SOFTWARE TOOL ARTICLE

ranacapa: An R package and Shiny web app to explore environmental DNA data with exploratory statistics and interactive visualizations [version 1; referees: 1 approved, 2 approved with reservations]

Gaurav S. Kandlikar [1], Zachary J. Gold[1], Madeline C. Cowen[1], Rachel S. Meyer[1], Amanda C. Freise[2], Nathan J.B. Kraft[1], Jordan Moberg-Parker [2], Joshua Sprague[3], David J. Kushner[3], Emily E. Curd[1]

[1]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, CA, 90095, USA
[2]Department of Microbiology and Microbial Genetics, University of California, Los Angeles, Los Angeles, CA, 90095, USA
[3]Channel Islands National Park, National Park Service, Ventura, CA, USA

v1 **First published:** 01 Nov 2018, **7**:1734 (https://doi.org/10.12688/f1000research.16680.1)
Latest published: 01 Nov 2018, **7**:1734 (https://doi.org/10.12688/f1000research.16680.1)

Abstract
Environmental DNA (eDNA) metabarcoding is becoming a core tool in ecology and conservation biology, and is being used in a growing number of education, biodiversity monitoring, and public outreach programs in which professional research scientists engage community partners in primary research. Results from eDNA analyses can engage and educate natural resource managers, students, community scientists, and naturalists, but without significant training in bioinformatics, it can be difficult for this diverse audience to interact with eDNA results. Here we present the R package ranacapa, at the core of which is a Shiny web app that helps perform exploratory biodiversity analyses and visualizations of eDNA results. The app requires a taxonomy-by-sample matrix and a simple metadata file with descriptive information about each sample. The app enables users to explore the data with interactive figures and presents results from simple community ecology analyses. We demonstrate the value of ranacapa to two groups of community partners engaging with eDNA metabarcoding results.

Keywords
environmental DNA, data visualization, citizen science, community science, shiny, metabarcoding, education, community ecology

This article is included in the RPackage gateway.

Open Peer Review

Referee Status: ? ? ✓

	Invited Referees 1	2	3
version 1 published 01 Nov 2018	? report	? report	✓ report

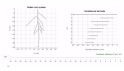
1 **Tristan Cordier**, University of Geneva, Switzerland

2 **Holly M. Bik**, University of California, Riverside, USA

3 **Niklaus J. Grünwald**, USDA Agricultural Research Service (United States Department of Agriculture), USA
Zach Foster, Oregon State University, USA

Discuss this article

Comments (0)



This article is included in the Interactive Figures collection.

Corresponding author: Gaurav S. Kandlikar (gkandlikar@ucla.edu)

Author roles: Kandlikar GS: Conceptualization, Data Curation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Gold ZJ**: Data Curation, Resources, Validation, Writing – Review & Editing; **Cowen MC**: Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Meyer RS**: Data Curation, Funding Acquisition, Project Administration, Resources, Validation, Writing – Review & Editing; **Freise AC**: Data Curation, Resources, Validation, Writing – Review & Editing; **Kraft NJB**: Funding Acquisition, Resources; **Moberg-Parker J**: Data Curation, Resources, Validation, Writing – Original Draft Preparation; **Sprague J**: Data Curation, Resources, Validation, Writing – Review & Editing; **Kushner DJ**: Data Curation, Resources, Validation, Writing – Review & Editing; **Curd EE**: Conceptualization, Data Curation, Project Administration, Resources, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: GSK and ZJG were supported by the US-NSF Graduate Research Fellowship [DEG No. 1650604]. NJBK was supported the National Science Foundation [DEB-1644641]. EEC, RSM, and the CALeDNA program are supported by the University of California President's Research Catalyst Award [CA-16-376437].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Kandlikar GS, Gold ZJ, Cowen MC *et al.* **ranacapa: An R package and Shiny web app to explore environmental DNA data with exploratory statistics and interactive visualizations [version 1; referees: 1 approved, 2 approved with reservations]** *F1000Research* 2018, **7**:1734 (https://doi.org/10.12688/f1000research.16680.1)

First published: 01 Nov 2018, **7**:1734 (https://doi.org/10.12688/f1000research.16680.1)

Introduction

The targeted amplification and sequencing of DNA that living organisms shed into their physical environment, termed "environmental DNA (eDNA) metabarcoding," is revolutionizing microbiology, ecology, and conservation research (Deiner *et al.*, 2017; Taberlet *et al.* 2012). Sequencing of eDNA extracted from field-collected soil, water, or sediment samples can yield insight into a range of questions, from profiling the composition of ancient plant and animal communities (Pedersen *et al.*, 2015), to monitoring populations of rare or endangered species (Balasingham *et al.*, 2018). As the cost of eDNA metabarcoding declines and sample collection techniques become more streamlined (e.g. Thomas *et al.* (2018)), professional research scientists are increasingly using eDNA metabarcoding as a platform to engage a diversity of community partners, including natural resource managers, undergraduate students, and citizen scientists in primary research. However, developing robust and impactful community science programs that engage community partners in all steps of the research process remains a challenge.

eDNA metabarcoding-based projects work well for programs that partner researchers with community scientists because non-experts can be quickly trained to collect samples in the field, and because eDNA metabarcoding is an exciting framework for research pertinent to disciplines such as medicine, agriculture, ecology, and geography (Deiner *et al.*, 2017). Community partners in such programs can have heterogeneous backgrounds, ranging from curious members of the public for whom collecting samples in the field is their first scientific research experience (e.g. University of California's CALeDNA program), to professional natural resource managers who regularly collaborate with research scientists (e.g. Center for Ocean Solutions' eDNA project). A key ingredient to promote sustained success of such programs is that community partners should be able to engage across multiple stages of the research project, not only in sample collection (European Citizen Science Association, 2015; Pandya, 2012). This can be a challenge for community science programs because although it is relatively easy to train community partners to collect eDNA samples, it is far more challenging to train them to independently visualize and analyze results from these studies. Indeed, learning the bioinformatic tools necessary for managing the large, multidimensional datasets generated in these studies can be difficult for professional researchers (Carey & Papin, 2018), let alone for the non-technical audience of some community science programs.

To address this challenge, we created the R package "`ranacapa`", at the core of which is a Shiny web app that can be used to visualize results from eDNA sequencing studies and perform simple community ecology analyses. `ranacapa` complements existing visualization platforms (e.g. Phinch (Bik & Phinch Interactive, 2014), Phyloseq-Shiny (McMurdie & Holmes, 2015), QIIME2 Viewer), because in addition to interactive visualizations, `ranacapa` includes brief explanations of several core analyses used in eDNA studies and includes links to additional educational resources. `ranacapa` works with community matrices generated via QIIME (Caporaso *et al.*, 2010) or the Anacapa sequence analysis pipeline, the latter being used extensively by the CALeDNA program.

Here, we describe the package and how it is used by two community science partnerships based at the University of California, Los Angeles (UCLA): first, a collaboration between eDNA researchers and resource managers at the National Park Service, and second, a partnership between community ecology researchers and an undergraduate microbiology course at UCLA. As we show in the Use cases, empowering community partners to interact with the data and perform simple but insightful community ecology analyses can help make these collaborations more enriching and valuable to both parties.

Implementation

At the core of `ranacapa` is a Shiny web app (Chang *et al.*, 2018), which is available at http://gauravsk.shinyapps.io/ranacapa or with `ranacapa::runRanacapa()`. The package also includes two categories of helper functions (Table 1) that transform user-uploaded taxonomy and metadata tables into `R` objects that can be visualized and analyzed using the Phyloseq (McMurdie & Holmes, 2013) and Vegan (Oksanen *et al.*, 2018) packages. `ranacapa` is available for installation from Github or CRAN:

```
devtools::install_github("gauravsk/ranacapa")
install.packages("ranacapa")
```

The `ranacapa` Shiny app allows users to interact with eDNA results through statistical summaries and interactive plots, displayed in the following tabs:

- **Sequencing depth**: Introduces the potential for variation in sequencing depth among samples and explains the basic logic behind rarefying samples in metagenomics studies (Figure 1). Users can rarefy the dataset

Table 1. Functions included within the ranacapa package.

Name	Description
scrub_seqNum_column	Removes any "xxx_seq_number" columns from the input taxonomy file if present (depends on which version of Anacapa was used to assign taxonomy)
scrub_taxon_paths	Replaces empty cells in input taxonomy tables with "Unknown"
validate_input_files	Verifies that the input taxonomy file and input mapping file meet specifications
convert_biom_to_taxon_table	Converts a phyloseq-imported biom table into an Anacapa-formatted taxonomy table
group_anacapa_by_taxonomy	Summarizes a site-abundance table from the Anacapa pipeline to each unique taxon
categorize_continuous_vector	Categorizes a continuous vector into low, medium, and high
convert_anacapa_to_phyloseq	Converts a site-abundance table from the Anacapa pipeline and the associated metadata file into a phyloseq object
vegan_otu	Creates a community matrix in the vegan package style using a phyloseq object and an otu_table object
custom_rarefaction	Rarefies a phyloseq object to a custom sample depth and with a given number of replicates
pairwise_adonis[1]	Wrapper function for multilevel pairwise comparison
ggrare[2]	Makes a rarefaction curve using ggplot2
runRanacapaApp	Runs the ranacapa Shiny app with tabs for interactive visualizations and statistical analyses

[1] adopted from https://github.com/pmartinezarbizu/pairwiseAdonis (GPL-3 License)

[2] adopted from https://github.com/mahendra-mariadassou/phyloseq-extended (GPL-3 License)

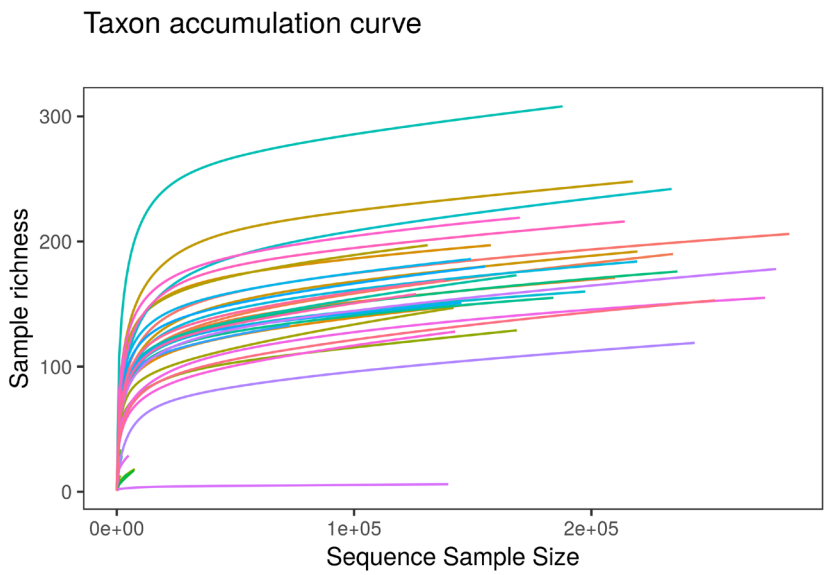


Figure 1. Taxon accumulation curve as shown in the ranacapa Shiny app. The online version of this figure is interactive.

to a sampling depth, or can proceed through the rest of the app without rarefying samples. The documentation acknowledges recent disagreement regarding the value of rarefying in metabarcoding and eDNA sequencing studies (McMurdie & Holmes, 2014).

- **Taxonomy heatmap**: Shows the taxon-by-sample matrix as an interactive heatmap made using `heatmaply::heatmaply()` (Galili *et al.*, 2018), where the color of each cell represents the number of times a given taxon was sequenced in a sample (Figure 2). Users can filter the taxon list by selecting or deselecting specific taxa.

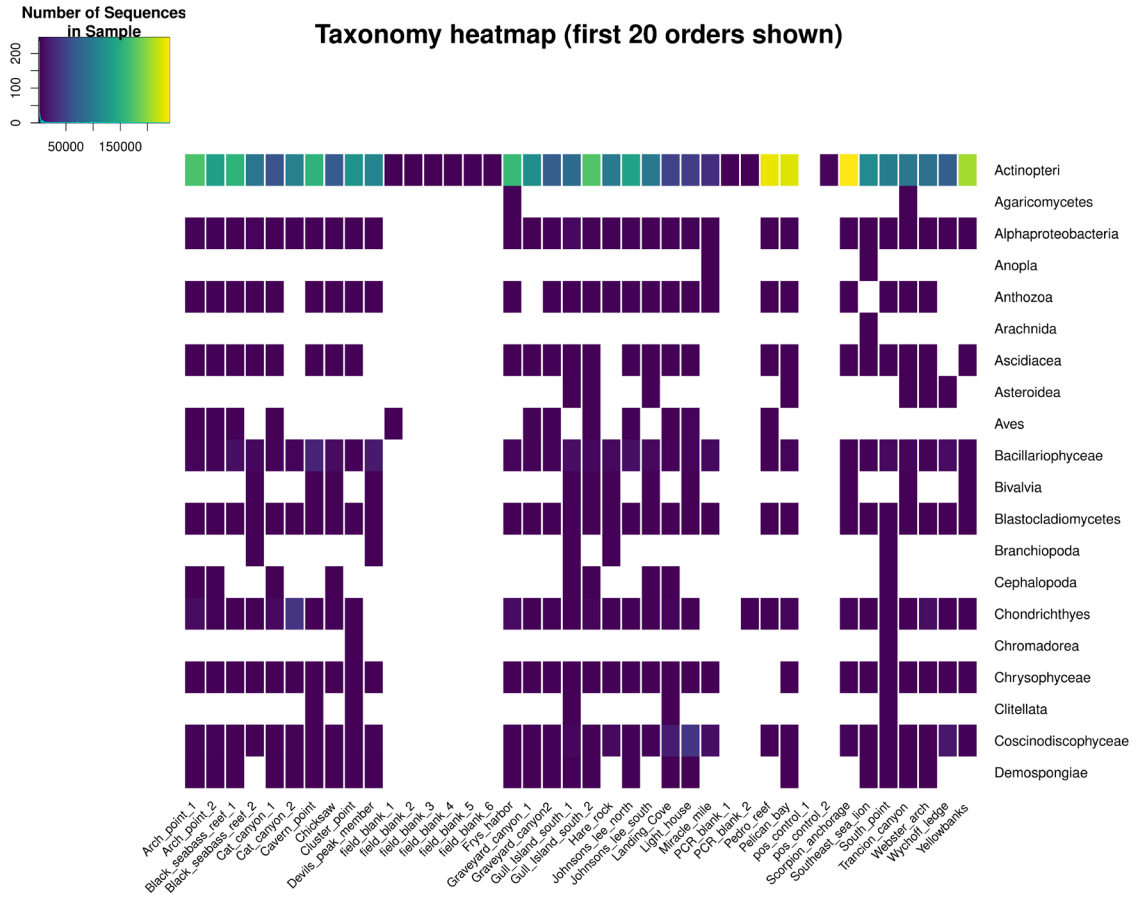


Figure 2. Taxonomy heatmap as shown in the `ranacapa` Shiny app. Taxonomy is shown at the Order level in this figure; in the app, users can choose the taxonomic level to show in the heatmap. Users can also select or deselect individual taxa to be shown in the heatmap. The online version of this figure is interactive.

- **Taxonomy barplot**: Shows the taxonomy-by-sample matrix as an interactive barplot (Figure 3).

- **Alpha diversity plots**: Introduces the concept of alpha diversity as the local diversity measured in a single habitat or sample. Users can plot alpha diversity as observed taxon richness or as Shannon diversity per sample, or can group samples according to a variable in the metadata file (Figure 4).

- **Alpha diversity statistics**: Allows users to choose a variable from the metadata, and generates an alpha diversity ANOVA table according to the user-selected variable. The tab also shows the output from a post-hoc Tukey test.

- **Beta diversity plots**: Introduces the concept of beta diversity as the turnover in species composition across habitats (or samples). The tab includes an ordination plot generated by `phyloseq::plot_ordination()`, which in turn uses an ordination object made with `phyloseq::ordinate(., method = "PCoA")`. Points on the PCoA plot are colored according to a user-selected metadata variable (Figure 5).

  The beta diversity plots tab also includes a dendrogram that groups sites based on Ward's cluster analysis (`stats::hclust(distance_object, method = "ward.d2")`), where `distance_object` is made using `phyloseq::distance()`. For both figures, users can toggle between using Jaccard and Bray-Curtis dissimilarity.

- **Beta diversity statistics**: Shows results from two statistical tests of species turnover across sites. The first test is a multivariate ANOVA of taxon turnover across sites, implemented with `vegan::adonis()`. The

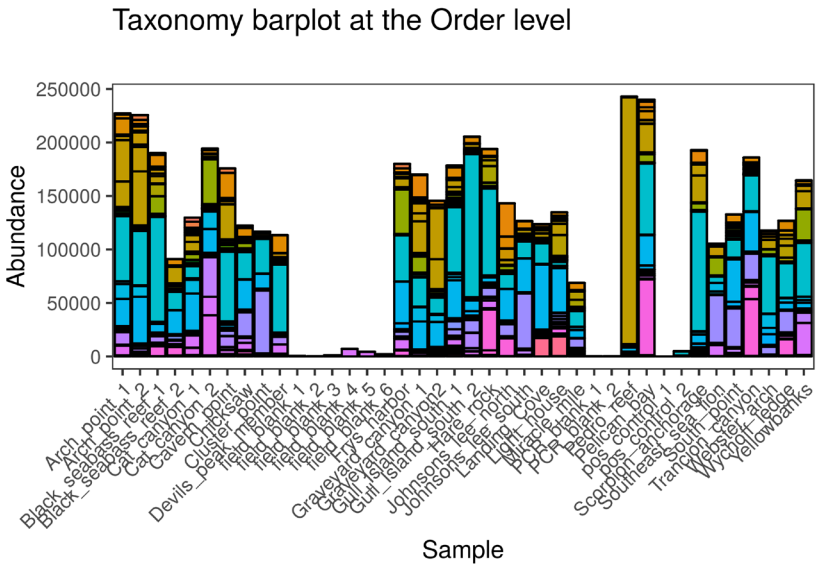


Figure 3. Taxonomy barplot as shown in the `ranacapa` Shiny app. Taxonomy is shown at the Order level in this figure; in the app, users can choose the taxonomic level to show in the barplot. The online version of this figure is interactive.

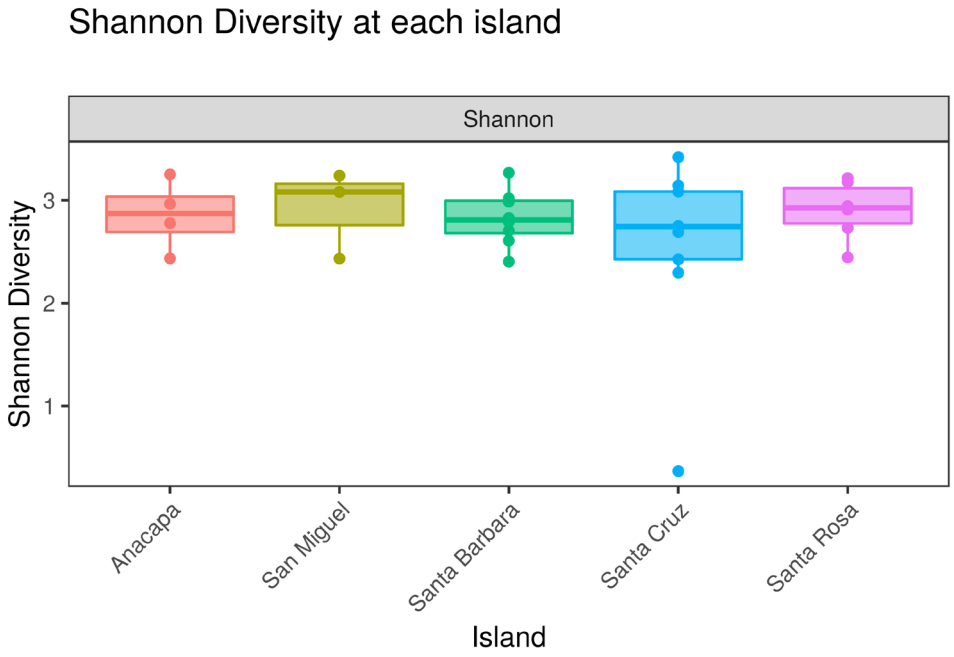


Figure 4. Alpha diversity boxplots as shown in the `ranacapa` Shiny app. Users can select the X-axis variable using a dropdown menu in the app. The online version of this figure is interactive.

second statistical test, which is implemented with `vegan::betadisper()`, is of heterogeneity of variances among samples. This test compares the degree of sample-to-sample variation within habitats (or within other user-selected groups).

Operation

ranacapa depends on Bioconductor v 3.7, which in turn relies on R v 3.5.0. The Shiny app has been tested on Chrome and Firefox on Windows, Mac-OSX, and Ubuntu.

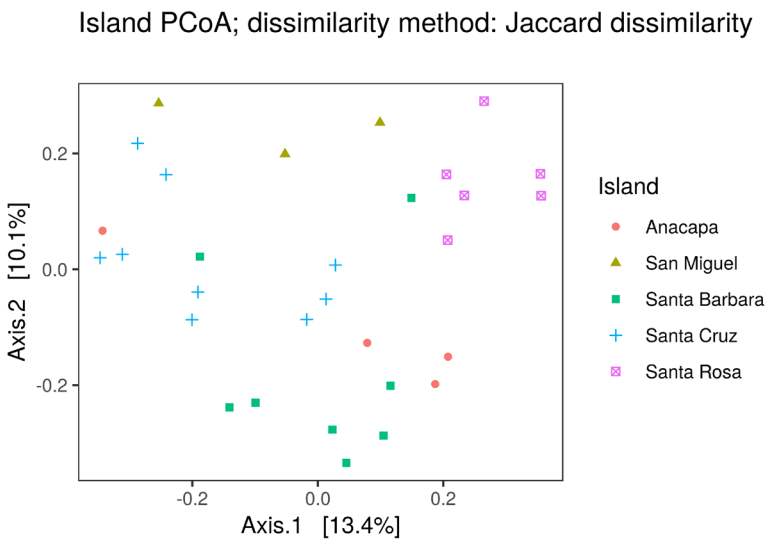


Figure 5. PCoA ordination of the samples as shown in the `ranacapa` Shiny app. Users can select the grouping variable with a dropdown menu in the app. The online version of this figure is interactive.

Input file structure

The *ranacapa* Shiny app requires two input files. The first requirement is a taxon-by-sample matrix, uploaded either as a rich, dense *.biom* table, or as a tab-separated *.txt* file. Qiime2-generated *.qza* files generated by QIIME2 are not immediately suitable for ranacapa, as they do not contain full taxonomy information. If the site-by-species matrix is uploaded as a *.txt* file, the file should match the specifications of the output files from the Anacapa eDNA sequence analysis pipeline. In Anacapa output, each row represents a taxonomic identification, and each column (save one) represents the number of times that taxon appears in each sequenced sample. One column, named *sum.taxonomy* must contain the taxonomic identification, with taxonomic rank separated by a semicolon, e.g. "*Chordata;Actinopteri;Chaetodontiformes;Chaetodontidae;Chaetodon; Chaetodon reticulatus*." A valid input file is structured as follows:

sum.taxonomy	Arch_point_1	Arch_point_2	Black_seabass_reef_1	Black_seabass_reef_2
<full path>	0	0	0	0
<full path>	0	0	43	87
<full path>	0	0	0	0
<full path>	0	0	0	0
<full path>	24	36	30	16
<full path>	0	0	0	0
<full path>	0	0	0	0
<full path>	0	0	16	177
<full path>	0	0	0	0
<full path>	0	0	0	0

The second requirement is a tab-separated *.txt* file that contains sample metadata. The first column in the metadata file should match the sample names in the taxonomy table; the remaining columns contain sample information for each of the samples in the taxon-by-site matrix. The metadata should contain categorical variables with two or more categories per variable. A valid metadata file for the taxonomy table above is structured as follows:

Sample	Sample_or_Control	Island	Protection	Locality
Black_seabass_reef_1	Sample	Anacapa	MPA	Black_seabass_reef
Arch_point_1	Sample	Santa Barbara	non-MPA	Arch_point
Arch_point_2	Sample	Santa Barbara	non-MPA	Arch_point
Black_seabass_reef_2	Sample	Anacapa	MPA	Black_seabass_reef

The `ranacapa` function `validate_input_files()` verifies that both the taxonomy table and the metadata files match structural requirements, which are documented in the function help files.

Use cases

We expect that researchers with expertise in bioinformatics will use the sequence analysis pipeline of their choice to assign taxonomy to eDNA datasets, and generate clean taxonomy and metadata files that can be visualized in `ranacapa`. Researchers can share these files with their partners, and emphasize the analyses or visualizations most appropriate to their use case. We now show how `ranacapa` can facilitate authentic communication between researchers and community partners in two settings.

Use case 1: Partnership between eDNA researchers and natural resource managers

A team of UCLA researchers partnered with resource managers at the Channel Islands National Park Service to assess the potential for eDNA as a biodiversity monitoring tool to supplement time-intensive visual biodiversity surveys in the Southern California Channel Islands (Deiner *et al.*, 2017; Lessios, 1996; Usseglio, 2015). For this partnership, resource managers collected and filtered 30 unique one-liter water samples for eDNA analysis at permanent monitoring sites inside and adjacent to protected areas, and research scientists at UCLA performed eDNA sequencing of the mitochondrial 12S (Miya *et al.*, 2015) and CO1 (Leray *et al.*, 2013) genes, targeting bony fishes, elasmobranches, and invertebrate taxa. The researchers processed sequences and assigned taxonomy using the Anacapa pipeline, and shared results with the resource managers using the `ranacapa` Shiny app.

The taxonomy heatmap of species detected using the 12S and CO1 metabarcodes (Figure 2) was the most valuable visualization to this collaboration, because it allowed the resource managers to filter the large observed species list down to a particular set of key taxa that they regularly monitor. The heatmap showed that this pilot study detected 36 of the 70 key metazoans at the species level, and the remaining 34 at the genus, family, or order level. This indicates that eDNA-based studies can provide critical information for ongoing management efforts and provide new insights into the spatial and temporal distributions of these species. The value of `ranacapa` in this scenario was to quickly sort through long species lists generated by eDNA sequencing to highlight the strengths and weaknesses in using eDNA to monitor diversity in the Channel Islands. The data from this study are packaged as the demo dataset for the `ranacapa` Shiny app and are available online (Kandlikar *et al.*, 2018a).

Use case 2: Partnership between eDNA researchers and an undergraduate microbiology course

A team of community ecology and environmental DNA researchers in the CALeDNA program collaborated with instructors of a research-based environmental microbiology course at UCLA (Shapiro *et al.*, 2015), in which students used eDNA metabarcoding to study the impact of a local wildfire on the plant and soil microbial community. The goal of this twenty-week course was to provide undergraduate students an authentic experience in basic microbiology and microbial community ecology research. Over the first ten weeks, eDNA researchers on the instructional team sequenced the ITS2 (Gu *et al.*, 2013) and 16S SSU RNA (Caporaso *et al.*, 2012) metabarcoding regions from student-collected soil samples and used the Anacapa pipeline to generate taxon-by-sample tables.

The course instructors used the `ranacapa` Shiny app to introduce students to the structure of eDNA sequencing results. The students were encouraged to explore data and perform the statistical analyses most pertinent to the hypotheses they had formed at the beginning of the course. A key benefit of using `ranacapa` was that despite having no prior bioinformatics experience, students could begin exploring the biodiversity in their samples in a matter of minutes by using the online instance of the Shiny app. This allowed the instructors to focus classroom time on biological questions rather than on troubleshooting bioinformatics problems, as had been the case in previous sessions of the course. The course instructors noted that visualizing eDNA data in `ranacapa` helped students understand the relationships between taxon-by-site matrices and the various metadata they had collected in the field. By significantly reducing the time and difficulty in visualizing basic biodiversity patterns, `ranacapa` helped students develop and pursue more sophisticated analyses during the remainder of the course, using tools such as STAMP (Parks *et al.*, 2014) and PICRUSt (Langille *et al.*, 2013). The taxonomy tables and metadata files used in this course are available online (Kandlikar *et al.*, 2018b).

Summary and future directions

Metabarcode sequencing of environmental DNA is becoming a key tool in a wide variety of ecological studies, and results from these studies are of interest to a broad audience. Our `R` package and Shiny app `ranacapa` helps users conduct exploratory analyses and visualizations on eDNA datasets, and is a step toward more fully engaging participants in all phases of eDNA sequencing-based community science projects.

We propose three avenues for future work with ranacapa. First, we plan to use `ranacapa` as the primary tool to present eDNA results from hundreds of samples sequenced by the CALeDNA community science program. Second, `ranacapa` is being integrated into the upcoming undergraduate curriculum module "Pipeline for Undergraduate Microbiome Analysis", which will be an open-source, comprehensive suite of analysis and data visualization tools for undergraduate researchers. Finally, in the long-term, we believe there is great promise in linking `ranacapa` with packages that connect with APIs of online biodiversity databases (e.g. Taxize (Chamberlain & Szöcs, 2013), rinat (Barve & Hart, 2017)). This will help users explore a much wider range of biodiversity questions, for example, by programmatically asking whether their samples include invasive species that are absent from other nearby sites. In sum, tools like `ranacapa` that allow non-technical audiences to easily interact with results from eDNA sequencing studies have great potential to engage community partners with a wide range of backgrounds and interests in primary research.

Software availability

- A Shiny app, including a dataset generated for demonstrations, is available at https://gauravsk.shinyapps.io/ranacapa
- `Source code is available from GitHub:` https://github.com/gauravsk/ranacapa
- Archived source code at time of publication: http://doi.org/10.5281/zenodo.1464285 (Kandlikar & Cowen, 2018)
- Software license (GPL-3)

Data availability

Datasets used for the Use cases are available from Figshare:

Dataset 1: Taxon table and metadata file for Channel Islands eDNA samples (mitochondrial 12S and CO1 metabarcodes sequenced) https://doi.org/10.6084/m9.figshare.7199477.v1 (Kandlikar *et al*., 2018a)

Dataset 2: Taxon table and metadata file for Santa Monica Mountains eDNA samples (16S and plant-ITS metabarcodes sequenced) https://doi.org/10.6084/m9.figshare.7199510.v1 (Kandlikar *et al*., 2018b)

Both datasets are available under a CC-BY 4.0 license

Grant information

GSK and ZJG were supported by the US-NSF Graduate Research Fellowship [DEG No. 1650604]. NJBK was supported the National Science Foundation [DEB-1644641]. EEC, RSM, and the CALeDNA program are supported by the University of California President's Research Catalyst Award [CA-16-376437].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We thank Sabrina Shirazi, Rachel Turba, Chris Dao, and Keith Mitchell for providing feedback on developmental versions of this package. We also thank Mahendra Mariadassau and Pedro Martinez Arbizu for making the `phyloseq-extended` and `pairwiseAdonis` packages openly available with a GPL-3 License.

References

Balasingham KD, Walter RP, Mandrak NE, *et al.*: **Environmental DNA detection of rare and invasive fish species in two Great Lakes tributaries.** *Mol Ecol.* 2018; **27**(1): 112–127.
PubMed Abstract | Publisher Full Text

Barve V, Hart E: **Rinat: Access iNaturalist data through apis.** 2017.
Reference Source

Bik, Phinch Interactive: **Phinch: An interactive, exploratory data visualization framework for –Omic datasets.** *bioRXiv.* 2014.
Publisher Full Text

Caporaso JG, Kuczynski J, Stombaugh J, *et al.*: **QIIME allows analysis of high-throughput community sequencing data.** *Nat Methods.* 2010; **7**(5): 335–336.
PubMed Abstract | Publisher Full Text | Free Full Text

Caporaso JG, Lauber CL, Walters WA, *et al.*: **Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq**

platforms. *ISME J.* 2012; **6**(8): 1621–1624.
PubMed Abstract | Publisher Full Text | Free Full Text

Carey MA, Papin JA: **Ten simple rules for biologists learning to program.** *PLoS Comput Biol.* 2018; **14**(1): e1005871.
PubMed Abstract | Publisher Full Text | Free Full Text

Chamberlain SA, Szöcs E: **taxize: taxonomic search and retrieval in R [version 1; referees: 3 approved].** *F1000Res.* 2013; **2**: 191.
PubMed Abstract | Publisher Full Text | Free Full Text

Chang W, Cheng J, Allaire J, *et al.*: **Shiny: Web application framework for r.** 2018.
Reference Source

Deiner K, Bik HM, Mächler E, *et al.*: **Environmental DNA metabarcoding: Transforming how we survey animal and plant communities.** *Mol Ecol.* 2017; **26**(21): 5872–5895.
PubMed Abstract | Publisher Full Text

European Citizen Science Association: **Ten principles of citizen science.** 2015.
Reference Source

Galili T, O'Callaghan A, Sidi J, *et al.*: **heatmaply: an R package for creating interactive cluster heatmaps for online publishing.** *Bioinformatics.* 2018; **34**(9): 1600–1602.
PubMed Abstract | Publisher Full Text | Free Full Text

Gu W, Song J, Cao Y, *et al.*: **Application of the ITS2 Region for Barcoding Medicinal Plants of Selaginellaceae in Pteridophyta.** *PLoS One.* 2013; **8**(6): e67818.
PubMed Abstract | Publisher Full Text | Free Full Text

Kandlikar G, Cowen M: **gauravsk/ranacapa: First release of ranacapa (Version v1.0.0).** *Zenodo.* 2018.
http://www.doi.org/10.5281/zenodo.1464285

Kandlikar GS, Gold ZJ, Cowen MC, *et al.*: **Taxon table and metadata file for Channel Islands eDNA samples (mitochondrial 12S and CO1 metabarcodes sequenced).** *Figshare.* 2018a.
http://www.doi.org/10.6084/m9.figshare.7199477.v1

Kandlikar GS, Gold ZJ, Cowen MC, *et al.*: **Taxon table and metadata file for Santa Monica Mountains eDNA samples (16S and plant-ITS metabarcodes sequenced).** *Figshare.* 2018b.
http://www.doi.org/10.6084/m9.figshare.7199510.v1

Langille MG, Zaneveld J, Caporaso JG, *et al.*: **Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.** *Nat Biotechnol.* 2013; **31**(9): 814–821.
PubMed Abstract | Publisher Full Text | Free Full Text

Leray M, Yang JY, Meyer CP, *et al.*: **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.** *Front Zool.* 2013; **10**: 34.
PubMed Abstract | Publisher Full Text | Free Full Text

Lessios HA: **METHODS for quantifying abundance of marine organisms.** In: *Methods and techniques of underwater research.* (eds. Lang, M. & Baldwin, C.). American Academy of Underwater Sciences (AAUS), 1996; 149–157.
Reference Source

McMurdie PJ, Holmes S: **phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.** *PLoS One.* 2013; **8**(4): e61217.
PubMed Abstract | Publisher Full Text | Free Full Text

McMurdie PJ, Holmes S: **Waste not, want not: why rarefying microbiome data is inadmissible.** *PLoS Comput Biol.* 2014; **10**(4): e1003531.
PubMed Abstract | Publisher Full Text | Free Full Text

McMurdie PJ, Holmes S: **Shiny-phyloseq: Web application for interactive microbiome analysis with provenance tracking.** *Bioinformatics.* 2015; **31**(2): 282–283.
PubMed Abstract | Publisher Full Text | Free Full Text

Miya M, Sato Y, Fukunaga T, *et al.*: **MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species.** *R Soc Open Sci.* 2015; **2**(7): 150088.
PubMed Abstract | Publisher Full Text | Free Full Text

Oksanen J, Blanchet FG, Friendly M, *et al.*: **Vegan: Community ecology package.** 2018.
Reference Source

Pandya RE: **A framework for engaging diverse communities in citizen science in the US.** *Front Ecol Environ.* 2012; **10**(6): 314–317.
Publisher Full Text

Parks DH, Tyson GW, Hugenholtz P, *et al.*: **STAMP: statistical analysis of taxonomic and functional profiles.** *Bioinformatics.* 2014; **30**(21): 3123–3124.
PubMed Abstract | Publisher Full Text | Free Full Text

Pedersen MW, Overballe-Petersen S, Ermini L, *et al.*: **Ancient and modern environmental DNA.** *Philos Trans R Soc Lond B Biol Sci.* 2015; **370**(1660): 20130383.
PubMed Abstract | Publisher Full Text | Free Full Text

Shapiro C, Moberg-Parker J, Toma S, *et al.*: **Comparing the Impact of Course-Based and Apprentice-Based Research Experiences in a Life Science Laboratory Curriculum.** *J Microbiol Biol Educ.* 2015; **16**(2): 186–197.
PubMed Abstract | Publisher Full Text | Free Full Text

Taberlet P, Coissac E, Hajibabaei M, *et al.*: **Environmental DNA.** *Mol Ecol.* 2012; **21**(8): 1789–1793.
PubMed Abstract | Publisher Full Text

Thomas AC, Howard J, Nguyen PL, *et al.*: **ANDe: A fully integrated environmental DNA sampling system.** *Methods Ecol Evol.* 2018; **9**(6): 1379–1385.
Publisher Full Text

Usseglio P: **Quantifying reef fishes: Bias in observational approaches.** In: *Ecology of fishes on coral reefs.*(ed. Mora, C.). Cambridge University Press, 2015; 270–273.
Publisher Full Text

Open Peer Review

Current Referee Status: ? ? ✓

Version 1

Referee Report 19 December 2018

https://doi.org/10.5256/f1000research.18233.r40128

✓ **Niklaus J. Grünwald** [1], **Zach Foster** [2]

[1] Horticultural Crops Research Laboratory, USDA Agricultural Research Service (United States Department of Agriculture), Corvallis, OR, USA

[2] Deptartment of Botany & Plant Pathology, Oregon State University, Corvallis, OR, USA

The authors present the R package rancapa that implements an interactive shiny web app available at https://gauravsk.shinyapps.io/ranacapa for exploration of environmental DNA characterized by amplicon sequencing. This app facilitates many of the standard tests and plots done for metabarcoding data. The app requires two dataset to run, namely a taxonomy-by-sample matrix file and a simple metadata file with descriptive information on the actual environmental samples. This package is particularly useful to anybody interested in exploring biome diversity but lacking extensive computational and bioinformatics skills. We appreciate the focus on open-source code and accessibility, in particular testing the app with Ubuntu and using the color-blind friendly viridis color palette. There are some minor changes in the style of plots and wording we would like to see, but these are a matter of opinion and do not affect the validity of the results. The authors admirably point out flaws and potential pitfalls in the analysis that often go unmentioned (e.g. the drawbacks of rarefaction). Although there is always room for improvement, we feel this app is quite useful as is and it is described well in the associated article. We suggest many improvements that could be made, but most of these are intended as constructive suggestions.

Specific comments:

- keywords: include “R package”

Introduction

- The sentence “A key ingredient to promote sustained success of such programs is that community partners should be able to engage across multiple stages of the research project, not only in sample collection” is a bit awkward. Perhaps something like the following would be more clear: “Such programs are more likely to succeed if community partners are able to be involved in multiple stages of the research project, not only sample collection”
- End of paragraph 2: “let alone for the non-technical audience of some community science programs.” We suggest “audience” be replaced with “members”, “partners”, or “collaborators”, since they are involved in the project.

Implementation

- Paragraph 1, sentence 1: The command “ranacapa::runRanacapa()” does not work. Should it be “ranacapa::runRanacapaApp()”?
- Figure 4: It would be nice if the results of the Tukey’s HSD test be included on the graph using letter codes.

- Sequencing depth, sentence 2: “Users can rarefy the data set to a sampling depth, or can proceed through the rest of the app without rarefying samples.” It's nice to have the option, but if sequences are not rarefied, perhaps they should be converted to proportions? The taxon abundance plots (e.g. the heat map) is biased without some correction for sampling depth and proportions would avoid the loss of data associated with rarefaction. While converting to proportions will not remove sample-size bias for alpha and beta diversity analyses, it will remove the bias for taxon abundance analyses for the most part. The conversion to proportions could also be an option in just the taxon abundance analyses.
- Beta diversity plots, sentence 1: “Introduces the concept of beta diversity as the turnover in species composition across habitats (or samples).” The word “turnover” seems to imply a change over time, rather than a comparison between communities. Perhaps something like: “Introduces the concept of beta diversity as a way to compare community composition across habitats (or samples).”
- Beta diversity statistics, sentence 1: “Shows results from two statistical tests of species turnover across sites.” See comment above on “turnover”.

Suggestions on the software

- Exploring ranacapa at https://gauravsk.shinyapps.io/ranacapa shows the interactive tools. We suggest that screenshots from the actual shinyapp be presented as figures as they provide the context for exploring the data. For example under Figure 2, the legend states that users can also select or deselect individual taxa’ but the figure 2 in the manuscript does not show this. This is confusing.
- We also suggest giving explicit examples for executing the shiny app on a desktop maybe right after the installation instructions.

Data import

- The proper format for input data is described in the article, but it would be nice if it was repeated in the “data import” section of the app as well. You could also include tips to convert common formats to the needed format here.

Sequencing depth

- Second paragraph, first sentence: “In eDNA sequencing, such variation in how deeply a given sample is sequenced can happen for a variety for reasons-”. Should the “-” be a “:”?
- Third paragraph, first sentence: You use double quotes for “deeply” in the first paragraph, but single quotes for ‘rarefy’.
- Add proportions as an alternative to rarefaction instead of “none”?
- The rarefied samples plot is a good idea, but hard to read in its current state. Perhaps change the faceting settings so that there is one column of plots or have a option to choose the sample shown?
- Taxonomy bar pot: The number of similar colors make this hard to read without selecting individual taxa. We know this is a limitation of stacked barcharts, but it might make it more readable to sort the taxa by average abundance and then stagger the colors like so: For example, if there are 30 taxa, then the first 10 get colors 1, 4, 7, 10, 13, 16, 19, 22, 25, 28 (seq(1,30, 30 / 10)). Also have the most abundant taxa be the lowest on the stacked bars. That way adjacent colors can be differentiated and the colors of rare taxa won't be confused with the colors of the abundant taxa.

Taxonomy heatmap

- How about faceting by sample metadata or clustering samples by similarity? You could use the results of the Ward’s Hierarchical Clustering like you do in the beta diversity section.
- Can you make the height of the plot depend on the number of samples plotted? Also, making the “Select taxa to visualize” box taller would be nice.

Alpha diversity plots

- You use the term “species richness” without explicitly defining it, although you implicitly define it. It might be a good idea to point out that “observed” is the same as “species richness”.
- Plotting the results of the Tukey’s HSD test on this plot would be great. Actually this tab could probably be combined with the “alpha diversity stats” tab.
- The color and the different point shapes are not needed, but that’s just a question of style.

Alpha diversity stats

- You say “ANOVAs only work when each group is represented by a few samples”, which is great to point out. How about just not allowing for the ANOVA and Tukey’s HSD if there are not multiple samples per treatment?
- How about not showing the Tukey’s test result if the ANOVA p-value is more than 0.1? You do say “the following results are meaningful only if the ANOVA table above suggests that sites are in fact unequal in terms of their diversity”, but it's very tempting to look at the results of the Tukey’s HSD test anyway. Alternatively, you could add a warning above the Tukey’s HSD results if the ANOVA p-value is more than 0.1.
- How about using the DT package to make the Tukey’s HSD results table sortable, like you do for other tables? The first thing most people will want to do is sort by p-value.

Beta diversity plots

- You say “samples that are distant in this plot have very different species lists”. That is not exactly true. Samples that are more distant from each other are more different from each other, but they all might be quite similar in terms of the species present; The PCoA will always show differences, no matter how small.
- An explanation of how to interpret the percent of the variation explained by each axis would be nice.
- How about confidence interval ellipsis for the sample types?
- How about allowing exploration of other dimensions (maybe in a 3d plot as well)?
- In the clustering dendrogram, how about coloring sample names by sample type? Ideally, it would be the same colors used in the PCoA plot, so you would not need a second legend.

Beta diversity stats

- How about displaying the results of the Adonis analysis in a table instead of a printed data.frame, like how it is done in the “Homogeneity of Variances” analysis? Tables from the DT package would be nice, so we can sort by p-value.

Data export

- It would be nice to provide an option to export the data in the taxmap format from the taxa package [1]. Zach Foster can help implement this.

References

1. Foster ZSL, Chamberlain S, Grünwald NJ: Taxa: An R package implementing data standards and methods for taxonomic data.*F1000Res*. 2018; **7**: 272 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: R programing, biology, bioinformatics

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 14 December 2018

https://doi.org/10.5256/f1000research.18233.r40129

? **Holly M. Bik**
Department of Nematology, University of California, Riverside, Riverside, CA, USA

This article presents a new R software package and shiny app called "ranacapa", aimed at producing rapid visualization and exploration of eDNA metabarcoding datasets. Overall I found this software tool to be exciting and user friendly, and the outlined use cases provided good examples of the need for such tools amongst end users including researchers and undergraduate students. The R shiny app was easy to explore using the demo dataset, and the plots were responsive and visually useful.

When trying to install ranacapa in RStudio (via install.packages), I am consistently getting an error message stating that "package ‘ranacapa’ is not available (for R version 3.3.2)". Furthermore, ranacapa does not appear to be listed as a package currently available in CRAN (https://cran.r-project.org/web/packages/available_packages_by_name.html#available-packages-R). Further investigation suggests that this package requires R 3.5.0 or higher to run, but this is not obvious for end users with older versions of R installed. Is it possible to add back compatibility to older versions of R? Especially for end users using HPC facilities who may be unable to update the base R version on their cluster.

"Taxonomy-by-sample matrix" - since ranacapa uses a non-standard input (essentially an OTU table summarized by taxonomy), it would be useful to provide specific instructions and/or a conversion script where users can take their QIIME BIOM files or tab-delimited OTU tables and convert them into a format that is suitable for uploading into ranacapa. Users with minimal technical expertise may struggle to convert their data in the correct way. Especially useful would be R scripts or package commands (e.g. phyloseq) that will output taxonomy-by-sample-matrixes in the correct format.

The R shiny web app does not provide much information regarding file formatting (users must read the F1000 article for specific information). The only documentation online states that the imported files should be in .biom or .txt formats. What format(s) are supported for the header row (commented out vs. plain text)? Are there any limitations on sparse vs. dense or JSON vs. HDF5 BIOM files (e.g. from QIIME 1 or

QIIME2)? Detailed information and documentation should be provided on the website itself in addition to the F1000 article.

For the metdatata file - can users directly import QIIME-formatted mapping files (e.g. with a header row beginning with #SampleID that has been validated via Keemi)? Or does this file need to be be edited further for ranacapa?

For the demo file on the web server - what study did this file come from? It would be useful to have some contextual information about the file, eDNA markers used, questions/hypotheses, and study design so that end users can explore the demo dataset in a more thoughtful way.

Some components of the R Shiny web app do not always seem to function correctly over a slow internet connection - it seems to freeze up and require re-loading, and/or users are presented with an error message. This error message sometimes goes away and the plots pop up as expected after a few minutes. Better error dialogs or "please wait" message boxes would be much less frustrating for the end user to figure out what is happening.

Sequencing Depth - the rarefaction depth slider bar is straightforward and should be familiar to most users, however, what impact do multiple rarefactions have on the displayed graphs? Are the rarefaction curves averaged somehow before being displayed? Further information on this parameter would be useful to include in the explanatory text.

Taxonomy heatmap - when choosing a lower taxonomic level, the taxonomy strings are overlapping and unreadable. Zooming in cuts off some of the categories on the x-axis. It would be better to be able to expand the heatmap along the y-axis (stretch it out) so you can read the taxonomy labels while still being able to view all of the heatmap information.

Taxonomy heatmap - when selecting a taxon to visualize, users might want to see more detail at lower taxonomic levels, however, selecting a taxon to visualize and then choosing a different taxonomic rank currently removes this filter choice and reverts back to showing all taxa at that level.

Alpha Diversity Plots - what is the difference between Observed vs. Shannon diversity visualizations? The explanatory text only partially explains the difference between these two options, and could be clarified to explicitly state the important differences in these two calculations. Furthermore, what do high vs. low numerical values mean for each of the two indices, in biological terms (e.g. lower number of species)? These kind of diversity statistics may be familiar for ecologists and microbiome researchers, but are likely to be confusing for citizen scientists, students, or scientists outside of these fields.

Alpha Diversity Stats - Unless you have a strong statistical background, it is impossible to interpret the numbers and outputs listed in this tab in the ranacapa shiny app. More explanation is needed to guide users in what they should be looking for - what columns are most important in the ANOVA table to see if samples are in fact unequal in terms of their diversity? In the post-hoc tests, should users be looking for P values <0.05 in the last column?

Beta Diversity - the Jaccard and Bray-Curtis indexes could be explained more plainly. Jaccard is the simplest measure of shared taxa/OTUs, while Bray-Curtis is an index focused on the most abundant taxa/OTUs in the dataset. Also some explanation of the PCoA axes (e.g. that they explain variation but do not relate to metadata) would be useful. Adding the Canberra index might also be useful here, as this is another useful index which ignores the most abundant taxa and looks at differences across rare species.

Beta Diversity Stats - same comment as Alpha Diversity stats (see above). Needs more context and explanation to be accessible to end users without a statistical background.

Use Case 1 - how was the UCLA eDNA dataset filtered down to the set of key taxa that resource managers were looking for? Was this done via R command line, or via the R shiny app interface? Custom filtering is likely to be a common use case, and it would be valuable to add in more specific information regarding how this type of taxon filtering can be accomplished for any dataset.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: metabarcoding, genomics, data visualization

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 03 December 2018

https://doi.org/10.5256/f1000research.18233.r40402

? **Tristan Cordier**
Department of Genetics and Evolution, University of Geneva, Geneva, Switzerland

The manuscript “ranacapa: An R package and Shiny web app to explore environmental DNA data with exploratory statistics and interactive visualizations” is describing an R package to visualize metabarcoding data and perform summary statistics.
I think that such application will be very useful for quick results sharing within any project that involves stakeholders with various background. Too many projects still rely solely on a printed (or pdf) report with key chosen figures, orienting the readers towards the main messages. I do believe that such application

will provide a nice complement to a project report for people that want to dive into the results, or explore the robustness of the conclusion that are highlighted in a report.
I also agree with the authors that such tools are particularly interesting for sharing results with non-expert enthusiasts and students. I appreciate the accompanying text to present the specificity of composition (count) data and the possible flaws when analyzing such data. The website “Gusta Me” (https://mb3is.megx.net/gustame) could maybe be mentioned as an external resource for guiding further the users for the interpretation of results, or to guide any additional analysis.
I wonder how such application is intended to be deployed. Do the authors plan to let research scientists deploy the app within their own computational resources or do they plan to make the application accessible (https://gauravsk.shinyapps.io/ranacapa/) opened to the community?
I made few comments on the GUI that the authors may or may not follow. The most important point to address would be to allow users to write their PERMANOVA model.

Data import:
- Maybe add the expected format, with a toy example, indicating the separator character (tab or comma or semi-comma?). This is mentioned in the GitHub page of the project (not the separator though), but I think this would be nice to have it here also. My guess is that many users would not even get to the GitHub page before using the app.

Sequencing depth:
- Would it be possible to add a “free” input field? As it is now, it is not easy to get a precise value. The slider does not allow this kind of precision, it increases or decreases by 1000 but on the demo example, it is stuck to 10006.
- Would it be possible to include normalization as an option? At least the relative abundance for a start, because as you mention on the accompanying text, we are far from having reach a consensus on that matter.
- What means the first field “select the variable”, do you mean to rarefy not samples but something else? I don’t understand.

Taxonomy barplot:
- I personally don’t like taxonomic barplots, but I recognize that they can be useful to easily illustrate some obvious differences (or similarities) between experimental treatments to an audience unfamiliar with metabarcoding data. The lowest taxonomic ranks return a (way) too many categories. In the presented example. Even the highest taxonomic rank (phylum), it yields too many categories to easily visualize what is actually in the samples, because colors are too close to each other. There is no solution, I just don’t like these plots. But it is nice to have this option in the app anyway.

Taxonomy heatmap:
- Is there an option to scale differently the X and Y axis? In the example dataset, it is actually a similar problem as for taxonomy barplot, lowest taxonomic rank gives too many rows on the plots. A way to better see is to zoom in, but then the horizontal axis is quite messed up.
- Maybe it would be nice to have input fields to manage those axes and font size.

Alpha-diversity plots:
- I guess the data behind is the rarefied version of the dataset, but this should be clear. Maybe the box can be there, but in grey to show that it would not make sense to use the raw dataset?

Alpha-diversity stats: I think this is fine.
Beta-diversity plots:
- The users should be able to refine the ordination plots, its size, the size of each axis.
- Why the ordination plots configuration (and the cluster analysis) is changing when we select different variables to group the dot shapes and colors?

Beta-diversity analyses:

- Would it be possible to let the user write his own model? As for now, it is not possible to have an interaction term, or a nested model.
- To be able to select the sample as variable does not make sense.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Referee Expertise: Molecular ecology, environmental genomics, bioinformatics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.