microViz: an R package for microbiome data visualization and statistics

David J.M. Barnett[1, 2], Ilja C.W. Arts[1], and John Penders[2, 3, 4]

1 Maastricht Centre for Systems Biology (MaCSBio) **2** Maastricht University Medical Center+, Department of Medical Microbiology **3** Maastricht University, Public Health Research Institute (CAPHRI) **4** Maastricht University, School of Nutrition and Translational Research in Metabolism (NUTRIM)

Summary

microViz is an R package for the statistical analysis and visualization of microbiota data. This package extends the functionality of popular microbial ecosystem data analysis R packages, including phyloseq (McMurdie & Holmes, 2013), vegan (Oksanen et al., 2020) and microbiome (Lahti & Shetty, 2012-2019). microViz provides a selection of powerful additions to the toolbox of researchers already familiar with phyloseq and microbiome, as well as assisting researchers with less R programming experience to independently explore and analyse their data and to generate publication-ready figures.

The tools offered by microViz include:

- A Shiny app (Chang et al., 2021) for interactive exploration of microbiota data within R, pairing ordination plots with abundance bar charts
- Easy to use functions for generating publication-ready ordination plots with ggplot2 (Wickham, 2016), accommodating constrained and partial ordination, bi-plots and tri-plots, and automatic captioning designed to promote methodological transparency and reproducibility
- A novel visualization approach pairing ordination plots with circular bar charts (iris plots) for comprehensive, intuitive and compact visualization of the similarity and composition of hundreds of microbial ecosystems
- Correlation and composition heatmaps for microbiome data annotated with plots showing each taxon's prevalence and/or abundance
- A compact cladogram visualization approach for intuitive comparison of numerous microbe-metadata associations derived from (multivariable) statistical models (taxonomic association trees)

Statement of need

Modern microbiome research typically involves the use of next-generation-sequencing methods to profile the relative abundance of hundreds of microbial taxa across tens, hundreds or thousands of samples. Alongside increasing sample sizes, the amount of relevant metadata collected is growing, particularly in human cohort studies. These trends all increase the size and complexity of the resulting dataset, which makes its exploration, statistical analysis, and presentation increasingly challenging.

Ordination methods like Principle Coordinates/Components Analyses (PCoA/PCA) are a staple method in microbiome research. The vegan R package implements the majority of distance

and ordination calculation methods and microViz makes available two further dissimilarity measures: Generalized UniFrac from the GUniFrac package, and the Aitchison distance. The former provides a balanced intermediate between unweighted and weighted UniFrac (Chen et al., 2012), and the latter is a distance measure designed for use with compositional data, such as sequencing read counts (Gloor et al., 2017).

The phyloseq R package provides an interface for producing ordination plots, with the ggplot2 R package. microViz streamlines the computation and presentation of ordination methods including the constrained analyses: redundancy analysis (RDA), distance-based RDA, partial RDA, and canonical correspondence analysis (CCA). microViz can generate highly customizable ggplot2 bi-plots and tri-plots, showing labelled arrows for microbial loadings and constraint variables when applicable. Furthermore, these figures are captioned automatically, by default. The captions are intended to promote better reporting of ordination methods in published research, where too often insufficient information is given to reproduce the ordination plot. To provide the automated captioning, microViz implements a simple S3 list class, ps_extra, for provenance tracking, by storing distance matrices and ordination objects alongside the phyloseq object they were created from, as well as relevant taxonomic aggregation and transformation information.

Moreover, microViz provides a Shiny app interface (Chang et al., 2021) that allows the user to interactively create and explore ordination plots directly from phyloseq objects. The Shiny app generates code that can be copy-pasted into a script to reproduce the interactively designed ordination plot. The user can click and drag on the interactive ordination plot to select samples and directly examine their taxonomic compositions on a customizable stacked bar chart with a clear colour scheme.

Alternatively, for a comprehensive and intuitive static presentation of both sample variation patterns and underlying microbial composition, microViz provides an easy approach to pair ordination plots with attractive circular bar charts (iris plots) by ordering the bar chart in accordance with the rotational position of samples around the origin point on the ordination plot, e.g. Figure 1. Bar charts do have limitations when visualizing highly diverse samples, such as the adult gut microbiome, at a detailed taxonomic level. This is why microViz also offers an enhanced heatmap visualization approach, pairing an ordered heatmap of (transformed and/or scaled) microbial abundances with compact plots showing each taxon's overall prevalence and/or abundance distribution. The same annotation can easily be added to metadata-to-microbe correlation heatmaps.

microViz provides a flexible wrapper around methods for the statistical modelling of microbial abundances, including e.g. beta-binomial regression models from the corncob R package (Martin et al., 2020), and compositional linear regression. To visualize metadata-to-microbiome associations derived from more complicated statistical models, microViz offers a visualization approach that combines multiple annotated cladograms to comprehensively and compactly display patterns of microbial associations with multiple covariates from the same multivariable statistical model. These "taxonomic association trees" facilitate direct comparison of the direction, strength and significance of microbial associations between covariates and across multiple taxonomic ranks. This visualisation also provides an intuitive reminder of the balancing act inherent in compositional data analysis: if one clade/branch goes up, others must go down. Other packages in R, such as ggtree (Yu et al., 2017) or metacoder (Foster et al., 2017), can be used to make annotated cladograms similar to the microViz taxonomic association tree visualizations, but the microViz style has a few advantages for the purpose of reporting multivariable model results: Firstly, microViz cladogram generating functions are directly paired with functions to compute the statistical model results for all taxa in a phyloseq. Secondly, the tree layouts are more compact, by default, for displaying multiple trees for easy comparison.

Finally, beyond the main visualization functionality, microViz provides a suite of tools for working easily with phyloseq objects including wrapper functions that bring approaches from

the popular dplyr package to phyloseq, to help the researcher easily filter, select, join, mutate and arrange phyloseq sample data. All microViz functions are designed to work with magrittr's pipe operator (%>%), to chain successive functions together and improve code readability (Bache & Wickham, 2020). Lastly, for user convenience, microViz documentation and tutorials are hosted online via a pkgdown (Wickham & Hesselberth, 2020) website on Github Pages, with extensive examples of code and output generated with example datasets.
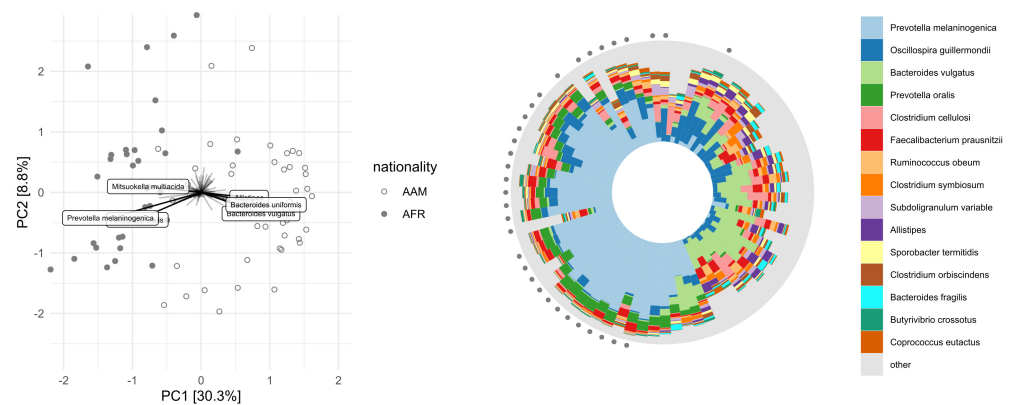
Figure 1: Simple example of a microViz figure pairing an ordination plot of microbial samples (left) with an "iris plot" (right): a circular stacked barchart showing the microbial compositions of samples ordered in accordance with the ordination plot. This figure is created with a subset of the "dietswap" dataset available within the microbiome R package. The ordination plot is a PCA bi-plot created using centered-log-ratio transformed species-like HITChip microbial features. The dark grey filled points on both plots indicate samples where the participant's nationality is AFR. AFR = African; AAM = African American.

Acknowledgements

References

Bache, S. M., & Wickham, H. (2020). *Magrittr: A forward-pipe operator for r.* https://CRAN.R-project.org/package=magrittr

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *Shiny: Web application framework for r.* https://CRAN.R-project.org/package=shiny

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman, R. G., Bushman, F. D., & Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, *28*(16), 2106–2113. https://doi.org/10.1093/bioinformatics/bts342

Foster, Z., Sharpton, T., & Grünwald, N. (2017). Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. *PLOS Computational Biology*, *13*(2), 1–15. https://doi.org/10.1371/journal.pcbi.1005404

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, *8*, 2224. https://doi.org/10.3389/fmicb.2017.02224

Lahti, L., & Shetty, S. (2012-2019). *Microbiome r package: Tools for microbiome analysis in r.* https://github.com/microbiome/microbiome

Martin, B. D., Witten, D., & Willis, A. D. (2020). Modeling microbial abundances and dysbiosis with beta-binomial regression. *The Annals of Applied Statistics*, *14*(1), 94–115. https://doi.org/10.1214/19-AOAS1283

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., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2020). *Vegan: Community Ecology Package.* https://CRAN.R-project.org/package=vegan

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4

Wickham, H., & Hesselberth, J. (2020). *Pkgdown: Make static HTML documentation for a package*. https://CRAN.R-project.org/package=pkgdown

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). Ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*, 28–36. https://doi.org/10.1111/2041-210X.12628