Preprint 10/2018

Microbiome data science

Sudarshan Shetty[1], Leo Lahti[2*]

[1]Laboratory of Microbiology, Wageningen University and Research, The Netherlands.
[2]Department of Mathematics and Statistics, University of Turku, Finland.
*Corresponding author <email: leo.lahti@iki.fi>

Abstract

The application of best practices of open data science is spreading across research fields, facilitating data sharing, collaborative methods development, research, and education. Microbiome bioinformatics is a rapidly developing area that can greatly benefit from this progress. The concept of microbiome data science refers to the application of open development model in microbiome bioinformatics. The increasing availability of microbiome profiling data, popularity of collaborative methods development, and the emergence of standard data formats are greatly facilitating the development of best practices in this field. A microbiome data science ecosystem combines experimental data sets with open research software, transparent and quality-controlled workflows, and reproducible tutorials that also serve as an educational resource. Here, we provide an overview of the current status of microbiome data science from a community developer perspective, discuss the prevailing gaps, and propose directions for future methods development.

Keywords Microbiome; Bioinformatics; Data science; Open science; Research Software

1. Introduction

Analysis of molecular profiling data obtained from high-throughput "-omics" approaches is essential for unravelling large-scale patterns in community composition, function and interactions between microbial organisms. The development of bioinformatics tools has been pivotal for understanding the importance of microbiome in human health (Erickson *et al.* 2012; Heintz-Buschart *et al.* 2017; Schirmer *et al.* 2018). Numerous tools from command line interfaces such as Mothur (Schloss *et al.* 2009) and the Python-based QIIME and QIIME2 (Bolyen E *et al.* 2018; Caporaso *et al.* 2010) to web-based tools such as Calypso (Zakrzewski *et al.* 2016) and MicrobiomeAnalyst (Dhariwal *et al.* 2017) have been designed to serve microbial bioinformaticians. The methods are developing rapidly, however, and the latest techniques are often not available even in actively maintained software projects and the quality and accessibility of published methods can vary widely (Mangul *et al.* 2018).

Community-driven data science ecosystems provide accelerated access to latest research algorithms. The emergence of open data science (Lahti 2018) has revolutionized collaborative research and is greatly facilitating the development and adoption of methods and best practices in data-intensive research fields. The availability of open data and research software, and open collaboration through distributed version control systems (Wilson *et al.* 2017) have created opportunities to transparently benchmark and criticize alternative approaches. Much of such development is currently focused on R and Python, where researchers share experimental software and reproducible notebooks that summarize complete data analytical procedures and provide practical guidance for research use. Users can further benefit from graphical interfaces (Venables and Smith 2006).

We provide a brief overview of the current status of microbiome data science from a community developer perspective. While the R ecosystem is one of the main platforms for current community-driven development efforts and our focus in this review, the key concepts apply more widely to other data science environments.

2. Microbiome data science

The route from processing of raw data to final analysis and reporting relies on a vast number of methods and basic concepts in microbial ecology (Figure 1). A single researcher is seldom able to fully master all relevant areas, and multi-disciplinary research can be supported by targeted data science ecosystems. These refer to well-designed combination of data, methods, and documentation that facilitate correct application of methods (Pollock *et al.* 2018; Knight *et al.* 2018; Schloss 2018b). Research software is best communicated in the context of experimental benchmarking data, combined with transparent workflows and reproducible online tutorials that serve as educational resources as well as open collaboration platform for methods development. The key elements enabling microbiome data science include open data, open methods, and open collaboration (Lahti 2018).

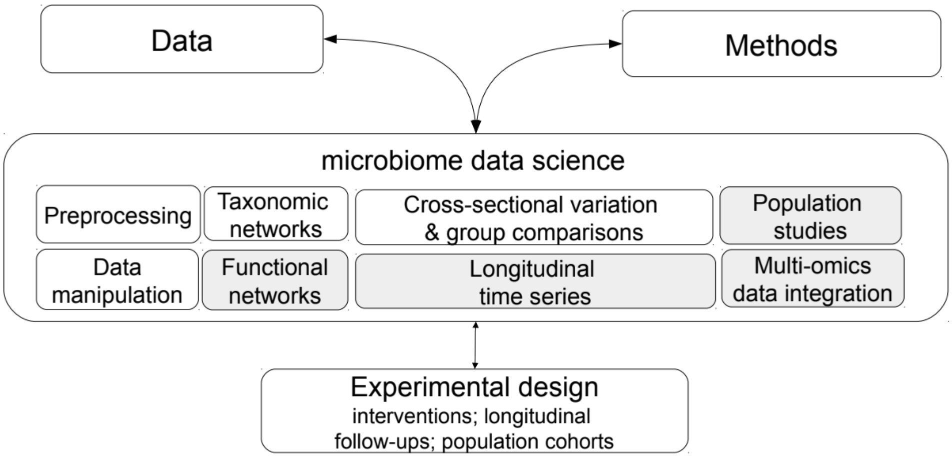

Figure 1: The current stage of microbiome data science ecosystem in R. The shaded boxes indicate research areas where the demand for new algorithmic tools is rapidly increasing.

2.1 Data

Convenient access to data is valuable for verification, meta-analysis, methods development and benchmarking. Availability of example data from published case studies in a readily accessible format can be highly convenient, and various R packages provide taxonomic and functional data from recent population and intervention studies of the human microbiome (Pasolli *et al.* 2017; Schiffer *et al.* 2018; Lahti and Shetty 2018).

2.2 Analysis

The contemporary R ecosystem for microbiome data science covers dozens of packages serving various analysis needs (Table 1). Most of the available methods focus on 16S rRNA amplicon sequencing or assume that OTU tables are readily available from metagenomic sequencing studies. Data summarization is facilitated by dedicated preprocessing algorithms such as *DADA2* (Callahan, McMurdie*, et al.* 2016), and class structures such as *phyloseq,* which is used to integrate OTU counts, taxonomic trees, and sample metadata into a single object that serves as a standardized starting point for various downstream methods (McMurdie and Holmes 2013). The *MultiAssayExperiment* provides utilities for parallel multi-omics experiments (Ramos *et al.* 2017), and further class structures are available for generic time series but these opportunities have not yet been fully exploited in the microbiome data science. Whereas Python has a more versatile set of algorithms for sequencing studies, R is well-suited for many interactive statistical analysis tasks. Estimation of alpha diversity and related ecological indices including richness, evenness, dominance, and rarity indices is a common task that has been implemented in various packages (Oksanen *et al.* 2011; Lahti and Shetty 2018) and can be complemented by phylogenetic trees (Kembel *et al.* 2010) or co-occurrence networks (Willis and Martin 2018). Community dissimilarity, or beta diversities, can be analysed using both phylogenetic (Chen 2012) and non-phylogenetic metrics (Beals 1984). Many methods are available for differential abundance analysis in individual taxa (Love, Huber, and Anders 2014; Robinson, McCarthy, and Smyth 2010; Paulson, Pop, and Bravo 2013; Fernandes *et al.* 2014), with varying performance (Weiss *et al.* 2017). Advanced approaches consider nested hierarchies in multiple testing scenarios (Sankaran and Holmes 2014). Community-level differences between sample groups with PERMANOVA and other methods (Oksanen *et al.* 2011; Anderson and Walsh 2013) can be complemented by unsupervised analyses (Sankaran and Holmes 2018b; Singh *et al.* 2018) such as Dirichlet Multinomial Mixtures (DMMs) (Harris *et al.* 2014; Ding and Schloss 2014).

Further tools are available for phylogenetic tree analysis (Paradis, Claude, and Strimmer 2004; Stevens *et al.* 2017; Washburne *et al.* 2017; Wright 2016), co-occurrence networks (Kurtz *et al.* 2015; Schwager *et al.* 2014), metabolic interactions (Cao *et al.* 2016), and microbiome function (Aßhauer *et al.* 2015). Visualization tools span from amplicon sequencing data (Andersen KSS *et al.* 2018) to unsupervised ordination by incorporating phylogenetic structure (Fukuyama 2017) to network analysis (Csardi and Nepusz 2006), phylogenetic trees (Paradis, Claude, and Strimmer 2004), taxonomic diversity (Foster, Sharpton, and Grünwald 2017), and geospatial analysis (Charlop-Powers and Brady 2015). Many generic utilities for microbiome profiling data are also available (Lagkouvardos *et al.* 2017; Chen, Simpson, and Levesque 2016; Lahti and Shetty 2018; Korpela 2016). R packages have also been created to access taxonomic information (Chamberlain *et al.* 2014) and to support interoperability with other systems such as the Python-based QIIME (Bittinger 2014). CRAN has strict technical checks for package consistency, and rOpenSci (Boettiger *et al.* 2015) and Bioconductor (Gentleman *et al.* 2004) have comprehensive software review procedures that signal good software quality.

2.3 Workflows

Sharing of technical knowledge and best practices can be greatly facilitated by transparent workflows, tutorials and online resources (Table1) that cover diverse aspects of microbiome data science (Schloss 2018a; Callahan, Sankaran*, et al.* 2016). Community-driven development can help to democratize microbiome data science and limit the monopoly of a few by facilitating free and open knowledge sharing. Good practices include routine application of automated unit tests and crowd-sourced quality control in the form of issue reports and case studies on reproducible notebooks (Wilson *et al.* 2017).

3. Discussion

Microbiome data science facilitates collaborative development and access to various concepts and methods in microbial ecology. Whereas we have provided a brief overview of the current microbiome data science ecosystem in R including data, methods, and educational resources, further methods are available in Python and other environments. The current R ecosystem is heavily focused on 16S analysis, and many packages contain overlapping functionality whose performance has not yet been comprehensively compared and benchmarked. Despite the progress in the field, the current microbiome data science ecosystem is specifically lacking dedicated methods for the analysis and integration of deep metagenomic and multi-omics profiling data and multivariate time series from targeted case studies and large population cohorts.

Pre-processing of raw reads to ASVs/OTUs BioC: dada2 (Callahan, McMurdie*, et al.* 2016)

Taxonomic classification and analysis BioC: rRDP (Hahsler and Nagar 2014), DECIPHER (IDTAXA algorithm) (Murali, Bhargava, and Wright 2018); CRAN: taxize (Chamberlain *et al.* 2014), microclass (Liland, Vinje, and Snipen 2017)

General data manipulation and visualisation BioC: Phyloseq (McMurdie and Holmes 2013), microbiome(Lahti and Shetty 2018); CRAN: vegan (Oksanen *et al.* 2011); theseus (Price *et al.* 2018), metacoder (Foster, Sharpton, Grünwald 2017); Github: mare (Korpela 2016), ampvis2 (Andersen KSS *et al.* 2018), microbiomeutilities (https://goo.gl/L4S5D6), microbiomeSeq (https://goo.gl/rfg5sA), yingtools2 (https://goo.gl/rfg5sA)

Diversity analysis CRAN: picante (Kembel *et al.* 2010), GUniFrac (Chen 2012), labdsv (Roberts 2007), breakaway (Willis and Bunge 2016), ape (Paradis, Claude, and Strimmer 2004), RAM (Chen, Simpson, and Levesque 2016); Github: DivNet (Willis and Martin 2018)

Community types BioC: DirichletMultinomial (Morgan 2017)

Network analysis BioC: CCREPE (Schwager *et al.* 2014); CRAN: igraph (Csardi and Nepusz 2006); Github: SPIEC-EASI (Kurtz *et al.* 2015)

Group-wise comparisons and association analysis BioC: structSSI, edgeR, DESeq2, metagenomeSeq; CRAN: mixOmics (Rohart *et al.* 2017), mixDIABLO (Singh *et al.* 2018), mixMC (Le Cao *et al.* 2016), Sigtree (Stevens *et al.* 2017), ALDEx2 (Fernandes *et al.* 2014)

Time series analysis Github: Seqtime (Faust *et al.* 2018), bootLong (https://goo.gl/jkXzQZ), treelapse (Sankaran and Holmes 2018a)

Pipelines/GUIs BioC: Pathostat (Manimaran *et al.* 2018), shiny-phyloseq (McMurdie and Holmes 2015), metavizr (Bravo HC *et al.* 2017); Github: Rhea (Lagkouvardos *et al.* 2017), DAME (Piccolo *et al.* 2018)

Interoperability CRAN: qiimer (Bittinger 2014)

Workflows and Tutorials

Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses (Callahan, Sankaran*, et al.* 2016)
The Riffomonas Reproducible Research Tutorial Series (Schloss 2018a),
Happy belly bioinformatics (https://astrobiomike.github.io/)
Microbiome package tutorial series (http://microbiome.github.io/microbiome/)
Open & Reproducible Microbiome Data Analysis (https://goo.gl/CPChhd)
Random Forest Modelling of the Lake Erie microbial community (https://tinyurl.com/ycz4rgfv) (Rpubs)

Table 1: Overview of the currently available online resources for microbiome data science in R. Bioconductor has the strictest software review procedure covering technical aspects as well as the package contents; CRAN requires comprehensive technical quality checks with minimal content review; and Github can host emerging or more established projects with no formal quality control. The indicated groupings are approximations as many packages span over multiple categories.

4. Acknowledgements

LL was funded by the Academy of Finland (grants 295741; 307127). Part of this research was by the SIAM Gravitation Grant 024.002.002 and UNLOCK project of the Netherlands Organization for Scientific Research (NWO).

5. References

Andersen KSS, Kirkegaard RH, Karst SM, and Mads A. 2018. 'ampvis2: an R package to analyse and visualise 16S rRNA amplicon data', *bioRxiv*: 299537.

Anderson, Marti J., and Daniel C. I. Walsh. 2013. 'PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing?', *Ecol. Monogr.*, 83: 557-74.

Aßhauer, Kathrin P., Bernd Wemheuer, Rolf Daniel, and Peter Meinicke. 2015. 'Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data', *Bioinformatics*, 31: 2882-84.

Beals, Edward W. 1984. 'Bray-Curtis Ordination: An Effective Strategy for Analysis of Multivariate Ecological Data.' in A. MacFadyen and E. D. Ford (eds.), *Advances in Ecological Research* (Academic Press).

Bittinger, K. 2014. 'qiimer: Work with QIIME output files in R', *R package version 0. 9*, 2.

Boettiger, Carl, Scott Chamberlain, Edmund Hart, and Karthik Ram. 2015. 'Building software, building community: lessons from the rOpenSci project', *Journal of Open Research Software*, 3.

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ, et al. 2018. 'QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science', *PeerJ Preprints*, 6:e27295v1.

Bravo HC, Chelaru F, Wagner J, Kancherla J, and Paulson J. 2017. "metavizr: R Interface to the metaviz web app for interactive metagenomics data analysis and visualization." In.: Bioconductor.

Callahan, Ben J., Kris Sankaran, Julia A. Fukuyama, Paul J. McMurdie, and Susan P. Holmes. 2016. 'Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses', *F1000Res.*, 5: 1492.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. 'DADA2: high-resolution sample inference from Illumina amplicon data', *Nature methods*, 13: 581.

Cao, Yang, Yuanyuan Wang, Xiaofei Zheng, Fei Li, and Xiaochen Bo. 2016. 'RevEcoR: an R package for the reverse ecology analysis of microbiomes', *BMC bioinformatics*, 17: 294.

Caporaso, J Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, et al. 2010. 'QIIME allows analysis of high-throughput community sequencing data', *Nature methods*, 7: 335.

Chamberlain, Scott, Eduard Szocs, Carl Boettiger, Karthik Ram, Ignasi Bartomeus, John Baumgartner, Zachary Foster, and James O'Donnell. 2014. 'taxize: Taxonomic information from around the web', *R package version*, 30.

Charlop-Powers, Zachary, and Sean F Brady. 2015. 'phylogeo: an R package for geographic analysis and visualization of microbiome data', *Bioinformatics*, 31: 2909-11.

Chen, J. 2012. 'GUniFrac: generalized UniFrac distances', *R package version*, 1: 2012.

Chen, W., J. Simpson, and C. Levesque. 2016. 'RAM: R for amplicon-sequencing-based microbial-ecology', *R package version*, 1.

Csardi, Gabor, and Tamas Nepusz. 2006. 'The igraph software package for complex network research', *InterJournal, Complex Systems*, 1695: 1-9.

Dhariwal, Achal, Jasmine Chong, Salam Habib, Irah L King, Luis B Agellon, and Jianguo Xia. 2017. 'MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data', *Nucleic Acids Research*: gkx295.

Ding, Tao, and Patrick D Schloss. 2014. 'Dynamics and associations of microbial community types across the human body', *Nature*, 509: 357.

Erickson, Alison R, Brandi L Cantarel, Regina Lamendella, Youssef Darzi, Emmanuel F Mongodin, Chongle Pan, Manesh Shah, Jonas Halfvarson, et al. 2012. 'Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease', *PLoS One*, 7: e49138.

Faust, Karoline, Franziska Bauchinger, Béatrice Laroche, Sophie de Buyl, Leo Lahti, Alex D Washburne, Didier Gonze, and Stefanie Widder. 2018. 'Signatures of ecological processes in microbial community time series', *Microbiome*, 6: 120.

Fernandes, Andrew D, Jennifer NS Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. 2014. 'Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis', *Microbiome*, 2: 15.

Foster, Zachary SL, Thomas J Sharpton, and Niklaus J Grünwald. 2017. 'Metacoder: An R package for visualization and manipulation of community taxonomic diversity data', *PLoS computational biology*, 13: e1005404.

Fukuyama, Julia. 2017. 'Adaptive gPCA: A method for structured dimensionality reduction', *arXiv [stat.ME]*.

Gentleman, Robert C., Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, et al. 2004. 'Bioconductor: open software development for computational biology and bioinformatics', *Genome Biol.*, 5: R80.

Hahsler, Michael, and Anurag Nagar. 2014. 'rRDP: Interface to the RDP Classifier'.

Harris, Keith, Todd L Parsons, Umer Z Ijaz, Leo Lahti, Ian Holmes, and Christopher Quince. 2014. 'Linking statistical and ecological theory: Hubbell's unified neutral theory of biodiversity as a hierarchical Dirichlet process'.

Heintz-Buschart, Anna, Patrick May, Cédric C Laczny, Laura A Lebrun, Camille Bellora, Abhimanyu Krishna, Linda Wampach, Jochen G Schneider, et al. 2017. 'Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes', *Nature microbiology*, 2: 16180.

Kembel, Steven W., Peter D. Cowan, Matthew R. Helmus, William K. Cornwell, Helene Morlon, David D. Ackerly, Simon P. Blomberg, and Campbell O. Webb. 2010.

'Picante: R tools for integrating phylogenies and ecology', *Bioinformatics*, 26: 1463-64.

Knight, Rob, Alison Vrbanac, Bryn C. Taylor, Alexander Aksenov, Chris Callewaert, Justine Debelius, Antonio Gonzalez, Tomasz Kosciolek, et al. 2018. 'Best practices for analysing microbiomes', *Nat. Rev. Microbiol.*, 16: 410-22.

Korpela, Katri. 2016. 'mare: Microbiota Analysis in R Easily. R package version 1.0.'.

Kurtz, Zachary D, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. 2015. 'Sparse and compositionally robust inference of microbial ecological networks', *PLoS Comput Biol*, 11: e1004226.

Lagkouvardos, Ilias, Sandra Fischer, Neeraj Kumar, and Thomas Clavel. 2017. 'Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons', *PeerJ*, 5: e2836.

Lahti, Leo. 2018. ' “Open Data Science”'.

Lahti, Leo, and Sudarshan A Shetty. 2018. 'Tools for microbiome analysis in R'.

Le Cao, Kim-Anh, Mary-Ellen Costello, Vanessa Anne Lakis, Francois Bartolo, Xin-Yi Chua, Remi Brazeilles, and Pascale Rondeau. 2016. 'MixMC: a multivariate statistical framework to gain insight into microbial communities', *PLoS One*, 11: e0160169.

Liland, Kristian Hovde, Hilde Vinje, and Lars Snipen. 2017. 'microclass: an R-package for 16S taxonomy classification', *BMC bioinformatics*, 18: 172.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15: 550.

Mangul, Serghei, Thiago Mosqueiro, Dat Duong, Keith Mitchell, Varuni Sarwal, Brian Hill, Jaqueline Brito, Russell Littman, et al. 2018. 'A comprehensive analysis of the usability and archival stability of omics computational tools and resources', *bioRxiv*.

Manimaran, S, M Bendall, SV Diaz, E Castro, T Faits, Y Zhao, and WE Johnson. 2018. 'PathoStat: PathoStat Statistical Microbiome Analysis Package. R package version 1.6.1, '.

McMurdie, Paul J, and Susan Holmes. 2013. 'phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data', *PLoS One*, 8: e61217.

McMurdie, PJ, and S Holmes. 2015. 'Shiny-phyloseq: Web application for interactive microbiome analysis with provenance tracking', *Bioinformatics*, 31: 282-83.

Morgan, M. 2017. 'DirichletMultinomial'.

Murali, Adithya, Aniruddha Bhargava, and Erik S. Wright. 2018. 'IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences', *Microbiome*, 6: 140.

Oksanen, Jari, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O’hara, Gavin L. Simpson, Peter Solymos, et al. 2011. 'vegan: Community ecology package', *R package version*: 117-18.

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. 'APE: Analyses of Phylogenetics and Evolution in R language', *Bioinformatics*, 20: 289-90.

Pasolli, Edoardo, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, et al. 2017. 'Accessible, curated metagenomic data through ExperimentHub', *Nat. Methods*, 14: 1023-24.

Paulson, Joseph Nathaniel, M. Pop, and H. C. Bravo. 2013. 'metagenomeSeq: Statistical analysis for sparse high-throughput sequencing', *Bioconductor package*, 1.

Piccolo, Brian D., Umesh D. Wankhade, Sree V. Chintapalli, Sudeepa Bhattacharyya, Luo Chunqiao, and Kartik Shankar. 2018. 'Dynamic assessment of microbial ecology (DAME): a web app for interactive analysis and visualization of microbial sequencing data', *Bioinformatics*, 34: 1050-52.

Pollock, Jolinda, Laura Glendinning, Trong Wisedchanwet, and Mick Watson. 2018. 'The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies', *Appl. Environ. Microbiol.*, 84.

Price, Jacob R, Stephen Woloszynek, Gail L Rosen, and Christopher M Sales. 2018. 'theseus-An R package for the analysis and visualization of microbial community data', *bioRxiv*: 295675.

Ramos, Marcel, Lucas Schiffer, Angela Re, Rimsha Azhar, Azfar Basunia, Carmen Rodriguez, Tiffany Chan, Phil Chapman, et al. 2017. 'Software for the Integration of Multiomics Experiments in Bioconductor', *Cancer Res.*, 77: e39-e42.

Roberts, David W. 2007. 'labdsv: Ordination and multivariate analysis for ecology', *R package version*, 1.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26: 139-40.

Rohart, Florian, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. 2017. 'mixOmics: An R package for 'omics feature selection and multiple data integration', *PLoS Comput. Biol.*, 13: e1005752.

Sankaran, K, and S Holmes. 2018a. 'Interactive Visualization of Hierarchically Structured Data', *J. Comput. Graph. Stat.*, 27: 553-63.

Sankaran, Kris, and Susan Holmes. 2014. 'structSSI: simultaneous and selective inference for grouped or hierarchically structured data', *Journal of Statistical Software*, 59: 1.

Sankaran, Kris, and Susan P. Holmes. 2018b. 'Latent variable modeling for the microbiome', *Biostatistics*.

Schiffer, Lucas, Rimsha Azhar, Lori Shepherd, Marcel Ramos, Ludwig Geistlinger, Curtis Huttenhower, Jennifer B. Dowd, Nicola Segata, et al. 2018. 'HMP16SData: Efficient Access to the Human Microbiome Project through Bioconductor', *bioRxiv*.

Schirmer, Melanie, Eric A. Franzosa, Jason Lloyd-Price, Lauren J. McIver, Randall Schwager, Tiffany W. Poon, Ashwin N. Ananthakrishnan, Elizabeth Andrews, et al. 2018. 'Dynamics of metatranscription in the inflammatory bowel disease gut microbiome', *Nat Microbiol*, 3: 337-46.

Schloss, Patrick D. 2018a. 'The Riffomonas Reproducible Research Tutorial Series', *Int. J. Occup. Saf. Ergon.*, 1: 13.

Schloss, Patrick D, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, et al. 2009. 'Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities', *Applied and Environmental Microbiology*, 75: 7537-41.

Schloss, Patrick D. 2018b. 'Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research', *MBio*, 9.

Schwager, Emma, George Weingart, Craig Bielski, and Curtis Huttenhower. 2014. 'CCREPE: Compositionality Corrected by Permutation and Renormalization'.

Singh, A., B. Gautier, C. P. Shannon, F. Rohart, M. Vacher, and others. 2018. 'DIABLO: from multi-omics assays to biomarker discovery, an integrative approach', *bioRxiv*.

Stevens, John R, Todd R Jones, Michael Lefevre, Balasubramanian Ganesan, and Bart C Weimer. 2017. 'SigTree: a microbial community analysis tool to identify and visualize significantly responsive branches in a phylogenetic tree', *Computational and structural biotechnology journal*, 15: 372-78.

Venables, William N, and David M Smith. 2006. 'The R development core team', *An Introduction to R. R Foundation for Statistical Computing, Vienna, Austria*.

Washburne, Alex D, Justin D Silverman, Jonathan W Leff, Dominic J Bennett, John L Darcy, Sayan Mukherjee, Noah Fierer, and Lawrence A David. 2017. 'Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets', *PeerJ*, 5: e2969.

Weiss, Sophie, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, et al. 2017. 'Normalization and microbial differential abundance strategies depend upon data characteristics', *Microbiome*, 5: 27.

Willis, A, and J Bunge. 2016. 'Species Richness Estimation and Modeling', *CRAN*.

Willis, Amy D., and Bryan D. Martin. 2018. 'DivNet: Estimating diversity in networked communities', *bioRxiv*.

Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. 2017. 'Good enough practices in scientific computing', *PLoS Comput. Biol.*, 13: e1005510.

Wright, Erik S. 2016. 'Using DECIPHER v2. 0 to analyze big biological sequence data in R', *The R Journal*, 8.

Zakrzewski, Martha, Carla Proietti, Jonathan J Ellis, Shihab Hasan, Marie-Jo Brion, Bernard Berger, and Lutz Krause. 2016. 'Calypso: a user-friendly web-server for mining and visualizing microbiome–environment interactions', *Bioinformatics*, 33: 782-83.