bioinformatic hurdles that need be addressed to improve the accuracy and reliability of eDNA methods.

Recent advancements in DNA metabarcoding have come from improved primer design, leading to the generation of ‘universal’ primers that target a broad diversity of life. However, no single barcode can both capture the taxonomic breadth in tandem with providing sufficient genetic differentiation needed for high taxonomic resolution. Thus, more studies have taken on multilocus metabarcoding to improve species identification, e.g. using 12 loci for detection of endangered species (Arulandhu *et al.* 2017). Multilocus metabarcoding helps reduce detection bias (Kelly *et al.* 2017) and can catalog much broader diversity within an ecosystem (Stat *et al.* 2017). As data collection through multilocus sequencing becomes more common, an important challenge is to efficiently and accurately assign taxonomy across all DNA barcodes. National Center for Biotechnology Information (NCBI) *BLAST*-based queries (Altschul *et al.* 2017) of public sequence databases are slow and inflated with redundant sequences and ‘environmental sample’ sequences that are not taxonomically informative (Yim & Cushman, 2017, Chen, Zobel, & Verspoor, 2017). This creates a need for data curation and alternative query methods.

*Big data challenges -- reference databases and bioinformatics*

Curated databases for barcoding loci offered a potential, but ultimately insufficient, solution to the taxonomy assignment challenge. *SILVA* or *greengenes* for rRNA (Quast *et al.* 2013, DeSantis *et al.* 2006), *UNITE* for the internal transcribed spacer, ITS (Kõljalg *et al.* 2013), and *MIDORI* for CO1 (Machida *et al.* 2017), have been successful. However, the expansion of sequence data has far outpaced the rate of database curation, such that published databases often lack the most current biodiversity data. User-generated custom databases present the solution to this wave of expanding the genomic loci of focus. *In silico* PCR software like *OBItools* *ecoPCR* (Boyer *et al.* 2016; Stoesser *et al.* 2002) allow researchers to generate references libraries from sources like NCBI Genbank (Benson *et al.* 2017) based on reads that contain target primers. However, most metabarcode sequences are deposited in reference databases without primer regions, and thus cannot be obtained via this method. Another custom database creation approach has been to use keyword searches (e.g., *MIDORI*; Machida *et al.* 2017), but this approach relies on accurate and complete metadata for uploaded reference sequences that is frequently missing or incomplete (Hinchcliff and Smith, 2014).

*A Sea Change in Taxonomic Classification*

eDNA metabarcoding methods for taxonomic classification have begun to move past cluster-based algorithms like *vsearch* (Rognes *et al.* 2016) that can inhibit the possibility of reaching species-level assignment, in favor of Bayesian probabilistic statistical approaches. Least common ancestor (LCA) classifiers are frequently used in programs like *MEGAN**LCA*(e.g., Huson, Auch, Qi, & Schuster, 2007), and recently, Bayesian LCA algorithms have further improved assignment by providing confidence estimation scores (Gao *et al.* 2017). In turn, the community has moved away from using operational taxonomic units (OTUs) that are essentially sequences constructed from clustering at arbitrary thresholds and embraced amplicon sequence variants (ASVs) that retain single nucleotide differences in the sequenced region (Callahann *et al.* 2016).

*Tradeoffs between read depth and amplicon length*

As eDNA metabarcoding studies expand in scope and scale, researchers are turning to sequencing platforms that offer more read depth at low cost, such as HiSeq and NovaSeq (both 150bp read lengths) (Liu *et al.* 2013; Alsos *et al.* 2018). However, HiSeq technology runs the risk of generating unmerged paired data for long amplicon targets that are incompatible with many existing metabarcoding toolkits (Boyer *et al.* 2016; Caporaso *et al.* 2012; Bolger, Lohse, & Usadel, 2014). In addition, current packages are conservative and inflexible to using partial sequence data where full length alignment with reference barcodes isn’t possible (Port *et al.* 2015); for instance, they can’t process an unpaired read. Furthermore, intolerance of variable length or partial data from metabarcodes could lead to selection bias against certain taxa (Deagle *et al.* 2014; Soergel *et al.* 2012).

*Needs for eDNA for community monitoring*

Ecosystems in the anthropocene are under extreme pressure from land use change, over-exploitation, climate change, and species invasions. Frequent and inexpensive biodiversity monitoring tools are critical for the maintenance of healthy environments and for the conservation of species. The University of California Conservation Genomics Consortium (UCCGC) formed in 2016 with a goal to leapfrog eDNA technology as a monitoring tool for California ecosystems. The Consortium created a community science program called CALeDNA ([www.ucedna.com](http://www.ucedna.com)) to deploy thousands of volunteers to collect soil, sediment, and water to serve as research material to pilot and optimize eDNA monitoring for the high biodiversity and endemism in the state (Bowman, 1990; Calsbeek, Thompson, & Richardson, 2003; Keeley & Swift, 1995; Parisi, 2003; Rissler, Hijmans, Graham, Moritz, & Wake, 2006) The CALeDNA program performs multilocus metabarcoding with HiSeq and MiSeq platforms to inventory algae, animals, archaea, bacteria, fungi, and plants. Given there was no optimal eDNA analysis method for both multiple loci and different sequence technology we have generated our own toolkit to meet our needs.

Anacapa ***--*** *an all-purpose eDNA metabarcoding analysis toolkit*

Here, we present *Anacapa*, an automated method to create custom reference databases and simultaneously analyze multiple metabarcoding reads produced by HiSeq and MiSeq Illumina sequence platforms, with a built-in exploration tool of the raw results output. *Anacapa* combines components of leading bioinformatics software to obtain quality reference databases and accurate eDNA results. Reference database libraries are prepared with a new approach: *Creating Reference libraries Using eXisting tools (CRUX)*. Quality control of raw sequences and inference of Amplicon Sequence Variants (ASVs) is performed using the smart error estimator, *Dada2* (Callahan *et al.* 2016). Taxonomic assignment leverages *Bowtie 2* andBayesian Least Common Ancestor algorithm (*BLCA*; Gao *et al.* 2017). Initial results exploration is achieved through an R package we created, called *ranacapa***,** which has an interactive graphical user interface (GUI) as well as command-line interface, to rarify samples and generate diversity statistics and plots. Here, we demonstrate performance of *Anacapa* tools relative to other existing methods, and showcase its performance compared to a leading alternative classifier method, Qiime2 v2018.2 sklearn, on real eDNA data collected through the CALeDNA program. *Anacapa* is designed to be optimal for low-cost, big data eDNA surveys such as CALeDNA. The *Anacapa* toolkit was built for use on standard academic high performance computers and can be easily installed using a container or research environment via *Singularity*. The software, code, Singularity container, and examples can be found at [www.ucedna.com/software](http://www.ucedna.com/software) and are permanently archived on DRYAD (link).

**Logic of the Anacapa Toolkit**

The toolkit is named for the iconic southern California island, Anacapa, that has significant cultural and biodiversity importance. The name derived from the Chumash word *Ennepah* or *Anyapakh* which translates to "mirage island" (Bright, 2004; Gudde, 2010). Much like the name, using eDNA to monitor biodiversity seems like an illusion on the horizon, but like the real island, the Anacapa toolkit can obtain true and quality results with full transparency of the caveats of eDNA. Here, we go through the steps of the toolkit that each build upon and connect leading software modules that will process DNA reads and assign taxonomy.

*CRUX: reference library generation*

To generate taxonomic assignments of metabarcoding reads, we developed a comprehensive reference library generating tool, *CRUX* (Fig. 1a). *CRUX* constructs custom reference libraries based on given a primer set and the publicly available databases *EMBL* and *NCBI* (Pruitt *et al.* 2005; Stoesser *et al.* 2002). First CRUX runs *in silico* PCR method *‘ecoPCR’* (Ficetola *et al.* 2010) against user-selected subsets of the EMBL standard nucleotide sequence database (Stoesser *et al.* 2002) to generate a seed library of reads with unique taxon identifiers. Other nucleotide databases can also be used in this step (e.g. *GenBank* or custom fasta files). *CRUX* then verifies that seed library reads match the specified amplicon by checking for correct primer regions and trimming them using cutadapt (Martin, 2011). Given that many sequencing records are deposited to *Genbank* (Benson *et al.* 2012) with the primer regions removed from the read, we then use *blastn* to probe (Camacho *et al.* 2009) the seed library against the *NCBI* non-redundant nucleotide database (Stoesser *et al.* 2002; Pruitt *et al.* 2005; Johnson *et al.* 2008) in order to obtain as many reference sequences as possible.

*CRUX* runs *blastn* twice where the first *blastn* query accepts up to 10,000 full-length reads (e.g. the same length as the seed reference sequence) and then de-replicates the resulting fasta files by NCBI accession version number (Camacho *et al.* 2009). The second *blastn* run accepts up to 10,000 sequences with at least 70% alignment in order to capture reference sequences that do not cover the entire seed sequence length. The resulting sequences are then sorted by length and de-replicated so that only the longest version of a sequence accession is retained. The user can tune these parameters. We caution that partial reads may bias taxonomic assignments.

*Entrez-qiime* (Baker, 2016)takes the retained *BLAST* output, the NCBI taxonomy reference library, and a map of association between taxonomy and accession version numbers (Camacho *et al.* 2009) to generate a file with full taxonomy paths for the CRUX library. Because NCBI taxonomy includes organisms with unknown taxonomy, *CRUX* generates two reference files: filtered and unfiltered. The filtered reference file excludes reads with equivocal taxonomic paths (i.e. taxonomic paths that include the phrase “uncultured”, “environmental”, “sample”, or four consecutive missing identifications, i.e. “NA;NA;NA;NA”). Next, *CRUX* generates a corresponding *Bowtie2* (Langmead and Salzberg, 2012) formatted data index library. The final outputs include a reference library fasta file, a file of taxonomic identifications, and corresponding *Bowtie2* formatted data index library for use in *Anacapa*metabarcode data processing. We note that a *CRUX* library can be formatted in this last step from any multifasta reference file with a corresponding taxonomy file.

*CRUX* performs these millions of BLAST queries efficiently by running multiple arrays of jobs, allowing users to generate reference libraries in merely a few hours to a few days.

*Sequence QC and ASV Parsing*

This step of the toolkit (Fig. 1b) conducts standard DNA sequence quality control (QC) and generates amplicon sequence variants (ASVs) using *Dada2* (Callahan *et al.* 2016). Anacapa simultaneously processes fastq reads for samples with single or multiple metabarcode targets generated on Illumina HiSeq and MiSeq platforms. It is also not required that all samples contain reads from the same metabarcodes, thus allowing users to combine multiple projects while only running the pipeline once. The user supplies two fasta files: one containing forward primer sequences, and one with corresponding reverse primer sequences. *Anacapa* is built to search for either Illumina Nextera or TruSeq adaptors, and can be easily modified to include others.

*Anacapa* takes raw Illumina fastq format reads and generates an md5sum for users to check the integrity of their files, then uncompresses fastq files and renames the file suffixes (e.g. R1\_001.fastq or R2\_001.fastq). *Cutadapt* (Martin, 2011) is called to trim the sequencing adapters from the 5' end of reads and both sequencing adapters and primers from the 3' end of reads. This is necessary when a barcode region may be shorter in length than the sequencing platform read length, which produces sequences extending through the end for the 3’ primer and sequence adapter. *Fastx-toolkit* (Gordon and Hannon, 2010) is then used to process for quality control. We include default quality score and length thresholds optimized for MiSeq 2x300 and Illumina 2x150 platforms, which can be modified by the user according to the project needs. Next, *cutadapt* (Martin, 2011) is used to sort reads by primer. Prior to running *Dada2*, the reads that pass QC are divided into three bins: one of read pairs, one of orphaned forward reads, and one of orphaned reverse reads. Each read bin is run separately through *Dada2*. *Dada2* denoises, dereplicates, merges paired reads into contigs, and removes chimeric sequences (Callahan *et al.* 2016). Here, ‘unmerged reads’ are retained as paired reads if a contig with a set amount of overlap (default: 10 bp) cannot be aligned, and if the total length of the pair is within the maximum contig size allowance set by the user for each metabarcode. If the pair exceeds maximum length settings for the metabarcode (default: expected maximum length of the barcoding region without the primer, plus 20 bp to account for minimum contig overlap), it is deemed likely problematic and removed from downstream processing. The outputs from this step are a table indicating the frequency of ASVs in each sample, and a fasta file of ASV sequences for each metabarcode.

*Assignment: Anacapa classifier assigns taxonomy with Bowtie2 and BLCA*

#### This next step (Fig. 1c) of the *Anacapa* toolkit assigns taxonomy to all ASVs. The best taxonomic hits for an ASV are chosen using *Bowtie 2* (Langmead and Salzberg, 2012) to query the *CRUX* reference database for the locus. *Anacapa* considers paired merged, paired unmerged, and unpaired sequencing reads, and thus a fast and flexible read aligner, such as *Bowtie2*, is required to handle all four read types. All reads are first globally aligned to reference sequences using *Bowtie 2* in global or “end-to-end” mode, which will match reads where all characters align. Any reads that fail to align globally are then aligned in “local” mode, which uses soft clipping to obtain the best alignment score. By default, up to 100 of the best *Bowtie 2* returns are retained, and stored in a SAM file. Taxonomy is then assigned with the Bayesian Least Common Ancestor (*BLCA*) method (Gao *et al.* 2017) that we modified to accept inputs of SAM files generated by *Bowtie 2*. *BLCA* uses pairwise sequence alignment to calculate sequence similarity between query sequences and reference library hits. The reference hits used for pairwise alignments must match a minimum proportion of the query. The default is 80, which means that at a minimum a hit from the “end-to-end” mode of *Bowtie 2* will have at least 80% identity to the query, or a hit from “local” mode can be 80% of the length of the query but must have a 100% identity. Taxonomy is assigned based on the lowest common ancestor of multiple reference library hits for each query sequence. The reliability of each taxonomic assignment is then evaluated through bootstrapping (default 100 times) and assigned a bootstrap classifier confidence (BCC).

#### *Anacapa* generates multiple outputs of ASV tables from this step, which contain assigned taxonomy for reads from each metabarcode and both “end-to-end” and “local” mode alignments. Important descriptors to help the user interpret the data are also retained in the output tables, including the length of the input ASV, the percent match of the best reference hit, notation of whether there were single or multiple best hits, the complete taxonomic lineage for the best assignment, the percent confidence for each level of taxonomy, and all the reference accession versions that were included in the *BLCA* step. In addition to this detailed table, summary tables are generated: there are summary tables that concatenate forward, reverse, merged and unmerged paired reads according to their assigned taxonomy, and summary tables with taxonomy reported to a hierarchical level having over a certain designated percentage of confidence.

*ranacapa: Exploratory results analysis with ecology methods*

*Anacapa* is accompanied by an R package *ranacapa* that helps the user perform exploratory results analysis that provides a first pass look at sequencing depth, taxonomic assignments, and generated data tables. Exploratory analysis is helpful in identifying potential glaring errors or contamination, and identifying patterns worth investigating through more robust analysis. We highly encourage data exploration before further analysis as the *Anacapa* user weighs different parameters. The *ranacapa* module is also useful for classroom education on eDNA approaches. We note results do not need to have been generated with *Anacapa* to be used in this extension tool.

The *ranacapa* script uses a variety of R packages, relying heavily on *phyloseq* (McMurdie & Holmes, 2013), *vegan* (Oksanen *et al.* 2007) and *ggplot2* (Wickham, 2016) and can be operated from the command line or via a *Shiny* app graphical interface (Chang *et al.* 2017). The inputs include an ASV table with assigned taxonomy, such as an Anacapa summary table where ASVs are rows and samples are columns, and a metadata table where the first column contains sample names, and following columns are categorical or continuous metadata. Standard QIIME-formatted OTU tables are compatible with *ranacapa.* The output is a series of figures and statistics including rarefaction curves, absolute and relative abundance bar plots, alpha diversity boxplots, and alpha diversity statistics. It automatically calculates Jaccard and Bray-Curtis distance NMDS ordination plots, network maps, heat maps, Ward-linkage maps, and two beta-diversity statistical tests: pairwise ‘adonis’ and ‘betadisp’ from the *vegan* package (Oksanen *et al.* 2007). Each of the above analyses are repeated with different groupings for each metadata column (e.g. habitat type or soil depth).

**Availability of *Anacapa* and custom reference databases**

All components of *Anacapa* can be downloaded through Github, with legacy versions available in DRYAD (link here). We also offer *Anacapa* in a *Singularity* container for easy installation and use. *Ranacapa* is already containerized in a *Shiny* app and not included in the *Singularity* container. All links to download pages containing full documentation and tutorials are available at [www.ucedna.com/software](http://www.ucedna.com/software).

For this study, we generated a series of *CRUX* reference libraries for common barcode markers to use in the benchmarking and demonstration tests. These are also provided on DRYAD in recognition that some users may prefer to not generate their own reference libraries. These include loci in the popular barcoding regions encoding 16S rRNA, 18S rRNA, CO1, 12S, Internal Transcribed Spacers (ITS) ITS1 and ITS2, that are publicly available on the Dryad database (here). Table S1 contains the specific loci, the primers used to target them, and their references.

**Anacapa tool benchmarking with mock data**

*CRUX reference databases*

Here we benchmark *CRUX*filtered and unfiltered databasesby generating both reference libraries corresponding to the same loci as published and well-curated reference libraries for barcoding markers 16S (*Greengenes* 2011 16S database: DeSantis *et al.* 2006; *Silva ‘*SILVA\_128\_QIIME\_release’ clustered at 99%: Pruesse *et al.* 2007, Quast *et al.* 2013), 18S V4 and V8-9 (*Silva,* same source as above), and CO1 (*Midori*: Machida *et al.* 2017), and then evaluating the taxonomic assignment output for sanger sequenced reads associated with mock eDNA datasets. The *Anacapa classifier* step of the toolkit was used in these analyses and all parameters other than reference database were held constant.

All *CRUX* databases had more reads and more unique taxonomic lineages than the corresponding published libraries (Table S2). One reason for this is that *CRUX* includes recently published NCBI data where others may be outdated. Another reason is *CRUX* reference libraries allow for the inclusion of partial reads down to 70% length of the barcode locus (default setting), and therefore may capture more deposited sequence data (Pruitt, Tatusova, & Maglott, 2006). And unlike *Silva* libraries that come clustered at several different percent similarities (e.g. 99% and 97%; Pruesse *et al.*, 2007), we do not cluster *CRUX* libraries and thus retain more reads.

*CRUX* filtered libraries had variable reduction in number of reads retained after filtering: only 17.5% of the original unfiltered set were retained in 16S, but 99% of the unfiltered set were retained in CO1. However, the number of unique taxonomic lineages was only at most reduced by 11.4% after filtering (for 16S; all others less than 7% reduced). Comparisons of *CRUX* filtered databases to published databases were as follows. 16S: *CRUX* contained 1.85x more reads and 1.92x more taxa than *Silva*, and 4.37x more reads and 8.68x more taxa than *Greengenes*. 18S V4 region: *CRUX* contained 1.84x more reads and 1.98x more taxa than this *Silva* database*.* 18S V8-9 region: *CRUX* had 1.50x more reads and 2.39x more taxa than *Silva*. CO1: *CRUX* had 12.69x more reads and 4.65x more taxa than *Midori*.

To test the performance of *CRUX* databases relative to published reference libraries, we iterated assigning taxonomy with each different published or *CRUX* database using the default assignment step of the Anacapa toolkit for three Sanger sequence isolate mock data sets (Kozich *et al.*, 2013; Bradley *et al.*, 2016; Callahan *et al.*, 2016; Krohn *et al.*, 2016; Leray *et al.*, 2017). The mock data were composed of 1) 75 sequenced isolates for 16S, 2) 12 organisms sequenced for 18S (V4 and V8-9), and 3) 32 organisms sequenced for CO1 (details and references in Table S3). Taxonomic assignment was iterated for the different reference databases for the appropriate markers and accuracy at the five BCC cut-offs from 60-95, for which we accepted all assignment at any level of the taxonomic path above the cut-off, was compared. Across all loci, CRUX databases yielded more correct assignments and more high resolution taxonomy (Figure 2a; Supplementary Figure 1).

There was clear improvement in performance when using the filtered CRUX database over the unfiltered database for 16S and 18S; however, the unfiltered CO1 database outperformed the filtered version. We discovered this incongruity to due to low resolution given taxonomic assignments in the mock CO1 dataset. We also found that correct species level assignment was frequently at 80% BCC or below in 16S, and clearly best at 60% confidence in 18S tests, suggesting that by and large, only the availability of more sequences, or use of phylogenetically informed methods, can improve assignment confidence. Based on these results, in subsequent benchmarking and demonstrations of Anacapa, we use the 60% BCC cutoff summary taxonomies. The exact taxonomic calls for all mock dataset results at this confidence level can be found in Tables S4-S7.

*Comparisons of taxonomic assignment methods*

The *Anacapa classifier* is a modification of the *BLCA* classifier (here called “BLAST *BLCA*”; Gao *et al.* 2017) adapted to accept *Bowtie2*-indexed inputs. In order to verify that the *Anacapa classifier* performs as well as the original *BLCA*that uses BLAST-indexed reference database inputs in queries for ASV matches, we ran both classifiers on the same Sanger read mock datasets and corresponding *CRUX* reference databases as the previous tests. The *Anacapa classifier* was run with default settings while BLAST *BLCA* was run with parameters of 80% minimum sequence similarity, 0% sequence overhang, 70% sequence coverage, and up to 100 query returns so as to have parallel settings. We generated matching *Bowtie2* and *BLAST* indexed libraries using the same *CRUX* generated reference libraries and taxonomy files. The accuracy of taxonomic assignment for mock data isolates was compared for the two methods. Out of the 798 total possibilities of levels of taxonomy across the 133 species in the mock data, we report only two differences in any level between the two classifiers, and in both cases our method performed slightly better: one 16S *Lactobacillus* ASV was called to correct species instead of only to genus, and one CO1 ASV was classified to the *Platyhelminthes* phylum instead of being unclassified (16S and CO1 shown in Figure 2b; 18S V4 and V8-9 shown in Supplementary Figure 2; Tables S8-11).

Additional tests were performed to compare the *Anacapa classifier* to three classifier approaches available in the latest version of *Qiime2* v 2018.2*:* naive Bayes pre-fitted sklearn-based taxonomy classifier (*sklearn*; Pedregosa *et al.*, 2011), BLAST+ consensus taxonomy classifier (*BLAST+ classifier*; Altschul *et al.*, 1990), and *VSEARCH* consensus taxonomy classifier (Rognes *et al.*, 2016). Settings for *sklearn* followed the feature classifier tutorial [(https://docs.qiime2.org/2018.2/tutorials/feature-classifier/](https://docs.qiime2.org/2018.2/tutorials/feature-classifier/))) with the exception that we accepted hits down to 60% confidence in taxonomy to match the *Anacapa classifier* BCC parameters. The *BLAST+* and *VSEARCH* classifiers were run with default settings with the exception of allowing 100 hit returns, and adjusting the percent required for consensus to 60%.

Assignments made with Bayesian methods performed nearly equally well and substantially outperformed consensus-based classifiers, (Figure 2b, and Supplementary Figure 2, Tables S8-S11). Interestingly, a fraction of results between Bayesian classifiers revealed discrepancies, where 8% of the results for 16S had species-level correct assignment in the *sklearn* result (only genus level in *Anacapa*), and in reverse, a different 8% were assigned to a correct species in the Anacapa classifier result but only to genus in *sklearn*. In the 18S V4 result, *sklearn* had called four more species-level assignments than *Anacapa*, but three were erroneous. In the CO1 result, our classifier assigned one more ASV to species than *sklearn*. In conclusion, it appears both classifiers provide competitive and valuable results, and that the concordance between them may suggest truly correct assignment. There are differences in the algorithms used by each classifier. *Anacapa* uses a bootstrapping method to assign taxonomy based on a Bayesian posterior probability (Gao *et al.* 2017) and sklearn uses a naive Bayes classifier trained on the reference library (<http://scikit-learn.org/stable/modules/naive_bayes.html#multinomial-naive-bayes>). Differences in the assumptions and methodology used to assign taxonomy are likely driving the discrepancies in species and genus level taxonomic resolution between classifiers.

*Anacapa functionalities with mock HiSeq and MiSeq data*

The short read lengths of paired-end NGS (75-300bp) frequently don’t have enough overlap after QC to produce a single contigs made of merged pairs. To our knowledge, no eDNA software exist that can process paired but unmerged reads as a unit. This is especially problematic if some metabarcode lengths are close to or longer than the paired read length on the NGS platform. Furthermore, Illumina sequencers often have poorer quality reverse reads that can require extensive trimming, and this can increase frequency of short metabarcodes that remain unable to merge (Bolger, Lohse, & Usadel, 2014). A small but sizeable fraction of NGS reads lose mates entirely due to low quality. The most substantial advantage of *Anacapa* is our ability to retain merged, unmerged paired, and unpaired sequences and assign taxonomy to them simultaneously. The *Anacapa classifier*allows users to both retain more data, and have sufficient detail in output tables to evaluate the benefits and caveats of partial or unideal data.

To demonstrate the ability of the Anacapa toolkit to handle a range of sequence lengths and types, we tested differences in assignment between MiSeq and HiSeq length datasets. Using MiSeq 2x250 bp or 2x300 bp generated mock community libraries of CO1, 16S, and 18S V4 and V8-9 pooled isolates, we trimmed reads to be 150 bp in length, so as to become a pseudo-HiSeq library set for comparison. The CO1 mock dataset was of evenly pooled DNA from 34 unique organisms (Tables S3 and S12; Leray and Knowlton, 2017). The 16S mock dataset consisted of DNA from 35 unique organisms pooled at different orders of magnitude concentrations (Tables S3 and S13; Callahan *et al.* 2016). Multiple datasets for 18S V4 and V8-9 consisted of pools of up to 12 isolates at different concentrations (Tables S3,S14-S15; Bradley *et al.* 2016). Using the same reference databases and BCC settings as prior benchmarking, we compared the relative abundance of reads assigned to each unique taxon in the MiSeq and HiSeq-length results for different library mock pools. Concordance between MiSeq taxonomic assignments and HiSeq taxonomic assignments was estimated using linear regression (Figure 3; full results in Tables S16-S19).

Concordance for a given taxonomy between HiSeq and MiSeq length sequences was high for 16s sequences to the level of species (Table S19, Figure 3) and identical if taxonomy was considered only to the level of genus (Supplementary Figure 3 ). However, for the lengthier metabarcode CO1, concordance between platforms was weaker, but driven by a few individual taxa with different resolution between platforms. For example, one species of Polychaete was identified correctly using the MiSeq length sequences (Table S12), however it was not assigned taxonomy using HiSeq length reads. Removing this organism from the analysis lead to high concordance in taxonomic assignment (Adj R2 = 0.9973, p < 0.001). The test of 18S mock community pools, representing different assemblages of the same taxa, reiterates that most taxa provide fine-scale resolution in HiSeq results, with seven exhibiting a concordance R2> 0.65, three of which were over R2>0.95 (Table S20). Nonetheless, one pool had no congruence between data types, suggesting for HiSeq data, species DNA barcodes should be checked synthetically. These results highlight the tradeoffs of using high capacity sequencing platforms like HiSeq that sacrifice read length for orders of magnitudes more sequences. Depending on the application of eDNA metabarcoding, the generation of more sequences may outweigh the need for species-level taxonomic assignment.

**Processing rate of Anacapa with different data characteristics**

Here we performed time benchmarking to help users estimate the different rates at which *Anacapa* will make assignments for ASVs for different metabarcodes. We provided this for MiSeq and pseudo-HiSeq generated ASVs, and also compared the *BLAST BLCA* (Gao *et al.*, 2017) approach to the *Bowtie 2* modification *Anacapa* employs. Figure 4 shows that the processing time of both methods increases as the number of reference database hits returned increases. Most 18S returned >50 hits, especially for forward reads, whereas CO1 and 16S results were more evenly distributed between 0 and 100 (the maximum allowance). This pattern mirrors the diagnostic capacity of the barcode loci. Processing time per read with 50 hits returned, using either method, ranged from 0.75-2.5 seconds for forward or reverse reads, regardless if 250 bp or 150 bp. ASVs from merged reads took as much as 1.5 seconds longer to process than those from orphaned reads (i.e., forward or reverse only). The *Anacapa classifier* was close in rate to *BLAST BLCA* at <30 hits retrned, but frequently slower than *BLAST BLCA* when more hits were returned, showing a variable increase in time of 0-4 seconds. However, for CO1 MiSeq data, which has the lengthiest ASVs, *Anacapa* outpaced *BLAST BCLA* in all categories, and significantly for orphan ASVs (Figure 4; Table S21).

**Demonstration of Anacapa with real eDNA data from CALeDNA**

Here we show the *Anacapa* toolkit on real eDNA data from the CALeDNA program. Samples were from different California habitats including three samples of sea water from kelp forests off Anacapa Island (from here in “Kelp Forest”), three samples of surface soil from coastal sage scrub in Laguna Coast Wilderness Park (“Laguna”), and surface soil from chaparral in Stunt Ranch UC Natural Reserve (“Stunt Ranch”) (Table S22). Multilocus metabarcoding libraries from Kelp Forest and Stunt Ranch were run on the Illumina HiSeq 2x150bp platform, and samples from Laguna were run on the IlIumina MiSeq 2x300bp platform (library preparation and sequencing details are in Supplementary Methods and unprocessed fastqs are archived here). These samples were part of different eDNA projects, and as such, different metabarcodes were amplified.

*The Anacapa Toolkit excels at processing real eDNA*

16S and CO1 results from Laguna samples (MiSeq) and Stunt Ranch samples (HiSeq) eDNA samples show the *Anacapa classifier* outperforms *Qiime2 sklearn* on California data. Unlike *Qiime2’s* QC and ASV plugins, the *Anacapa* toolkit allows unmerged paired-end reads to be processed. The sklearn classifier cannot successfully taxonomy for these ASVs. We show this in a comparison of biodiversity metrics generated from both classifier outputs, using the visualization and statistics using default *ranacapa*, the results exploration tool (Figure 5).

The *Anacapa* classifier assigned taxonomy to more ASVs than *sklearn* for 16S (2.12x for Stunt Ranch and 1.09x for Laguna, Figure 5, Table S23) and for CO1 (235.54x for Stunt Ranch and 1.87x for Laguna, Table 24). The 16S V4 marker amplifies approximately 291 base pairs and the CO1 marker targets 313 bp, thus the majority of paired HiSeq processed reads were unmerged (Tables S23-S24). It was expected that more ASVs would be classified with *Anacapa* because it process unmerged reads, and indeed we see this for the Stunt Ranch 16S data where over half (52%) of the taxonomy identified for unmerged reads are not detected by sklearn in the merged, forward, or reverse ASVs (Table S23). The ASVs generated with CO1 for the Stunt samples were primarily comprised of unmerged sequence pairs (98.8% of the total sequences; Table S24). Only 350 reads were assignable by *sklearn* for the Stunt CO1 HiSeq samples, compared to 19,000, which was the lowest number in the *Anacapa* processed samples. We chose to drop the *sklearn* assignments from the analyses of Stunt Ranch CO1 (Figure 5) so that we could proceed with rarefaction and diversity estimates in *ranacapa*. Rarefied, we found that genus and species alpha diversity was significantly higher using the *Anacapa**classifier*. Classifier type produced significant (Table S27) differences in biodiversity estimates, most pronounced along Axis 2 in the Jaccard NMDS plot for 16S that included assignments down to species (Figure 5).

We next explored the applicability of the *Anacapa*toolkit for identifying organisms of interest for aquatic and terrestrial species management. The Mifish 12S (Miya *et al.* 2015) metabarcode region is frequently used to survey marine fish populations. We explored the data generated for the Kelp Forest samples. For each sample we retained only taxonomy with high confidence (over 300 sequences, based on the average number of reads given a taxonomic assignment for 12S found in the sample blanks). Across the three samples, we identified 15 species, 29 genera, and 26 families of vertebrates (Figure 6a). Many California taxa have not been sequenced for 12S, and in several instances we sequenced the correct genus or the sister species that is not present on the California coast. The Anacapa pipeline to capture a broad diversity of vertebrate taxa including keystone kelp forest species (e.g. California Sheephead, *Semicossyphus pulcher)*, species of management concern (e.g. California sea lion, *Zalophus californianus*), and rare and transient species (Boobies, *Sula sp.*). This highlights the ability of eDNA to detect a wide diversity of marine life from largely unmerged HiSeq data, supporting the potential application for monitoring of marine ecosystems.

We used the plant ITS2 locus (Gu *et al.* 2013) to survey plant populations in Laguna samples. For each sample we retained only taxonomy with high confidence (over 150 sequences, based on the criteria above). We found 15 families of green algae, charyophytes, non-vascular and vascular plants, representing 24 genera and 23 species (Figure 6b). Of the macroscopic organisms, most have been identified by iNaturalist (<https://www.inaturalist.org>) users in similar nearby habitats of Southern California (e.g *Funaria hygrometrica, Artemisia californica, Erodium moschatum, Crassula aquatica*, and the genera *Chrysanthemum, Gemmabryum* and *Bryum*). Neither *Leucobryum* sp. or *Chrysanthemum* *indicum* have been identified in California. These eDNA uncovered local diversity that we expected to find in this habitat. One major benefit of this method is the rich inventory of the algal organisms living in association with plants and soil crust microbes, which may be used to understand small eukaryote ecosystem function, which have been historically difficult to characterize (Büdel *et al.* 2016).

**Conclusion**

The *Anacapa* toolkit meets the needs of the metabarcoding community and can be broadly applied to microbiome and eDNA projects. By encouraging users to generate their own high quality and customizable reference databases, people can look at broader sets of barcodes and leverage the myriad sequence data being published *in perpetuum*, from local initiatives such as the Diversity Initiative for the Southern California Ocean (<https://research.nhm.org/disco/>), to networks such as the Barcode of Life (<http://www.barcodeoflife.org/>), to whole genomes (e.g., <https://phytozome.jgi.doe.gov>). The toolkit is packaged so as to easily simultaneously process multiple barcodes and retain all quality data, which lets users consider HiSeq approaches as well as lengthier barcodes for their projects. Robust taxonomic assignment and ample information about how assignments were made (as metadata in output tables) give the decision power back to users so they can decide what data to put on the chopping block. The capacity to automatically perform biodiversity statistical analyses also equips users with guides to optimize settings and catch potential errors in the library preparation steps, which could save time. In addition, the modularity and transparency of the tools make it possible to incorporate other software packages including *MACSEQ* (Ranwez, Harispe, Delsuc, & Douzery 2011), *DeBlur* (Amir *et al.* 2017), etc. at various steps.

We created *Anacapa* to make eDNA research more accessible, from performing the research to teaching the approach: modules of Anacapa can be used in educational settings and can be aligned with metabarcoding outputs such as from the CALeDNA program or the Earth Microbiome Project (www.earthmicrobiome.org/), allowing even novice eDNA explorers to engage with data and perform sophisticated and publishable analyses. eDNA is excellent at providing an inventory of a wide diversity of taxa (e.g, Deiner *et al.*, 2017) that address spatial or temporal questions in many disciplines. eDNA analysis with Anacapa is currently being used in undergraduate science classes in the University of California system. By containerizing the toolkit, *Anacapa* can be adopted by other institutions and agencies. As biodiversity monitoring initiatives by government and conservation and restoration agencies are increasingly looking at whole community dynamics and not only single species, exploring the capabilities of eDNA should be focused mainly on the biology, not the informatics details.

*et al.*

**Acknowledgements**

We would like to acknowledge the many beta- testers who contributed to the development of the toolkit including Rachel Turba, Dannise Ruiz Ramos, and Sabrina Shirazi, and thank Beth Shapiro for providing helpful advice. We acknowledge Dr. Raffaella D'Auria and the IDRE Statistical Consulting at UCLA for assistance in implementing the pipeline on the UCLA cluster. In addition, thank Sarvani Chadalapaka for assistance implementing Anacapa on the UC Merced cluster. We thank The Bruin Naturalist Club, Gary Bucciarelli, Marianne Maki and her AP Environmental Science Class with for assisting with sample collection at the UCLA Stunt Ranch Reserve.  We thank Rachel Chock, Shauna King and Debra Shier for assisting with sample collection at the Laguna Coast Wilderness Park. We acknowledge David Kushner and Joshua Sprague from the Channel Islands National Park Service for sample collection. This work was made possible by the University of California President’s Research Catalyst Award (CA-16-376437).

**References [will update upon final draft]**

About us — scikit-learn 0.19.1 documentation. (n.d.). Retrieved March 17, 2018, from<http://scikit-learn.org/stable/about.html>

Alsos, I. G., Lammers, Y., Yoccoz, N. G., Joergensen, T., Sjoegren, P., Gielly, L., & Edwards, M. E. (2018). Metabarcoding lake sediments: taphonomy and representation of contemporary vegetation in environmental DNA (eDNA) records. *BioRxiv*, 264903. doi:[10.1101/264903](https://doi.org/10.1101/264903)

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.

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., … Knight, R. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *MSystems*, *2*(2), e00191-16. doi:[10.1128/mSystems.00191-16](https://doi.org/10.1128/mSystems.00191-16)

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. doi:[10.1093/gigascience/gix080](https://doi.org/10.1093/gigascience/gix080)

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2017). GenBank. *Nucleic Acids Research*, *45*(Database issue), D37.

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. doi:[10.1016/j.tree.2014.04.003](https://doi.org/10.1016/j.tree.2014.04.003)

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Bowman, R. I. (1990). Evolution and biodiversity in California. *California’s Wild Heritage: Threatened and Endangered Animals in the Golden State*.

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.

Büdel, B., Dulić, T., Darienko, T., Rybalka, N. and Friedl, T., 2016. Cyanobacteria and algae of biological soil crusts. In *Biological Soil Crusts: An Organizing Principle in Drylands* (pp. 55-80). Springer, Cham.

Bradley, I. M., Pinto, A. J., & Guest, J. S. (2016). Design and evaluation of Illumina MiSeq-compatible, 18S rRNA gene-specific primers for improved characterization of mixed phototrophic communities. *Applied and Environmental Microbiology*, *82*(19), 5878–5891.

Bright, W. (2004). *Native American Placenames of the United States*. University of Oklahoma Press.

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. doi:[10.1038/nmeth.3869](https://doi.org/10.1038/nmeth.3869)

Calsbeek, R., Thompson, J. N., & Richardson, J. E. (2003). Patterns of molecular evolution and diversification in a biodiversity hotspot: the California Floristic Province. *Molecular Ecology*, *12*(4), 1021–1029.

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

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Gordon, J. I. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., … Knight, R. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal*, *6*(8), 1621–1624. doi:[10.1038/ismej.2012.8](https://doi.org/10.1038/ismej.2012.8)

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2017). shiny: Web application framework for R [Computer software]. *URL Http://CRAN. R-Project. Org/Package= Shiny (R Package Version 1.0. 0)*.

Chen, Q., Zobel, J., & Verspoor, K. (2017). Benchmarks for measurement of duplicate detection methods in nucleotide databases. *Database*, baw164.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., … Leon, C. (2010). Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species. *PLOS ONE*, *5*(1), e8613. doi:[10.1371/journal.pone.0008613](https://doi.org/10.1371/journal.pone.0008613)

Chris Baker. (2016). *bakerccm/entrez\_qiime: entrez\_qiime v2.0*. Zenodo. doi:[10.5281/zenodo.159607](https://doi.org/10.5281/zenodo.159607)

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). doi:[10.1098/rsbl.2014.0562](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.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., … Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072.

Ficetola, G. F., 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. doi:[10.1186/1471-2164-11-434](https://doi.org/10.1186/1471-2164-11-434)

Fujikura, K. (Japan A. for M.-E. S. and T., Nishida, S., & Shirayama, Y. (2009). Census of marine life. *Bulletin of the Plankton Society of Japan (Japan)*. Retrieved from<http://agris.fao.org/agris-search/search.do?recordID=JP2009007057>

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.

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

Gudde, E. G. (2010). *California Place Names: The Origin and Etymology of Current Geographical Names*. Univ of California Press.

Hinchliff, C. E., & Smith, S. A. (2014). Some Limitations of Public Sequence Data for Phylogenetic Inference (in Plants). *PLOS ONE*, *9*(7), e98986. doi:[10.1371/journal.pone.0098986](https://doi.org/10.1371/journal.pone.0098986)

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, *36*(suppl\_2), W5–W9.

Keeley, J. E., & Swift, C. C. (1995). Biodiversity and Ecosystem Functioning in Mediterranean-Climate California. In *Mediterranean-Type Ecosystems* (pp. 121–183). Springer, Berlin, Heidelberg. doi:[10.1007/978-3-642-78881-9\_3](https://doi.org/10.1007/978-3-642-78881-9_3)

Kelly, R. P., Closek, C. J., O’Donnell, J. L., Kralj, J. E., Shelton, A. O., & Samhouri, J. F. (2017). Genetic and Manual Survey Methods Yield Different and Complementary Views of an Ecosystem. *Frontiers in Marine Science*, *3*. doi:[10.3389/fmars.2016.00283](https://doi.org/10.3389/fmars.2016.00283)

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*. doi:[10.1371/journal.pone.0086175](https://doi.org/10.1371/journal.pone.0086175)

Kõljalg, U., Larsson, K.-H., Abarenkov, K., Nilsson, R. H., Alexander, I. J., Eberhardt, U., … Vrålstad, T. (2005). UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist*, *166*(3), 1063–1068. doi:[10.1111/j.1469-8137.2005.01376.x](https://doi.org/10.1111/j.1469-8137.2005.01376.x)

Kress, W. J., García-Robledo, C., Uriarte, M., & Erickson, D. L. (2015). DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology & Evolution*, *30*(1), 25–35.

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

Leray, M., & Knowlton, N. (2017). Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. *PeerJ*, *5*, e3006. doi:[10.7717/peerj.3006](https://doi.org/10.7717/peerj.3006)

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. doi:[10.1186/1742-9994-10-34](https://doi.org/10.1186/1742-9994-10-34)

Liu, S., Li, Y., Lu, J., Su, X., Tang, M., Zhang, R., … Zhou, X. (2013). SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods in Ecology and Evolution*, *4*(12), 1142–1150. doi:[10.1111/2041-210X.12120](https://doi.org/10.1111/2041-210X.12120)

MacConaill, L. E., Burns, R. T., Nag, A., Coleman, H. A., Slevin, M. K., Giorda, K., … Thorner, A. R. (2018). Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics*, *19*. doi:[10.1186/s12864-017-4428-5](https://doi.org/10.1186/s12864-017-4428-5)

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. doi:[10.1038/sdata.2017.27](https://doi.org/10.1038/sdata.2017.27)

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

Oksanen, J., Kindt, R., Legendre, P., O’Hara, B., Stevens, M. H. H., Oksanen, M. J., & Suggests, M. (2007). The vegan package. *Community Ecology Package*, *10*, 631–637.

Parisi, M. (2003). Atlas of the Biodiversity of California. *California Department of Fish and Game, Sacramento*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. (n.d.). Retrieved March 16, 2018, from<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061217>

Pinherio, J., Bates, D., DebRoy, S., & Sarkar, D. (n.d.). *Core Team R (2017). nmle: linear and Nonlinear Mixed Effects Models. R package version 3.1-131*.

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*.

Proença, V., Martin, L. J., Pereira, H. M., Fernandez, M., McRae, L., Belnap, J., … Gregory, R. D. (2017). Global biodiversity monitoring: from data sources to essential biodiversity variables. *Biological Conservation*, *213*, 256–263.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, *35*(21), 7188–7196.

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.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2006). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*(suppl\_1), D61–D65.

*qiime-default-reference: Default reference data files for use with QIIME*. (2018). Python, biocore. Retrieved from<https://github.com/biocore/qiime-default-reference> (Original work published 2014)

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

Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding [PeerJ]. (n.d.). Retrieved March 16, 2018, from<https://peerj.com/articles/3006/?utm_source=TrendMD&utm_campaign=PeerJ_TrendMD_0&utm_medium=TrendMD>

Ranwez, V., Harispe, S., Delsuc, F., & Douzery, E. J. P. (2011). MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLOS ONE*, *6*(9), e22594. doi:[10.1371/journal.pone.0022594](https://doi.org/10.1371/journal.pone.0022594)

Rissler, L. J., Hijmans, R. J., Graham, C. H., Moritz, C., & Wake, D. B. (2006). Phylogeographic Lineages and Species Comparisons in Conservation Analyses: A Case Study of California Herpetofauna. *The American Naturalist*, *167*(5), 655–666. doi:[10.1086/503332](https://doi.org/10.1086/503332)

Ritari, J., Salojärvi, J., Lahti, L., & de Vos, W. M. (2015). Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics*, *16*, 1056. doi:[10.1186/s12864-015-2265-y](https://doi.org/10.1186/s12864-015-2265-y)

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ Preprints*, *4*. doi:[10.7287/peerj.preprints.2409v1](https://doi.org/10.7287/peerj.preprints.2409v1)

Soergel, D. A. W., Dey, N., Knight, R., & Brenner, S. E. (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *The ISME Journal*, *6*(7), 1440–1444. doi:[10.1038/ismej.2011.208](https://doi.org/10.1038/ismej.2011.208)

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.

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.

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

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Yim, W. C., & Cushman, J. C. (2017). Divide and Conquer (DC) BLAST: fast and easy BLAST execution within HPC environments. *PeerJ*, *5*, e3486.

**Data Accessibility**

[www.ucedna.com/software](http://www.ucedna.com/software)

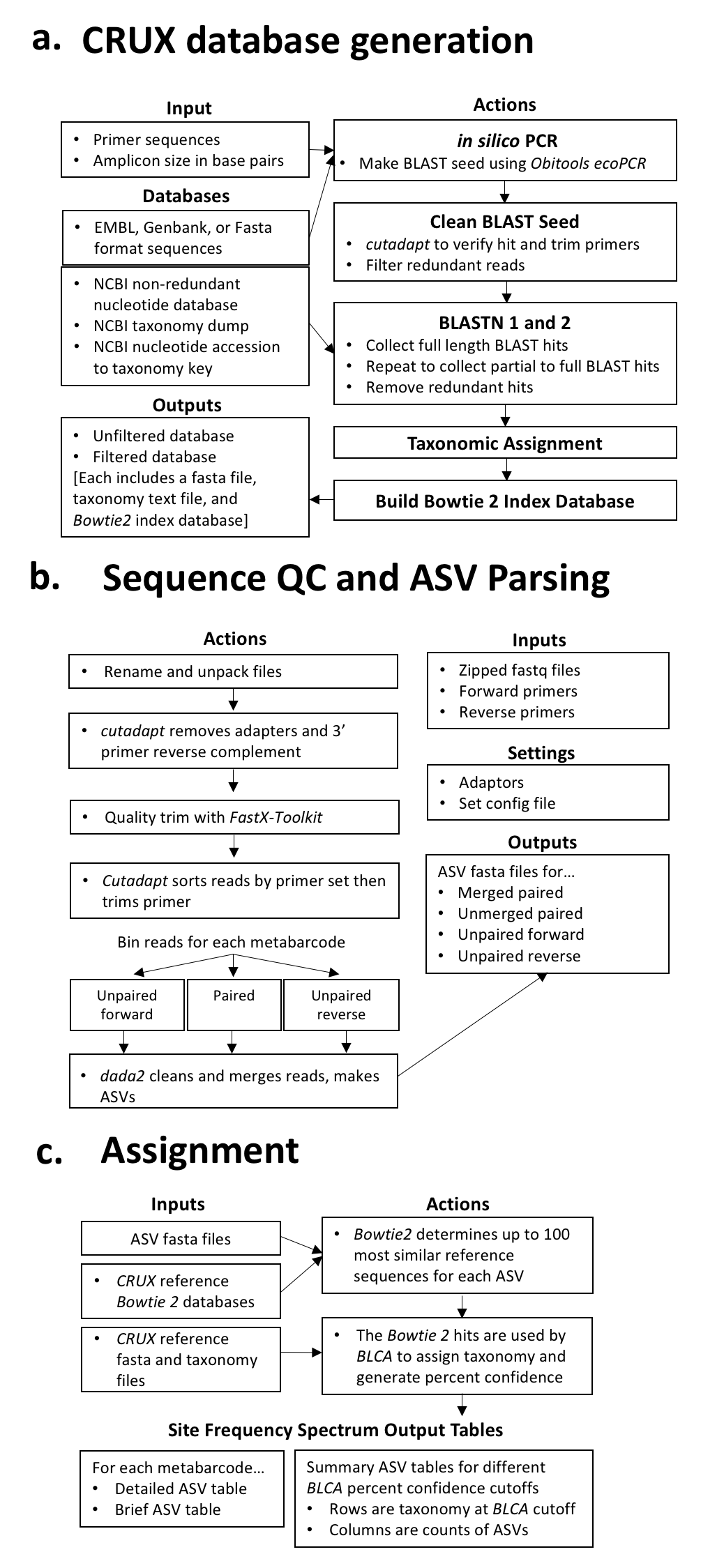
DRYAD

NCBI SRA

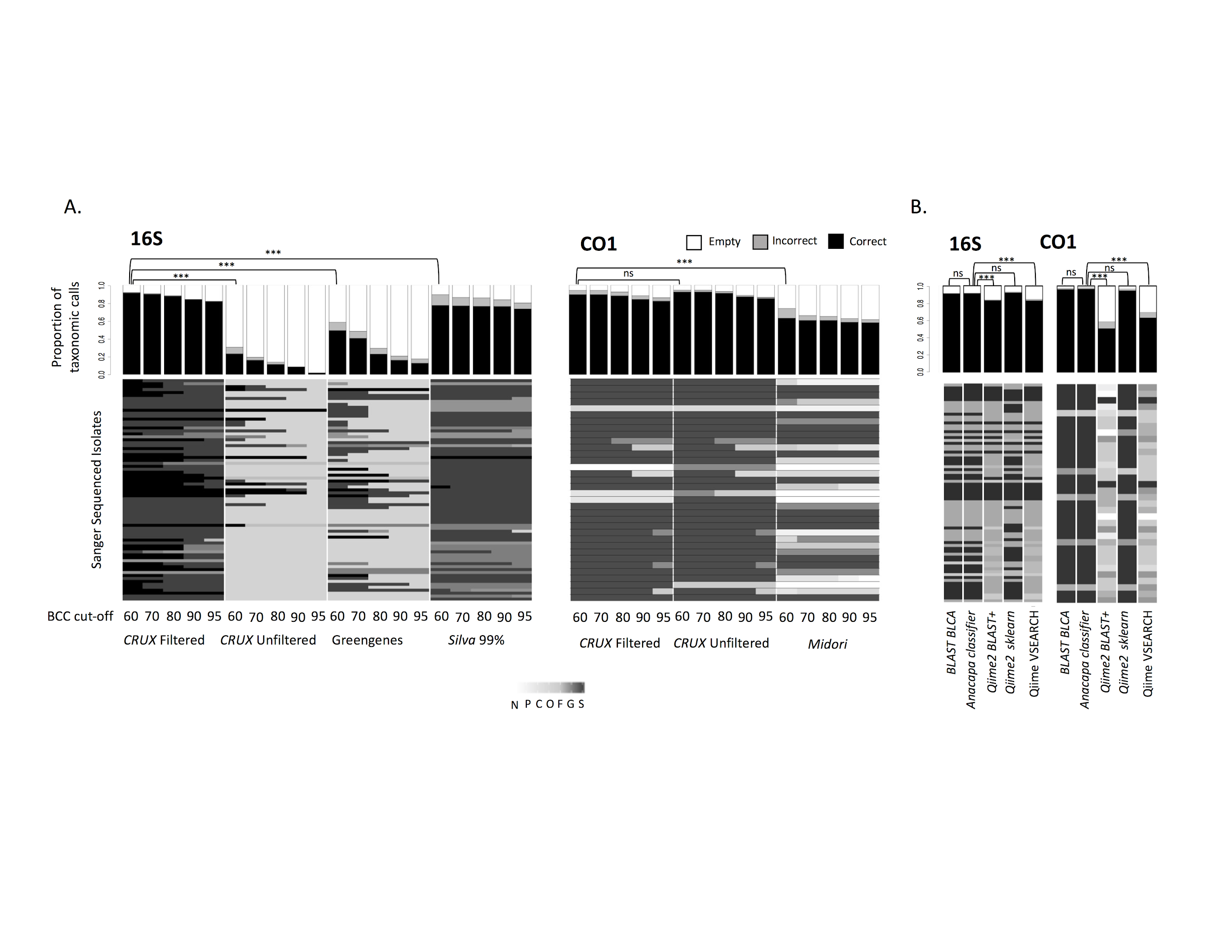
Github direct links

**Author Contributions**

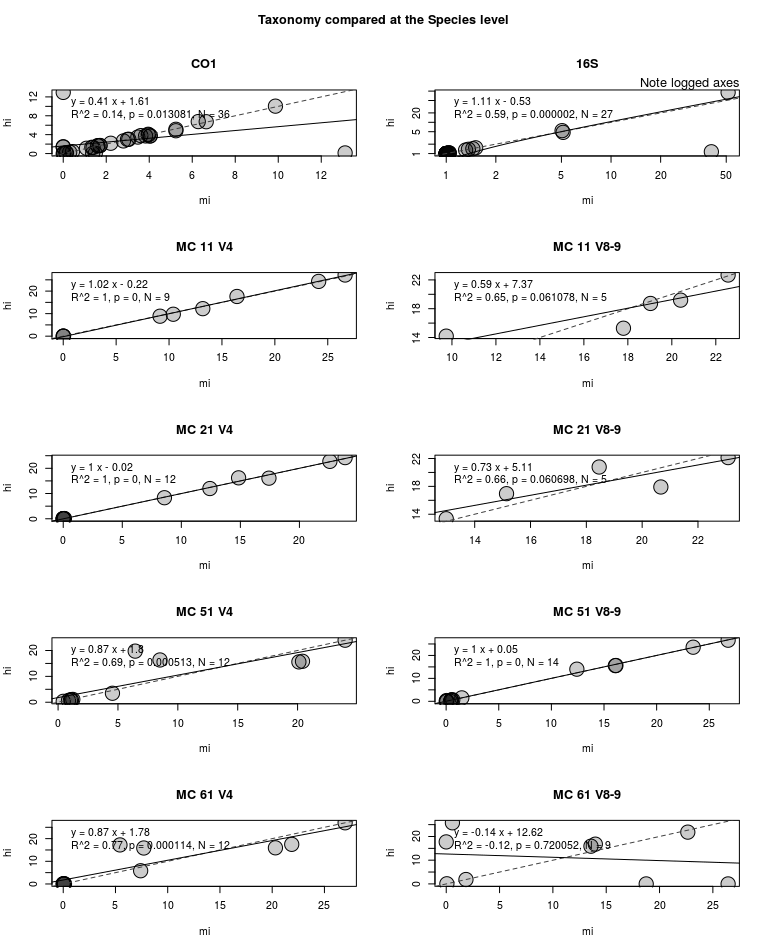
**Figure 1.** Flowchart of *Anacapa* tools. *Ranacapa* tool not included here.

****

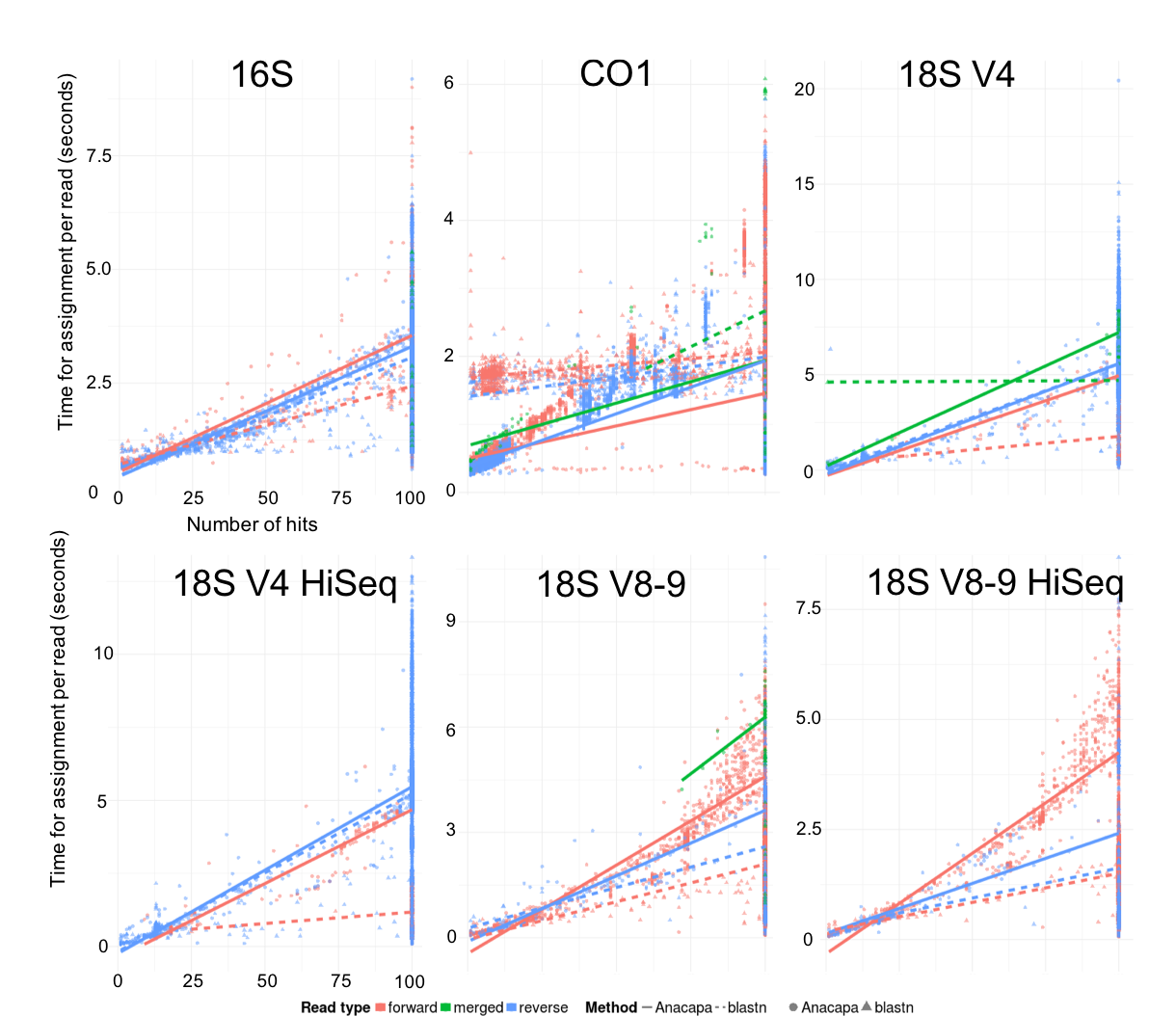
**Figure 2. a.** Database Comparisons for 16S and CO1 primer regions. 16S or CO1 reads generated by Sanger sequencing organisms were assigned taxonomy using *CRUX* filtered and unfiltered libraries, *Silva* 99% percent clustered database (16S), *Greengenes* database (16S), and the *Midori* CO1 database. **b.** The Anacapa classifier uses *BLCA* with a modification that implements *Bowtie 2*. Here, the Anacapa classifier is compared to alternative classifiers including the original BLCA BLAST classifier, the Bayesian *Qiime2* approach *sklearn*, and *Qiime2* consensus classifiers *BLAST+* and *Vsearch*. The bar charts indicate the proportion of times a step in the taxonomic path was identified correctly, incorrectly, or not at all. Assignments were limited in resolution to only path levels above a certain bootstrap of classifier confidence (BCC) cut-off (out of 100; here ‘cut-off’). The heat maps indicate the best resolution to correct taxonomy assigned to a given isolate for each database. The colors indicate taxonomy level: not correct (N) or assigned correctly to phylum (P), class (C), order (O), family (F), genus (G), or species (S). Z-scores between BCC 60 cut-off results are reported to a Bonferroni significance of \*\*\*=<0.001, ns = not significant



**Figure 3.** Expected 1:1 versus linear regressions of MiSeq vs pseudo-HiSeq read abundance for each unique taxon in a sample. R2 values show concordance. Each graph includes a dotted line with a slope of 1 that would indicate where a sample would land if proportional reads were assigned to the same full taxonomic path.

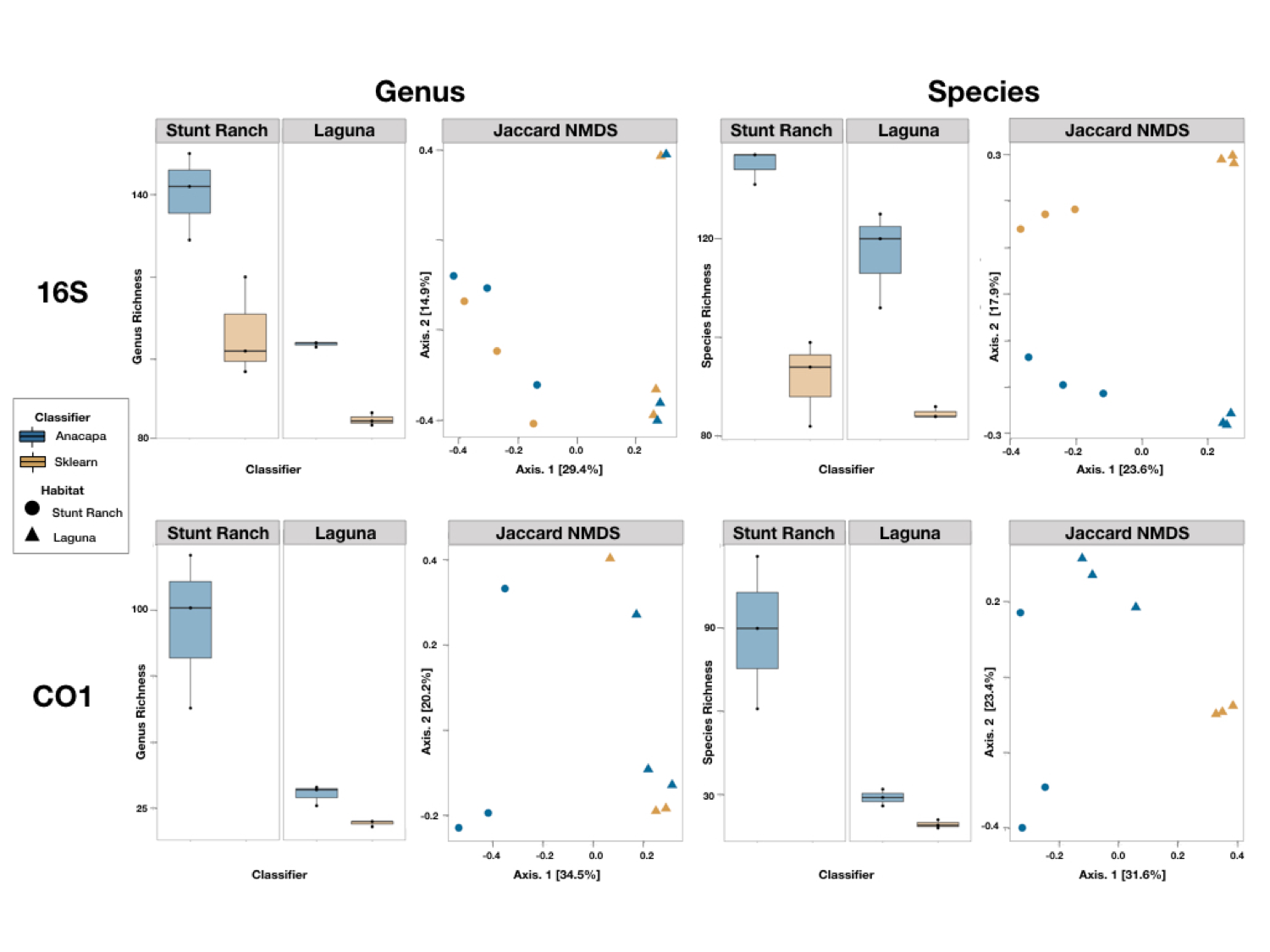
****

**Figure 4.** Differences in sequence handling times for taxonomic assignment for BLCA methods comparison between the Anacapa classifier (solid lines, circles) and the original BLAST BLCA method (dotted lines, triangles). Time benchmarking was run on four Miseq (16S, CO1, 18S V4 and V8-9) and two pseudo-HiSeq (18S V4 and V8-9) mock datasets using forward (red), merged (green), and reverse (blue) ASV types. 'Number of hits' refers to the number of query alignments generated by the *Anacapa classifier* or *BLAST BLCA*.

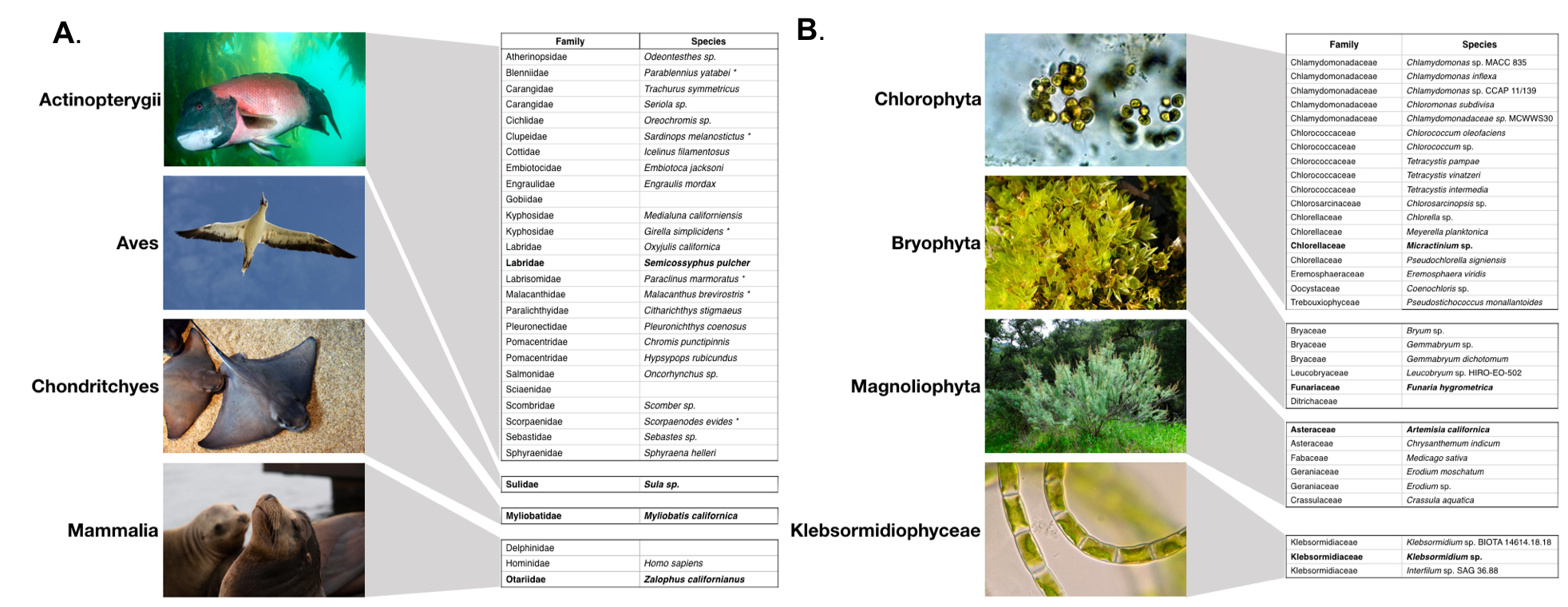
****

**Figure 5.** Comparison of *Anacapa* and *sklearn* classifiers on California eDNA Samples.

Here we compare the *Anacapa* and *sklearn* classifiers on three eDNA soil samples collected from Stunt Ranch Nature Reserve and the Laguna Coast Wilderness Park for two amplicon markers 16S and CO1. We found that the *Anacapa* *classifier* identifies significantly more genera and species level assignments than the *sklearn* classifier for the 16S marker (ANOVA, Genera: p <0.034, Species: p<0.009) (Supplemental Results).

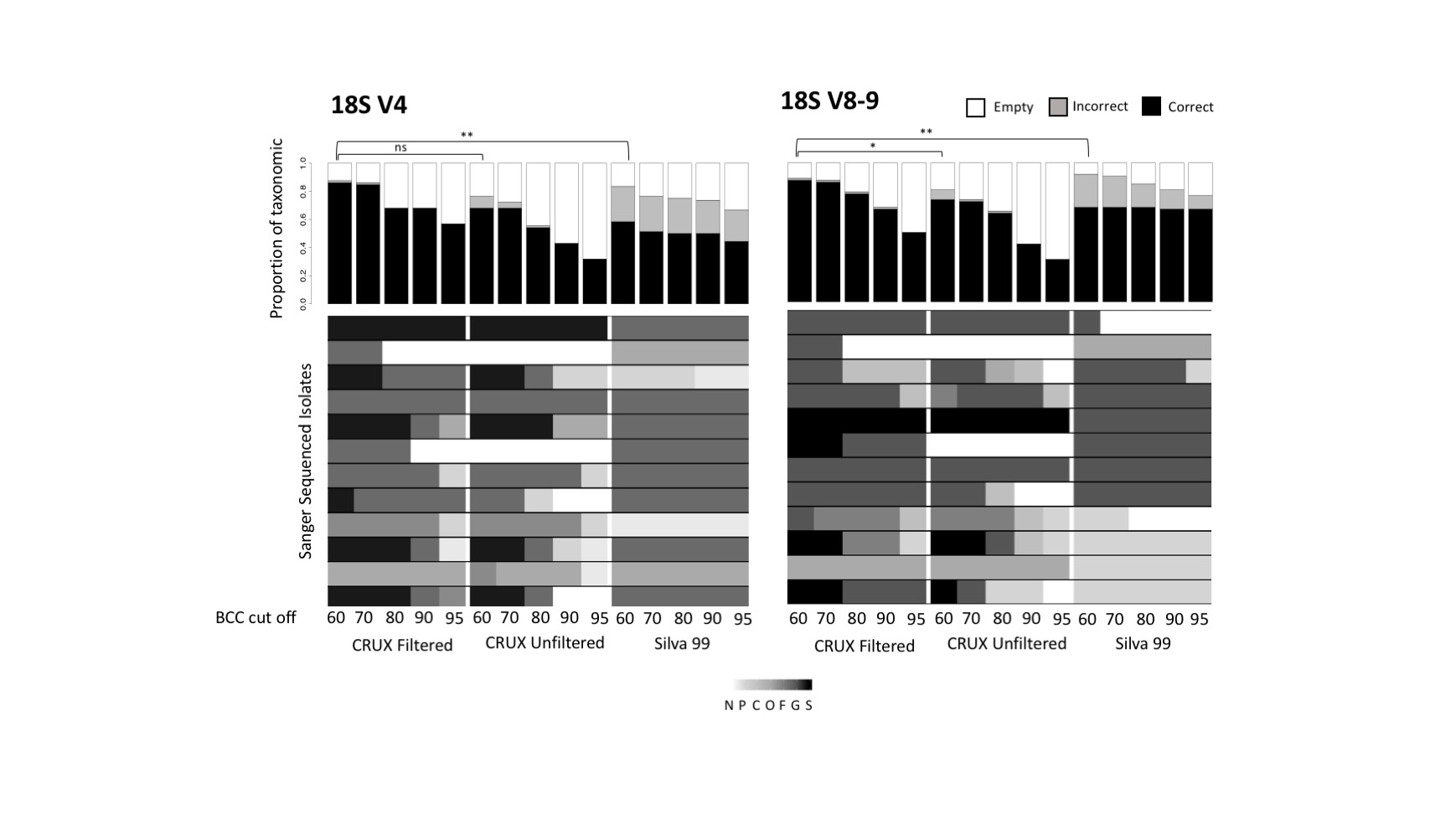


**Figure 6. a.** Highlight of ASV assignments from California environmental samples. Taxonomic level reported is limited to a BCC at over greater than 60% confidence in assignment. Bolded names are featured in images to the left. **a.** Anacapa Island vertebrates assigned to ASVs from 12s marker metabarcoding on three samples of 1L water collected at 10m depth. eDNA libraries were sequenced on an Illumina 2x150bp HiSeq.While most species are expected to be found in the environment sampled, some species marked with \* are not native, naturalized, or reported exotics in California. These are examples for which native species within the same genus have no NCBI Genbank reference sequences, suggesting the taxonomy is the closest sister species sequenced, not the actual species. **b.** Plants and algae assigned to ASVs from plant ITS2 marker metabarcoding on three 0.25 g surface soil samples from Laguna, CA, which is a coastal sage scrub habitat characterized by the presence of *Artemisia californica*.

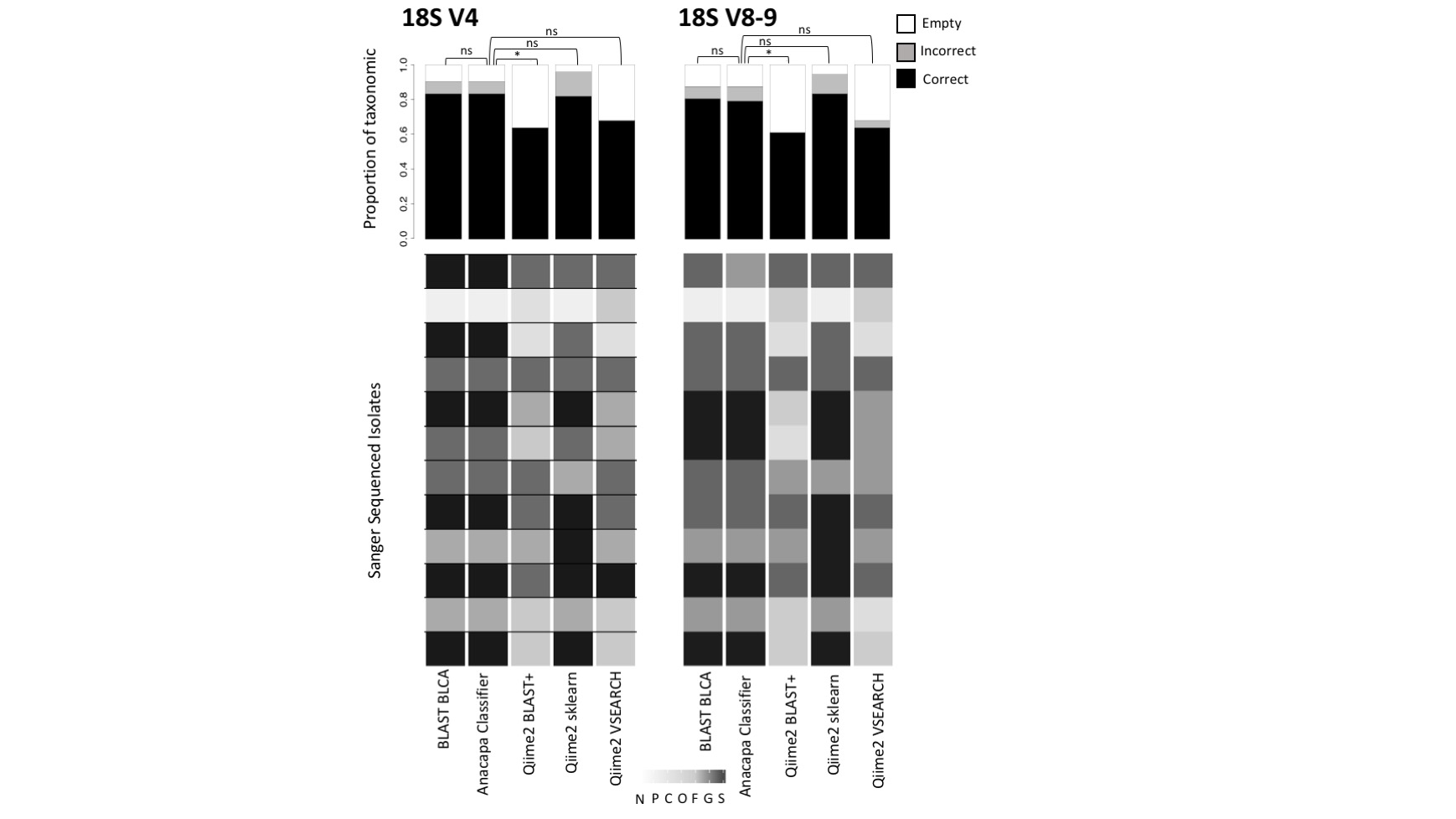


**Tables and Figures (with captions or the legend text)**

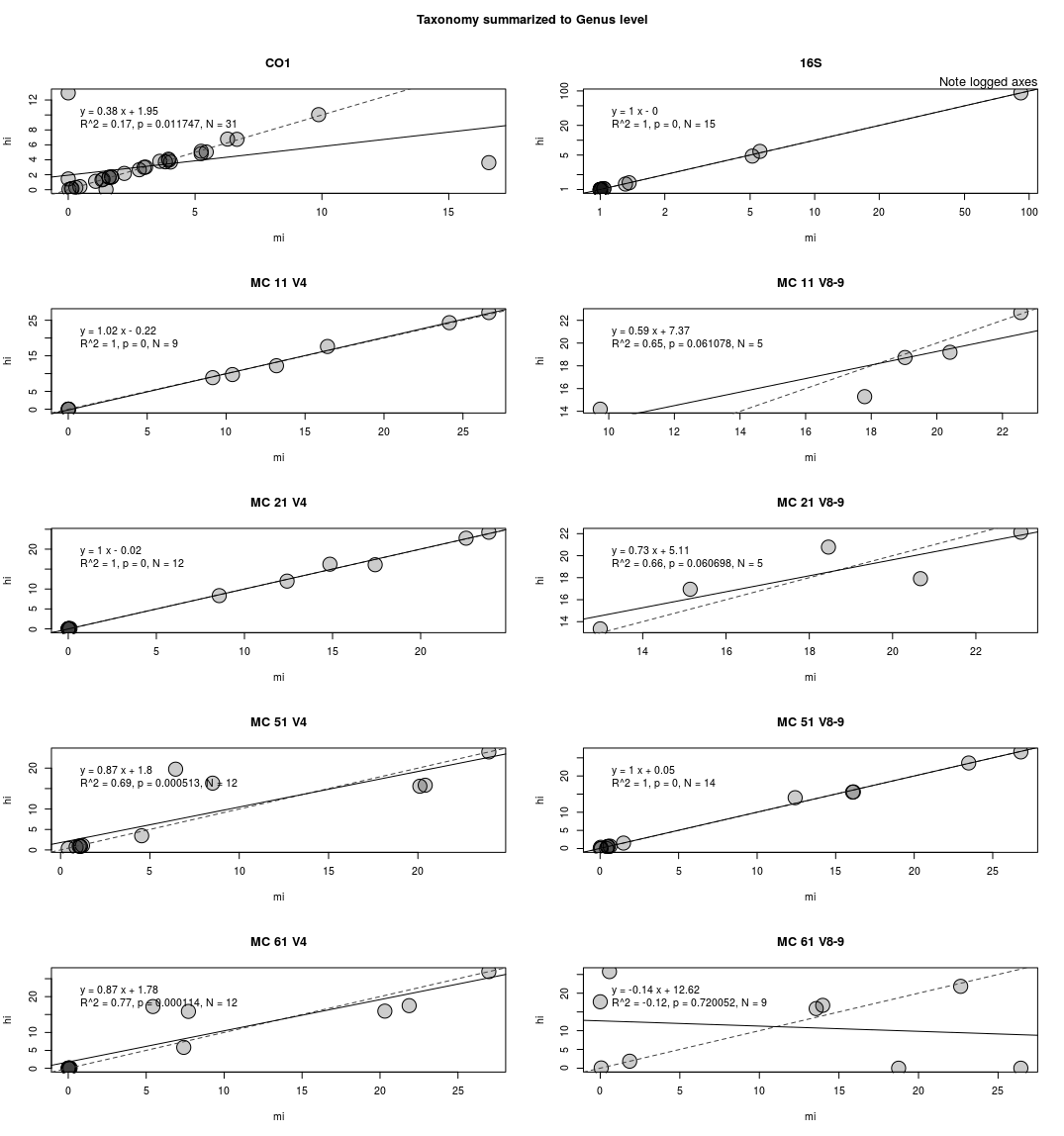
**Supplementary Figure 1**. Comparisons for 18S V4 and 18S V8-9 primer regions between CRUX filtered and unfiltered 18S libraries made for each region, and the Silva’s percent clustered library for 12 Sanger sequenced Isolates. The bar charts indicate the proportion of times a step in the taxonomic path was identified as correct, incorrect, or empty across all isolates at a given bootstrap of classifier confidence (BCC) cut-off (out of 100 bootstraps). The heat maps indicate the highest correct taxonomy assigned to a given isolate for each database, and at a given BCC. The colors indicate that taxonomy was not correct (N) or assigned correctly to phylum (P), class (C), order (O), family (F), genus (G), or species (S).



**Supplementary Figure 2.** Differences between the original *BLAST BLCA* algorithm, the Anacapa classifier, and classifiers implemented in Qiime2 (sklearn, BLAST+, and Vsearch). The bar charts indicate the proportion of times a step in the taxonomic path was identified correctly for a given classifier. The heat maps indicate the highest correct taxonomy assigned to a given isolate for each classifier. The colors indicate that taxonomy was not correct (N) or assigned correctly to phylum (P), class (C), order (O), family (F), genus (G), or species (S).



**Supplementary Figure 3.** Expected 1:1 versus linear regressions of MiSeq vs pseudo-HiSeq read abundance for each unique taxon in a sample down to the level of genus. R2 values show concordance. Each graph includes a dotted line with a slope of 1 that would indicate where a sample would land if proportional reads were assigned to the same full taxonomic path.



**Supplementary Tables Index**

|  |
| --- |
| Table S1. Primers used to amplify eDNA in this study and used to generate CRUX reference databases published with this study in DRYAD.  Table S2. Number of reads and unique taxon paths (TP) included in the databases used for comparisons.  Table S3. Mock datasets of Sanger sequenced isolates used (Isolates...) for tests shown in Fig 2 and Supplemental Figs 1 and 2. MiSeq pools of mock communities used in benchmarking tests are also shown (Miseq…).  Table S4. 16S V4 taxonomic assignment for reads generated from known Sanger sequence mock data from different databases, CRUX Filtered and Unfiltered libraries, Silva 99%, and Greengenes 2011 libraries, using assignments at or above BCC 60% cut-off.  Table S5. CO1 taxonomic assignment for known sequences given different databases, CRUX Filtered and Unfiltered libraries, Silva 99%, and Greengenes 2011 libraries, using assignments at or above the BCC 60% cut-off.  Table S6. 18S V4 taxonomic assignment for known sequences given different databases, CRUX Filtered and Unfiltered libraries, Silva 99%, and Greengenes 2011 libraries, using assignments at or above BCC 60% cut-off.  Table S7. 18S V8-9 taxonomic assignment for known sequences given different databases, CRUX Filtered and Unfiltered libraries, Silva 99%, and Greengenes 2011 libraries, using assignments at or above BCC 60% cut-off.  Table S8. 16S V4 taxonomic assignment given BLAST BLCA, Anacapa classifier, and three of the Qiime2 classifiers, using the CRUX filtered reference library, reported at a summary taxonomy at or above the BCC 60% cut-off.  Table S9. 18S V4 taxonomic assignment given BLAST BLCA, Anacapa classifier, and three of the Qiime2 classifiers, using the CRUX filtered reference library, reported at a summary taxonomy at or above the BCC 60% cut-off.  Table S10. 18S V8-9 taxonomic assignment given BLAST BLCA, Anacapa classifier, and three of the Qiime2 classifiers, using the CRUX filtered reference library, reported at a summary taxonomy at or above the BCC 60% cut-off.  Table S11. CO1 taxonomic assignment given BLAST BLCA, Anacapa classifier and three of the Qiime2 classifiers, using the CRUX filtered reference library, reported at a summary taxonomy at or above the BCC 60% cut-off.  Table S12. Taxonomy assigned for CO1 mock data set sequenced at MiSeq length (2 x 250) and truncated to HiSeq length (2 x 150) at summary taxonomy at or above the BCC 60% cut-off.  Table S13. Taxonomy assigned for 16S mock data set sequenced at MiSeq length (2 x 250) and truncated to HiSeq Length (2 x 150) at or above the BCC 60% cut-off.  Table S14. Taxonomy assigned for 18S V4 mock data set sequenced at MiSeq length (2 x 250) and truncated to HiSeq length (2 x 150) at summary taxonomy assignment at or above the BCC 60% cut-off.  Table S15. Taxonomy assigned for 18S V8-9 mock data set sequenced at MiSeq length (2 x 250) and pseudo-HiSeq length (2 x 150) at or above the BCC 60% cut-off.  Table S16. Comparison of the taxonomic assignment (BCC cut-off 60%) for the CO1 data set of 34 pooled \* species at MiSeq length and pseudo - HiSeq length.  Table S17. Comparison of the taxonomic assignment (BCC cut-off 60%) for the 16S data set of 35 pooled isolates at MiSeq length and pseudo-HiSeq length.  Table S18. Comparison of the taxonomic assignment (BCC cut-off 60%) for the 18S V4 data set of up to 12 pooled isolates (see supplementary methods) at MiSeq length and pseudo-HiSeq length.  Table S19. Comparison of the taxonomic assignment (BCC cut-off 60%) for the 18S V8-9 data set of up to 12 pooled isolates (see supplementary methods) at MiSeq length and pseudo-HiSeq length.  Table S20. Mock data from MiSeq runs was truncated to typical HiSeq length (2x150bp) and both MiSeq and pseudo-HiSeq reads were processed with Anacapa. Here are linear regressions depicting the concordance between MiSeq vs HiSeq length runs from the same samples. The relative abundance of the reads for a full taxonomic path assigned to them for MiSeq is plotted as the x-axis and for HiSeq is plotted as the y-Axis (see Fig 4).  Table S21. Significance of different BLCA method comparisons (Anacapa classifier method using Bowtie2 vs. blastn) on time for mock datasets.  Table S22. CALeDNA Sample data  Table S23. List of ASVs for taxonomic assignments (BCC cut-off 60%) for the 16S Stunt Ranch HiSeq samples and the Laguna Coast Wilderness Park MiSeq samples, displayed by processed read type.  Table S24. List of ASVs for taxonomic assignments (BCC cut-off 60%) for the CO1 Stunt Ranch HiSeq samples and the Laguna Coast Wilderness Park MiSeq samples, displayed by processed read type.  Table S25. Taxonomy table for 16S sample data after averaging 100 rarefactions to 8500 sequences.  Table S26. Taxonomy table for CO1 sample data after averaging 100 rarefactions to 19000 sequences.  Table S27. ANOVA tables of 16S and CO1 environmental DNA samples to compare species and genus richness between classifiers.  Table S28. PERMANOVA (Adonis) tables for 16S and CO1 environmental DNA samples to community composition between classifiers or habitats.  Table S29. Homogeneity of dispersion between classifiers or habitat for 16S and CO1 eDNA at the level of species or genus.  Table S30. Number of ASVs assigned to taxa in three Kelp Forest eDNA samples from CALeDNA.  Table S31. Number of ASVs assigned to taxa in three Laguna eDNA samples from CALeDNA.  Table S32. Number of ASVs assigned to taxa in three Stunt Ranch HiSeq eDNA samples from CALeDNA. |