ranacapa: An R package to explore environmental DNA data with exploratory statistics and interactive visualizations

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 Kushner (3), and Emily E. Curd (1)

### Affiliations and Contact Information

1. Department of Ecology and Evolutionary Biology, University of California – Los Angeles
2. Department of Microbiology and Molecular Genetics, University of California – Los Angeles
3. National Park Service

(+) Corresponding Author

* Email: [gkandlikar@ucla.edu](mailto:gkandlikar@ucla.edu)
* Phone: (+1) 952-288-7351
* Mailing Address: Dept. of Ecology & Evolutionary Biology, 621 Charles E. Young Drive S., Los Angeles, CA 90095

**Keywords**: environmental DNA; data visualization; citizen science; community science; shiny; metabarcoding; education; community ecology

## Abstract

Environmental DNA (eDNA) metabarcoding is becoming a core tool in ecology and conservation biology, and is being used in a growing number of community science programs that engage non-professional scientists 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 these analyses. 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 scientists engaging with eDNA metabarcoding results.

## 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 (Taberlet *et al.* 2012; Deiner *et al.* 2017). Sequencing of environmental DNA extracted from field-collected soil, water, or sediment samples can yield insight into a range of questions, ranging from profiling the composition of ancient plant and animal communities (Pedersen *et al.* 2014), to motoring populations of rare or endangered species (Balasingham *et al.* 2017). 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, <http://www.ucedna.com/>), to professional natural resource managers who regularly collaborate with research scientists (e.g. Center for Ocean Solutions’ eDNA project, <https://oceansolutions.stanford.edu/project-environmental-dna>). A key ingredient to promote sustained engagement of community partners in these programs is that they should be able to participate across multiple stages of the research project, not only in sample collection (Pandya 2012; European Citizen Science Association 2015). This can be a challenge for eDNA metabacoding-based 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 community science programs.

To address this challenge, we created an R package “ranacapa”, at the core of which is a Shiny webapp 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, Phyloseq-Shiny, QIIME2 Viewer), because in addition to interactive visualizations, ranacapa includes brief explanations of several core analyses used in eDNA studies core and includes links to additional educational resources. ranacapa works with community matrices generated via QIIME [Caporaso\_2010] or the Anacapa sequence analysis pipeline (<https://github.com/limey-bean/Anacapa>), which is used extensively by the CALeDNA program.

In the remainder of this manuscript, 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 webapp (Chang *et al.* (2018)), which is avialable 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 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(), 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.
* **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 as ordination made with phyloseq::ordinate(., method = "PCoA"). Points on the PCoA plot are colored according to a user-selected metadata variable (Figure 5).

This 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 site. The first statistical test is a multivariate ANOVA implemented with vegan::adonis()), which shows results from a statistical test of taxon turnover across sites. The second statistical test is of heterogeneity of variances among samples implemented with vegan::betadisper(), dthat 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.

### 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 (<https://github.com/limey-bean/Anacapa>). 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. A valid input file is structured as follows:

## sum.taxonomy  
## 413 Chordata;Actinopteri;Chaetodontiformes;Chaetodontidae;Chaetodon;Chaetodon reticulatus  
## 1133 NA;Florideophyceae;Ceramiales;Delesseriaceae;Membranoptera;Membranoptera platyphylla  
## 588 Chordata;Actinopteri;Lutjaniformes;Lutjanidae;Aphareus;Aphareus furca  
## 1269 NA;Oomycetes;Peronosporales;Peronosporaceae;Calycofera;Calycofera operculata  
## 1351 Phaeophyceae;NA;Ectocarpales;Chordariaceae;NA;Chordariaceae sp. 2 AP-2014  
## 66 Arthropoda;Insecta;Lepidoptera;NA;;  
## 759 Chordata;Actinopteri;Perciformes;Scorpaenidae;Sebastapistes;Sebastapistes tinkhami  
## 1282 NA;Oomycetes;Pythiales;Pythiaceae;Pythium;Pythium aff. diclinum  
## 792 Chordata;Actinopteri;Perciformes;Zaniolepididae;Zaniolepis;Zaniolepis latipinnis  
## 656 Chordata;Actinopteri;NA;Pomacanthidae;Pomacanthus;Pomacanthus semicirculatus  
## Arch\_point\_1 Arch\_point\_2 Black\_seabass\_reef\_1 Black\_seabass\_reef\_2  
## 413 0 0 0 0  
## 1133 0 0 43 87  
## 588 0 0 0 0  
## 1269 0 0 0 0  
## 1351 24 36 30 16  
## 66 0 0 0 0  
## 759 0 0 0 0  
## 1282 0 0 16 177  
## 792 0 0 0 0  
## 656 0 0 0 0  
## Cat\_canyon\_1  
## 413 0  
## 1133 0  
## 588 0  
## 1269 0  
## 1351 13  
## 66 0  
## 759 0  
## 1282 0  
## 792 0  
## 656 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  
## 16 Black\_seabass\_reef\_1 Sample Anacapa MPA  
## 17 Arch\_point\_1 Sample Santa Barabara unprotected  
## 18 Arch\_point\_2 Sample Santa Barabara unprotected  
## 19 Cat\_canyon\_1 Sample Santa Barabara unprotected  
## 35 Black\_seabass\_reef\_2 Sample Anacapa MPA  
## Locality  
## 16 Black\_seabass\_reef  
## 17 Arch\_point  
## 18 Arch\_point  
## 19 Cat\_canyon  
## 35 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 bioinformatic expertise 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 expensive and time-intensive visual biodiversity surveys in the Southern California Channel Islands (Lessios 1996; Usseglio 2015; Deiner *et al.* 2017). For this partnership, resource managers collected and filtered thirty-1L 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 and CO1 gene, 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 (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 likely supplement 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 in 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 web-app and are available online at XXX.

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

A team of community ecology and environmental DNA researchers collaborate 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 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 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 most sophisticated 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 at XXXX.

## 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 web-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\_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 web-app, including a dataset generated for demonstrations, is available at <https://gauravsk.shinyapps.io/ranacapa>
* ranacapa is available for installation at <https://github.com/gauravsk/ranacapa>
* Link to source code as at time of publication (F1000Research TO GENERATE)
* Link to archived source code as at time of publication (F1000Research TO GENERATE)
* Software license (GPL-3)
* Data availability
* Figshare: [DOI]
* License: CC-BY 4.0

## Author contributions

GSK led the development of ranacapa, with help from MCC. ZJG and EEC provided feedback regarding which analyses and visualization options to include. ZJG performed the MPA eDNA study in collaboration with JS and DK. NK, GSK, EC, and RM collaborated with AF and JMP, who used ranacapa in their microbiology undergraduate course. GSK wrote the first draft of this manuscript; all authors contributed to revisions.

## Competing interests

No competing interests were disclosed

## Grant information

GSK and ZJG were supported by the US-NSF Graduate Research Fellowship (DEG No. 1650604). EEC is supported by the CALeDNA program, with funds from University of California President’s Research Catalyst Award (CA-16-376437).

## 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 their phyloseq-extended (<https://github.com/mahendra-mariadassou/phyloseq-extended>) and pairwiseAdonis (<https://github.com/pmartinezarbizu/pairwiseAdonis>) packages openly available with a GPL-3 License.

## Figures

#### Figure 1

## PhantomJS not found. You can install it with webshot::install\_phantomjs(). If it is installed, please make sure the phantomjs executable can be found via the PATH variable.

Figure 1: Taxon accumulation curve as shown in the first tab of ranacapa.

#### Figure 2

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 filter select or deselect individual taxa to be shown in the heatmap.

#### Figure 3

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

#### Figure 4

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.

#### Figure 5

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.

## References

Balasingham, K.D., Walter, R.P., Mandrak, N.E. & Heath, D.D. (2017). Environmental DNA detection of rare and invasive fish species in two great lakes tributaries. *Molecular Ecology*, 27, 112–127.

Barve, V. & Hart, E. (2017). *Rinat: Access iNaturalist data through apis*.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J. & Fierer, N. *et al.* (2012). Ultra-high-throughput microbial community analysis on the illumina HiSeq and MiSeq platforms. *The ISME Journal*, 6, 1621–1624.

Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. (2018). *Shiny: Web application framework for r*.

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

European Citizen Science Association. (2015). Ten principles of citizen science. [*https://ecsa.citizen-science.net/sites/default/files/ecsa\_ten\_principles\_of\_citizen\_science.pdf*](https://ecsa.citizen-science.net/sites/default/files/ecsa_ten_principles_of_citizen_science.pdf).

Gu, W., Song, J., Cao, Y., Sun, Q., Yao, H. & Wu, Q. *et al.* (2013). Application of the ITS2 region for barcoding medicinal plants of selaginellaceae in pteridophyta. *PLoS ONE*, 8, e67818.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D. & Reyes, J.A. *et al.* (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31, 814–821.

Lessios, H.A. (1996). 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), pp. 149–157.

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

McMurdie, P.J. & Holmes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, 10, e1003531.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P. & McGlinn, D. *et al.* (2018). *Vegan: Community ecology package*.

Pandya, R.E. (2012). A framework for engaging diverse communities in citizen science in the US. *Frontiers in Ecology and the Environment*, 10, 314–317.

Parks, D.H., Tyson, G.W., Hugenholtz, P. & Beiko, R.G. (2014). STAMP: Statistical analysis of taxonomic and functional profiles. *Bioinformatics*, 30, 3123–3124.

Pedersen, M.W., Overballe-Petersen, S., Ermini, L., Sarkissian, C.D., Haile, J. & Hellstrom, M. *et al.* (2014). Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370, 20130383.

Shapiro, C., Toma, S., Roth-Johnson, E.A., Hancock, S.P., Ayon, C. & Zimmerman, H. *et al.* (2015). Comparing the impact of course-based and apprentice-based research experiences in a life science laboratory curriculum. *Journal of Microbiology & Biology Education*, 16, 186–197.

Taberlet, P., Coissac, E., Hajibabaei, M. & Riesberg, L.H. (2012). Environmental DNA. *Molecular Ecology*, 21, 1789–1793.

Thomas, A.C., Howard, J., Nguyen, P.L., Seimon, T.A. & Goldberg, C.S. (2018). ANDe : A fully integrated environmental DNA sampling system. *Methods in Ecology and Evolution*, 9, 1379–1385.

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