




1 rbiom: An R Package for Streamlined Microbiome 2 Analysis and Automated Statistical Visualization

3 Daniel P Smith ^{1,2}, Sara J Javornik Cregeen ^{1,2}, and Joseph F
4 Petrosino ^{1,2}

5 **1** The Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and
6 Microbiology, Baylor College of Medicine, Houston, TX 77030, USA **2** Department of Molecular Virology
7 and Microbiology, Baylor College of Medicine, Houston, TX, 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Published: unpublished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](#))

8 Summary

9 rbiom is a software toolkit for the R programming language that simplifies the analysis of
10 complex ecological data. Its main purpose is to help researchers uncover hidden trends and
11 patterns by connecting what they observe in a biological sample (like the types and numbers
12 of bacteria) with information they know about that sample (like where it was collected or
13 the health of the individual). rbiom makes this process easy by focusing on creating clear,
14 publication-quality figures that include statistical test results right on the chart. This allows
15 scientists to quickly visualize and understand the important relationships within their data.

16 Statement of Need

A core challenge in microbiome analysis is identifying meaningful associations between a
community's composition and its associated metadata. While established packages like
phyloseq ([McMurdie & Holmes, 2013](#)) exist for handling the large and complex datasets often
stored in Biological Observation Matrix (BIOM) files, there is a persistent need for a tool
that streamlines the entire workflow from data ingestion to the creation of publication-quality
figures with integrated statistical results. rbiom addresses this need by providing a unified
interface that connects the analysis of taxonomic abundance and diversity with powerful
statistical methods and visualizations. Its design prioritizes speed and reproducibility, making
it particularly suitable for both routine data exploration and rigorous statistical reporting.

26 Key Features

27 The core design principles of rbiom focus on providing a seamless and reproducible workflow
28 for microbiome data analysis. Key features include:

- 29 ▪ **Integrated statistical visualization:** Statistical tests are integrated directly into plotting
30 functions, automatically annotating figures with results like p-values.
- 31 ▪ **Reproducibility:** The package provides the underlying R code used to generate figures
32 and statistics, allowing for easy customization and extension.
- 33 ▪ **Workflow simplification:** A unified interface connects data import, manipulation, analysis,
34 and visualization in a single toolkit.

Related Works

A recent review by Wen et al. (2023) highlighted the vast number of R packages available for microbiome analysis. While many of these tools offer some level of statistical capability, few make a concerted effort to display statistical results directly on generated figures. For example, the foundational phyloseq package lacks built-in significance testing. The closest peer to rbiom is microeco (Liu et al., 2020), which offers a comprehensive array of statistical tests and integrates their output into figures through the ggpubr package (?). rbiom extends this functionality by automatically positioning significance brackets, displaying the statistical method on the plot, and providing the underlying R code used to generate the figures and statistics. This approach enhances reproducibility and allows users to easily customize or extend the analysis.

Functionality

The rbiom package offers the following key functionalities:

- **Data Management and Manipulation:** rbiom seamlessly imports various data formats, including BIOM files, QIIME2 (Bolyen et al., 2019) and mothur (Schloss et al., 2009) outputs, and objects from other popular R packages using the phyloseq or SummarizedExperiment classes. The package includes a robust set of tools for rarefaction, filtering, and summarization, enabling users to prepare their data for downstream analysis.
- **Diversity and Composition Analysis:** The package focuses on three key community features: **alpha diversity**, **beta diversity**, and **taxa abundance**. For each, it can compute statistics against categorical or continuous metadata using appropriate methods such as **Kruskal-Wallis**, **Mann-Whitney**, **PERMANOVA**, and **estimated marginal means** (Anderson, 2001; Kruskal & Wallis, 1952; Lenth, 2025; Mann & Whitney, 1947).
- **Statistical Visualization:** A core feature of rbiom is its ability to directly overlay statistical test results onto ggplot2-based figures (Wickham, 2016). Functions like `adiv_boxplot()` and `stats_corrplot()` generate visualizations with automated annotations for p-values, trend lines, confidence intervals, and methodology. Customization is supported through parameters like `p.label` to display only significant results, ensuring clean, publication-ready graphics.
- **High-Level Plotting:** The package provides a diverse set of highly customizable plot types, including **boxplots**, **heatmaps**, **stacked bar charts**, and **ordination plots**. Users have extensive control over aesthetics, with options for specifying metadata variables for the x-axis, statistical groups, and plot faceting.

Example Figures

```
library(rbiom)
biom <- rarefy(hmp50) # hmp50 dataset is included with rbiom
bdiv_ord_plot(biom, bdiv = "UniFrac", stat.by = "Body Site", facet.by = "Sex")
```

Weighted UniFrac PCoA (Genus)

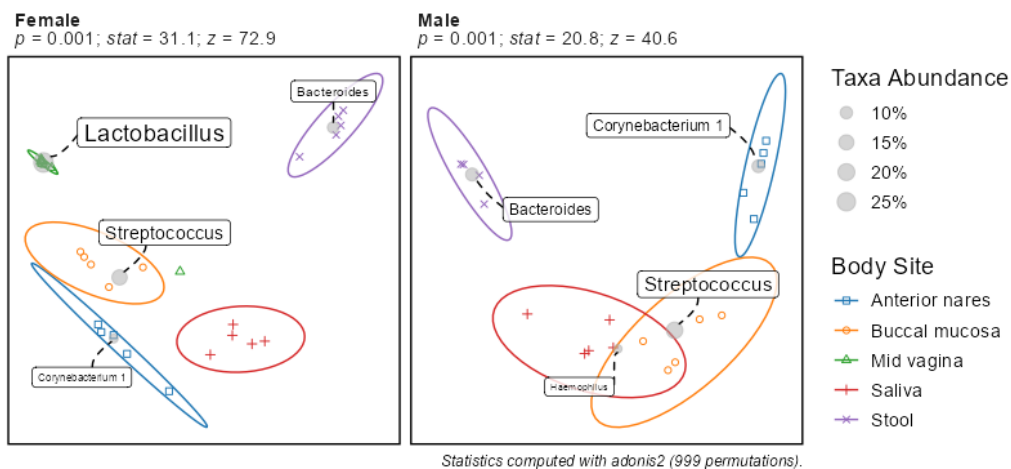


Figure 1: A beta diversity ordination plot. Samples cluster significantly by body site ($p = 0.001$) and are characterized by different bacterial genera.

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

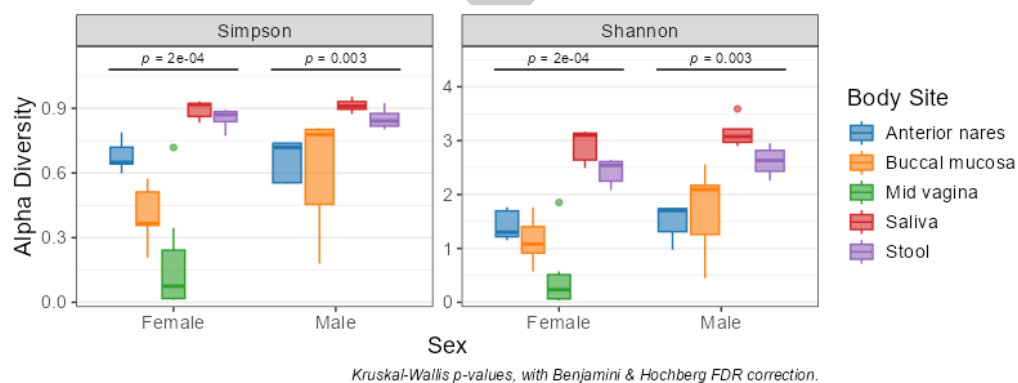


Figure 2: An alpha diversity box plot. Observed OTUs and Shannon diversity indices vary significantly by body site for both males ($p = 2e-04$) and females ($p = 0.003$).

```
subset(biom, `Body Site` == 'Buccal mucosa') %>%
  taxa_corrplot("Age", taxa = 2, layers = 'ptc', fit = 'lm', test = 'emtrends') +
  ggplot2::theme_classic()
```

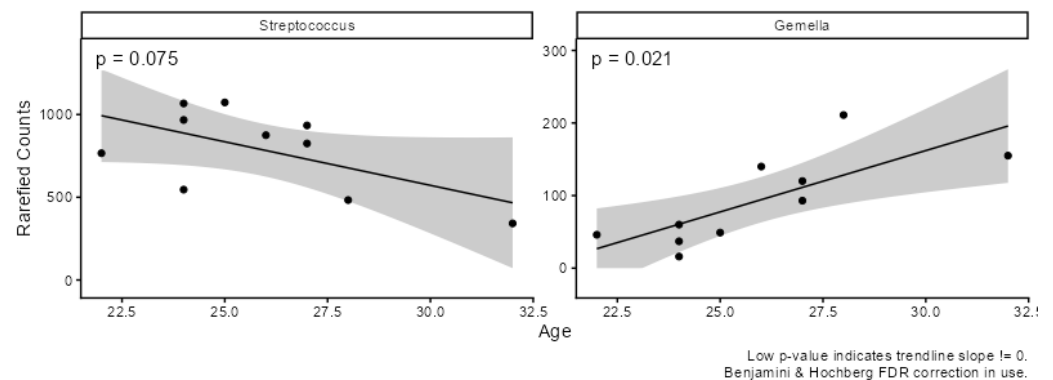


Figure 3: A taxa correlation plot using an alternative theme from ggplot2. The two most abundant buccal mucosa-associated genera show weak correlations with age.

Conclusion

By unifying data management, statistical analysis, and automated visualization, rbiom provides a powerful and accessible tool for microbiome research. The package's focus on reproducible, publication-ready graphics with built-in statistical results makes it a valuable addition to the ecosystem of R packages for ecological and biological data science.

Acknowledgements

This study was supported by NIH/NIAD (Grant number U19 AI44297), and Baylor College of Medicine and Alkek Foundation Seed.

The authors would like to thank Gemini for its assistance in drafting and refining this manuscript.

References

- Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1), 32–46. <https://doi.org/10.1046/j.1442-9993.2001.01070.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>
- 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>
- Lenth, R. V. (2025). *Emmeans: Estimated marginal means, aka least-squares means*. <https://doi.org/10.32614/CRAN.package.emmeans>
- 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>
- 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>

- 99 McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An r package for reproducible interactive
100 analysis and graphics of microbiome census data. *PloS One*, 8(4), e61217. <https://doi.org/10.1371/journal.pone.0061217>
101
- 102 Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B.,
103 Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., & others. (2009).
104 Introducing mothur: Open-source, platform-independent, community-supported software for
105 describing and comparing microbial communities. *Applied and Environmental Microbiology*,
106 75(23), 7537–7541. <https://doi.org/10.1128/AEM.01541-09>
- 107 Wen, T., Niu, G., Chen, T., Shen, Q., Yuan, J., & Liu, Y.-X. (2023). The best practice for
108 microbiome analysis using r. *Protein & Cell*, 14(10), 713–725. <https://doi.org/10.1093/procel/pwad024>
109
- 110 Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.
111 <https://doi.org/10.1007/978-3-319-24277-4>

DRAFT