




Microbiome.jl and BiobakeryUtils.jl - Julia packages for working with microbial community data

Kevin S. Bonham^{*1}, Annelle Abatoni Kayisire¹, Anika S. Luo¹, and Vanja Klepac-Ceraj^{†1}

¹ Department of Biological Sciences, Wellesley College

DOI: [10.21105/joss.03876](https://doi.org/10.21105/joss.03876)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Will Rowe](#) 

Reviewers:

- [@adRn-s](#)
- [@aguang](#)

Submitted: 20 October 2021

Published: 17 November 2021

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

Microbiome.jl is a Julia package to facilitate analysis of microbial community data. BiobakeryUtils.jl is built on top of Microbiome.jl, and provides utilities for working with a suite of command line tools (the bioBakery) that are widely used for converting raw metagenomic sequencing data into tables of taxon and gene function counts. Together, these packages provide an effective way to link microbial community data with the power of Julia's numerical, statistical, and plotting libraries.

Statement of need

Complex microbial communities exist everywhere, including in and on the human body, and have profound effects on the environment and human health ([Lloyd-Price et al., 2017](#)). Common methods for analyzing microbial communities (eg. 16S amplicon or metagenomic sequencing) generate a large quantity of numerical data (eg. count or relative abundance data) as well as metadata associated with biological samples (eg. locations, human subject data) and microbial features (eg. taxa, gene functions) ([Mallick et al., 2017](#)).

The Julia programming language ([Bezanson et al., 2017](#)) is gaining increasing prominence in biological research due to its speed and flexibility ([Roesch et al., 2021](#)), and has a growing ecosystem of packages for working with biological and ecological data, as well as libraries for Bayesian statistical analysis ([Ge et al., 2018](#)), scientific machine learning ([Rackauckas & Nie, 2017](#)), and plotting ([Danisch & Krumbiegel, 2021](#)). Julia's type system makes it incredibly easy for packages to interoperate, making Microbiome.jl and BiobakeryUtils.jl an effective bridge between microbial community data and Julia's package ecosystem, while remaining agnostic to downstream analysis.

Functionality

At its most basic, microbial community data can be represented as a sparse matrix, where one dimension is indexed by microbial features (eg. species), and the other is indexed by biological samples or observations (eg. a stool sample). Together, the measured abundances of each feature in each sample make up the taxonomic or function "profile." Typically, additional information (metadata) about each sample is also needed for downstream statistical analysis, such as the location or human subject it was collected from, data about that

^{*}Corresponding author

[†]Corresponding author

environment (salinity, temperature, etc. for environmental samples; clinical covariates for human subjects), and storage or processing details. While the observed values for microbial features are uniformly numeric, and can be efficiently stored in a sparse matrix of floating point numbers, metadata can take many forms. Further, `CommunityProfiles` may have hundreds to hundreds of thousands of features, while typically only a few dozen metadata variables are necessary for a given analysis.

`Microbiome.jl` provides a convenient set of types and type constructors to store and access this information ([Figure 1](#)).

- The `MicrobiomeSample` type contains name and metadata fields, and methods for efficiently adding and extracting stored metadata
- The `Taxon` type stores name and taxonomic rank (eg. genus, phylum) fields
- The `GeneFunction` type stores name and taxon fields, the later of which may be a `Taxon` (allowing taxonomically stratified gene functions).
- The `CommunityProfile` type is a wrapped `SparseMatrixCSC`, with `MicrobiomeSamples` as columns and features (`Taxons` or `GeneFunctions`) as rows.
- `CommunityProfiles` can be indexed like normal julia arrays with integers, or with strings and regular expressions that will search on the name fields of the sample or feature dimensions.

Further, the `CommunityProfile` type implements the `Tables.jl` interface, making it trivial to convert to other tabular representations, in particular enabling round-tripping to and from column separated values (`.csv`) files using `CSV.jl`. Feature (`Taxon` and `GeneFunction`), `MicrobiomeSample`, and `CommunityProfile` types are also implemented with the interface of `EcoBase.jl`, potentially enabling integration with the wider `EcoJulia` family of packages.

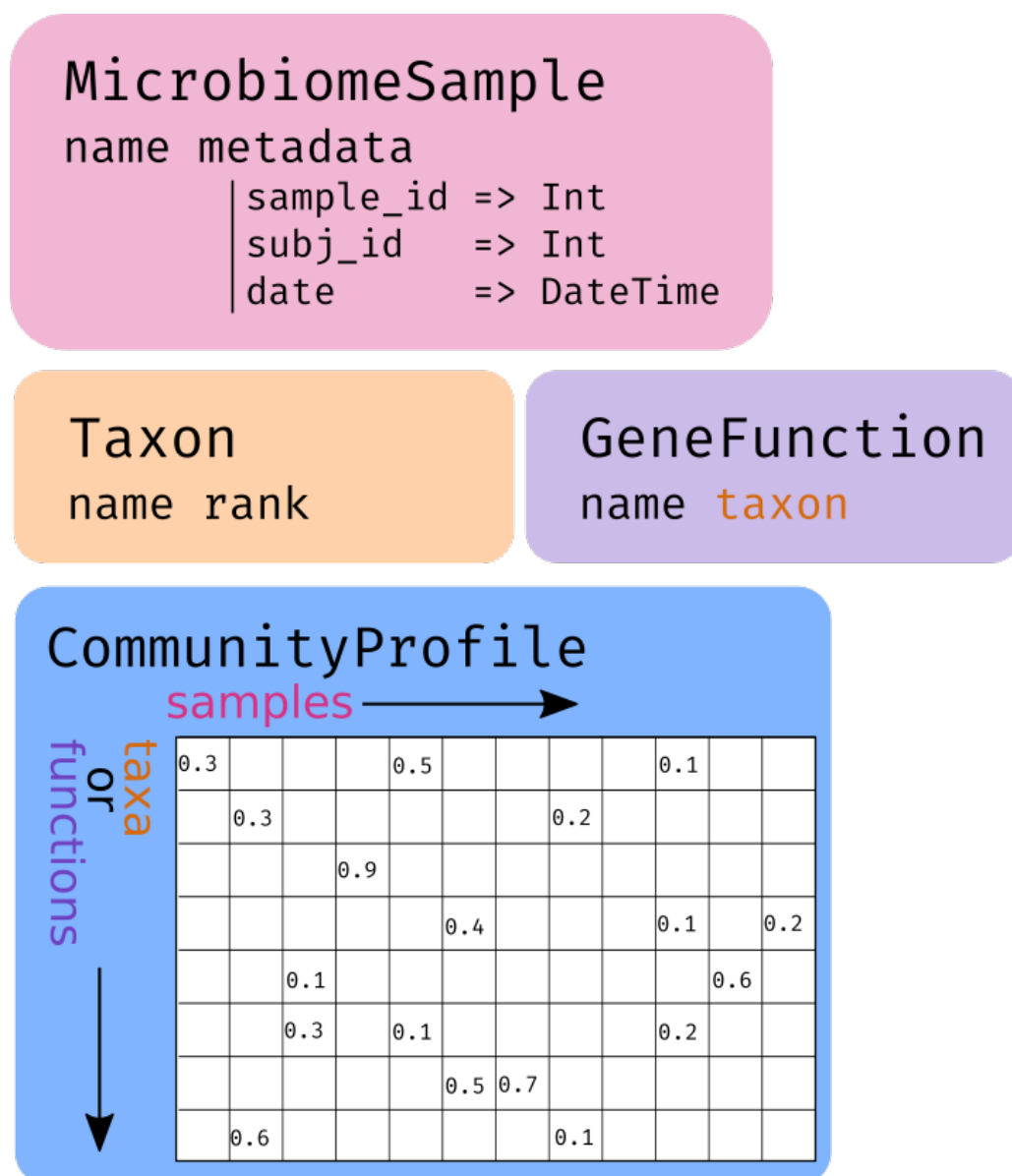


Figure 1: Concrete types provided by `Microbiome.jl` for storing information about features, samples, and whole communities.

`BiobakeryUtils.jl` provides a julia interface for the command line utilities from `HUMANn` and `MetaPhlAn`, two widely-used tools for using metagenomic sequencing reads to generate functional and taxonomic profiles, respectively. It also provides functionality to simplify installation of the tools and I/O for the common file types used and produced by those tools. Together, `Microbiome.jl` and `BiobakeryUtils.jl` make it easy to load, manipulate, and analyze microbial community data (Figure 2).

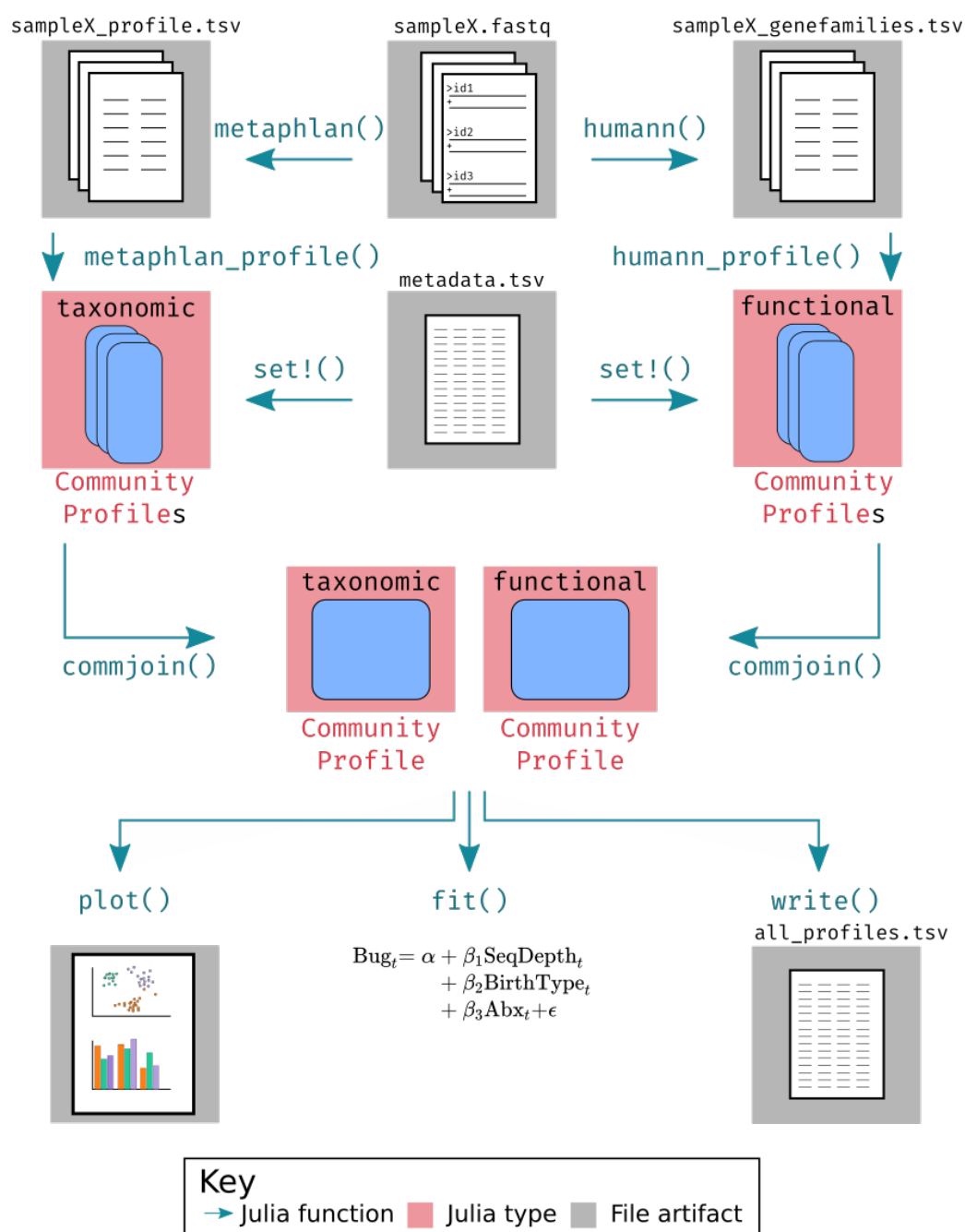


Figure 2: Microbial community analysis workflow using Microbiome.jl and BiobakeryUtils.jl

Limitations and future work

While Microbiome.jl and BiobakeryUtils.jl are already functional and being used for research (Lewis et al., 2021; Peterson et al., 2021; Tso et al., 2021), there are several avenues for further development.

First, there are many additional tools in the bioBakery whose interface and outputs could be incorporated into BiobakeryUtils.jl. In particular, StrainPhlAn and PanPhlAn (Beghini et al., 2021), which have tabular output distinct from but quite similar to that of HUMAnN and

MetaPhlAn could be supported.

Second, two of the largest plotting packages in the julia ecosystem, `Plots.jl` and `Makie.jl` (Breloff, 2021; Danisch & Krumbiegel, 2021) share a common “recipes” system, enabling package authors to provide instructions for how to plot their types. `Microbiome.jl` currently contains convenience functions to facilitate the generation of easy-to-plot data structures, but including plot recipes for things like ordinations (PCoA), abundance bar plots, and other commonly used microbial community visualizations would make it even easier to generate publication-quality figures.

Finally, better integration with `EcoJulia` would carry a host of benefits. For example, `Diversity.jl` (Reeve et al., 2016) provides a wide array of alpha and beta diversity metrics that could be beneficial for investigations of microbial diversity. There are also several packages that provide functionality around phylogenies and taxonomic information that could enhance or replace `Taxon`, making it easier to gain insight into the relationships between different microbial taxa found in communities.

Acknowledgements

The authors would like to thank the families participating in the RESONANCE cohort. This work was funded in part by the NIH UG3 OD023313 (VK-C).

References

- Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A., & Segata, N. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife*, 10. <https://doi.org/10.7554/eLife.65088>
- Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Rev.*, 59(1), 65–98. <https://doi.org/10.1137/141000671>
- Breloff, T. (2021). *Plots.jl* (Version v1.22.6) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.5566503>
- Danisch, S., & Krumbiegel, J. (2021). Makie.jl: Flexible high-performance data visualization for julia. *Journal of Open Source Software*, 6(65), 3349. <https://doi.org/10.21105/joss.03349>
- Ge, H., Xu, K., & Ghahramani, Z. (2018). Turing: A language for flexible probabilistic inference. *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, 1682–1690. <http://proceedings.mlr.press/v84/ge18b.html>
- Lewis, C. R., Bonham, K. S., McCann, S. H., Volpe, A. R., D'Sa, V., Naymik, M., De Both, M. D., Huentelman, M. J., Lemery-Chalfant, K., Highlander, S. K., Deoni, S. C. L., & Klepac-Ceraj, V. (2021). Family SES is associated with the gut microbiome in infants and children. *Microorganisms*, 9(8), 1608. <https://doi.org/10.3390/microorganisms9081608>
- Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., Brady, A., Creasy, H. H., McCracken, C., Giglio, M. G., McDonald, D., Franzosa, E. A., Knight, R., White, O. R., & Huttenhower, C. (2017). Strains, functions and dynamics in the expanded human microbiome project. *Nature*, 550, 61–66. <https://doi.org/10.1186/s13059-017-1359-z>

- Mallick, H., Ma, S., Franzosa, E. A., Vatanen, T., Morgan, X. C., & Huttenhower, C. (2017). Experimental design and quantitative analysis of microbial community multiomics. *Genome Biology*, 18.
- Peterson, D., Bonham, K. S., Rowland, S., Pattanayak, C. W., RESONANCE Consortium, & Klepac-Ceraj, V. (2021). Comparative analysis of 16S rRNA gene and metagenome sequencing in pediatric gut microbiomes. *Front. Microbiol.*, 12, 670336. <https://doi.org/10.3389/fmicb.2021.670336>
- Rackauckas, C., & Nie, Q. (2017). Differentialequations.jl—a performant and feature-rich ecosystem for solving differential equations in julia. *Journal of Open Research Software*, 5(1). <https://doi.org/10.5334/jors.151>
- Reeve, R., Leinster, T., Cobbold, C. A., Thompson, J., Brummitt, N. A., Mitchell, S. N., & Matthews, L. (2016). How to partition diversity. *ArXiv e-Prints*. <http://arxiv.org/abs/1404.6520>
- Roesch, E., Greener, J. G., MacLean, A. L., Nassar, H., Rackauckas, C., Holy, T. E., & Stumpf, M. P. H. (2021). *Julia for biologists*. <http://arxiv.org/abs/2109.09973>
- Tso, L., Bonham, K. S., Fishbein, A., Rowland, S., & Klepac-Ceraj, V. (2021). Targeted High-Resolution taxonomic identification of bifidobacterium longum subsp. Infantis using human milk oligosaccharide metabolizing genes. *Nutrients*, 13(8), 2833. <https://doi.org/10.3390/nu13082833>