The Journal of Open Source Software

MiscMetabar : an R packages to facilitate visualization and reproducibility in metabarcoding analysis

Adrien Taudière [1]

1 IdEst, Saint-Bonnet-de-Salendrinque, 30460 France

DOI: 10.21105/joss.06038

Software

- Review
- Repository
- Archive

Editor: Kelly Rowland

Reviewers:

- @tkchafin
- @iimog

Submitted: 23 October 2023
Published: 13 December 2023

Summary

Describing communities of living organisms increasingly relies on massive DNA sequencing from environmental samples (e-DNA). The analysis of these large amounts of sequences is well established in the R ecosystem, especially for metabarcoding, i.e. the massive sequencing of one or several given DNA regions, called markers. The `MiscMetabar` package aims to facilitate the *description*, *transformation*, *exploration*, and *reproducibility* of metabarcoding analysis. Several tutorials are available online.

Statement of Need

Biological studies, especially in ecology, health sciences and taxonomy, need to describe the biological composition of samples. During the last twenty years, the development of (i) high-throughput DNA sequencing, (ii) reference databases and (iii) bioinformatics resources have allowed the description of biological communities through metabarcoding. Metabarcoding involves the sequencing of millions (*meta-*) of short regions of specific DNA (*-barcoding*, Valentini et al. (2009)) often from environmental samples (eDNA, Taberlet et al. (2012)) such as human stomach contents, lake water, soil and air.

Several platforms (referenced in Tedersoo et al. (2022)) such as QIIME2 (Bolyen et al., 2019), mothur (Schloss, 2020), and Galaxy (Jalili et al., 2020) allow complete analysis from raw fastq sequences to statistical analysis and visualization. However, the R ecosystem (R Core Team, 2023), is very rich (fig. 1) and more flexible than these platforms.

`MiscMetabar` aims to facilitate the **description**, **transformation**, **exploration** and **reproducibility** of metabarcoding analysis using R. The development of `MiscMetabar` relies heavily on the R packages `dada2`, `phyloseq` and `targets`.

State of the Field in R

The metabarcoding ecosystem in the R language is mature, well-constructed, and relies on a very active community in both the bioconductor and cran projects. The bioconductor even creates specific task views in Metagenomics and Microbiome.

R package dada2 (Callahan et al., 2016) provides a highly cited and recommended clustering method (Pauvert et al., 2019). `phyloseq` (McMurdie & Holmes, 2013) facilitate metagenomics analysis by providing a way to store data (the `phyloseq` class) and provides graphical and statistical functions. `MiscMetabar` is based on the `phyloseq` class from `phyloseq`, the most cited package in metagenomics (Wen et al., 2023). For a description and comparison of other integrated packages competing with phyloseq, see Wen et al. (2023). Some packages already extend the phyloseq packages, in particular `microbiome` package collection (Ernst et al., 2023), the `speedyseq` package (McLaren, 2020) and the package phylosmith (Smith, 2019).

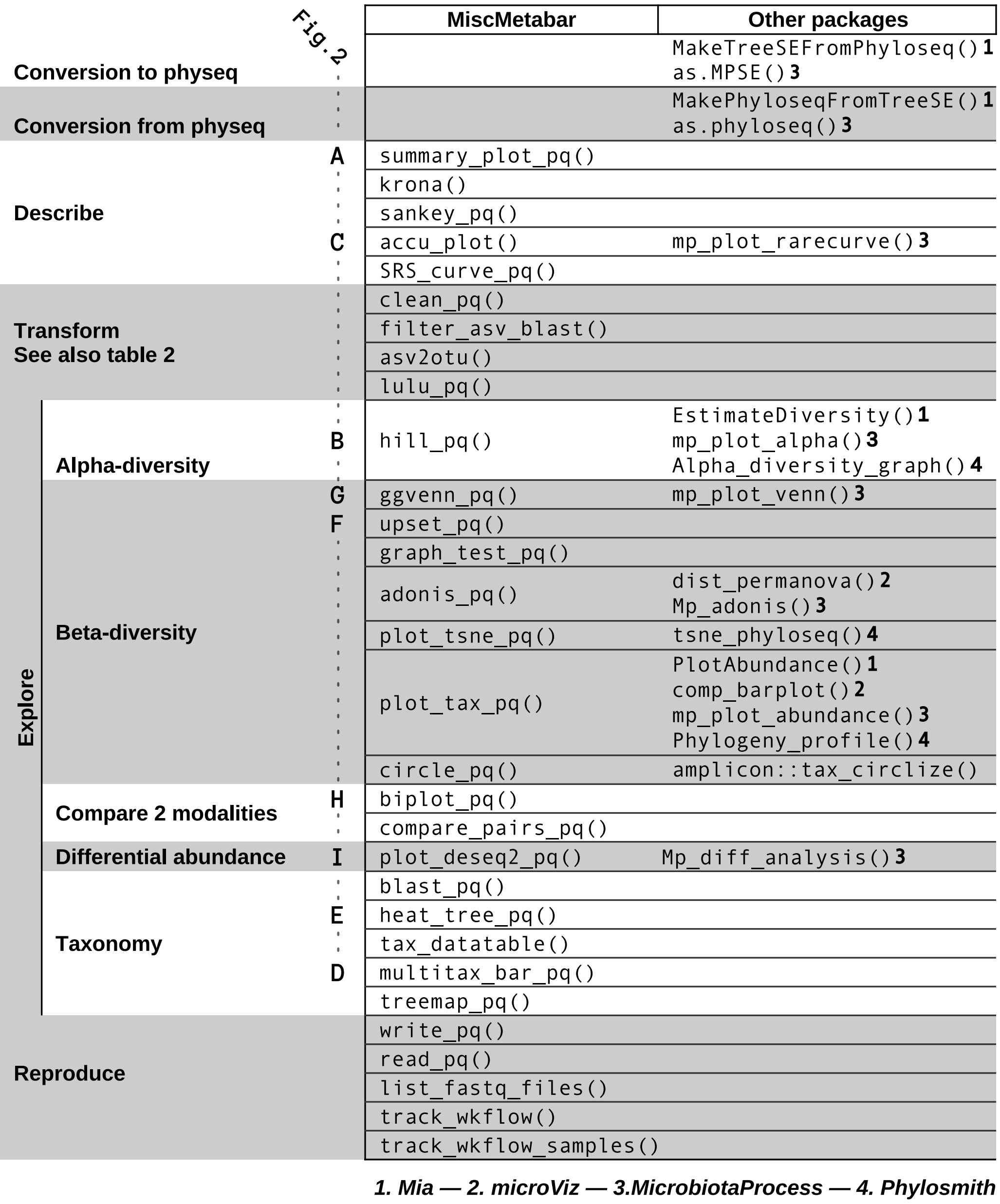

		Fig. 2	MiscMetabar	Other packages
Conversion to physeq				MakeTreeSEFromPhyloseq() **1** as.MPSE() **3**
Conversion from physeq				MakePhyloseqFromTreeSE() **1** as.phyloseq() **3**
Describe		A	summary_plot_pq()	
			krona()	
			sankey_pq()	
		C	accu_plot()	mp_plot_rarecurve() **3**
			SRS_curve_pq()	
Transform See also table 2			clean_pq()	
			filter_asv_blast()	
			asv2otu()	
			lulu_pq()	
Explore	Alpha-diversity	B	hill_pq()	EstimateDiversity() **1** mp_plot_alpha() **3** Alpha_diversity_graph() **4**
	Beta-diversity	G	ggvenn_pq()	mp_plot_venn() **3**
		F	upset_pq()	
			graph_test_pq()	
			adonis_pq()	dist_permanova() **2** Mp_adonis() **3**
			plot_tsne_pq()	tsne_phyloseq() **4**
			plot_tax_pq()	PlotAbundance() **1** comp_barplot() **2** mp_plot_abundance() **3** Phylogeny_profile() **4**
			circle_pq()	amplicon::tax_circlize()
	Compare 2 modalities	H	biplot_pq()	
			compare_pairs_pq()	
	Differential abundance	I	plot_deseq2_pq()	Mp_diff_analysis() **3**
	Taxonomy		blast_pq()	
		E	heat_tree_pq()	
			tax_datatable()	
		D	multitax_bar_pq()	
			treemap_pq()	
Reproduce			write_pq()	
			read_pq()	
			list_fastq_files()	
			track_wkflow()	
			track_wkflow_samples()	

1. Mia — 2. microViz — 3.MicrobiotaProcess — 4. Phylosmith

Figure 1: Important functions of MiscMetabar with their equivalent when available in other R packages: 1. Mia (Ernst et al., 2023); 2. microViz (Barnett et al., 2021); 3. MicrobiotaProcess (Xu et al., 2023); 4 Phylosmith (Smith, 2019).

`MiscMetabar` enriches this R ecosystem by providing functions to (i) **describe** your dataset visually, (ii) **transform** your data, (iii) **explore** biological diversity (alpha, beta, and taxonomic diversity), and (iv) simplify **reproducibility**. `MiscMetabar` is already used by the scientific community in several teams (Bouilloud et al., 2023; Mark McCauley et al., 2022; M. McCauley et al., 2023; Pleić et al., 2022; Vieira et al., 2023; Vieira & Pecchia, 2021).

Features

Description

A quick graphical representation of the phyloseq object is available using the `summary_plot_pq()` function (fig. 2A). This plot provides an information-rich structural overview of the phyloseq object. The functions `krona()` and `tax_datatable()` describe the taxonomy of organisms using krona interactive pie chart (Ondov et al., 2011) and datatable libraries, respectively.

Transformation

Post-clustering

Several pipelines use at least two step of clustering. The function `asv2otu()`, using either the `DECIPHER::Clusterize()` function from R or the vsearch software allow to recluster existing groups such as **ASV** (stands for *Amplicon Sequence Variant*) obtained by the `dada2::dada()` function (see the vignette reclustering). Another transformation method is implemented in `lulu_pq()`, which uses Frøslev et al. (2017)'s method for post-clustering curation of DNA amplicon data. Note that a fast and robust C++ re-implementation of lulu called mumu (Mahé, 2023) is also available through the function `mumu_pq()`.

Cleaning and filtering

The function `clean_pq()` validates a phyloseq object, mainly by removing empty taxa and samples, and checking for discrepancies between taxa and sample names in different slots.

The filter functions `subset_samples_pq()` and `subset_taxa_pq()` complement `subset_samples()` and `subset_taxa()` from the `phyloseq` package, allowing the use of a boolean vector to filter samples or taxa from a phyloseq object.

I also implement a function to filter taxa based on their blast to a custom database (`filter_asv_blast()`). This function uses the blastn software (Altschul et al., 1990) to compare ASV sequences to a database and filter out species that are below a given threshold of e-value and/or bit-score.

Exploration

`MiscMetabar` provides a large number of facilities to explore the biological diversity in a phyloseq object. In most functions, a parameter enables the effect of the number of reads (sampling depth) to be controlled by rarefaction or other statistical methods, depending on the function. For example, the alpha diversity analysis (function `hill_pq()`) uses the HSD-Tuckey test on a linear model that includes the square roots of the number of reads as the first explanatory variable.

To illustrate the effect of sample variables on the taxonomy, `MiscMetabar` provides the functions `treemap_pq()`, `multitax_bar_pq()` (fig. 2D) and `heat_tree_pq()` (fig. 2E). The effect of an environmental variable (beta-diversity) on a biological organism can be explored by upset plot (`pset_pq()`; fig. 2F), venn diagram (`ggvenn_pq()`; fig. 2G), and circle plot (`circle_pq()`). This effect can be tested with PERMANOVA (`adonis_pq()`) and the network test (`graph_test_pq()`). If only two modalities are compared, `biplot_pq()` is very useful (fig. 2H). Differential abundance analysis can be performed directly using the `plot_deseq2_pq()` function (fig. 2I).

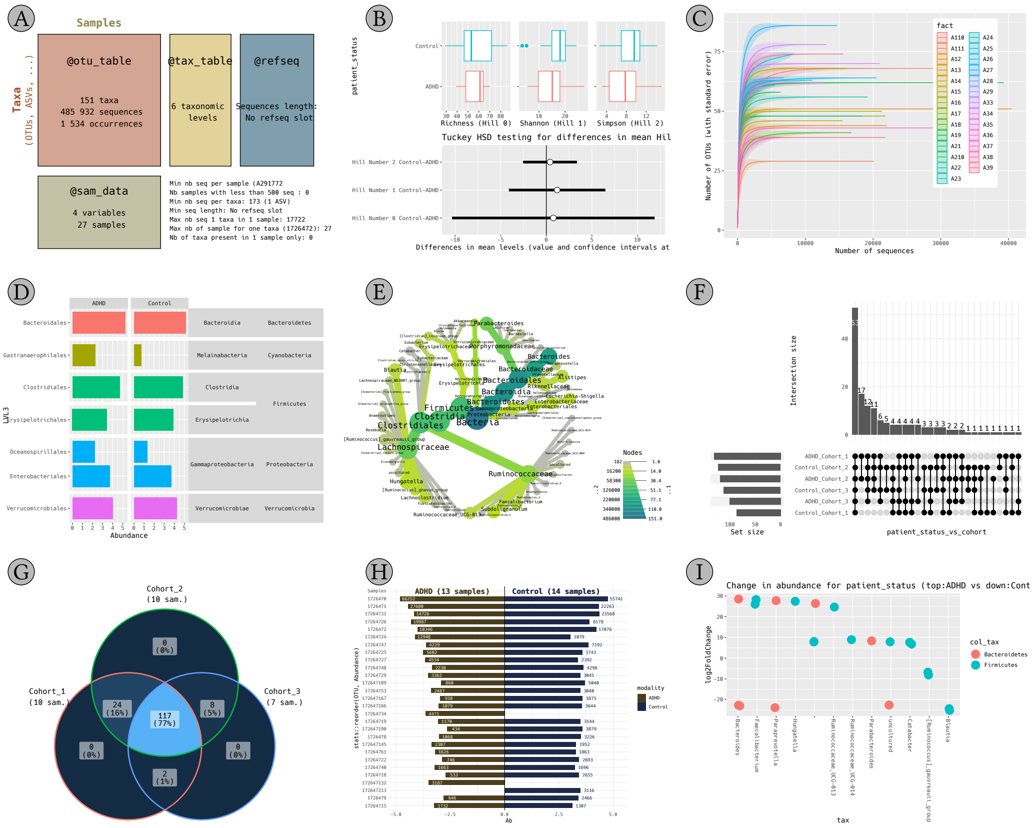

Figure 2: Some illustrations from MiscMetabar with the tengeler dataset from mia R package

Reproducibility

The targets R package (Landau, 2021) improves the efficiency and reproducibility of the pipeline in R. It orchestrates the stage of the pipeline and stores the objects to skip tasks that are already up to date. Given the complexity, runtime, and parameter sensitivity of bioinformatic analysis, the use of targets is particularly relevant for metabarcoding. I developed functions to list fastq files in a directory (`list_fastq_files()`) and to track the number of sequences, clusters and samples through the pipeline (`track_wkflow()`) for a variety of objects. Function `write_pq()` save an object of class phyloseq and `read_pq()` read a phyloseq object from files.

Acknowledgements

I thank Will Landau, Paul McMurdie, and Benjamin Callahan for their excellent R packages on which `MiscMetabar` rests. I also want to acknowledge Franck Richard, Lise Roy, Élisa Taschen and the Mycea team for the discussion and work around metabarcoding.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.32388/rhq6vj

Barnett, D. J. M., Arts, I. C. W., & Penders, J. (2021). microViz: An r package for microbiome data visualization and statistics. *Journal of Open Source Software*, *6*(63), 3201. https://doi.org/10.21105/joss.03201

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., & others. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Bouilloud, M., Galan, M., Pradel, J., Loiseau, A., FERRERO, J., Gallet, R., Roche, B., & Charbonnel, N. (2023). Effects of forest urbanization on the interplay between small mammal communities and their gut microbiota. *bioRxiv*, 2023–2009. https://doi.org/10.21105/joss.03201

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Ernst, F. G. M., Shetty, S. A., Borman, T., & Lahti, L. (2023). *Mia: Microbiome analysis*. https://doi.org/10.18129/B9.bioc.mia

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, *8*(1), 1188. https://doi.org/10.1038/s41467-017-01312-x

Jalili, V., Afgan, E., Gu, Q., Clements, D., Blankenberg, D., Goecks, J., Taylor, J., & Nekrutenko, A. (2020). The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Research*, *48*(W1), W395–W402.

Landau, W. M. (2021). The targets r package: A dynamic make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, *6*(57), 2959. https://doi.org/10.21105/joss.02959

Mahé, F. (2023). *Mumu: Post-clustering curation tool for metabarcoding data* (Version 1.0.2). https://github.com/frederic-mahe/mumu/

McCauley, Mark, Goulet, T., Jackson, C., & Loesgen, S. (2022). Meta-analysis of cnidarian microbiomes reveals insights into the structure, specificity, and fidelity of marine associations. *Research Square*. https://doi.org/10.21203/rs.3.rs-2011054/v1

McCauley, M., Goulet, T., Jackson, C., & Loesgen, S. (2023). Systematic review of cnidarian microbiomes reveals insights into the structure, specificity, and fidelity of marine associations. *Nature Communications*, *14*(1), 4899. https://doi.org/10.21203/rs.3.rs-2011054/v1

McLaren, M. (2020). *Mikemc/speedyseq: Speedyseq v0.2.0* (Version v0.2.0). Zenodo. https://doi.org/10.5281/zenodo.3923184

McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, *8*(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a web browser. *BMC Bioinformatics*, *12*(1), 1–10.

Pauvert, C., Buée, M., Laval, V., Edel-Hermann, V., Fauchery, L., Gautier, A., Lesur, I., Vallance, J., & Vacher, C. (2019). Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. *Fungal Ecology*, *41*, 23–33. https://doi.org/10.1016/j.funeco.2019.03.005

Pleić, I. L., Bušelić, I., Messina, M., Hrabar, J., Žuvić, L., Talijančić, I., Žužul, I., Pavelin, T., Anđelić, I., Pleadin, J., Puizina, J., Grubišić, L., Tibaldi, E., & Šegvić-Bubić, T. (2022). A plant-based diet supplemented with hermetia illucens alone or in combination with poultry by-product meal: One step closer to sustainable aquafeeds for european seabass. *Journal of Animal Science and Biotechnology*, *13*(1), 77. https://doi.org/10.1186/s40104-022-00725-z

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Schloss, P. D. (2020). Reintroducing mothur: 10 years later. *Applied and Environmental Microbiology*, *86*(2), e02343–19. https://doi.org/10.1128/aem.02343-19

Smith, S. D. (2019). Phylosmith: An r-package for reproducible and efficient microbiome analysis with phyloseq-objects. *Journal of Open Source Software*, *4*(38), 1442. https://doi.org/10.21105/joss.01442

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental dna. In *Molecular ecology* (No. 8; Vol. 21, pp. 1789–1793). Wiley Online Library. https://doi.org/10.1002/(issn)2637-4943

Tedersoo, L., Bahram, M., Zinger, L., Nilsson, R. H., Kennedy, P. G., Yang, T., Anslan, S., & Mikryukov, V. (2022). Best practices in metabarcoding of fungi: From experimental design to results. *Molecular Ecology*, *31*(10), 2769–2795. https://doi.org/10.22541/au.163430390.04226544/v1

Valentini, A., Pompanon, F., & Taberlet, P. (2009). DNA barcoding for ecologists. *Trends in Ecology & Evolution*, *24*(2), 110–117. https://doi.org/10.1016/j.tree.2008.09.011

Vieira, F. R., Di Tomassi, I., O'Connor, E., Bull, C. T., Pecchia, J. A., & Hockett, K. L. (2023). Manipulating agaricus bisporus developmental patterns by passaging microbial communities in complex substrates. *Microbiology Spectrum*, e01978–23.

Vieira, F. R., & Pecchia, J. A. (2021). Bacterial community patterns in the agaricus bisporus cultivation system, from compost raw materials to mushroom caps. *Microbial Ecology*, *84*(1), 20–32. https://doi.org/10.1007/s00248-021-01833-5

Wen, T., Niu, G., Chen, T., Shen, Q., Yuan, J., & Liu, Y.-X. (2023). The best practice for microbiome analysis using r. *Protein & Cell*, pwad024.

Xu, S., Zhan, L., Tang, W., Wang, Q., Dai, Z., Zhou, L., Feng, T., Chen, M., Wu, T., Hu, E., & others. (2023). MicrobiotaProcess: A comprehensive r package for deep mining microbiome. *The Innovation*, *4*(2). https://doi.org/10.1016/j.xinn.2023.100388