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 biodiversity monitoring and conservation, and is a promising way to go beyond species inventory to systems-level analyses of community ecological dynamics. Results from eDNA analyses can inform and inspire research scientists, natural resource managers, students, community scientists, and naturalists; however, there is a dearth of easily accessible data exploration tools for this diverse audience. 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 web-app accepts multiple formats of taxonomy tables, and requires a simple metadata file with descriptive information about each sample. The app allows users to explore the data with interactive figures for instant community ecology analysis. We demonstrate the usability of ranacapa by multiple user groups, including the National Park Service, a public community science program, and an undergraduate microbiology course.

## Introduction

The targeted amplification and sequencing of DNA that living organisms shed into theie physical environment ,termed “environmental DNA metabarcoding”, or “eDNA sequencing”, 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 shed light on 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 sequencing declines and sample collection techniques become more streamlined (e.g. Thomas *et al.* (2018)), professional research scientists are increasingly using eDNA sequencing as a platform to partner with members of the community, including natural resource managers, undergraduate students, and citizen scientists (collectively referred to in this manuscript as “community partners”), 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 sequencing-based projects work well for community science partnerships because non-experts can be quickly trained to collect samples in the field, and because eDNA sequencing 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 the 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>). As in any community science program, one important ingredient to ensure long-term success of eDNA-based community science programs is that the participants should be able to participate across multiple stages of the project, not only in sample collection (Pandya 2012; European Citizen Science Association 2015). This can be a challenge for eDNA sequencing-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 interact with and visualize the results from these studies. Indeed, learning the bioinformatic tools necessary for visualizing and analyzing the large, multidimensional datasets generated in these studies can be difficult for professional researchers (Carey & Papin 2018), let alone for the diverse participants in community science programs.

To address this challenge, we created an R package “ranacapa”, at the core of which is a Shiny webapp with which users can visualize and perform simple community ecology analyses on results from eDNA sequencing studies. ranacapa complements existing visualization platforms (e.g. Phinch, Phyloseq-Shiny, QIIME2 Viewer), because in addition to interactive visualizations, ranacapa includes brief explanations and links to additional educational resources to provide users with an overview of basic data analyses used in eDNA studies. ranacapa works with community matrices generated via QIIME () 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 ranacapa and demonstrate its use 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 available at <http://gauravsk.shinyapps.io/ranacapa> or launched with `ranacapa::runRanacapaApp()`. The package also includes two categories of helper functions that are used to transform the user-uploaded taxonomy and metadata tables into R objects that can be visualized and analyzed using the Phyloseq and Vegan packages (Table 1).

The Shiny app (<http://gauravsk.shinyapps.io/ranacapa> or rancapa::runRanacapaApp()) allows users to interact with eDNA results through statistical summaries and interactive plots, displayed in the following tabs:

* **Sample rarefying**: This tab explains the basic logic behind rarefying samples in metagenomics studies, and allows users to choose the sampling depth (Figure 1). Users can control the number of times the samples are rarified. The documentation on this tab also acknowledges recent disagreement regarding the value of rarefying in metabarcoding and eDNA sequencing studies (McMurdie & Holmes (2014)), and allows users to proceed through the rest of the app without rarefying samples.

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

* **Taxonomy heatmap**: This tab shows the taxonomy-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).

#### 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.

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

#### 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

* **Alpha diversity plots**: This tab 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).

#### Figure 4

Figure 4: Alpha diveristy boxplots as shown in the ranacapa Shiny app. Users can select the X-axis variable using a dropdown menu in the app.

* **Alpha diversity statistics**: This tab 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**: This tab 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() based on an ordination made with phyloseq::ordinate(., method = "PCoA"). Points on the PCoA plot are colored according to a user-selected metadata variable (Figure 5). The tab also shows dissimilarity among sites according in a dendrogram generated with 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.

#### 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.

* **Beta diversity statistics**: This tab shows results from two statistical tests of species turnover across site: first, a multivariate ANOVA implemented with vegan::adonis()), which shows users results from a statistical test of taxon turnover across sites, and second, a statistical test of heterogeneity of variances among samples implemented with vegan::betadisper(), which shows results from a statistical test that compares the degree of sample-to-sample variation within habitats (or within other user-selected groups).
* **More references**

## 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. The package can be installed using the command devtools::install\_github("gauravsk/ranacapa"), and the Shiny app is available at <http://gauravsk.shinyapps.io/ranacapa>.

### Input file structure

The ranacapa shiny app requires two input files. The first requirement is a site-by-species matrix, uploaded either as a rich, dense biom table or as a tab-separated .txt file. If the site-by-species matrix is uploaded as a .biom file, the file should contain taxonomy and abundance information; thus, .qza files generated by QIIME2 are not immediately suitable for ranacapa. 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 (). Specifically, each row in the .txt file must represent the taxonomic identification for one Amplicon Sequence Variant (ASV), and each column, save one, must represent the number of times that ASV appears in each sequenced sample. One column, named sum.taxonomy must contain the taxonomic identification, with taxonomic rank separated by a semicolon, as follows:

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

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 best-practices to assign taxonomy to eDNA datasets using the pipeline of their choice, and generate clean taxonomy and metadata files that they can share with their community partners. Researchers should 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: How ranacapa facilitated a collaboration between eDNA researchers and managers at the National Park Service

We expect the interactive taxonomy heatmap to be especially useful for researchers collaborating with natural resource managers whose efforts are focused a targetted list of rare or invasive taxa. For example, UCLA researchers who partner 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) use ranacapa to share eDNA sequencing results. 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 toolkit, 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 metazoan taxa 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 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: How ranacapa helps undergraduate environmental microbiology students pursue sophisticated microbiome analyses

We expect ranacapa’s exploratory visualizations and the accompanying explanations to help researchers introducing ecology concepts and analyses in classrooms or other informal education settings. For example, ranacapa was used by 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 the instructional team, which included eDNA research scientists, extracted total DNA from student-collected soil samples and sequenced the ITS2 (Gu *et al.* 2013) and 16S SSU RNA (Caporaso *et al.* 2012) metabarcoding region to characterize the plant, bacterial, and archaeal communities. The instructional team processed sequences and assigned taxonomy using the Anacapa toolkit.

Students were introduced to the structure of eDNA results and were taught to explore data and perform simple statistical analyses using ranacapa. 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 (<http://gauravsk.shinyapps.io/ranacapa>). This allowed the instructors to focus their time with the students 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 community profile 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 [] and PiCrust []. 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 making data and analyses from eDNA sequencing-based studies more accessible and understandable for a wide range of community research partners.

We propose three avenues for future work in 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 will be a key tool in the upcoming undergraduate curriculum module “Pipeline for Undergraduate Microbiome Analysis”, which is being built as a complete suite of analysis and data visualization tools which will be made openly available to undergraduate researchers. Finally, in the long-term, we believe that there is great promise in connecting ranacapa to packages that connect with APIs of online biodiversity databases (e.g. Taxize, rinat). 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 thank Mahendra Mariadassau and Pedro Martinez Arbizu for making their phyloseq-extended and pairwiseAdonis packages openly available with a GPL-3 License.

## 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.

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.

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.

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.

Props, R., Kerckhof, F.-M., Rubbens, P., Vrieze, J.D., Sanabria, E.H. & Waegeman, W. *et al.* (2016). Absolute quantification of microbial taxon abundances. *The ISME Journal*, 11, 584–587.

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.