





1 rbiom: An R package for microbiome analysis

2 Daniel P Smith ¹ and Joseph F Petrosino ¹

3 ¹ Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and
4 Microbiology, Baylor College of Medicine, Houston, Texas, USA  Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

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

Editor: [Open Journals](#) 

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

License

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

5 Summary

6 Microbes live all around us, on us, and even inside us. Their impacts can be as personal as
7 protecting or predisposing us to disease, or as global as regulating planetary biogeochemical
8 cycles. Modern DNA sequencing technology allows studies to characterize thousands of
9 microbial communities at a time. The bottleneck is no longer generating data, but rather
10 analyzing and interpreting the results.

11 rbiom is an R package for analyzing microbial community datasets. Its features include:

- 12 1. **Preprocessing** - Merges, subsets, and rarefies data from multiple sources.
- 13 2. **Calculation** - Diversity, similarity, and abundance metrics using fast multithreaded C
14 code.
- 15 3. **Statistics** - Identifies significant correlations with sample metadata.
- 16 4. **Visualization** - Customizable box plots, ordinations, heatmaps, and stacked bar charts.

Statement of Need

18 Working with microbiome datasets is challenging. Analyses must integrate observed counts,
19 sample metadata, taxonomic mappings, and phylogenetic trees into data structures compatible
20 with statistical testing and visualization. rbiom makes it simple to turn these complex datasets
21 into informative figures, bringing together the speed of C with the powerful statistics and
22 elegant graphics of R.

23 State of the Field

24 Current Tools

25 Wen et al. ([2023](#)) provides an excellent review of R packages for microbiome analysis. The list
26 below includes those mentioned by Wen et al as and others that are actively maintained.

27 R Packages

- 28 ▪ [ampvis2](#) ([Andersen et al., 2018](#))
- 29 ▪ [animalcules](#) ([Zhao et al., 2021](#))
- 30 ▪ [mia](#) ([Borman et al., 2025](#))
- 31 ▪ [microbiome](#) ([Lahti & Shetty, 2012-2019](#))
- 32 ▪ [MicrobiomeAnalystR](#) ([Dhariwal et al., 2017](#))
- 33 ▪ [MicrobiomeStat](#) ([Yang et al., 2025](#))
- 34 ▪ [MicrobiotaProcess](#) ([Xu et al., 2023](#))
- 35 ▪ [Microeco](#) ([C. Liu et al., 2020](#))
- 36 ▪ [microViz](#) ([Barnett et al., 2021](#))
- 37 ▪ [phyloseq](#) ([McMurdie & Holmes, 2013](#))

- 38 ▪ phyloSMITH ([Smith, 2019](#))
- 39 ▪ TidyTacos (?)
- 40 ▪ vegan ([Oksanen et al., 2025](#))

41 **Command-line Tools**

- 42 ▪ EasyAmplicon ([Y. Liu et al., 2023](#))
- 43 ▪ mothur ([Schloss et al., 2009](#))
- 44 ▪ QIIME2 ([Bolyen et al., 2019](#))

45 **Advancement**

46 rbiom's speed and usability sets it apart from this crowd.

47 **Speed**

48 The R packages listed above rely on GUniFrac ([Chen et al., 2023](#)), vegan, phyloseq, and/or
49 ampvis2 for calculating beta diversity metrics, namely bray-curtis, euclidean, manhattan,
50 jaccard, and UniFrac. The speed of these $O(n^2)$ operations determines how many samples
51 can be cross-compared in a feasible amount of time. The exception is mia, which uses rbiom
52 for UniFrac calculations.

53 Benchmarks show that rbiom calculates UniFrac dissimilarities 10 times faster than GUniFrac,
54 50 times faster than phyloseq, and 400 times faster than ampvis2¹. Additionally, rbiom
55 calculates bray-curtis, euclidean, manhattan, and jaccard metrics four times faster than vegan².

56 rbiom has managed these improvements by implementing many central algorithms in C and
57 ensuring full utilization of multi-CPU core systems. This brings the speed of rbiom in line with
58 compiled tools such as mothur and QIIME2.

59 **Usability**

60 rbiom can import data from a variety of sources - BIOM files, R data frames, phyloseq objects,
61 and more. It can export to all these formats as well.

62 Pre-built binaries are available from CRAN and Anaconda, making installation easy on Windows,
63 MacOS, and Linux. As much effort has been put into documentation as the code itself. Users
64 will find that all functions are clearly documented with examples.

65 These programs offer a plethora of features and are fast even on large datasets, however, they
66 come with a steep learning curve that may be discouraging to new users. Additionally, as
67 command-line programs, it is cumbersome to interact with them from R.

68 rbiom sets itself apart from in two ways. First, it uses compiled C code to speed up calculations,
69 thereby enabling processing of much larger datasets.

70 but it is the first to

71 Several packages are currently available for working with microbiome datasets.

72 QIIME2 ([Bolyen et al., 2019](#)), mothur ([Schloss et al., 2009](#)), and Phyloseq ([McMurdie &](#)
73 [Holmes, 2013](#)) offer overlapping functionality with rbiom, but with important distinctions.
74 QIIME2 and mothur are designed for command-line interaction, making them difficult to
75 integrate into R projects. Phyloseq has been a staple of R bioinformatics for a decade, but is
76 frustratingly slow for studies with thousands of samples.

77 This package is designed for users of all experience levels. Novice R users will appreciate that
78 a couple commands will produce publication-ready figures. Advanced R users can use rbiom
79 to complement their existing pipelines with faster and more flexible functions.

¹Based on 100 replications of a 50-sample dataset (rbiom::HMP50) on six CPU cores.

²Based on 100 replications of a 1006-sample dataset (rbiom::GEMS) on six CPU cores.

Implementation

rbiom is an R package for working with abundance datasets, such as OTU or ASV counts from 16S amplicon sequencing. It enables importing/exporting all BIOM formats, subsetting, rarefying, manipulation of metadata/taxonomy/phylogeny, computation of alpha and beta diversity metrics, and summarizing counts per taxonomic rank. Computationally intensive tasks (including UniFrac (Lozupone & Knight, 2005)) have been implemented with multithreaded C to greatly reduce calculation time.

Visualization is a key component of rbiom. Rarefaction curves, taxa abundances, alpha diversity, and beta diversity can all be plotted in a variety of graphical formats, including correlation, heatmap, ordination, stacked bar, and box plots. In rbiom, box plots can be any combination of box, bar, violin, dot, strip, and/or range layers. Each plot includes provenance and modification history as attributes, as well as the ggplot2 (Wickham, 2016) call used to render it to encourage downstream user customization.

Correlations between sample metadata and microbiome structure can be identified by mapping one or more metadata variables of interest to a plot's axes, facets, and/or aesthetics. These mappings can optionally define color/shape/pattern assignments, category ordering, or subsetting parameters. When metadata is associated with a axis or aesthetic, rbiom will automatically run the appropriate statistical test, correct for multiple comparisons, and display significant differences on the plot, captioning it with a brief methodology.

Currently, rbiom can perform four types of significance testing. On correlation plots with a numeric metadata variable on the x-axis (e.g., Age, BMI), linear regression will be computed with R's lm linear model function. For plots with two categories (e.g. Male vs Female), a Mann-Whitney test (Mann & Whitney, 1947) is run with R's wilcox.test. When three or more categories are compared, the Kruskal-Wallis rank sum test (Kruskal & Wallis, 1952) is used instead via R's kruskal.test function. P-values for ordinations are derived using the adonis2 function from the vegan R package (Oksanen et al., 2025), which randomly re-categorizes samples 1,000 times to estimate the significance of the observed clustering. P-values are corrected for multiple comparisons using the method described by (Benjamini & Hochberg, 1995) via R's p.adjust function to control for the false discovery rate.

Usage

Installation

rbiom can be installed using R's default package manager.

```
install.packages('rbiom')
```

It can also be installed directly to a Conda environment.

```
conda install conda-forge::r-rbiom
```

Tutorials

Documentation and examples are available on the [rbiom website](#) and in R with:

```
help(package = 'rbiom')
```

Example Analysis

Import the bundled hmp50.bz2 biological observation matrix (BIOM) file.

```
library(rbiom)
infile <- system.file(package = "rbiom", "extdata", "hmp50.bz2")
biom <- as_rbiom(infile)
```

```
biom
#> ══ Human Microbiome Project - 50 Sample Demo ═══
#>
#> Oral, nasal, vaginal, and fecal samples from a
#> diverse set of healthy volunteers. Source: Human
#> Microbiome Project (<https://hmpdacc.org>).
#>
#> 50 Samples: HMP01, HMP02, HMP03, ...
#> 490 OTUs:   Unc01yki, Unc53100, ...
#> 7 Ranks:    .otu, Kingdom, Phylum, ...
#> 5 Fields:   .sample, Age, BMI, ...
#> Tree:      <present>
#>
#> — 182 - 22k reads/sample ————— 2023-09-22 —
```

117 Rarefy abundance counts and explore associations with metadata..

```
biom <- rarefy(biom)
bdiv_ord_plot(biom, stat.by = "Body Site", facet.by = "Sex")
```

Weighted Bray-Curtis PCoA (Genus)

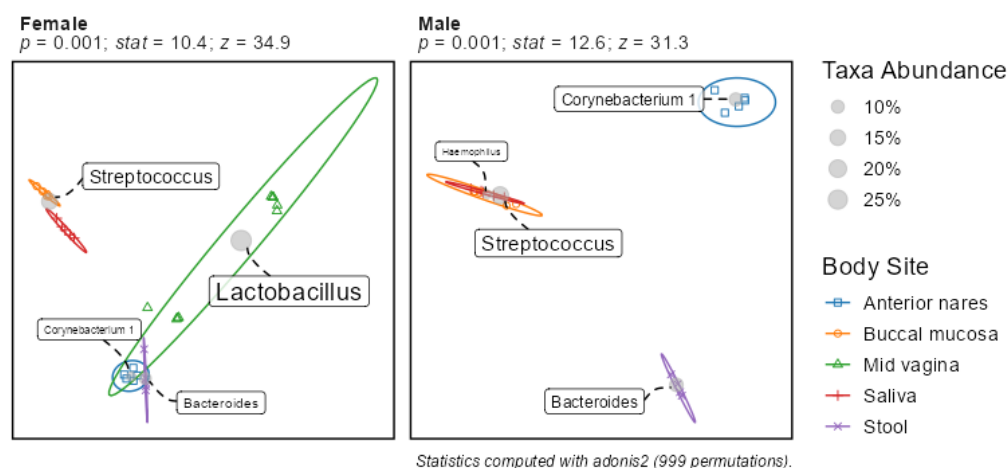


Figure 1: A beta diversity ordination plot. Samples cluster significantly by body site ($p = 0.001$) and are characterized by different bacterial genera.

```
adiv_boxplot(biom, x = "Sex", adiv = c("otu", "shan"), stat.by = "Body Site")
```

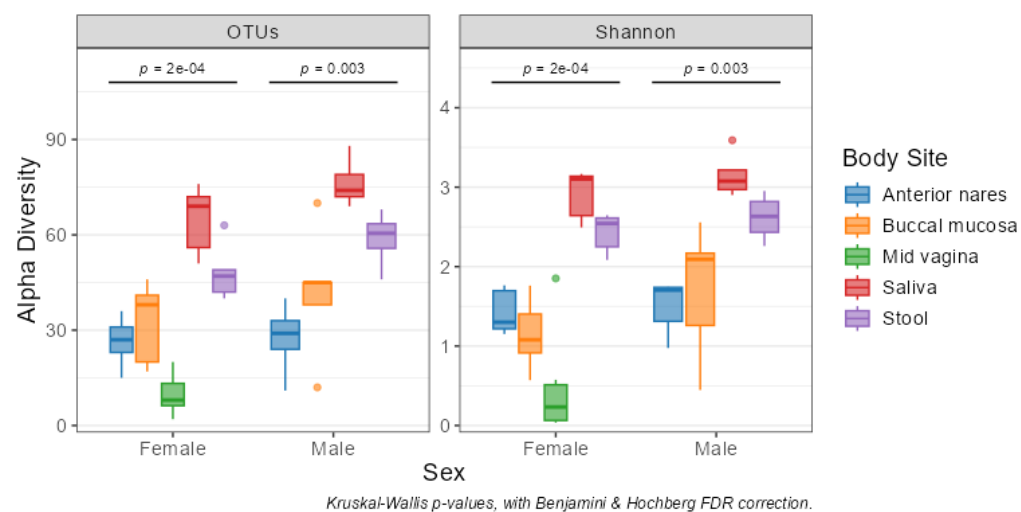


Figure 2: An alpha diversity box plot. Observed OTUs and shannon diversity indices vary significantly by body site for both males ($p = 2e-04$) and females ($p = 0.003$).

```
subset(biom, `Body Site` == 'Buccal mucosa') %>%
  taxa_corrplot("Age", taxa = 2, layers = 'ptc', fit = 'lm', test = 'emtrends')
```

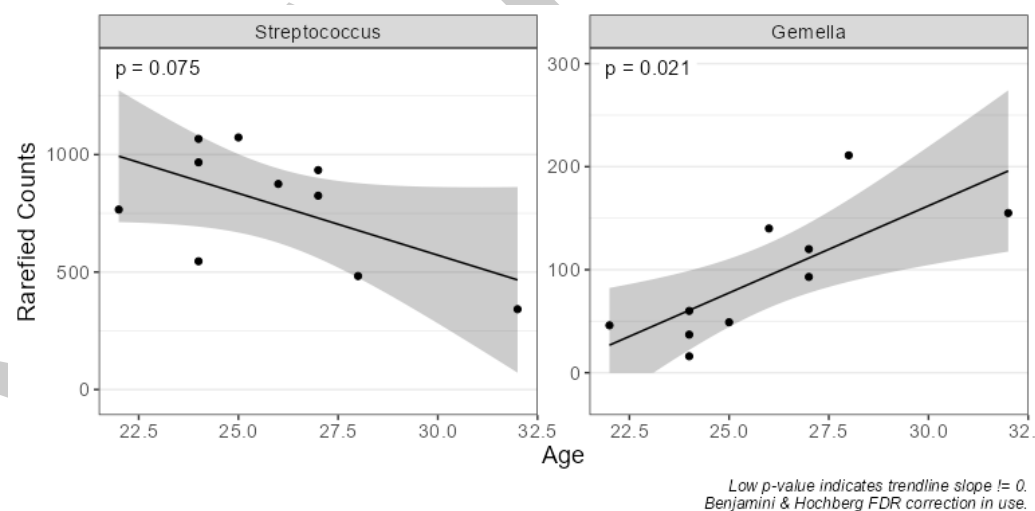


Figure 3: A taxa correlation plot. The two most abundant buccal mucosa-associated genera show weak correlations with age.

```
taxa_heatmap(biom, taxa = 10, tracks = c("body", "age"))
```

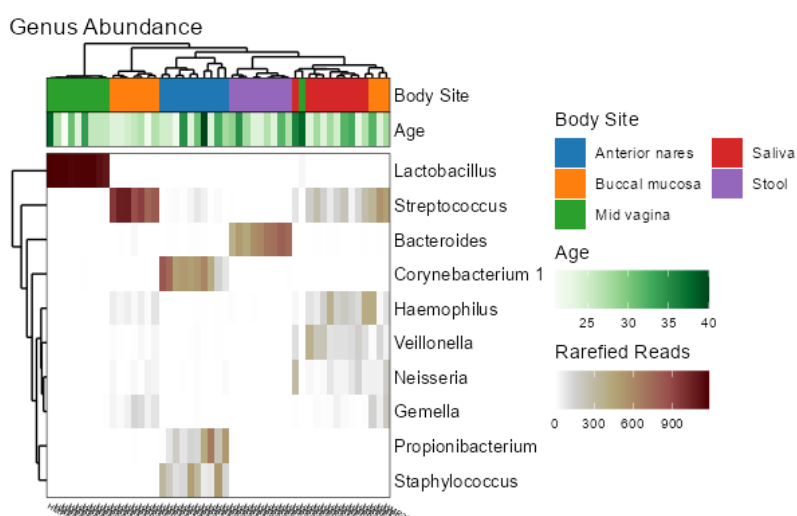


Figure 4: A taxa heatmap plot. The 10 most abundant genera are primarily found on a single specific body site.

```
taxa_stacked(biom, rank = "Phylum")
```

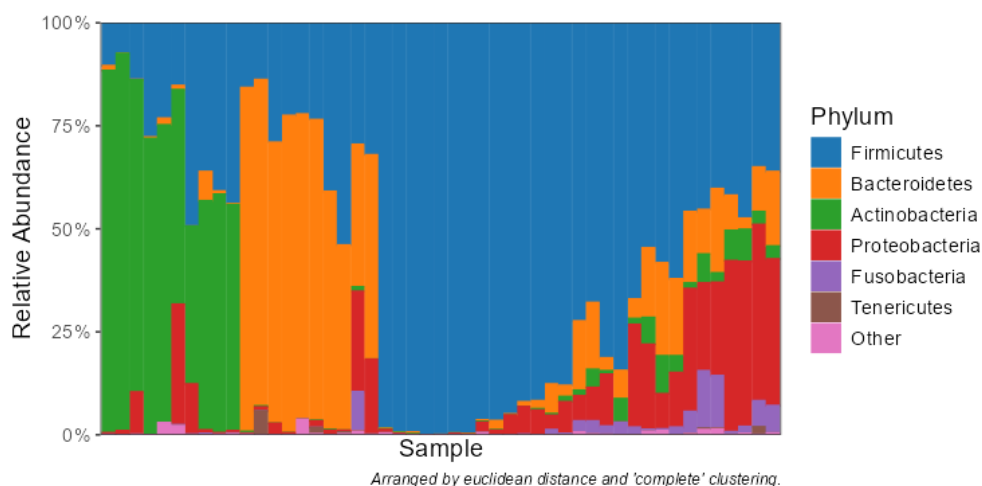


Figure 5: A taxa stacked bar plot. This dataset contains well-defined phylum-level ecotypes.

Acknowledgements

References

- Andersen, K. S., Kirkegaard, R. H., Karst, S. M., & Albertsen, M. (2018). ampvis2: An R package to analyse and visualise 16S rRNA amplicon data. *bioRxiv*. <https://doi.org/10.1101/299537>
- Barnett, D. J., Arts, I. C., & 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>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1),

- 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- 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., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., ... Caporaso, J. G. (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>
- Borman, T., Ernst, F. G. M., Shetty, S. A., & Lahti, L. (2025). *Mia: Microbiome analysis*. <https://doi.org/10.18129/B9.bioc.mia>
- Chen, J., Zhang, X., Yang, L., & Zhang, L. (2023). *GUniFrac: Generalized UniFrac distances, distance-based multivariate methods and feature-based univariate methods for microbiome data analysis*. <https://doi.org/10.32614/CRAN.package.GUniFrac>
- Dhariwal, A., Chong, J., Habib, S., King, I. L., Agellon, L. B., & Xia, J. (2017). MicrobiomeAnalyst: A web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Research*, 45(W1), W180–W188. <https://doi.org/10.1093/nar/gkx295>
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621. <https://doi.org/10.1080/01621459.1952.10483441>
- Lahti, L., & Shetty, S. (2012–2019). *Microbiome r package*. <https://doi.org/10.18129/B9.bioc.microbiome>
- Liu, C., Cui, Y., Li, X., & Yao, M. (2020). Microeco: An r package for data mining in microbial community ecology. *FEMS Microbiology Ecology*, 97(2). <https://doi.org/10.1093/femsec/fiaa255>
- Liu, Y., Chen, L., Ma, T., Li, X., Zheng, M., Zhou, X., Chen, L., Qian, X., Xi, J., Lu, H., Cao, H., Ma, X., Bian, B., Zhang, P., Wu, J., Gan, R., Jia, B., Sun, L., Ju, Z., ... Chen, T. (2023). EasyAmplicon: An easy-to-use, open-source, reproducible, and community-based pipeline for amplicon data analysis in microbiome research. *iMeta*, 2(1). <https://doi.org/10.1002/imt2.83>
- Lozupone, C., & Knight, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12), 8228–8235. <https://doi.org/10.1128/AEM.71.12.8228-8235.2005>
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1), 50–60. <https://doi.org/10.1214/aoms/1177730491>
- 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>
- Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., ... Borman, T. (2025). *Vegan: Community ecology package*. <https://doi.org/10.32614/CRAN.package.vegan>
- Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., & others. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23), 7537–7541. <https://doi.org/10.1128/AEM.01541-09>
- Smith, S. D. (2019). Phylosmith: An r-package for reproducible and efficient microbiome

- 176 analysis with phyloseq-objects. *Journal of Open Source Software*, 4(38), 1442. <https://doi.org/10.21105/joss.01442>
177
- 178 Wen, T., Niu, G., Chen, T., Shen, Q., Yuan, J., & Liu, Y.-X. (2023). The best practice for
179 microbiome analysis using r. *Protein & Cell*, 14(10), 713–725. [https://doi.org/10.1093/](https://doi.org/10.1093/procel/pwad024)
180 [procel/pwad024](https://doi.org/10.1093/procel/pwad024)
- 181 Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.
182 ISBN: 978-3-319-24277-4
- 183 Xu, S., Zhan, L., Tang, W., Wang, Q., Dai, Z., Zhou, L., Feng, T., Chen, M., Wu, T., Hu,
184 E., & Yu, G. (2023). MicrobiotaProcess: A comprehensive r package for deep mining
185 microbiome. *The Innovation*, 4(2), 100388. <https://doi.org/10.1016/j.xinn.2023.100388>
- 186 Yang, C., Chen, J., Zhang, X., & Zhou, H. (2025). *Comprehensive statistical and visualization*
187 *methods for microbiome and multi-omics data*. Zenodo. [https://doi.org/10.5281/zenodo.](https://doi.org/10.5281/zenodo.14881492)
188 [14881492](https://doi.org/10.5281/zenodo.14881492)
- 189 Zhao, Y., Federico, A., Faits, T., Manimaran, S., Segrè, D., Monti, S., & Johnson, W. E.
190 (2021). Animalcules: Interactive microbiome analytics and visualization in r. *Microbiome*,
191 9(1). <https://doi.org/10.1186/s40168-021-01013-0>

DRAFT