Microbiome Research Reports

Technical Note

Open Access

A tool to assess the mock community samples in 16S rRNA gene-based microbiota profiling studies

Sudarshan A. Shetty[1,2], **Jolanda Kool**[1], **Susana Fuentes**[1]

[1]Center for Infectious Disease Control, National Institute for Public Health and the Environment, Antonie van Leeuwenhoeklaan 9, Bilthoven 3721 MA, Netherlands.
[2]Department of Medical Microbiology and Infection Prevention, Virology and Immunology Research Group, University Medical Center Groningen, Hanzeplein 1, Groningen 9713 GZ, Netherlands.

Correspondence to: Dr. Sudarshan A. Shetty. Center for Infectious Disease Control, National Institute for Public Health and the Environment, Antonie van Leeuwenhoeklaan 9, Bilthoven 3721 MA, Netherlands. E-mail: sudarshanshetty9@gmail.com; Dr. Susana Fuentes. Center for Infectious Disease Control, National Institute for Public Health and the Environment, Antonie van Leeuwenhoeklaan 9, Bilthoven 3721 MA, Netherlands. E-mail: susana.fuentes@rivm.nl

How to cite this article: Shetty SA, Kool J, Fuentes S. A tool to assess the mock community samples in 16S rRNA gene-based microbiota profiling studies. *Microbiome Res Rep* 2023;2:14. https://dx.doi.org/10.20517/mrr.2022.18

Received: 23 Nov 2022 **First Decision:** 13 Feb 2023 **Revised:** 24 Mar 2023 **Accepted:** 17 Apr 2023 **Published:** 27 Apr 2023

Academic Editors: Jennifer Mahony, Marco Ventura **Copy Editor:** Ke-Cui Yang **Production Editor:** Ke-Cui Yang

Abstract

Inclusion and investigation of technical controls in microbiome sequencing studies is important for understanding technical biases and errors. Here, we present *chkMocks*, a general R-based tool that allows researchers to compare the composition of mock communities that are processed along with samples to their theoretical composition. A visual comparison between experimental and theoretical community composition and their correlation is provided for researchers to assess the quality of their sample processing workflows.

Keywords: Mock community, microbiome profiling, positive control

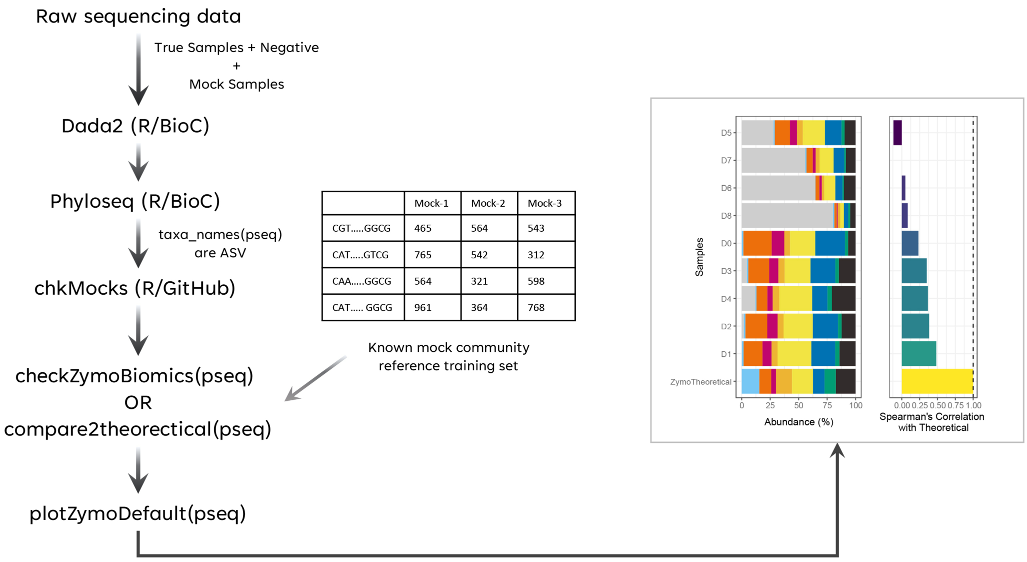
INTRODUCTION

Microbiota profiling of diverse environments is widely done using 16S rRNA gene sequencing. Preparation of samples for microbiota profiling consists of sampling, storage, DNA extraction, PCR, library preparation, sequencing, and downstream bioinformatics analysis[1-4]. At every step, technical variability is a major factor that can ultimately affect the observed microbiota profiles[5-8]. Including negative and positive controls, especially mock communities with known microbial composition, is suggested to help identify technical variability and improve protocols if required[9]. Mock communities with known composition can be included at the step of DNA extraction (mixture of different cells) or at the PCR step (mixture of DNA from different cells). This allows for evaluating where technical variation is introduced. For example, it is known that DNA extraction methods can differently bias certain cell types, e.g., Gram-positive and Gram-negative bacteria, and that primer choice at the PCR step can neglect or favor some organisms[5,8]. In addition, these mock communities allow for identifying potential reagent contamination, well-to-well contamination, and to some extent, cross-sample contamination[10-13]. Therefore, every microbiota profiling study should include both positive and negative controls during sample processing.

Analyzing the mock community profiles and comparing them to the theoretical composition is, however, not straightforward, especially for novice microbiome scientists. A very limited number of tools are available for analyzing and comparing mock communities. The QIIME2 consists of a plugin called q2-quality-control[14,15]. The ZymoBIOMICS research team provides a tool called FIGARO for ZymoBIOMICS™ Microbial Community Standard[16]. Here, we present an R-based tool, *chkMocks*, specifically designed for outputs from the R-based dada2 pipeline. The *chkMocks* R package provides a slightly different approach for investigating mock communities (see below). This tool provides support for ZymoBIOMICS™ Microbial Community Standard and offers the ability to use it for custom mock communities.

IMPLEMENTATION AND FEATURES

The *chkMocks* tool is implemented in R and depends on the following R packages/tools: *dada2*, *DECIPHER*, *tidyverse* tools, *microbiome*, *phyloseq* and *patchwork*[17-22]. An overview of the workflow/steps is depicted in Figure 1. The *chkMocks* tool requires data that completed the *dada2* workflow, from raw reads to obtaining the taxonomy assigned *phyloseq* object. The *phyloseq* object should have sequences of variants as taxa names and not be converted to text ID's like ASV:1, etc. The *chkMocks* tool can be used by two different approaches, distinguished by the type of mock sample that is used. If users have sequenced the ZymoBIOMICS™ Microbial Community Standard (Catalog No. D6300), they can use the default *checkZymoBiomics*. For this, we have created a taxonomic training set using the FASTA files for full-length 16S rRNA gene sequences of expected microbes provided by ZymoBiomics. To demonstrate the *chkMocks* utility, we used data from a study investigating reagent contamination using the ZymoBIOMICS™ Microbial Community Standard[10]. Here, the Microbial Community Standard was subjected to 8 series of a 3-fold dilution (D0 to D8) and processed for 16S rRNA gene-based microbiota profiling. The outputs of *checkZymoBiomics* are (a) A *phyloseq* object with input ASVs, their abundances and taxonomic assignments; (b) A *phyloseq* object with input ASVs aggregated to species level and their abundances; and (c) A correlation table with Spearman's correlation (rho) values of positive controls compared to theoretical composition. The user can simply plot the results with *plotZymoDefault*; this function visualizes the composition of positive controls and theoretical composition as a stacked bar plot [Figure 2A]. This is accompanied by a bar plot of Spearman's correlation (rho) between positive controls and theoretical composition [Figure 2B]. The user can also compare the abundances of individual taxa for a clearer understanding of biases towards specific taxa [Figure 2C and D]. Here, the percentage of 'unknown' taxa, i.e., not matching any of the expected taxa included in the mock community, increases as dilution increases and is in agreement with values reported by the original study. All these plots provide first-hand insights to the user about the quality of their sample processing by directly comparing positive controls with expected observations.



	Mock-1	Mock-2	Mock-3
CGT.....GGCG	465	564	543
CAT.....GTCG	765	542	312
CAA.....GGCG	564	321	598
CAT..... GGCG	961	364	768

Figure 1. Overview of the workflow for comparing experimental mock samples with the theoretically expected composition.

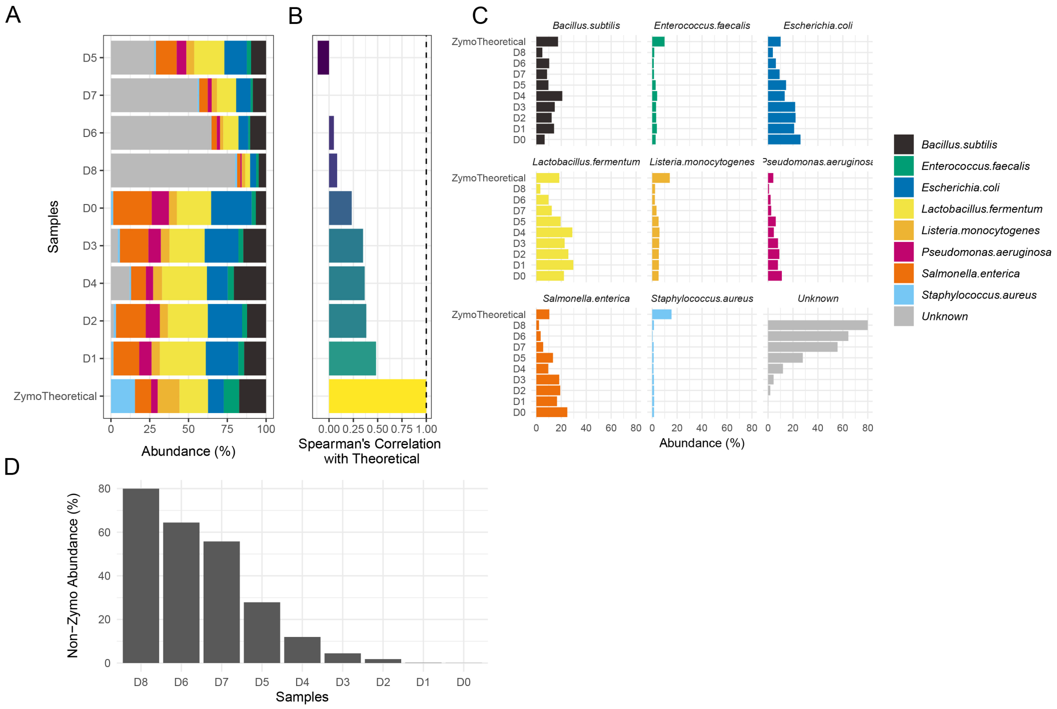


Figure 2. Overview of the key results generated by *chkMocks*.(A) Community composition of positive controls and expected composition of ZymoBIOMICS^{TM} Microbial Community Standard; (B) spearman's correlation (rho) values of positive controls compared to theoretical composition; (C) percent abundances of individual taxa; (D) percent abundances of "unknown" taxa, i.e., not matching any of the standard expected taxa.

For researchers using a custom mock community or mock communities from a different vendor, we provide a step-by-step guide on preparing the training set as a FASTA file for full-length 16S rRNA gene sequences of expected microbes using the *DECIPHER* R/BioC package. To this end, the taxonomic assignment can be done using the *assignTaxonomyCustomMock*. We provide this tutorial on the package website (https://microsud.github.io/chkMocks/) and include an example of how to compare the custom mocks with their theoretical composition. Of note, we rely on the *DECIPHER:IdTaxa* function for taxonomic assignments and *chkMocks* only supports bacteria and archaea[23].

To demonstrate the application for custom mock communities, we used data from a study investigating an ASV profiling tool, NG-Tax[24] and experimental samples from a previous synthetic microbiome study[25]. Additionally, we also provide training sets for the ZymoBIOMICS™ Microbial Community Standard (Catalog No. D6331) which consists of 19 of the 21 microbes. The two fungi, *Candida albicans* and *Saccharomyces cerevisiae*, are excluded from this training set.

CONCLUSION

The *chkMocks* was developed for the comparison of experimental mock communities with their expected compositions. The wet-lab protocols are often standardized depending on the target ecosystem that is investigated. Standardization requires analysis of positive controls, which are often microbial communities of known composition. Furthermore, a comparison of mock communities between batches when processing a large number of samples can help identify any technical variability. We developed a simple-to-use R package to ease the process of standardization and general quality check.

DECLARATIONS

Authors' contributions

Conceptualized the work: Shetty SA, Fuentes S
Wrote the code: Shetty SA
Provided technical assistance: Kool J

Availability of data and materials

The *chkMocks* is implemented in the R statistical language and is released under the MIT license. The source code and associated example data are available at: https://github.com/microsud/chkMocks/.

Financial support and sponsorship

This work was supported by the Dutch Ministry of Health, Welfare and Sport, and the Strategic Program of the National Institute for Public Health and the Environment (RIVM).

Conflicts of interest

All authors declared that there are no conflicts of interest. While we use ZymoBiomics data, we, the developers of *chkMocks,* are not associated with the manufacturers and this work should not be considered as an endorsement for the said product.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

REFERENCES

1. Sorbie A, Delgado Jiménez R, Weiler M, Benakis C. Protocol for microbiota analysis of a murine stroke model. *STAR Protoc* 2023;4:101969. DOI PubMed PMC
2. Love CJ, Gubert C, Kodikara S, Kong G, Lê Cao KA, Hannan AJ. Microbiota DNA isolation, 16S rRNA amplicon sequencing, and bioinformatic analysis for bacterial microbiome profiling of rodent fecal samples. *STAR Protoc* 2022;3:101772. DOI PubMed PMC
3. Ghosh TS, Das M. Chapter two - emerging tools for understanding the human microbiome. *Prog Mol Biol Transl Sci* 2022;191:29-51. DOI
4. Amir A. Microbiome analysis using 16S amplicon sequencing: from samples to ASVs. In: Shomron N, editor. Deep sequencing data analysis. Methods in molecular biology. New York; 2021. pp. 123-41. DOI
5. Hornung BVH, Zwittink RD, Kuijper EJ. Issues and current standards of controls in microbiome research. *FEMS Microbiol Ecol* 2019:95. DOI PubMed PMC
6. Kim D, Hofstaedter CE, Zhao C, et al. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* 2017;5:52. DOI PubMed PMC
7. Bokulich NA, Ziemski M, Robeson MS 2nd, Kaehler BD. Measuring the microbiome: best practices for developing and benchmarking microbiomics methods. *Comput Struct Biotechnol J* 2020;18:4048-62. DOI PubMed PMC
8. Parada AE, Needham DM, Fuhrman JA. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* 2016;18:1403-14. DOI PubMed
9. Quince C, Lanzén A, Curtis TP, et al. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 2009;6:639-41. DOI
10. Karstens L, Asquith M, Davin S, et al. Controlling for contaminants in low-biomass 16S rRNA gene sequencing experiments. *mSystems* 2019:4. DOI PubMed PMC
11. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 2014;12:87. DOI PubMed PMC
12. Minich JJ, Sanders JG, Amir A, Humphrey G, Gilbert JA, Knight R. Quantifying and understanding well-to-well contamination in microbiome research. *mSystems* 2019:4. DOI PubMed PMC
13. Minich JJ, Zhu Q, Janssen S, et al. KatharoSeq enables high-throughput microbiome analysis from low-biomass samples. *mSystems* 2018:3. DOI PubMed PMC
14. Bokulich NA, Kaehler BD, Rideout JR, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* 2018;6:90. DOI PubMed PMC
15. Bolyen E, Rideout JR, Dillon MR, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 2019;37:852-7. DOI
16. Weinstein MM, Prem A, Jin M, Tang S, Bhasin JM. FIGARO: an efficient and objective tool for optimizing microbiome rRNA gene trimming parameters. *bioRxiv* ;2019:610394. DOI
17. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016;13:581-3. DOI PubMed PMC
18. Wright ES. Using DECIPHER v2.0 to analyze big biological sequence data in R. Available from: https://pdfs.semanticscholar.org/687f/973e9b1416a1289a86e58474e7259bdb57f1.pdf [Last accessed on 26 Apr 2023].
19. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. *J Open Res Softw* 2019;4:1686. DOI
20. Lahti L, Shetty SA. Tools for microbiome analysis in R. Available from: https://bioconductor.org/packages/release/bioc/html/microbiome.html [Last accessed on 26 Apr 2023].
21. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 2013;8:e61217. DOI PubMed PMC
22. Pedersen TL. Patchwork: the composer of plots. Available from: https://github.com/thomasp85/patchwork [Last accessed on 26 Apr 2023].
23. Murali A, Bhargava A, Wright ES. IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome* 2018;6:140. DOI PubMed PMC
24. Ramiro-Garcia J, Hermes GDA, Giatsis C, et al. NG-Tax, a highly accurate and validated pipeline for analysis of 16S rRNA amplicons from complex biomes. *F1000Res* 2016;5:1791. DOI PubMed PMC
25. Shetty SA, Kostopoulos I, Geerlings SY, Smidt H, de Vos WM, Belzer C. Dynamic metabolic interactions and trophic roles of human gut microbes identified using a minimal microbiome exhibiting ecological properties. *ISME J* 2022;16:2144-59. DOI PubMed PMC