Effect of Internal Standard Normalization of Microbiome Data on Outcomes of a Controlled Feeding Study and a Longitudinal Study in a Multiethnic Cohort

Jacob Edward Fong-Gurzinsky

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2021

Committee:

Johanna W. Lampe, Chair

Meredith A. J. Hullar

Jennifer E. Balkus

Program Authorized to Offer Degree:

Department of Epidemiology

University of Washington

Abstract

Effect of Internal Standard Normalization of Microbiome Data on Outcomes of a Controlled Feeding Study and a Longitudinal Study in a Multiethnic Cohort

Jacob Edward Fong-Gurzinsky

Chair of the Supervisory Committee:

Johanna W. Lampe

Department of Epidemiology

Standardization would benefit the interpretability of human microbiome data because unintended variability can be introduced at each level of data production and processing. One way to bring standardization to microbiome studies is with internal standards (IS). In three microbiome sequencing methods—16S rRNA gene sequencing, and metagenomic and metatranscriptomic sequencing—this standardization can involve the addition of nucleic acid sequences into a sample. We assessed how internal standards would change the interpretation of the data in a controlled feeding study and longitudinal study in a multiethnic cohort. We compared both coefficients of variation (CV) and intraclass correlation coefficients (ICC) for both IS-normalized data and non-normalized data for both studies. The effect of IS-normalization on the ICCs and CVs was inconsistent in our results and did not improve data interpretation.

Acknowledgments

Thank you to Johanna Lampe and Meredith Hullar for their guidance on this project. I have learned a great deal from this process and in being a part of your research group. Thank you to Jennifer Balkus for her feedback on this thesis. Thank you to Keith Curtis, Sepideh Moghadam, Orsalem Kahsai, Benjamin Fu, and Lisa Levy for their technical assistance and advice. Thank you to my friends and family for their support throughout my time in graduate school.

Chapter 1: Use of Internal Standards with Microbiota Data

Introduction

The human gut microbiota makes up a complex microbial community whose disruption has been associated with diseases of the digestive system, but also broadly with the nervous, cardiovascular, immune, endocrine, and respiratory systems.[1] Changes in the composition of the human gut microbiota has implications for therapies; one such example is the finding that certain bacteria within the gut can affect how well cancer treatment works on tumors.[1–3] The microbiota can be studied with biomarker sequencing (such as the 16S rRNA gene) to determine the species diversity of a microbial community, metagenomics sequencing to assemble genomes and infer the function of the genes, metatranscriptomics sequencing to determine the RNA being made by the cell, metaproteomics to determine the proteins being made from the RNA, and metabolomics to determine the metabolites being produced.[1] This paper will focus on the 16s rRNA gene, metagenomic, and metatranscriptomic sequencing. In order to be effective for use in human health, standardization is essential in human microbiota studies, but studies implementing such standardization have not been common.[4] Specifically, standardization allows for studies to be repeated and compared to each other, in addition to improving the strength of the analysis, which allows for the microbiome to be relevant for epidemiologists and for clinicians.[4–6] The Microbiome Quality Control project baseline study had the goal of studying microbiome protocols to help standardize this methodology by estimating the possible variability introduced by the steps of a microbiome study and to offer guidance on how to select a protocol given an analysis's unique constraints.[4] The above concerns about standardization have been brought up in the context of 16S rRNA gene sequencing. The purpose of this paper is to review the use of internal standards in all microbiota studies, though with an interest in applying them to human microbiota, and how they can be used to correct for any biases that arise from the laboratory and analytical methods of the studies. Specifically, the internal standards that will be reviewed are nucleic acids that are added to samples and software that serves as internal checks.

16S rRNA Gene Sequencing

High-throughput 16S ribosomal RNA (rRNA) gene amplicon sequencing (referred to as 16S sequencing) has allowed for more microbiome research to take place due to its cost effectiveness and the development of

better ways to bioinformatically process data.[7–10] In general, this method involves the steps of DNA extraction, polymerase chain reaction (PCR) amplification or DNA shearing, library preparation, sequencing, bioinformatic processing, and analysis. Variability can be unintentionally introduced at every level of analysis resulting in inconsistency in 16S rRNA sequencing data.[7] In fact, each step could produce the amount of heterogeneity that one would expect from biological effects, thus making it difficult to detect biological differences and effects due to interventions.[4] Using 16S sequencing also only allows for the relative abundance of microbes, while absolute abundances allow for a more comprehensive analysis because relative abundances of bacteria species can be shifted by an increase in the number of another bacteria species[7,11,12] and limit the ability to compare studies that use communal rRNA gene data sets.[13] There are mathematical transformations and analytical techniques to overcome this issue, but they make it harder to interpret the biological or ecological meaning of the data.[13–15]

Incorrectly identifying or cross-contaminating samples during 16S sequencing experiments has been a significant problem, with previous research suggesting that 0.3 to 3% of high-throughput sequencing data could have issues relating to sample contamination and mix-ups in the lab.[16–18] Internal standards are able to address some of the variability that can arise in these data. Internal amplification control quantitative PCR (IAC qPCR) has been used to determine when qPCR inhibition occurs by adding a control DNA sequence to the experimental sample and checking for a delay in the sequencing process, which would require re-doing the qPCR.[19] For qPCR, using primers and probe that target an essential gene, such as 18S rRNA for humans, can be used as an extraction control.[19] The addition of DNA sequences into a sample is referred to as spiking in the sequence and the sequences themselves are referred to as spike-ins. Spiking in synthetic DNA sequences has been used to quantify absolute abundances of microbes (gene copies per ng of DNA or mg of sample), measure differences in how sequence reads are processed (comparing Pearson's correlation coefficients, dose-response curves, and proportion of read counts taken out of the data for different amounts of processing compared to no processing), determine sequencing accuracy (comparing error rate of assigned bases for different libraries), and determine the quantity of operational taxonomic units (OTUs) generated through sequencing (comparing sample OTUs to internal standard OTUs).[7] These human-made spike-in standards allow for multiple opportunities for quality control since they can be added to a microbiome sample at different points in the DNA extraction to library preparation process.[7] Multiple synthetic standards can also be used to

create sample tracking mixes which enable researchers to identify cross-contamination and the mis-labelling of sequences by comparing the number of reads from each sample tracking mix in the experimental sample.[18] Tourlousse *et al.* noted that the use of DNA sequence spike-in standards does not always give a reliable count of the cells in a sample because the 16S rRNA gene copy number can vary for different bacterial cells.[7] Additionally, DNA from prokaryotic and eukaryotic organisms has been used as internal standards to determine absolute abundances (number of gene copies per ml of seawater) and its accuracy has been confirmed with various methods, including pigment markers and flow cytometry.[13] In addition to the 16S rRNA gene copy problem mentioned above, PCR bias has also been identified as a significant problem contributing to bias in relative abundances[13,20,21] and is a particular concern when the guanine and cytosine (GC) content is inconsistent.[13]

Microbial communities with a known composition have also been used as an internal standard for verifying the accuracy of 16S sequencing methods, such as with spike-in standards and a new library preparation method.[7,22] de Muinck and colleagues showed that the relative abundances measured were affected by how much of the DNA template they used; it was not clear if this was specific for the mock community used or if it is a universal issue.[22] In order to successfully make templates from amplicon libraries, a spike-in library of a known composition, called PhiX DNA, is used[22] and has been used in other studies.[7,18] These spike-ins can lead to better quality sequencing reads, though half of the reads become part of a non-targeted template for the company Illumina's MiSeq sequencing platform;[23] heterogeneity spacers can be added to correct for unequal numbers of nucleotides and requires less spike-in to sequence effectively.[22,23]

Metagenomics and Metatranscriptomics

Metagenomics generally refers to analyzing multiple genomes in a sample, and allows researchers to profile the functional genes within the sample.[24] 16S sequencing allows for the researcher to infer evolutionary relationships in a microbial community, but it is only based on the sequence of the 16S rRNA gene.[24] Metagenomics allows for evolutionary inference to come from many genes and allows for questions to be asked about both a gene's function and evolutionary relationship.[24] A metagenomics pipeline with notes on where internal standards or checks might be implemented is in Figure 1.

To analyze the metatranscriptome, all the RNA from a bacterial sample is extracted which can be a challenge from a technical standpoint, since mRNA specifically provides information on which genes are being

transcribed in the cell.[25] Messenger RNA (mRNA) transcripts only make up around 1 to 5% of the RNA in a prokaryotic cell so the mRNA needs to be isolated;[25,26] it is a particular challenge to isolate these bacterial mRNAs because prokaryotic mRNA do not have a poly-A tail which would aid in isolating it when converting the RNA to complementary DNA (cDNA).[25] To alleviate this challenge, one method is to isolate rRNA and enrich for mRNA from total RNA, which involves using magnetic beads that are connected to a probe which can anneal to sections of the rRNA.[25]

Another important challenge of metatranscriptomics is selection of reverse transcriptases (RT), which are used to make cDNA from RNA. The RT should be able to handle high temperatures, which are needed to break down the folding of the RNA molecule.[27,28] Group II intron RTs are a good candidate for this type of work because they synthesize accurate cDNA, can handle secondary and tertiary structure, and are thermostable.[27] Researchers have expressed fusion proteins of these RTs with the above benefits, in addition to template switching, which makes this RT useful for cloning microRNA (miRNA) or RNA within proteins.[27] However, the high levels of bias introduced when making cDNA,[29,30] make the methods for semi-direct RNA sequencing more attractive.[29,31–33] Other issues that need to be considered when conducting a metatranscriptomic analysis include appropriate preservation of samples, extracting sufficient high-quality RNA, the quick degradation of mRNA, incomplete transcriptome databases, and host RNA in the experimental sample.[29] Different methods for cDNA and metatranscriptome library production have been compared and one of the findings is that different methods yield different results as to which genes are active in the sample and which affects comparison of data generated using different library production methods.[34]

Metagenomics and Metatranscriptomics Short Reads

Using Next Generation Sequencing (NGS) for metagenomics brings with it the limitations that come with using short reads (i.e. reads from 75 to 800 base pairs long),[35] particularly the possibility of the reads being assembled incorrectly or creating gaps into the consensus sequence, which can be referred to as a genome assembly.[36] Additionally, significant structural variations, such as insertions or deletions, are not accurately recognized by short read sequencers, and methods that rely on PCR have problems with sequencing high-GC% sequences.[36] NGS (short-reads) has previously been reviewed, and that review provides information on important considerations when choosing sequencing instruments such as error rate, read length, runtime, and costs.[35] Additionally, it is important to point out that all meta-omics with microbial community samples cannot

provide information on the absolute abundance of a nucleic acid molecule.[37] Particularly in the area of viral metagenomics in clinical settings, the use of quality control in those studies has not been well-evaluated.[38] In metagenomics analysis, an important issue is whether the reads should be assembled first or whether the raw reads should be used instead.[39] Use of raw reads requires a lot of computing power due to the use of expansive reference databases,[39] and the reference genome must be complete to reduce the number of errors in how a bacteria's taxonomy is determined.[39,40] Assembly allows for more of the genome to be recovered and for the sequence to more readily and accurately be analyzed for taxonomic, along with functional, information; however, chimeras (in this context, an assembly that contains the DNA from multiple source organisms) can be formed due to not being able to assemble reads with a smaller number of abundances and the combination of sections of genome sequences that do not belong together, therefore creating assemblies which inaccurately reflect the genomes of the microbes in the community.[39] Additionally, new metagenomics analytical tools are continuously being developed, which brings the need for them to be "benchmarked" against other tools currently used; one category of such tools is taxonomic classifiers.[41]

The approach to addressing unintentional variability in metagenomics and metatranscriptomics has begun with internal standard techniques that are similar to those used with 16S rRNA sequencing. One method uses DNA from a microbe that is foreign to the microbial community of interest or synthetic mRNA as a spike-in standard, allowing for the quantification of absolute differences in microbes.[37] For metagenomics, internal standards allow one to calculate how many genes and how many molecules of a gene of interest are in a sample; for metatranscriptomics, internal standards allow one to calculate how many RNA transcripts and how many molecules of a RNA transcript of interest are in the sample.[37] For standards with a length of greater than 500 nucleotides, there was little difference in the % recovery of the standards.[37] However, the DNA or RNA in the sample is enclosed in cells, while standards are not, which could cause inaccurate estimates.[37] These include underestimation of the sample, since not all cells may get lysed properly and underestimation of the standard since it could be more exposed to a shearing step or to RNAse.[37] Similarly, a method for using an internal standard sequence consisting of MS2 bacteriophage RNA was used for quality control to identify problems with the laboratory reagents or equipment, inhibitors in the samples, and verify the results.[38] For the internal standard, this validation involved monitoring the semi-quantitative PCR cycles.[38] However, the results of such standards should be understood in the context of the percentage of the viral reads because the MS2

bacteriophage internal standard reads can be drowned out when there is more virus present.[38] The bioinformatic processing of these reads has led to the development of methods for checking how well the taxonomy is assigned for contigs and the bins that are made from these contigs.[39] Additionally, the bins can be checked to determine if each one contains all of its contigs, has any sequence contamination, and the variability of the strains within it.[39] Together these methods serve as an internal check for the analysis pipeline. The analytical software that incorporates these methods for these internal checks, SqueezeMeta, is also able to run without much computing power, though that can limit how large and diverse a microbiome researchers can analyze.[39] The authors note that their software is also able to handle long reads from using the MinIon technology from Oxford Nanopore.[39] In order to inform how sequences are binned and profiled during the analysis of metagenomics, reference databases are used, but a standard reference database needs to be used to truly compare different taxonomic identification software.[41]

Metagenomics and Metatranscriptomics Long Reads

Long-read sequencing, or third-generation sequencing, is a newer type of sequencing that gives more resolution to genome studies.[36] Various companies have created their own versions of such technologies (Table 1), which have been reviewed in recent years.[35,36,42] The company Pacific Biosciences has single-molecule, real-time sequencing instruments that can produce reads that are on average 3,000 to 10,000 base pairs long.[35] As these technologies are used more and as new methods are developed for different studies, ways for comparing the different methods are needed to move the field forward. Long-read sequencing can have similar issues to 16S sequencing and short reads that can be addressed with spike-in standards and these will be more fully described below. Long and short reads have been combined in a variety of ways to make better genome assemblies, since both can complement each other, depending on the sequencing technology being used. One such combination that was found to produce high-quality genomes is Oxford Nanopore's MinIon technology and the company Illumina's MiSeq sequencing technology.[43]

In long-read sequencing metagenomics/metatranscriptomics, microbial communities of known composition or human assembled microbial communities or other previously analyzed data, are used to verify how different technologies and methods are working or to compare different technologies to each other.[44–48] As noted above, the amount of DNA sequenced can bias the results, which could impact these verification experiments. Human assembled microbial communities also are not able to be added to samples since they would cause

contamination of the experimental sample.[44] In a similar way to 16S rRNA gene sequencing, a mix of synthetic DNA sequences (sequins) can be used as a spike-in standard for metagenomics sequencing, which allows for data normalization (comparing samples to sequins), quantitative accuracy (running 3 trials and linear regression of depth of sequin coverage as a function of input concentration), coverage (comparing between sequins), and de novo assembly (% sequin assembly as a function of input concentration and comparing coverage to sequin assembly).[44] Another way to that long and short reads have been used to complement each other uses long read sequences made from Illumina's True-Seq technology to serve as references for short reads of the same microbial community sample.[47] Illumina's TruSeqNano construction libraries were the best at generating internal reference genome bins, but this method is expensive, requires a technician to do a lot of work, and requires a lot of input material which makes it not widely applicable to all research pipelines .[47] Other researchers have combined sequencing techniques by sequencing with both Nanopore long reads and Illumina short reads to generate almost full genome sequences using their assembler OPERA-MS, then comparing their assembler with others by using data made from artificial microbial communities.[46] The authors describe that their technology's ability to determine the genome sequence below the species level will depend on Illumina sequencing's error rates and sequencing coverage, though this will change depending on the manufacturer and technology used.[46] The software IDP-denovo was tested using human cell line genome data and the plant *D. officinale's* genome to use short reads by aligning them to long reads to create pseudo-references, which are used to annotate the exons of the sequence; the results are then compared to the data's previously generated annotation libraries.[48] This approach, which works for metagenomics and metatranscriptomics, was not sufficient when there are less than 43 base pair changes (i.e. small changes) in exon composition.[48]

External Controls

While this paper focuses primarily on internal standards, another method for determining if there is bias in a microbiome study is with external controls. These control samples are run concurrently with the study samples. One area in which these controls are useful is in DNA contamination from other bacteria that can be found in the reagent kits used for DNA extraction, DNA amplification, and library preparation; this can prevent the actual composition of a sample from being determined when there are few microbes in the sample.[49,50] Contamination signals are thus seen in batches of DNA samples that are processed together and can be

identified by the particular method used for the steps of DNA extraction and PCR, technical kit box, or reagent lot that resulted in the contamination.[49–51] Contamination in reagents causes inaccurate microbiome analyses[49–51] and requires methods to determine if a reagent is contaminated.[49] Specifically in regards to external controls, it is recommended that negative controls be used for each laboratory data-generating step in addition to positive controls to find contamination; positive controls also allow for the amount of microbial biomass in the sample to be determined.[49] It is also recommended to be aware of when one is making batches so that they can be statistically analyzed to check for how the results are changing for different batches and for samples within the same batch.[49] Positive controls (a microbe community separate from the one under investigation) and negative controls can be used to identify the effects of data drift[6] because longitudinal variation is an important consideration when collecting microbiome samples for answering a specific question.[51]

Conclusions

The use of internal standards for laboratory procedures are important for producing meaningful data for analysis of the microbiota. Investigators have shown the utility of internal standards in laboratory procedures like various types of spike-in standards and the use of microbial communities as standards. There has also been computer software or protocols that incorporate internal quality control checks for the data analysis pipelines for metagenomics and metatranscriptomics, an example is in the recent establishment of technologies to better combine short and long reads. These quality control checks allow the researcher to assess the accuracy of their own project, but it is not clear that they will allow for comparisons between studies. It will thus be important to continue to test their strengths and limitations as they are applied to health settings. However, there does not appear to be any official guidelines on how to use internal standards or other standards. A recent review has still called for standards that researchers can agree upon, analysis techniques that can be agreed upon, and reference databases that are well-curated.[35] Universal standardization may not be possible since there is no regulatory body for researchers around the world, but there have been smaller scale efforts, such as the Microbiome Quality Control project, which could be applied to metagenomics and metatranscriptomics.

References

1. Lynch S V., Pedersen O. The human intestinal microbiome in health and disease. *N Engl J Med*. 2016;375(24):2369-2379. doi:10.1056/NEJMra1600266
2. Iida N, Dzutsev A, Stewart CA, et al. Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment. *Science (80-)*. 2013;342(6161):967-970. doi:10.1126/science.1240527
3. Viaud S, Saccheri F, Mignot G, et al. The intestinal microbiota modulates the anticancer immune effects of cyclophosphamide. *Science (80-)*. 2013;342(6161):971-976. doi:10.1126/science.1240537
4. Sinha R, Abu-Ali G, Vogtmann E, et al. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nat Biotechnol*. 2017;35(11):1077-1086. doi:10.1038/nbt.3981
5. Huttenhower C, Knight R, Brown CT, et al. Advancing the microbiome research community. *Cell*. 2014;159(2):227-230. doi:10.1016/j.cell.2014.09.022
6. Brooks JP, Edwards DJ, Harwich MD, et al. The truth about metagenomics: Quantifying and counteracting bias in 16S rRNA studies Ecological and evolutionary microbiology. *BMC Microbiol*. 2015;15(1):1-14. doi:10.1186/s12866-015-0351-6
7. Tourlousse DM, Yoshiike S, Ohashi A, Matsukura S, Noda N, Sekiguchi Y. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. *Nucleic Acids Res*. 2017;45(4):e23. doi:10.1093/nar/gkw984
8. Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. 2009;75(23):7537-7541. doi:10.1128/AEM.01541-09
9. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high- throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing. *Nat Methods*. 2010;7(5):335-336. doi:10.1038/nmeth0510-335
10. Edgar RC. UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nat Methods*. 2013;10(10):996-998. doi:10.1038/nmeth.2604
11. Tsilimigras MCB, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann Epidemiol*. 2016;26(5):330-335. doi:10.1016/j.annepidem.2016.03.002
12. Props R, Kerckhof FM, Rubbens P, et al. Absolute quantification of microbial taxon abundances. *ISME J*. 2017;11(2):584-587. doi:10.1038/ismej.2016.117
13. Lin Y, Gifford S, Ducklow H, Schofield O, Cassar N. Towards Quantitative Microbiome Community Profiling Using Internal Standards. *Appl Environ Microbiol*. 2019;85(5):e02634-18. doi:10.1128/AEM .02634-18.
14. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11:R106. doi:https://doi.org/10.1186/gb-2010-11 -10-r106.
15. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):1-21. doi:10.1186/s13059-014-0550-8
16. Hu H, Liu X, Jin W, Ropers HH, Wienker TF. Evaluating information content of SNPs for sample-tagging in re-sequencing projects. *Sci Rep*. 2015;5(10247). doi:10.1038/srep10247
17. Sehn JK, Spencer DH, Pfeifer JD, et al. Occult specimen contamination in routine clinical next-generation sequencing testing. *Am J Clin Pathol*. 2015;144(4):667-674. doi:10.1309/AJCPR88WDJJLDMBN
18. Tourlousse DiM, Ohashi A, Sekiguchi Y. Sample tracking in microbiome community profiling assays using synthetic 16S rRNA gene spike-in controls. *Sci Rep*. 2018;8(1):1-9. doi:10.1038/s41598-018-27314-3
19. Khot PD, Ko DL, Hackman RC, Fredricks DN. Development and optimization of quantitative PCR for the diagnosis of invasive aspergillosis with bronchoalveolar lavage fluid. *BMC Infect Dis*. 2008;8:1-13. doi:10.1186/1471-2334-8-73
20. Polz MF, Cavanaugh CM. Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol*. 1998;64(10):3724-3730. doi:10.1128/aem.64.10.3724-3730.1998
21. Suzuki MT, Giovannoni SJ. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol*. 1996;62(2):625-630. doi:10.1128/aem.62.2.625-630.1996
22. de Muinck EJ, Trosvik P, Gilfillan GD, Hov JR, Sundaram AYM. A novel ultra high-throughput 16S rRNA

gene amplicon sequencing library preparation method for the Illumina HiSeq platform. *Microbiome*. 2017;5(1):68. doi:10.1186/s40168-017-0279-1
23. Fadrosh DW, Bing Ma PG, Sengamalay N, Ott S, Brotman RM, Ravel J. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*. 2014;2(6).
24. Thomas T, Gilbert J, Meyer F. Metagenomics: A guide from sampling to data analysis. *Microb Inform Exp*. 2012;2(3).
25. Bashiardes S, Zilberman-Schapira G, Elinav E. Use of metatranscriptomics in microbiome research. *Bioinform Biol Insights*. 2016;10:19-25. doi:10.4137/BBI.S34610
26. Peano C, Pietrelli A, Consolandi C, et al. An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microb Inform Exp*. 2013;3(1):1-11. doi:10.1186/2042-5783-3-1
27. Mohr S, Ghanem E, Smith W, et al. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *Rna*. 2013;19(7):958-970. doi:10.1261/rna.039743.113
28. Mayer G, Müller J, Lünse CE. RNA diagnostics: Real-time RT-PCR strategies and promising novel target RNAs. *Wiley Interdiscip Rev RNA*. 2011;2:32-41. doi:10.1002/wrna.46
29. Bikel S, Valdez-Lara A, Cornejo-Granados F, et al. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: Towards a systems-level understanding of human microbiome. *Comput Struct Biotechnol J*. 2015;13:390-401. doi:10.1016/j.csbj.2015.06.001
30. Liu D, Graber JH. Quantitative comparison of EST libraries requires compensation for systematic biases in cDNA generation. *BMC Bioinformatics*. 2006;7:1-10. doi:10.1186/1471-2105-7-77
31. Ozsolak F, Platt AR, Jones DR, et al. Direct RNA sequencing. *Nature*. 2009;461(7265):814-818. doi:10.1038/nature08390
32. Ozsolak F, Milos PM. Single-molecule direct RNA sequencing without cDNA synthesis. *Wiley Interdiscip Rev RNA*. 2011;2(4):565-570. doi:10.1002/wrna.84
33. Hickman SE, Kingery ND, Ohsumi TK, et al. The microglial sensome revealed by direct RNA sequencing. *Nat Neurosci*. 2013;16(12):1896-1905. doi:10.1038/nn.3554
34. Alberti A, Belser C, Engelen S, et al. Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics*. 2014;15(1):1-13. doi:10.1186/1471-2164-15-912
35. MacCannell D. Platforms and Analytical Tools Used in Nucleic Acid Sequence-Based Microbial Genotyping Procedures *. *Microbiol Spectr*. 2019;7(1):1-17. doi:10.1128/microbiolspec.ame-0005-2018
36. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. *Trends Genet*. 2018;34(9):666-681. doi:10.1016/j.tig.2018.05.008
37. Satinsky BM, Gifford SM, Crump BC, Moran MA. Use of Internal Standards for Quantitative Metatranscriptome and Metagenome Analysis. *Methods Enzymol*. 2013;531:237-50. doi:10.1016/B978-0-12-407863-5.00012-5
38. Bal A, Pichon M, Picard C, et al. Quality control implementation for universal characterization of DNA and RNA viruses in clinical respiratory samples using single metagenomic next-generation sequencing workflow. *BMC Infect Dis*. 2018;18(1):1-10. doi:10.1186/s12879-018-3446-5
39. Tamames J, Puente-Sánchez F. SqueezeMeta, a highly portable, fully automatic metagenomic analysis pipeline. *Front Microbiol*. 2019;10(JAN):1-10. doi:10.3389/fmicb.2018.03349
40. Pignatelli M, Aparicio G, Blanquer I, Hernández V, Moya A, Tamames J. Metagenomics reveals our incomplete knowledge of global diversity. *Bioinformatics*. 2008;24(18):2124-2125. doi:10.1093/bioinformatics/btn355
41. Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*. 2019;178(4):779-794. doi:10.1016/j.cell.2019.07.010
42. Sedlazeck FJ, Lee H, Darby CA. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet*. 2018;19:329-346. https://doi.org/10.1038/%0As41576-018-0003-4
43. Goldstein S, Beka L, Graf J, Klassen JL. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics*. 2019;20(1):1-17. doi:10.1186/s12864-018-5381-7
44. Hardwick SA, Chen WY, Wong T, et al. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. *Nat Commun*. 2018;9(1). doi:10.1038/s41467-018-05555-0

45. Callahan BJ, Wong J, Heiner C, et al. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res*. 2019;47(18):e103. doi:10.1093/nar/gkz569
46. Bertrand D, Shaw J, Kalathiyappan M, et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat Biotechnol*. 2019;37(8):937-944. doi:10.1038/s41587-019-0191-2
47. Sanders JG, Nurk S, Salido RA, et al. Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. *Genome Biol*. 2019;20(1):1-14. doi:10.1186/s13059-019-1834-9
48. Fu S, Ma Y, Yao H, et al. IDP-denovo: De novo transcriptome assembly and isoform annotation by hybrid sequencing. *Bioinformatics*. 2018;34(13):2168-2176. doi:10.1093/bioinformatics/bty098
49. de Goffau MC, Lager S, Salter SJ, et al. Recognizing the reagent microbiome. *Nat Microbiol*. 2018;3(8):851-853. doi:10.1038/s41564-018-0202-y
50. Salter SJ, Cox MJ, Turek EM, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol*. 2014;12(1):1-12. doi:10.1186/s12915-014-0087-z
51. Kim D, Hofstaedter CE, Zhao C, et al. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome*. 2017;5(1):1-14. doi:10.1186/s40168-017-0267-5
52. Li R, Hsieh CL, Young A, Zhang Z, Ren X, Zhao Z. Illumina synthetic long read sequencing allows recovery of missing sequences even in the “Finished” C. elegans genome. *Sci Rep*. 2015;5(June):1-15. doi:10.1038/srep10814

Figures and Tables:

Figure 1. Schematic of Metagenomic Pipeline. Internal standards refer to physical standards added to samples in the laboratory. Internal checks are done as part of the analysis pipeline.

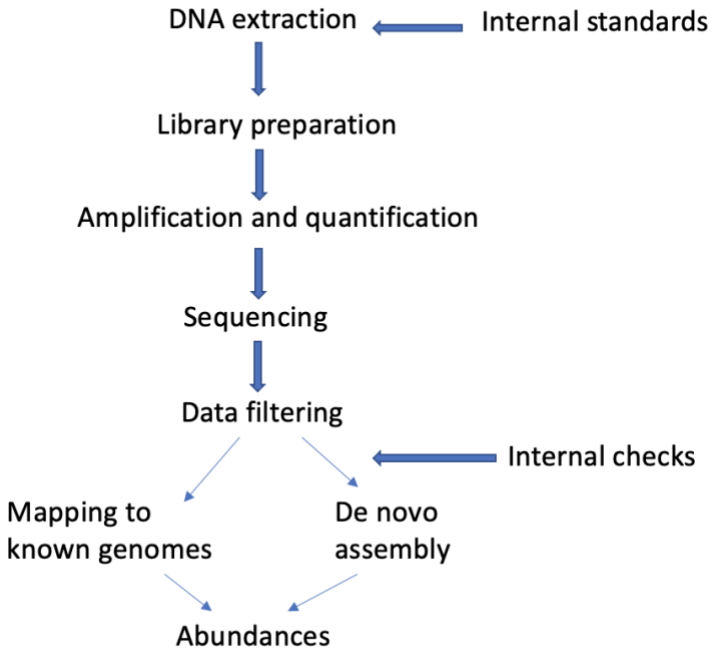


Table 1. Long Read Sequencing Technologies for Microbiota Analysis.[35,36,42,52]

Platform	Average Read Length (kb)	Maximum Read Length (kb)	Error rate	Applications	Bioinformatics challenges
Pacific Biosciences single-molecule real time sequencing	10 - 15	Greater than 80	10 – 15%[a] (indel)	De novo genome assembly	High error rates; needs to be newly aligned and have errors corrected
Oxford Nanopore sequencing	Limited by the DNA's length	Limited by the DNA's length	5 – 15%[a] (indel and homopolymer)	De novo genome assembly	High error rates; needs to be newly aligned and have errors corrected
10X Genomics Chromium	NA[c]	Up to 100	Sequencing is done on a NGS sequencer such as Illumina. Illumina's error rate is 0.1%[a] (substitutions)	De novo genome assembly	Sparse sequencing, not actual long reads; harder to align and bad resolution of repetitive segments of sequences
Illumina synthetic long reads	NA[c]	Approx. 10	[b]SNV: 0.011% Insertion: 0.019% Deletion: 0.101%	De novo genome assembly	Needs a new type of algorithm to enable accurate genome assembly, not actual long reads.

[a]single-pass error rate
[b]all percentages are the ratio of sequence variations comparing the long reads to a reference genome.
[c]Reads are made up of combined short reads, so only the maximum length is given.

Table 2. Use of External Controls in Microbiota studies.

External Control	What it Addresses	Uses
Positive Control	Contamination	Compare between batches and within batches
	Data drift	Looking for longitudinal differences
Negative Control	Contamination	Compare between batches and within batches
	Data drift	Looking for longitudinal differences
	Quantification	Determine the amount of microbial biomass in the sample

Chapter 2: Effect of Internal Standard Normalization of Gut Microbiome Data on Outcomes of a Controlled Feeding Study

Introduction

Standardization would benefit the interpretability of human microbiome data because unintended variability can be introduced at each level of data production and processing.[1] One way to bring standardization to microbiome studies is with internal standards. In three microbiome sequencing methods—16S rRNA gene sequencing, and metagenomic and metatranscriptomic sequencing--this standardization can involve the addition of nucleic acid sequences into a sample. This is referred to as “spiking” in the sequences and the sequences themselves are referred to as “spike-ins.” In 16S rRNA gene sequencing, spiking in synthetic DNA sequences has been used to quantify absolute abundances of microbes (gene copies per ng of DNA or mg of sample).[2] This internal standard method is useful because 16S rRNA gene sequencing without the use of an internal standard only allows for the relative abundance of microbes.[2] Absolute abundances allow for a more comprehensive analysis because relative abundances of bacteria species can be shifted by an increase in the number of another bacteria species.[2–4] Additionally, all meta-omics studies that use microbial community samples cannot provide information on the absolute abundance of a nucleic acid molecule.[5] Similarly to 16S rRNA gene data, DNA from a microbe that is foreign to the microbial community of interest or synthetic mRNA can be used as a spike-in standard, allowing for the quantification of absolute differences in microbes.[5] For metagenomics, these internal standards allow one to calculate how many genes and how many molecules of a gene of interest are in a sample. For metatranscriptomics, these internal standards allow one to calculate how many RNA transcripts and how many molecules of a RNA transcript of interest are in the sample.[5]

In this study, we aimed to assess the degree of standardization in the gut microbiome data from the Carbohydrate and Related Biomarkers (CARB) study,[6] a controlled feeding study, by incorporating internal standards, which will also give us a more quantitatively accurate analysis, as mentioned above.[2–4] We aimed to use data from the following microbiome sequencing methods: biomarker sequencing (the biomarker used for this study will be the 16S rRNA gene) to determine the species diversity and taxonomic relationships within a microbial community, metagenomics sequencing to assemble genomes and infer the function of the genes, and metatranscriptomics sequencing to determine the RNA being made by the cell.[7]

Methods

Study design:

This study used data from the CARB study,[6] which was a randomized, crossover controlled feeding study. The CARB study used investigator-designed low- and high-glycemic load diets as the interventions of interest. The low-glycemic load diet consisted of foods that contained whole grains, legumes, fruits, vegetables, and low-glycemic index carbohydrates. The other diet included refined grains and high-glycemic index carbohydrates. These two diets were the interventions of interest in the current test of the internal standards. The participants were randomized to eat either of the two diets for study period 1 (28 days), which was followed by a washout period of at least 28 days during which they ate their normal diet, and which was then followed by study period 2 (28 days) where the participant ate the diet to which they were not initially randomized.

The goal of the CARB study was to test the hypothesis that the low-glycemic load compared to high-glycemic load diet would result in the participants having optimal levels of metabolic and inflammation biomarkers—biomarkers associated with cancer risk. Another aim of the CARB study was to assess how the diets would affect the participants' gut microbiomes. During the study, stool samples were collected by the participants at three points. The samples were sequenced using the 16S rRNA gene, metagenomic, and metatranscriptomic methods.

Study setting:

The CARB study was conducted during the period of June 2006 and July 2009 under the Principal Investigators Drs. Johanna Lampe and Marian Neuhouser. The CARB study was a part of the Seattle Transdisciplinary Research on Energetics and Cancer (TREC) Center under the PI Anne McTiernan. Recruitment for the study was done in the Seattle area and all foods for the study were prepared and served at the Human Nutrition Laboratory at the Fred Hutchinson Cancer Research Center (Fred Hutch) in Seattle, Washington. The stool samples from the participants were stored and processed at Fred Hutch. The 16S rRNA gene, metagenomic, and metatranscriptomic sequencing was done at Molecular Research, LP in Shallowater, Texas. The sequencing data was sent to Fred Hutch for analysis.

Study participants:

The participants of the CARB study were healthy adults between the ages of 18 and 45 years. The recruitment criteria were that half of the participants had a normal level of adiposity (body mass index (BMI) 18.5 – 25.0 kg/m^2) and the other half were overweight/obese (BMI 28.0 – 40.0 kg/m^2). The gap in BMI between the two weight groups allowed for there to be a meaningful difference between them. The exclusion criteria included: fasting blood glucose >100 mg/dl, taking any prescription medications, tobacco use, actively trying to lose or gain weight, diagnosis of a medical condition that required dietary restrictions, and a refusal to eat the study foods. There were 82 participants (41 cisgender men and 41 cisgender women) recruited for the CARB study and 80 completed the two study periods.

Data collection:

As part of the CARB study, participants collected a stool sample before they were randomized to either of the diets and then at the end of each feeding period. The participants placed the stool samples into RNAlater, which protects bacterial DNA and RNA from degradation (http://www.ambion.com). CARB participants were not required to give stool samples as part of the study activities and ultimately there were stool sample pairs for 70 participants: one sample provided at the end of each feeding period. The number of pairs of stool samples that were sequenced for each type of data are: 69 for the 16S rRNA gene, 64 for the metagenomics, and 66 for the metatranscriptomics. For the 16S rRNA gene data, the fecal samples were homogenized and the bacterial DNA was extracted.[8] Library preparation and sequencing with paired-end reads on the Illumina Mi-Seq platform was done using standard protocols, producing ~250 bp sequences using the V4 region of the 16S rRNA gene. As an internal standard, 0.5% of total genomic DNA from *Ruegeria pomeroyi*[9,10] (ATCC® 700808™) was added to each sample after extraction. The ATCC® 700808™ genome is cross referenced on the ATCC website as GenBank Accessions CP000031 and CP000032, which are both included in the GenBank Assembly Accession GCA_000011965.2 in their current version (CP000031.2 and CP000032.1). For the metagenomics data, genomic DNA was extracted from feces samples.[8] The same internal standard that was used for the 16S rRNA gene data was added, but was limited to 3 Fred Hutch quality control (FHQC) samples; internal standard was not added to the participant metagenomic samples. The FHQC samples were created by combining multiple independent stool samples that were not from the present study and extracting the DNA. Each sample library was prepared using the Nextera DNA Sample preparation kit (Illumina) using the recommendations of the manufacturer's user guide. Shotgun metagenomic sequencing

was done with paired-end reads on the Illumina Mi-Seq platform.[11] This method produces 10 million reads per sample and each read is 600 bp long.

For the metatranscriptomics data, RNA was extracted,[12] followed by the bacterial mRNA enrichment process.[13] Then, the mRNA was amplified with MessageAmpII (Ambion, Foster City, CA). For this dataset's internal standard, 0.5% of the RNA from the nirS gene, which encodes the nitrite reductase cytochrome cd_1-nitrite reductase (EC 1.7.2.1), was added to each sample after extraction. The nirS gene used for this study was from *Nitrobacter winogradskyi*[14] (ATCC® 25391™). The ATCC® 25391™ genome is cross referenced on the ATCC website as GenBank Accessions as CP000115, which is included in the GenBank Assembly Accession GCA_000012725.1. The RNA internal standard was produced from the full-length nirS gene with in-vitro transcription using Novagen plasmids.[5] cDNA was generated from the RNA. The library and sequencing for the metatranscriptomics data followed the same methods as the metagenomics data.

Internal Standard Characterization

Initial bioinformatic processing of the data was completed by Dr. Meredith Hullar and Mr. Keith Curtis. The 16S rRNA gene data were processed with QIIME v.1.8[15] while the metagenomic and metatranscriptomic data underwent quality control checks and were processed with HUMAnN2 v.0.11.2[16], which included MetaPhlAn2.[17,18]

16S rRNA Gene Data:

To determine how *Ruegeria pomeroyi* would be classified in our dataset, we first determined that the only Alphaproteobacteria that were present in the data were in the order Rhodospirillales. These were *Azospirillum* sp. 47_25, Gut Metagenome, Uncultured Bacterium, and Uncultured Organism. To determine how similar *Ruegeria pomeroyi* DSS-3 is to the four Rhodospirillales bacteria, we downloaded those 16S rRNA sequences (Table 1) from the ARB SILVA website's 138.1 SSU database.[19,20] The listed sequences are the ones that had a 16S rRNA gene V4 region, which was extracted using a Python program, which also reverse-transcribed them into DNA. The 16S rRNA gene V4 regions were first aligned using SINA,[21] which transcribed them back into RNA. The alignments were used to reconstruct a phylogenetic tree using the ARB SILVA website's Alignment, Classification, and Tree Service.[22,23] In SILVA, the parameter "variability profile" in the advanced alignment parameters was changed to "Bacteria. Additionally, the parameter "domain" in the advanced tree computation parameters was changed to "Bacteria". Otherwise, default parameters were used,

including the exclusion of highly and moderately variable positions in the sequence. We also aligned each of the 3 *Ruegeria pomeroyi* DSS-3 V4 regions from SILVA to each of the Rhodospirillales V4 regions from SILVA using NCBI's BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Lastly, the V4 region of the *Ruegeria pomeroyi* DSS-3 16S rRNA gene sequences from SILVA (Table 1) was run through QIIME in Python to verify that it was identified as the correct species. These bioinformatic techniques allowed us to determine whether *Ruegeria pomeroyi* DSS-3 was present in our dataset.

Metagenomics Data:

To identify the internal standards in the FHQC samples, we searched the MetaPhlAn2 output sample ids for *Ruegeria pomeroyi* DSS-3. We calculated the expected number of *Ruegeria pomeroyi* DSS-3 genomes in the sample using the formula (amount of *Ruegeria pomeroyi* DSS-3 genome in the sample in ng x 6.022×10^{23} molecules/mol) / (4601048 bp in the *Ruegeria pomeroyi* DSS-3 genome x 660 g/mol in DNA x 1×10^{9} ng/g). The genome size is from the statistics listed for GenBank Assembly Accession GCA_000011965.2.

Once the internal standards were identified, we first determined the counts of each taxa in each control sample by multiplying the total number of reads by the percentage of each taxon generated by MetaPhlAn2 divided by 100. We then used two different methods to transform and normalize the data. One was using the centered log-ratio transformation (CLR) which is used for compositional data.[24] After merging the data from each sample, the CLR transformation was taken for each sample at different taxonomic levels after the internal standard was removed from the dataset and after 1 was added to each taxon's count to account for missing values. Secondly, after merging the data from each sample, we used a normalization method using a DNA internal standard that was developed by previous researchers.[5] Then, the normalized dataset at different taxonomic levels was transformed with the interquartile log-ratio transformation (IQLR), which is used for compositional data and gives it a near-normal districbution.[25,26] The internal standard was removed from the dataset before transforming the data with the IQLR transformation. These two methods were done in R version 4.0.2, using the "compositions" package[27] for the CLR transformation and "propr" package[28] for the IQLR transformation. We then calculated the mean, standard deviation (SD), and coefficient of variation (CV) in Microsoft Excel (using the AVERAGE and STDEV formulas, then dividing the standard deviation by the mean and multiplying by 100 respectively) for the ten most abundant taxa across all three samples. Our internal

standard was only identifiable up to the class level, as shown by the fact that every taxon falling under Alphaproteobacteria had the same count. Therefore, we could not use our internal standard to normalize our data at the phylum level. We then determined the percentage of the organisms in each of the three samples for both CLR transformed data and the IQLR-transformed internal standard-normalized (IS) data at the species level. We presented the 20 most abundant species in each sample in pie charts, which were made in Microsoft Excel.

Metatranscriptomics Data:

To identify the internal standards, we used Diamond v0.9.22.123[29] to search for the *Nitrobacter winogradskyi* NB-255 (GenBank: CP000115.1) nirS amino acid sequence (GenBank: ABA05901.1) in the metatranscriptomics raw reads with BLAST+ in Linux. To verify that the sequence was correct, we BLASTed the nirS DNA sequence against NCBI and identified six similar sequences. The amino acid sequences that corresponded to the DNA sequences that matched nirS (ABE64015.1, BAO96037.1 BAO95901.1, ARR54947.1, QTH21414.1, and ABQ68154.1) were then run with Diamond to search for them in one metatranscriptomics read. The expected number of nirS transcripts was calculated using the formula (amount of mRNA transcript in the sample in ng x 6.022×10^{23} molecules/mol) / (972 bp in the nirS transcript x 340 g/mol in mRNA x 1×10^{9} ng/g). The length of nirS transcript was determined by searching for the E.C. number 1.7.2.1 with the Google Chrome search function in the GenBank Accession CP000115.1.

Results

16S rRNA Gene Sequencing:

The phylogenetic tree (Figure 1) shows that the five sequences that are in the Rhodospirillales order cluster together and that they cluster next to the three *Ruegeria pomeroyi* DSS-3 sequences. Together those two clusters form an Alphaproteobacteria monophyletic group. The *Bacteroides fragilis* species in the Proteobacteria, Bacteroidota, Firmicutes, or Verracomicrobiota phyla all clustered outside of the Alphaproteobacteria monophyletic group. The Ruegeria pomeroyi DSS-3 16S rRNA gene V4 regions were around 82% identical to the Rhodospirillales 16S rRNA gene V4 regions. Additionally, for the five sequences that were identified as the *Azospirillum* sp. 47_25, Gut Metagenome, Uncultured Bacterium, and Uncultured organism taxa, the mean proportion of total reads for each sample was 0.002. The expected proportion is 0.005. Lastly, after running the 16S rRNA gene V4 regions of *Ruegeria pomeroyi* DSS-3 through QIIME, we

found that those regions were identified as Rhodobacteracea *Pseudophaeobacter* bacterium enrichment culture clone 12(2013).

Metagenomic Sequencing:

When we started this analysis, we did not realize that internal standard had not been added to CARB participant metagenomic samples. Data were available for the 3 FHQC samples to which internal standard had been added. The internal standard was identified in MetaPhlAn2 as *Ruegeria pomeroyi* taxon GCF_000011965 and it accounted for 0.04921% of the reads in FHQC Sample 3, 0.04295% of the reads in FHQC Sample 2, 0.04263% of the reads in the FHQC Samples 1, and a mean of 0.04493% of the reads for all 3 control samples. Based on the formula given above we estimate that we added 49577 copies of the *Ruegeria* genome to the 3 FHQC samples.

In Table 2 we can see how our transformations and normalization affected the data. Comparing the counts and IS-normalized counts, we see that the IS-normalized counts are generally larger than the counts and both have similar CV values. Comparing the counts to the CLR-transformed counts, we see that the count CV values are generally further from zero than the CLR-transformed CV for the high-level taxa, but the CV values for both counts are generally more similar for the lower-level taxa. For example, the Verracomicrobia count CV is 3.9 while the CLR-transformed count CV is -56.4 and the Prevotella count CV is 8.3 while the CLR-transformed count CV is 3.9 Comparing the IS-normalized counts to the IQLR-transformed IS-normalized counts, we see that the CV values are generally smaller for the IQLR-transformed IS-normalized counts. For example, the Prevotella IS-normalized count CV is 9.0 while the IQLR-transformed IS-normalized count CV is 1.1.

The 20 most abundant species in both the IQLR-transformed IS-normalized and CLR-transformed 3 FHQC samples are shown in Figures 2 through 4. When looking at Table 2, the mean counts and mean IS-normalized counts differ by several degrees of magnitude, while the CLR-transformed counts and IQLR-transformed IS-normalized counts differ by only one degree of magnitude. This is shown visually in Figures 2 through 4. For the CLR transformed data (Figures 2A, 3A, and 4A), each species makes up between 4% and 8% of the 20 most abundant species. For the IS-normalized, then IQLR-transformed data (Figures 2B, 3B, and 4B), each species make up between 3% and 9% of the 20 most abundant species.

Metatranscriptomic Sequencing:

The expected number of copies of the nirS transcript is 227,774,752 as calculated by the equation listed above. There were 264 reads for 132 samples produced during the metatranscriptomics sequencing. After running Diamond with the nirS amino acid sequence, we found that of the 264 reads, 92 of them had 0 nirS transcripts, 56 had 1 nirS transcript, 39 had 2 nirS transcripts, 20 had 3 nirS transcripts, and various other shorter lengths. The minimum number of nirS transcripts found was 0 and the maximum found was 33 transcripts. Only 1 read had 33 nirS transcripts. One of the reads for the IFC2 sample had 56 nirS transcripts and one of the reads for the IFC3 sample had 53 nirS transcripts. Diamond was run with the six amino acid sequences that matched the nirS gene through NCBI's BLAST using a sample where one read had 5 matches with the nirS amino acid sequence and one read with 1 match to the nirS gene. For this set of six amino acid sequences, there were only 3 hits in total for the read with 5 original hits and 0 hits for the other read.

Discussion

Our ability to test our hypothesis that addition of internal standards would improve quantitation of 16S rRNA gene abundance and metagenomic and metatranscriptomic data was hindered by several methodologic challenges. Below we discuss our findings for the three sets of microbiome data.

16S rRNA Gene Sequencing:

We discuss here our reasoning for determining that the 16S rRNA gene dataset could not be normalized using the *Ruegeria pomeroyi* DSS-3 internal standard. Previously, we had used this internal standard when we did 16S rRNA gene sequencing with the V1 to V3 regions of the gene. For the 16S rRNA gene sequencing done in this study, we used the V4 region of the gene, and had not tested whether it would give us the same taxonomic resolution as the V1 to V3 region. Below we describe how our findings indicated that we were not able to recognize our internal standard sequences in 16S rRNA gene sequencing dataset.

We will discuss first the sequences that we downloaded from ARB SILVA. The *Azospirillum* sp. 47_25 sequence was from a human host (https://www.arb-silva.de/). The genus *Azospirillum* are plant growth promoting rhizobacteria[6,7,8] and can be found in the soil around plant roots.[9,10,8] They are Gram-negative bacteria that fix nitrogen in the rhizosphere.[8] Azospirillum bacteria were seen in the gut microbiomes of children 1-5 years old in Zimbabwe.[11] While our population is from an urban area, they might have less contact with soil, but they could still contain *Azospirillum* in their gut microbiomes. The other 4 Rhodospirillales sequences are all uncultured organisms that have not been classified beyond the order level. The

Rhodospirillales order has 389 genomes included under its taxonomy ID (204441) on NCBI, so it is a broad taxonomic group. Since, the *Azospirillum* sequence could be part of the normal human gut microbiome, and the fact that the Rhodospirillales sequences did not cluster with the *Ruegeria pomeroyi* DSS-3 sequences, the Rhodospirillales sequences should not be counted as the *Ruegeria pomeroyi* DSS-3 internal standard.

Furthermore, based on running the *Ruegeria pomeroyi* DSS-3 16S rRNA gene V4 region sequences from SILVA through the QIIME pipeline used for our dataset, we found that they were identified as Rhodobacteracea *Pseudophaeobacter* bacterium enrichment culture clone 12(2013). Our dataset did not contain this OTU, which indicates that it did not reach a level of detection in our samples, even though we added it.

A possible explanation for this lack of detection is that we added only 0.1 ng of our internal standard to the PCR done before our sequencing. Lin et al., 2019 added 14.85 ng of genomic DNA was spiked into their samples as an internal standard, which was then amplified with a 16S rRNA gene V4 primer set.[30] This suggests that we may have added too little of our internal standard DNA to our samples. Additionally, PCR-based sequencing has lower coverage at around 50% GC content and greater,[31] which could influence if our internal standard was correctly amplified. The GC content was 56.01% for each of the three 16S rRNA sequences annotated in GenBank Accession CP000031.2, using the website https://jamiemcgowan.ie/bioinf/gc.html.

Overall, for best results when using a 16S rRNA gene internal standard, before we began our study we would first determine if our internal standard is distinguishable from other microbes with our variable regions. We would then optimize the amount of internal standard that is spiked into the samples so that is detectable in our sample. Lastly, we would make sure our bioinformatic pipeline is using reference databases that contain the internal standard. We were not able to use this approach because we wanted to see if this internal standard method would work with a less intensive approach.

Metagenomic Sequencing:

The differences between the counts and the IS-normalized counts show that normalization can make a substantial difference in the counts. This finding echoes previous researchers who have noted that relative abundance measures can lead to incorrect inferences since one species can affect another's relative abundance, as noted above.[2–4] When the data were transformed, the IS normalization did not drastically

change the composition of the sample. The IQLR transformation did appear to reduce the correlation of variation more than the CLR transformation, indicating that there was more variation in the IS-normalized data. While IS normalization has appealing benefits, it has its limitations, which are mentioned below in the General Considerations section.

It should be noted that the internal standard normalization method that we employed was designed for metagenome gene abundances[5] while we used it to calculate taxon abundances. MetaPhlAn2 uses marker genes to taxonomically identify sequences,[17,18,32] which can be considered the same as protein-encoding genes in this context since they are both genes within the sequence library. Therefore, this method for internal standard normalization is accurate for our dataset. Our findings indicate that we can add this internal standard to all of the metagenomic samples to normalize the resulting dataset.

Metatranscriptomic Sequencing:

Since we expect that we added 227,774,751.9 nirS transcripts to our samples, this means that our percent recovery for the sample with the greatest number of internal standards, 56 in IFC2, the percent yield is 2.46×10^{-5} %. Satinsky et al. found that a 1000 bp internal standard had a percent recovery that was greater than 0.001%. We therefore found a much lower percent recovery than has previously been reported. It is also a problem that we did not recover any internal standard in many of the metranscriptomic samples.

A possible explanation is that we were searching for the wrong nirS gene. This idea is what motivated us to BLAST our nirS gene and search for the six matches within one of our metatranscriptomic samples. We found fewer sequences with those six sequences than we did with our original nirS sequence, so we did not identify a sequence that is more likely to be the nirS sequence that was added to our samples. A possible alternative to normalizing metagenomic data with internal standards would be to use marker genes instead. Marker genes have been used in metagenome assembly and taxonomy assignment.[33]

There are also possible technical explanations for our low nirS transcript recovery. In vitro transcription using the T7 promoter is still being researched to increase its transcription rate,[34] so perhaps a low transcription rate reduced our nirS transcript recovery. The protocol we followed used a T7 promoter.[5] Additionally, Illumina sequencing has been found to produce errors that are dependent on the sequence and causes changes in sequence coverage,[35] which could also affect our internal standard recovery.

General Considerations:

Even without the technical challenges of identifying internal standards, our study design itself had some limitations. For 16S rRNA gene sequencing, Tourlousse et al. noted that the use of DNA sequence spike-in standards does not always give a reliable count of the cells in a sample because the 16S rRNA gene copy number can vary for different bacterial cells.[2] For metatranscriptomics sequencing, there is a bias towards longer RNA transcripts, which can affect how much RNA is recovered.[5]

Conclusion

Through conducting this study, we learned that there are challenges to implementing the addition of internal standards into our samples. To overcome these challenges, we need to conduct lab experiments to optimize our protocols. We also demonstrated the importance of bioinformatic techniques since they are needed to determine whether we can use our standardization techniques.

References

1. Leigh Greathouse K, Sinha R, Vogtmann E. DNA extraction for human microbiome studies: The issue of standardization. *Genome Biol*. 2019;20(1):1-4. doi:10.1186/s13059-019-1843-8
2. Tourlousse DM, Yoshiike S, Ohashi A, Matsukura S, Noda N, Sekiguchi Y. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. *Nucleic Acids Res*. 2017;45(4):e23. doi:10.1093/nar/gkw984
3. Tsilimigras MCB, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann Epidemiol*. 2016;26(5):330-335. doi:10.1016/j.annepidem.2016.03.002
4. Props R, Kerckhof FM, Rubbens P, et al. Absolute quantification of microbial taxon abundances. *ISME J*. 2017;11(2):584-587. doi:10.1038/ismej.2016.117
5. Satinsky BM, Gifford SM, Crump BC, Moran MA. Use of Internal Standards for Quantitative Metatranscriptome and Metagenome Analysis. *Methods Enzymol.* 2013;531:237-50. doi:10.1016/B978-0-12-407863-5.00012-5
6. Neuhouser ML, Schwarz Y, Wang C, et al. A low-glycemic load diet reduces serum C-reactive protein and modestly increases adiponectin in overweight and obese adults. *J Nutr*. 2012;142(2):369-374. doi:10.3945/jn.111.149807
7. Lynch S V., Pedersen O. The human intestinal microbiome in health and disease. *N Engl J Med*. 2016;375(24):2369-2379. doi:10.1056/NEJMra1600266
8. Li F, Hullar MAJ, Lampe JW. Optimization of terminal restriction fragment polymorphism (TRFLP) analysis of human gut microbiota. *J Microbiol Methods*. 2007;68(2):303-311. doi:10.1016/j.mimet.2006.09.006
9. González JM, Covert JS, Whitman WB, et al. Silicibacter pomeroyi sp. nov. and Roseovarius nubinhibens sp. nov., dimethylsulfoniopropionate-demethylating bacteria from marine environments. *Int J Syst Evol Microbiol*. 2003;53(5):1261-1269. doi:10.1099/ijs.0.02491-0
10. Yi H, Lim YW, Chun J. Taxonomic evaluation of the genera Ruegeria and Silicibacter: A proposal to transfer the genus Silicibacter Petursdottir and Kristjansson 1999 to the genus Ruegeria Uchino et al. 1999. *Int J Syst Evol Microbiol*. 2007;57(4):815-819. doi:10.1099/ijs.0.64568-0
11. Olafson PU, Lohmeyer KH, Dowd SE. Analysis of expressed sequence tags from a significant livestock pest, the stable fly (Stomoxys calcitrans), identifies transcripts with a putative role in chemosensation and sex determination. *Arch Insect Biochem Physiol*. 2010;74(3):179-204. doi:10.1002/arch.20372
12. Zoetendal EG, Booijink CCGM, Klaassens ES, et al. Isolation of RNA from bacterial samples of the human gastrointestinal tract. *Nat Protoc*. 2006;1(2):954-959. doi:10.1038/nprot.2006.143
13. Stewart FJ, Ottesen EA, Delong EF. Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J*. 2010;4(7):896-907. doi:10.1038/ismej.2010.18

14. Winslow C, Broadhurst J, Buchanan R, Krumwiede C, Rogers L, Smith G. The Families and Genera of the Bacteria: Preliminary Report of the Committee of the Society of American Bacteriologists on Characterization and Classification of Bacterial Types. *J Bacteriol*. 1917;2:505-566.
15. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high- throughput community sequencing data. *Nat Publ Gr*. 2010;7(5):335-336. doi:10.1038/nmeth0510-335
16. Franzosa EA, McIver LJ, Rahnavard G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods*. 2018;15(11):962-968. doi:10.1038/s41592-018-0176-y
17. Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods*. 2015;12(10):902-903. doi:10.1038/nmeth.3589
18. Truong DT, Franzosa EA, Tickle TL, et al. Erratum: MetaPhlAn2 for enhanced metagenomic taxonomic profiling (Nature Methods (2015) 12 (902-903)). *Nat Methods*. 2015;13(1):101. doi:10.1038/nmeth0116-101b
19. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res*. 2013;41(D1):590-596. doi:10.1093/nar/gks1219
20. Yilmaz P, Parfrey LW, Yarza P, et al. The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res*. 2014;42(D1):643-648. doi:10.1093/nar/gkt1209
21. Pruesse E, Peplies J, Glöckner FO. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*. 2012;28(14):1823-1829. doi:10.1093/bioinformatics/bts252
22. Price MN, Dehal PS, Arkin AP. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol*. 2009;26(7):1641-1650. doi:10.1093/molbev/msp077
23. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010;5(3). doi:10.1371/journal.pone.0009490
24. Aitchison J. *The Statistical Analysis of Compositional Data*. Chapman and Hall; 1986.
25. Wu JR, Macklaim JM, Genge BL, Gloor GB. Finding the centre: Corrections for asymmetry in high-throughput sequencing datasets. *arXiv*. Published online 2017.
26. Aitchison J. The Statistical Analysis of Compositional Data Author (s): J . Aitchison Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 44 , No . 2 (1982), Published by : Wiley for the Royal Statistical Society Stable URL : htt. *J R Stat Soc*. 1982;44(2):139-177.
27. van den Boogaart KG, Tolosana-Delgado R. "compositions": A unified R package to analyze compositional data. *Comput Geosci*. 2008;34(4):320-338. doi:10.1016/j.cageo.2006.11.017
28. Quinn TP, Richardson MF, Lovell D, Crowley TM. Propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Sci Rep*. 2017;7(1):1-9. doi:10.1038/s41598-017-16520-0
29. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2014;12(1):59-60. doi:10.1038/nmeth.3176
30. Lin Y, Gifford S, Ducklow H, Schofield O, Cassar N. Towards Quantitative Microbiome Community Profiling Using Internal Standards. *Appl Environ Microbiol*. 2019;85(5):e02634-18. doi:10.1128/AEM.02634-18.
31. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. *Trends Genet*. 2018;34(9):666-681. doi:10.1016/j.tig.2018.05.008
32. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods*. 2012;9(8):811-814. doi:10.1038/nmeth.2066
33. Frank JA, Pan Y, Tooming-Klunderud A, et al. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci Rep*. 2016;6(May):1-10. doi:10.1038/srep25373
34. Conrad T, Plumbom I, Alcobendas M, Vidal R, Sauer S. Maximizing transcription of nucleic acids with efficient T7 promoters. *Commun Biol*. 2020;3(1):1-8. doi:10.1038/s42003-020-01167-x
35. Nakamura K, Oshima T, Morimoto T, et al. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res*. 2011;39(13). doi:10.1093/nar/gkr344

Figures and Tables:

Table 1. Organisms used for Phylogenetic Analysis. The organism's name, lineage, and SILVA accessions are included. The Bacteroides fragilis sequences were used as an outgroup and the accession numbers were not listed because of the large quantity of sequences.

Organism	Lineage	SILVA Sequences
Ruegeria pomeroyi DSS-3	cellular organisms; Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Ruegeria; Ruegeria pomeroyi	CP000031.3466174.3467628 CP000031.261910.263364 CP000031.4014510.4015964
Azospirillum sp. 47_25	Bacteria;__Proteobacteria;__Alphaproteobacteria;__Rhodospirillales;__uncultured;__Azospirillum_sp._47_25	MNTU01000015.14.1361
Gut Metagenome	Bacteria;__Proteobacteria;__Alphaproteobacteria;__Rhodospirillales;__uncultured;__gut_metagenome	CDYF01019559.281.1764 CEAR01043182.5169.6651
Uncultured Bacterium	Bacteria;__Proteobacteria;__Alphaproteobacteria;__Rhodospirillales;__uncultured;__uncultured_bacterium	DQ905714.1.1544
Uncultured Organism	Bacteria;__Proteobacteria;__Alphaproteobacteria;__Rhodospirillales;__uncultured;__uncultured_organism	HQ774881.1.1442
Bacteroides fragilis	Bacteria; _Proteobacteria or Bacteroidota or Firmicutes or Verracomicrobiota	416 sequences from SILVA

Figure 1. Segment of Phylogenetic Tree. This segment of the tree shows where the *Ruegeria* and Rhodospirillales sequences cluster in comparison to the *Bacteroides fragilis* species. The *Ruegeria* and Rhodospirillales cluster separately from each other indicating that they should be classified separately.

Table 2. Ten most abundant taxa at the phylum, class, order, family, genus, and species level for the three FHQC samples. SD is standard deviation and CV is coefficient of variation. There are NA values for the phylum Proteobacteria under IS-Normalized Counts and IQLR-Transformed IS-Normalized Counts because our internal standard could not be distinguished from other taxa in this phylum.

	Counts			CLR-Transformed Counts			IS-Normalized Counts			IQLR-Transformed IS-Normalized Counts		
	Mean	SD	CV	Mean	SD	CV	Mean	SD	CV	Mean	SD	CV
Phylum												
Firmicutes	2399308	180500	7.5	3.08	0.25	8.0	51602739	4409522	8.5	1.25	0.02	1.5
Bacteroidetes	1988814	121880	6.1	2.89	0.25	8.6	42756806	2796760	6.5	1.06	0.01	0.8
Actinobacteria	637186	68873	10.8	1.75	0.21	12.1	13695168	1438932	10.5	-0.08	0.05	-63.4
Verrucomicrobia	67859	2631	3.9	-0.48	0.27	-56.4	1459454	84473	5.8	-2.31	0.02	-0.8
Proteobacteria	40288	7010	17.4	-1.02	0.14	-13.8	NA	NA	NA	NA	NA	NA
Euryarchaeota	16242	2994	18.4	-1.93	0.36	-18.7	347756	52524	15	-3.75	0.22	-5.8
Viruses_noname	4037	6038	149.6	-4.30	1.42	-33.1	84681	125889	149	-6.13	1.68	-27.4
Class												
Clostridia	2174143	178604	8.2	4.32	0.44	10.1	46751305	4119687	8.8	2.53	0.02	1.0
Bacteroidia	1988814	121880	6.1	4.23	0.44	10.4	42756806	2796760	6.5	2.45	0.01	0.2
Actinobacteria	637186	68873	10.8	3.09	0.41	13.4	13695168	1438932	10.5	1.31	0.04	3.3
Negativicutes	216099	11700	5.4	2.01	0.53	26.3	4656212	480180	10.3	0.23	0.09	40.9
Verrucomicrobiae	67859	2631	3.9	0.85	0.46	54.2	1459454	84473	5.8	-0.93	0.03	-3.1
Betaproteobacteria	19164	2626	13.7	-0.42	0.37	-88.0	411275	48004	11.7	-2.20	0.07	-3.2
Methanobacteria	16242	2994	18.4	-0.59	0.47	-79.7	347756	52524	15.1	-2.37	0.21	-8.9
Deltaproteobacteria	14225	2305	16.2	-0.72	0.35	-48.2	305139	42681	14.0	-2.50	0.09	-3.8
Bacilli	8790	1663	18.9	-1.20	0.47	-39.1	189454	39114	20.6	-2.99	0.16	-5.3
Gammaproteobacteria	4591	2088	45.5	-1.91	0.05	-2.7	97657	41474	42.5	-3.69	0.41	-11.0
Order												
Clostridiales	2174143	178604	8.2	4.80	0.65	13.6	46751305	4119687	8.8	2.86	0.03	0.9
Bacteroidales	1988814	121880	6.1	4.72	0.66	14.1	42756806	2796760	6.5	2.77	0.002	0.1
Bifidobacteriales	546610	61422	11.2	3.42	0.62	18.2	11746778	1261575	10.7	1.48	0.05	3.4
Selenomonadales	216099	11700	5.4	2.50	0.75	30.1	4656212	480180	10.3	0.55	0.09	15.9
Coriobacteriales	90576	7715	8.5	1.63	0.66	40.6	1948390	188283	9.7	-0.32	0.03	-10.9
Verrucomicrobiales	67859	2631	3.9	1.34	0.69	51.3	1459454	84473	5.8	-0.61	0.02	-3.9
Burkholderiales	19164	2626	13.7	0.07	0.59	844.4	411275	48004	11.7	-1.88	0.08	-4.1
Methanobacteriales	16242	2994	18.4	-0.10	0.72	-708.8	347756	52524	15.1	-2.05	0.21	-10.3
Desulfovibrionales	14225	2305	16.2	-0.23	0.56	-244.9	305139	42681	14.0	-2.18	0.10	-4.6
Lactobacillales	8790	1663	18.9	-0.72	0.65	-91.2	189454	39114	20.6	-2.66	0.16	-5.9
Family												
Prevotellaceae	892472	73562	8.2	4.15	0.27	6.5	19193192	1731186	9.0	2.14	0.03	1.6
Lachnospiraceae	802502	73753	9.2	4.05	0.27	6.8	17264646	1796602	10.4	2.03	0.05	2.3
Ruminococcaceae	783406	45659	5.8	4.02	0.28	6.9	16832509	801235	4.8	2.01	0.03	1.3
Bacteroidaceae	754178	45339	6.0	3.99	0.28	6.9	16209500	939734	5.8	1.97	0.02	0.9
Bifidobacteriaceae	546610	61422	11.2	3.66	0.24	6.5	11746778	1261575	10.7	1.65	0.06	3.8
Eubacteriaceae	387210	41420	10.7	3.32	0.26	8.0	8330071	971355	11.7	1.30	0.06	4.8
Rikenellaceae	253037	13525	5.3	2.89	0.30	10.2	5433616	56342	1.0	0.88	0.06	6.9
Veillonellaceae	139506	14243	10.2	2.30	0.43	18.8	3009095	440391	14.6	0.28	0.14	49.6
Clostridiaceae	110729	8736	7.9	2.07	0.26	12.8	2380237	188127	7.9	0.05	0.03	56.6
Coriobacteriaceae	90576	7715	8.5	1.87	0.27	14.7	1948390	188283	9.7	-0.15	0.04	-26.8
Genus												
Prevotella	795165	65607	8.3	4.65	0.18	3.9	17100539	1543804	9.0	3.17	0.04	1.1
Bacteroides	754178	45339	6.0	4.60	0.19	4.2	16209500	939734	5.8	3.12	0.01	0.3
Bifidobacterium	546610	61422	11.2	4.27	0.15	3.5	11746778	1261575	10.7	2.79	0.05	1.8
Faecalibacterium	403548	43073	10.7	3.97	0.17	4.2	8678607	970585	11.2	2.49	0.06	2.3
Eubacterium	387210	41420	10.7	3.93	0.18	4.5	8330071	971355	11.7	2.45	0.06	2.6
Coprococcus	386140	59534	15.4	3.92	0.18	4.5	8312047	1369721	16.5	2.44	0.12	4.7
Subdoligranulum	302492	18860	6.2	3.68	0.23	6.3	6493908	129404	2.0	2.20	0.08	3.5
Alistipes	253037	13525	5.3	3.50	0.22	6.2	5433616	56342	1.0	2.02	0.06	2.7
Butyrivibrio	197661	9970	5.0	3.26	0.20	6.2	4246538	143465	3.4	1.78	0.03	1.5
Clostridium	110729	8736	7.9	2.68	0.18	6.6	2380237	188127	7.9	1.20	0.02	1.8
Species												
Prevotella_copri	795165	65607	8.3	4.75	0.21	4.4	17100539	1543804	9.0	3.66	0.04	1.0
Bifidobacterium_adolescentis	413505	47593	11.5	4.09	0.18	4.4	8882618	928796	10.5	3.01	0.03	1.2
Faecalibacterium_prausnitzii	403548	43073	10.7	4.07	0.19	4.6	8678607	970585	11.2	2.98	0.05	1.7
Subdoligranulum_unclassified	302492	18860	6.2	3.78	0.30	7.8	6493908	129404	2.0	2.70	0.09	3.2
Coprococcus_eutactus	262456	41727	15.9	3.63	0.16	4.4	5650791	966324	17.1	2.55	0.12	4.6
Butyrivibrio_crossotus	197661	9970	5.0	3.36	0.25	7.5	4246538	143465	3.4	2.27	0.04	1.7
Bacteroides_sp_4_3_47FAA	196249	15574	7.9	3.35	0.21	6.4	4217914	323238	7.7	2.27	0.01	0.7
Eubacterium_rectale	179619	28507	15.9	3.25	0.15	4.7	3865717	645361	16.7	2.17	0.11	5.1
Alistipes_putredinis	161296	10620	6.6	3.15	0.30	9.5	3462421	85301	2.5	2.07	0.09	4.4
Bacteroides_stercoris	133746	6926	5.2	2.97	0.24	8.1	2875501	168065	5.8	1.88	0.03	1.7

Figure 2. Pie Chart of FHQC Sample Number 1's 20 most abundant species after **A)** CLR transformation **B)** IS normalization, then IQLR transformation. The composition of the 20 most abundant species is relatively even after both transformations.

A. **B.**

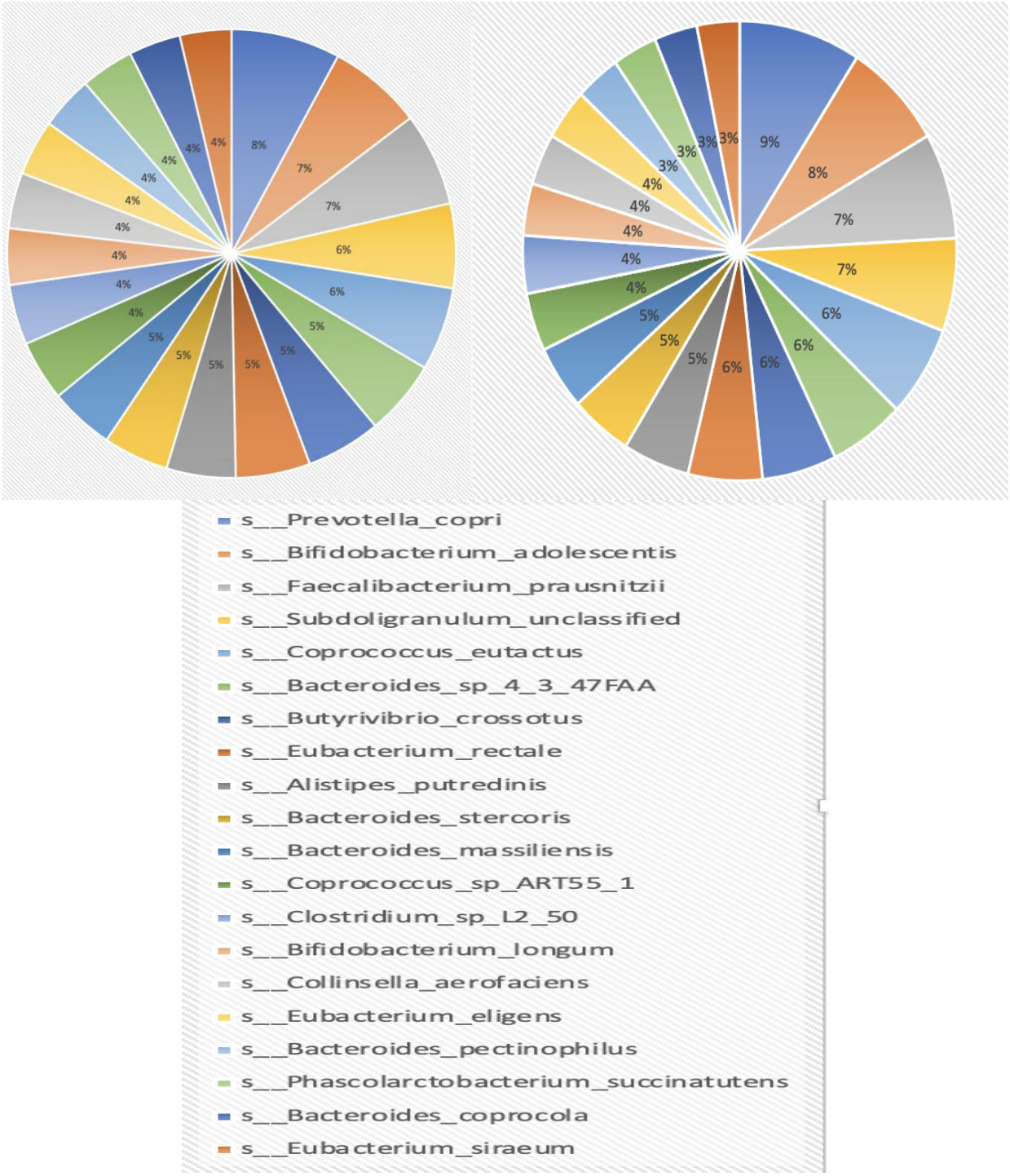


Figure 3. Pie Chart of FHQC Sample Number 2's 20 most abundant species after **A)** CLR transformation and **B)** after IS normalization then IQLR transformation. The composition of the 20 most abundant species is relatively even after both transformations.

A. **B.**

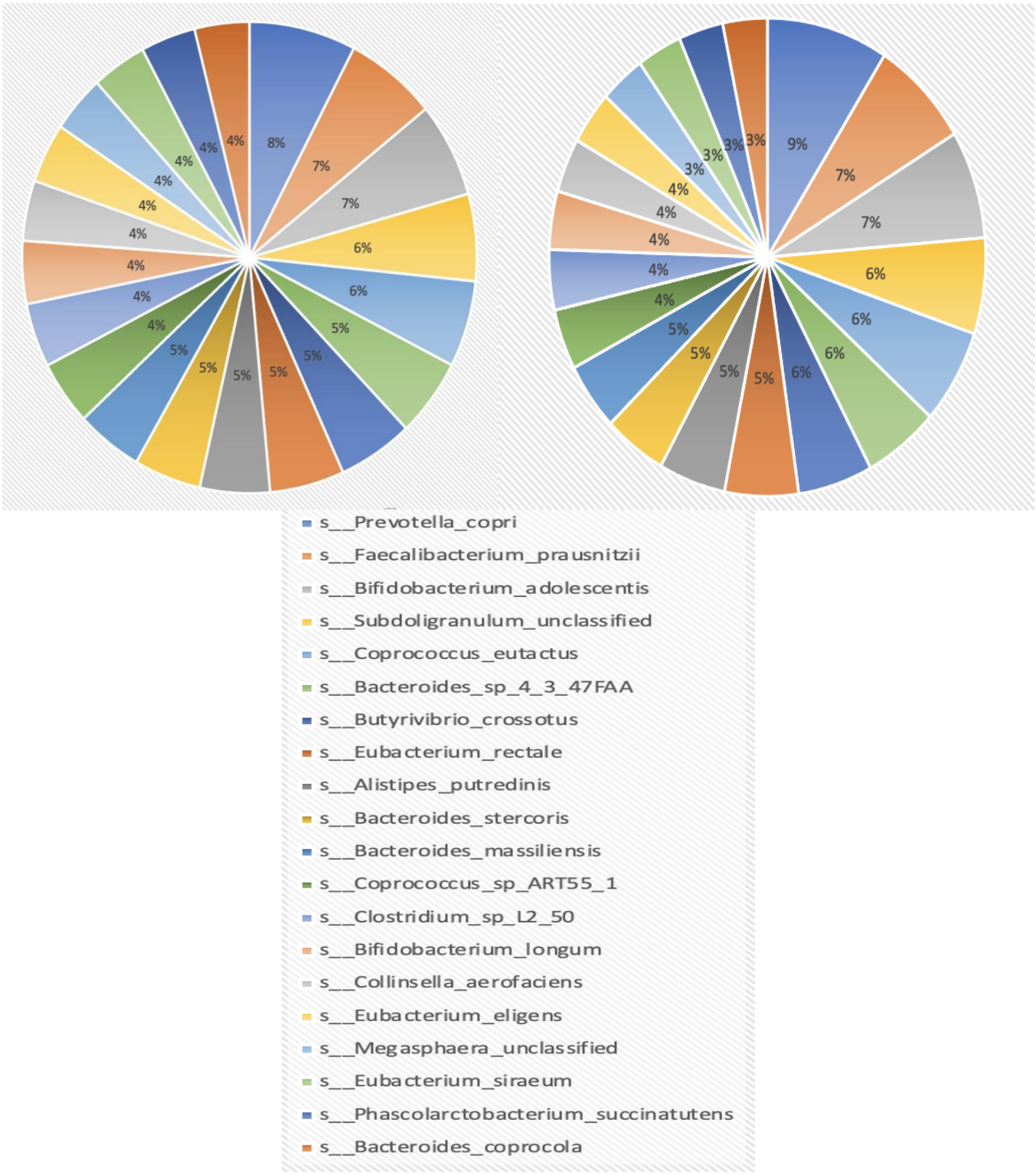
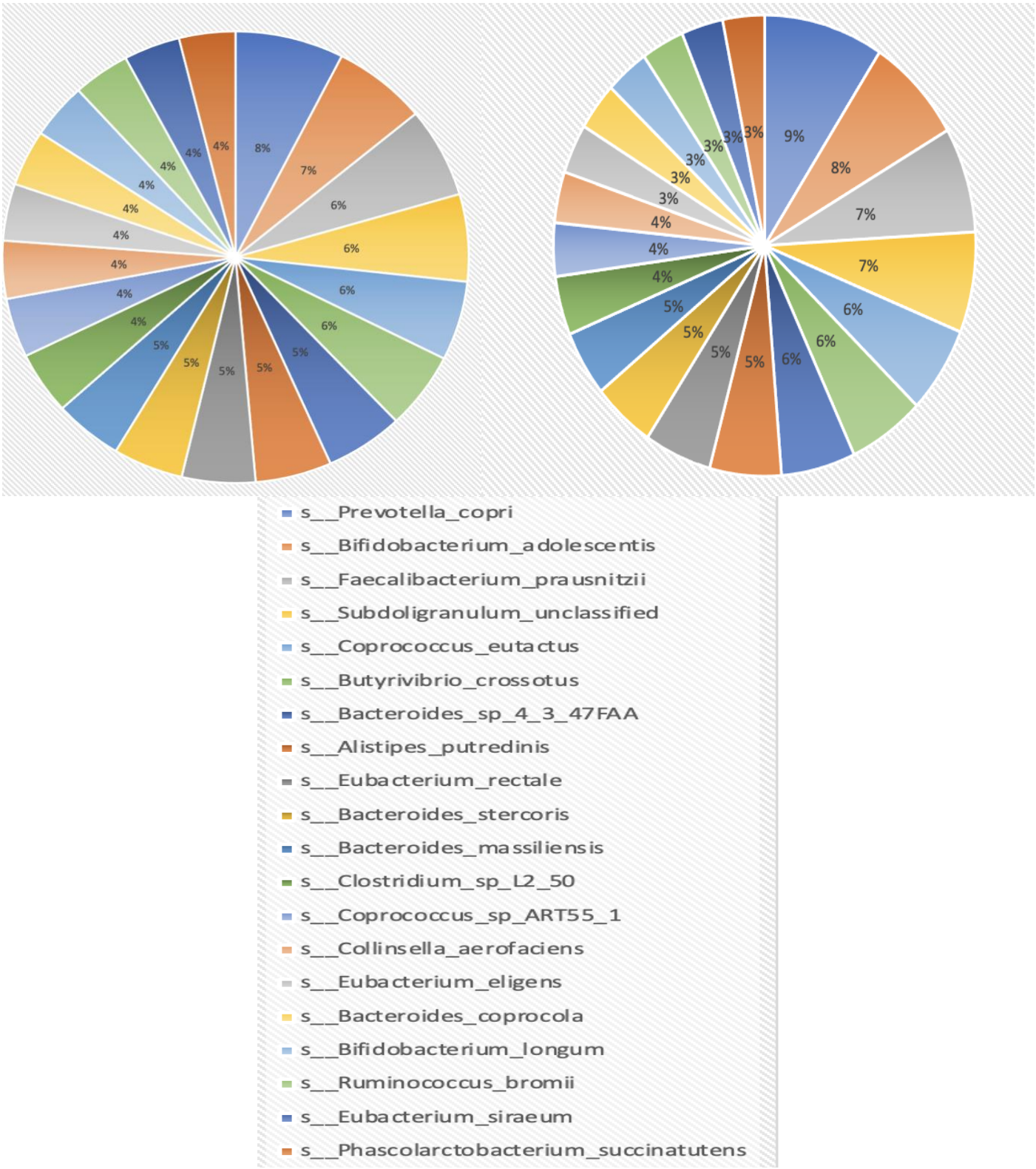


Figure 4. Pie Chart of FHQC Sample Number 3's 20 most abundant species after **A)** CLR transformation and **B)** after IS normalization, then IQLR transformation. The composition of the 20 most abundant species is relatively even after both transformations.

A. **B.**



Chapter 3: Effect of Internal Standard Normalization of Fecal Microbiome Data on Outcomes of a Longitudinal Study in a Multiethnic Cohort

Introduction

Standardization approaches have the potential to improve the interpretability of human microbiome data because unintended variability can be introduced at each level of data production and processing.[1] One way to bring standardization to microbiome studies is with internal standards. In three microbiome sequencing methods—16S rRNA gene sequencing, and metagenomic and metatranscriptomic sequencing--this standardization can involve the addition of nucleic acid sequences into a sample. This is referred to as "spiking" in the sequences and the sequences themselves are referred to as "spike-ins." In 16S rRNA gene sequencing, spiking in synthetic DNA sequences has been used to quantify absolute abundances of microbes (gene copies per ng of DNA or mg of sample).[2] This internal standard method is useful because 16S rRNA gene sequencing without the use of an internal standard only allows for the relative abundance of microbes.[2] Absolute abundances allow for a more comprehensive analysis because relative abundances of bacteria species can be shifted by an increase in the number of another bacteria species.[2–4] Additionally, all meta-omics studies that use microbial community samples cannot provide information on the absolute abundance of a nucleic acid molecule.[5] Similarly to 16S rRNA gene data, DNA from a microbe that is foreign to the microbial community of interest or synthetic mRNA can be used as a spike-in standard, allowing for the quantification of absolute differences in microbes.[5]

In this study, we aimed to assess the degree of standardization in a subset of the gut microbiome data from the Multiethnic Cohort (MEC) study,[6] a prospective cohort study, by incorporating internal standards, which will also give us a more quantitatively accurate analysis, as mentioned above.[2–4] Previous researchers used intraclass correlation coefficients (ICC) to determine the reliability of microbiome measures from the MEC study.[7] We aimed to use the ICCs to determine the reliability of newly bioinformatically processed genus abundances from the MEC study with and without an internal standard normalization.

Material and Methods

Study design:

This study used data from a sample of participants in the MEC study,[6] which was a prospective cohort study. In the study, 215,251 men and women aged 45 to 75 were enrolled from 1993 to 1996 and the participants were predominantly members of the following five racial/ethnic groups: African American, Japanese American, Latino, Native Hawaiian, and white. Of the participants who were aged 60 to 77, more than 1,800 were recruited to the MEC Adiposity Phenotype Study from 2013 to 2017. The aim of the MEC-APS study was to assess associations between body fat distribution and the exposome, genome, microbiome, and metabolome. Of the MEC-APS participants, 50 participants were chosen to be in the longitudinal fecal microbiome study; their data were used in the present analysis.[7] The aim of this longitudinal fecal microbiome study was to determine how the fecal microbiome changes over a 2-year time period in the study population.

Study setting:

The MEC study was conducted in Hawaii and Los Angeles County, with study centers in Honolulu, HI and Los Angeles, CA. Stool samples were shipped from the study centers to Fred Hutchinson Cancer Research Center (Fred Hutch) in Seattle, Washington for processing. The 16S rRNA gene sequencing was done at Molecular Research, LP in Shallowater, Texas. The sequencing data was sent to Fred Hutch for analysis.

Study participants:

From the MEC study, 1,800 participants between 60 and 70 years old were recruited for MEC-APS. The exclusion criteria for this study included BMI beyond 18.5 – 40 kg/m^2; using oral or injectable antibiotics in the past 3 months; smoking within the past 2 years; vaccinations in the past month; 20 lbs or greater weight change in the past 6 months; soft/metal implants; ileostomy/colectomy; dialysis; insulin/thyroid medication; any of these procedures/treatments in the last 6 months: chemotherapy, radiotherapy, corticosteroid hormones, prescription drugs for weight-loss, endoscopy/irrigation of large intestine.

The 50 participants from MEC-APS for the longitudinal fecal microbiome study were randomly selected for an even distribution of the following characteristic: sex assigned at birth (25 male, 25 female), five MEC racial/ethnic groups (10 African American, 10 Japanese American, 10 Native Hawaiian, 10 Latino, and 10 white), and BMI (each sex-ethnic group had one individual from 22 – 24.9, 25 – 26.9, 27 – 29.9, 30 – 34.9 kg/m^2 and one individual from either 18.5 – 21.9 or 35 – 40 kg/m^2).

Data collection:

The participants were requested to collect one stool sample every 6 months for 2 years, so that each participant would have 5 samples. The participants collected the stool samples in a collection tube with 5-mL RNAlater (Fisher Scientific) and sterile 5-mm glass beads (Ambion) to ensure the sample is incorporated into the RNAlater. The participants filled out a questionnaire each time they submitted their sample that asked about diet, collection time, any probiotic foods eaten in the past 6 months, and any antibiotic use in the past 6 months. Those who took antibiotics in the past 6 months at baseline delayed sample collection by 6 months and filled out the baseline questionnaire again.

Samples were processed and prepared for sequencing at Fred Hutch, before being shipped for 16S rRNA gene sequencing with the V1-V3 region. As an internal standard, 0.5% of total genomic DNA from *Ruegeria pomeroyi*[8,9] (ATCC® 700808™) was added to each sample after extraction. The ATCC® 700808™ genome is cross referenced on the ATCC website as GenBank Accessions CP000031 and CP000032, which are both included in the GenBank Assembly Accession GCA_000011965.2 in their current version (CP000031.2 and CP000032.1). Internal standard was also added to Fred Hutch control (FHC) samples run as part of the MEC analysis (as described in detail below). These were used to measure variability in library preparation and sequencing batches. These FHC samples were made by combining stool samples from 6 individuals who did not take antibiotics in the 3 months prior to collection and were not part of the MEC study.

Data Processing:

Initial bioinformatic processing of the data was completed by Dr. Meredith Hullar and Mr. Keith Curtis. The 16S rRNA gene data were processed with QIIME v.1.8[10] and the SILVA database (release 111).

MEC Sample Data Analysis:

To identify the internal standards in the MEC participants' samples, we searched the QIIME output samples for the genus *Ruegeria*. We calculated the expected number of *Ruegeria pomeroyi* DSS-3 16S rRNA genes

in the sample using the formula (3 rrn operons in the *Ruegeria pomeroyi* DSS-3 genome x amount of *Ruegeria pomeroyi* DSS-3 genome in the sample in ng x 6.022×10^{23} molecules/mol) / (4601048 bp in the *Ruegeria pomeroyi* DSS-3 genome x 660 g/mol in DNA x 1×10^{9} ng/g). Using this formula, the expected number of copies of the internal standard was 59,492. The genome size is from the statistics listed for GenBank Assembly Accession GCA_000011965.2.

Once the internal standard was identified in the dataset, we used a normalization method using a DNA internal standard that was developed by previous researchers.[11] The abundance of a taxa using this method is calculated as (number of reads of the taxa) x (number of 16S rRNA genes spiked into the sample) / (number of 16S rRNA genes sequenced in the sample) x (volume of the sample).[11] We did not include the volume of the sample because we were not using a liquid sample. Outliers were determined by plotting the library size by the internal standards recovered and defined as points not clustered with the others. The median was used to replace the outliers because doing so decreased the size of mean, while replacing the outliers with the mean increased the size of the median. It was preferable to possibly underestimate the size of the mean than overestimate the size of the median. Then, for both the phyla and genera, the abundances and the internal standard (IS)-normalized abundances were IQLR transformed after adding 1 to each abundance and the ICCs were calculated with a linear mixed-effects model using the "lmer" function in the lme4 package in R[12] and the "performance::icc" function in the "sjstats" package in R. For the phyla, the IS-normalized abundances were also CLR-transformed after adding 1 to each abundance using the "compositions" package in R.[13] ICCs for the genera were plotted against the mean abundance. Then, for both the genera and OTU abundances and the internal standard (IS)-normalized abundances, ICCs were calculated using the first PCoA axis (PC1) for the Bray-Curtis measure using the R packages "vegan"[14] and "sjstats". The IS-normalized abundances were rounded before generating the Bray-Curtis measure because integers are needed for the calculation. All analysis was done in R version 4.0.2 and was based off the R code used by Fu et al., 2019.[7] Reliability was categorized as excellent when ICC ≥ 0.75, good when 0.74 ≥ ICC ≥ 0.60, fair when 0.59 ≥ ICC ≥ 0.40, and poor when ICC ≤ 0.39.[15]

FHC Sample Analysis:

To identify the internal standards in the FHC samples (n=11), we searched the QIIME output samples for the genus *Ruegeria*. We calculated the expected number of *Ruegeria pomeroyi* DSS-3 16S rRNA genes in the sample to be 59,492 copies using the same formula as above for the MEC samples.

Once the internal standard was identified in the FHCs and the outliers were replaced with median as described for the MEC samples, we used a normalization method using a DNA internal standard that was developed by previous researchers.[11] Then, for both the genera abundances and the internal standard (IS)-normalized abundances, the mean (AVERAGE formula in Microsoft Excel), standard deviation (SD) (STDEV formula in Microsoft Excel) and the coefficient of variation (CV) (SD/mean x 100) were calculated for the phyla, orders, and genera in the FHCs. The mean, SD, and CV was taken at each of those 3 levels across all 11 samples and for the samples that were within the same sequencing plate. The means, SDs, and CVs are presented for 3 phyla and for the 10 most abundant orders and genera. The order of the abundances was determined by mean abundance for all 11 FHC samples. For the orders and genera, the CVs of the abundances and IS-normalized abundances were plotted against each other. When an order or genus had an undefined CV, the CV was not plotted. A CV was undefined when the SD was 0, which

happened when the samples had the same abundance for a genus. These plots were also done across all 11 samples and for the samples that were within the same sequencing plate. The R-squared was calculated for those CVs as well (RSQ function in Microsoft Excel).

Results

Participant characteristics:

The 50 participants in this study had equal representation for both sexes and for each of the MEC ethnic groups (Table 1). Of the 50 total participants, 23 took antibiotics at least once over the two-year sampling period.

Analysis of microbiome measures:

MEC Sample Analysis. After replacing the outlier abundances with the median values, the internal standard had a mean ± SD of 91 ± 87 and a median (range) of 62 (1-549) reads for the study samples. At the phyla level, the ICCs for the IQLR-transformed phyla abundances were larger than the IS-normalized abundances (Table 2). At the genus level, more of genera have an ICC above 0.40 for the IQLR-transformed abundances compared to the IQLR-transformed IS-normalized abundances (Figure 1). The ICC and mean abundance used for each of the plots in Figure 1 are presented in Tables 3 to 5. For the ICCs, an adjusted ICC, which uses the mean random effect variance, and a conditional ICC, which uses a full model with covariates, were calculated. Since both the adjusted ICC and conditional ICC were the same value, only one ICC value was presented and plotted. For both the genus and OTU ICCs using the Bray-Curtis PC1, the abundance and IS-normalized values were the same.

FHC Sample Analysis. After replacing the outlier abundances with the median values, the internal standard had a mean ± SD of 25 ± 13 and a median (range) of 23 (11-58) reads for the FHC samples. Three of the most abundant phyla presented by Fu et al.[7] are presented for the present study. At the phylum level, the abundances across the plates and within the plates are generally similar (Table 8). The IS-normalized abundances across the plates are more drastically different (Table 8). Additionally, the IS-normalized phyla count CVs are larger than the count CVs for plate 2, but the IS-normalized phyla count CVs are smaller than the count CVs for plate 3 (Table 8).

The ten most abundant orders as determined by the means of each order across all 11 FHC plates are presented. At the order level, the CVs across the plates are sometimes larger and sometimes smaller than the CVs within each plate without a general pattern (Table 9). Similarly to the phyla, for the orders the IS-normalized count CVs were larger than the count CVs for plate 2 but the IS-normalized count CV were smaller than the count CVs for plate 3 (Table 9).

The ten most abundant genera as determined by the means of each order across all 11 FHC plates are presented. At the genus level, the CVs across the plates are sometimes larger and sometimes smaller than the CVs within each plate without a general pattern (Table 10). Similarly to the phyla and orders, for the genera the IS-normalized count CVs were larger than the count CVs for plate 2 but the IS-normalized count CV were smaller than the count CVs for plate 3 (Table 10).

Discussion

In this study, we examined the effect of IS-normalization on IQLR-transformed taxon and beta-diversity ICCs in a sample of longitudinal repeat samples from the MEC and CVs in a pooled FHC quality control tested within 4 plates and across all plates. We found that IS-normalized phylum abundances generated

lower ICCs than IQLR-transformed abundances. We also found that more of the IQLR-transformed IS-normalized genus abundances were below 0.40 compared to the IQLR-transformed genus abundances. When using the Bray-Curtis PC1, the IS-normalized genus and OTU abundances ICCs were the same as the abundance ICCs. The Bray-Curtis index describes the proportion of the total counts of the species that differs between two sites.[16] Therefore, the Bray-Curtis index would change if the relative abundance of a species changes. Perhaps the IS-normalization process does not change the relative abundance of each species.

At the phyla level, in the FHC samples, the CVs were only lower for the IS-normalized abundances for the comparison within the plate 3. At the order level, there were FHC sample plate comparisons that where the IS-normalized abundance CVs were larger than the abundances and comparisons where the opposite was true. At the genus level, there two FHC sample plate comparisons where the IS-normalized count CVs were smaller than the abundances and the rest of CVs were larger for IS-normalized abundances. Lin et al. normalized 16S rRNA and 18S rRNA sequencing data with an IS and found that they had a better statistical and ecological understanding of their data.[11] Lin et al. tested the reproducibility of their IS-normalization method by sequencing duplicates of a sample and found an average estimated taxon abundance CV of 2.8% and a maximum of 12.3%, while ours were much higher.[11] Since the FHC samples were replicates of the same sample, we would expect small CVs, so high CVs indicates that substantial technical variation was introduced into our data.

It should also be noted that the use of internal standards does come with its own biases. For 16S rRNA gene sequencing, Tourlousse et al. noted that the use of DNA sequence spike-in standards does not always give a reliable count of the cells in a sample because the 16S rRNA gene copy number can vary for different bacterial cells.[2] Tourlousse et al. recommend cell biology techniques like fluorescence *in situ* hybridization to get accurate cell counts.[2] Lin et al. verified their IS-normalized counts using CHEMTAX pigment analysis and counts of bacterial cells with flow cytometry and found good correlations, though they note that PCR bias limits any PCR-based sequencing.[11] So, cell biology techniques should still be used to double-check the accuracy of this technique.

We hypothesized that IS-normalized abundances would have higher ICCs and lower CVs than the abundances because both the ICCs[7] and CVs analyses were meant to account for technical variation. We found that this was not the case with all of the analyses. When looking at means and SDs for phyla, orders, and genera in the FHC samples, the values are larger for the IS-normalized abundances than for the non-IS-normalized abundances. While this did not always equate to larger CVs for the IS-normalized abundances, it might mean that the IS-normalization process is introducing more variation into the data. This increase in means and SDs could then also be making the ICCs lower in the analysis. Future research could investigate the ICCs and CVs for IS-normalized data with a mock community to better understand how the normalization process works.

References

1. Leigh Greathouse K, Sinha R, Vogtmann E. DNA extraction for human microbiome studies: The issue of standardization. *Genome Biol*. 2019;20(1):1-4. doi:10.1186/s13059-019-1843-8

2. Tourlousse DM, Yoshiike S, Ohashi A, Matsukura S, Noda N, Sekiguchi Y. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. *Nucleic Acids Res*. 2017;45(4):e23. doi:10.1093/nar/gkw984
3. Tsilimigras MCB, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann Epidemiol*. 2016;26(5):330-335. doi:10.1016/j.annepidem.2016.03.002
4. Props R, Kerckhof FM, Rubbens P, et al. Absolute quantification of microbial taxon abundances. *ISME J*. 2017;11(2):584-587. doi:10.1038/ismej.2016.117
5. Satinsky BM, Gifford SM, Crump BC, Moran MA. Use of Internal Standards for Quantitative Metatranscriptome and Metagenome Analysis. *Methods Enzymol*. 2013;531:237-50. doi:10.1016/B978-0-12-407863-5.00012-5
6. Kolonel LN, Henderson BE, Hankin JH, et al. A multiethnic cohort in Hawaii and Los Angeles: Baseline characteristics. *Am J Epidemiol*. 2000;151(4):346-357. doi:10.1093/oxfordjournals.aje.a010213
7. Fu BC, Randolph TW, Lim U, et al. Temporal variability and stability of the fecal microbiome: The multiethnic cohort study. *Cancer Epidemiol Biomarkers Prev*. 2019;28(1):154-162. doi:10.1158/1055-9965.EPI-18-0348
8. González JM, Covert JS, Whitman WB, et al. Silicibacter pomeroyi sp. nov. and Roseovarius nubinhibens sp. nov., dimethylsulfoniopropionate-demethylating bacteria from marine environments. *Int J Syst Evol Microbiol*. 2003;53(5):1261-1269. doi:10.1099/ijs.0.02491-0
9. Yi H, Lim YW, Chun J. Taxonomic evaluation of the genera Ruegeria and Silicibacter: A proposal to transfer the genus Silicibacter Petursdottir and Kristjansson 1999 to the genus Ruegeria Uchino et al. 1999. *Int J Syst Evol Microbiol*. 2007;57(4):815-819. doi:10.1099/ijs.0.64568-0
10. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high- throughput community sequencing data. *Nat Publ Gr*. 2010;7(5):335-336. doi:10.1038/nmeth0510-335
11. Lin Y, Gifford S, Ducklow H, Schofield O, Cassar N. Towards Quantitative Microbiome Community Profiling Using Internal Standards. *Appl Environ Microbiol*. 2019;85(5):e02634-18. doi:10.1128/AEM.02634-18.
12. Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1). doi:10.18637/jss.v067.i01
13. van den Boogaart KG, Tolosana-Delgado R. “compositions”: A unified R package to analyze compositional data. *Comput Geosci*. 2008;34(4):320-338. doi:10.1016/j.cageo.2006.11.017
14. Dixon P. Computer program review VEGAN , a package of R functions for community ecology. *J Veg Sci*. 2003;14(6):927-930. http://doi.wiley.com/10.1111/j.1654-1103.2002.tb02049.x
15. Cicchetti D V. Guidelines , Criteria , and Rules of Thumb for Evaluating Normed and. *Psychol Assess*. 1994;6(4):284-290.
16. Ricotta C, Podani J. On some properties of the Bray-Curtis dissimilarity and their ecological meaning. *Ecol Complex*. 2017;31:201-205. doi:10.1016/j.ecocom.2017.07.003

Figure and Tables

Table 1. Characteristic of the MEC participants who were part of this study. The characteristics are stratified by any antibiotic use during the two-year sampling period. Mean ± SD is reported for the continuous variables and n (%) is reported for the categorical variables.

	No antibiotic use (n = 27)	Antibiotic use (n = 23)	Total (n = 50)
Age, years	68.6 ± 2.7	69.0 ± 2.3	68.6 ± 2.7
Female	15 (55.6)	10 (43.5)	25 (50)
Race/ethnicity			
African American	5 (18.5)	5 (21.7)	10 (20)
Japanese American	7 (25.9)	3 (13.0)	10 (20)
Native Hawaiian	3 (11.1)	7 (30.4)	10 (20)
Latino	6 (22.2)	4 (17.4)	10 (20)
White	6 (22.2)	4 (17.4)	10 (20)
Education, years	15.0 ± 2.5	13.8 ± 3.4	14.4 ± 3.0
Smoking status			
Never	19 (70.4)	17 (73.9)	36 (72.0)
Former	8 (29.6)	6 (26.1)	14 (28.0)
Body fat %	33.2 ± 6.8	32.2 ± 9.2	32.8 ± 7.9

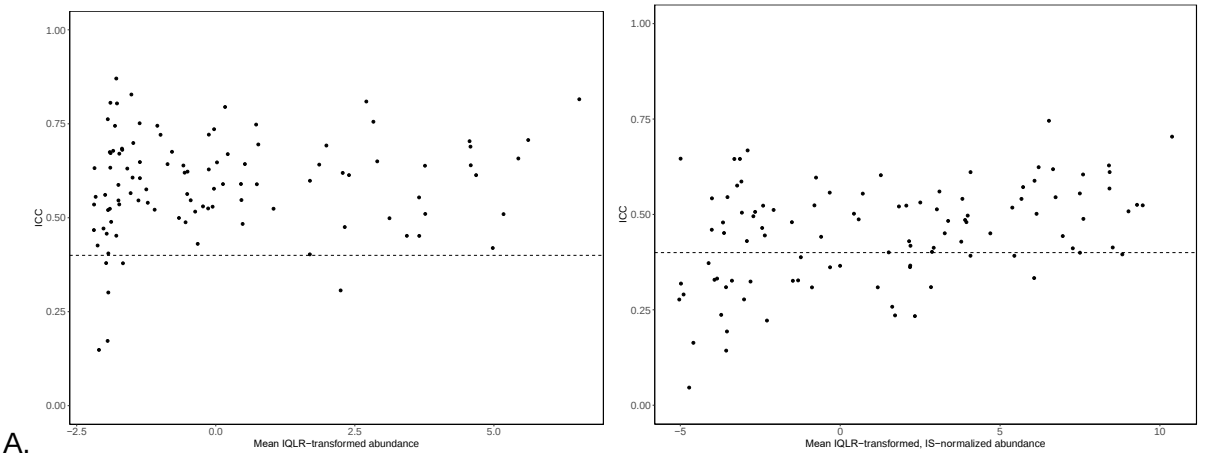
Table 2. ICCs of three of the most abundant phyla for the MEC participants. ICCs are presented for IQLR-transformed abundances, IS-normalized abundances, IQLR-transformed IS-normalized abundances, and CLR-transformed IS-normalized abundances. There are NAs for the two of the IS-normalized phyla because the ICCs could not be calculated for those phyla due to some variances equaling zero, which indicates that there might be singularity.

Total	IQLR-Transformed Abundances		IS-Normalized Abundances		IQLR-Transformed IS-Normalized Abundances		CLR-Transformed IS-Normalized Abundances	
Phylum	ICC	Mean Abundance	ICC	Mean Abundance	ICC	Mean Abundance	ICC	Mean Abundance
Firmicutes	0.44	9.27	0.02	22511330	0.22	10.30	0.28	9.73
Bacteroidetes	0.52	9.11	NA	20108900	0.24	10.13	0.28	9.56
Actinobacteria	0.41	4.70	0.11	424336	0.36	5.71	0.36	5.14

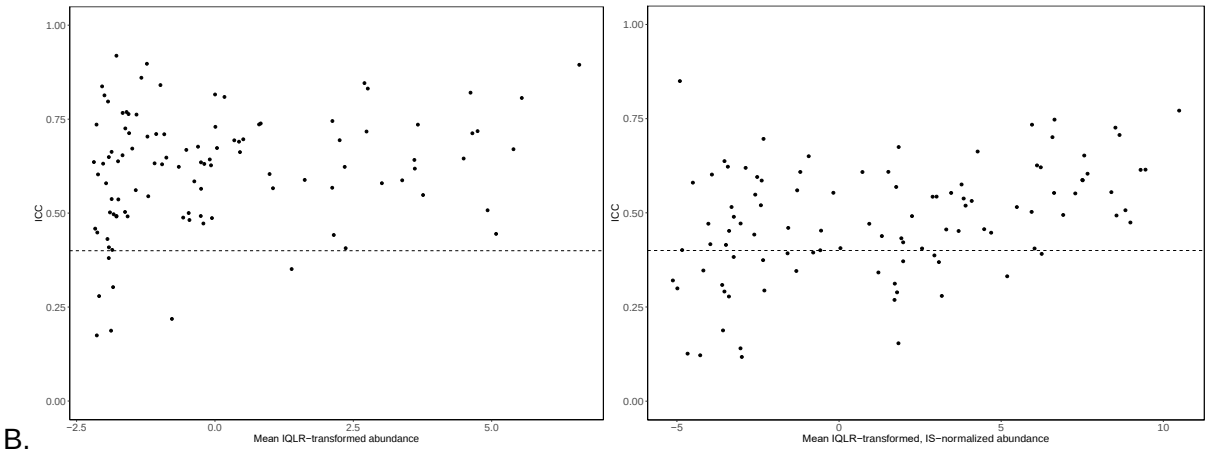
No Antibiotics	IQLR-Transformed Abundances		IS-Normalized Abundances		IQLR-Transformed IS-Normalized Abundances		CLR-Transformed IS-Normalized Abundances	
Phylum	ICC	Mean Abundance	ICC	Mean Abundance	ICC	Mean Abundance	ICC	Mean Abundance
Firmicutes	0.49	9.22	0.01	26105650	0.24	10.37	0.30	9.72
Bacteroidetes	0.53	9.14	NA	25637220	0.25	10.29	0.23	9.64
Actinobacteria	0.53	4.93	0.08	544879	0.41	6.06	0.46	5.41

Antibiotics	IQLR-Transformed Abundances		IS-Normalized Abundances		IQLR-Transformed IS-Normalized Abundances		CLR-Transformed IS-Normalized Abundances	
Phylum	ICC	Mean Abundance	ICC	Mean Abundance	ICC	Mean Abundance	ICC	Mean Abundance
Firmicutes	0.38	9.34	0.11	18417790	0.20	10.22	0.27	9.73
Bacteroidetes	0.52	9.07	0.07	13812770	0.22	9.95	0.31	9.47
Actinobacteria	0.26	4.44	0.20	287050	0.24	5.31	0.23	4.82

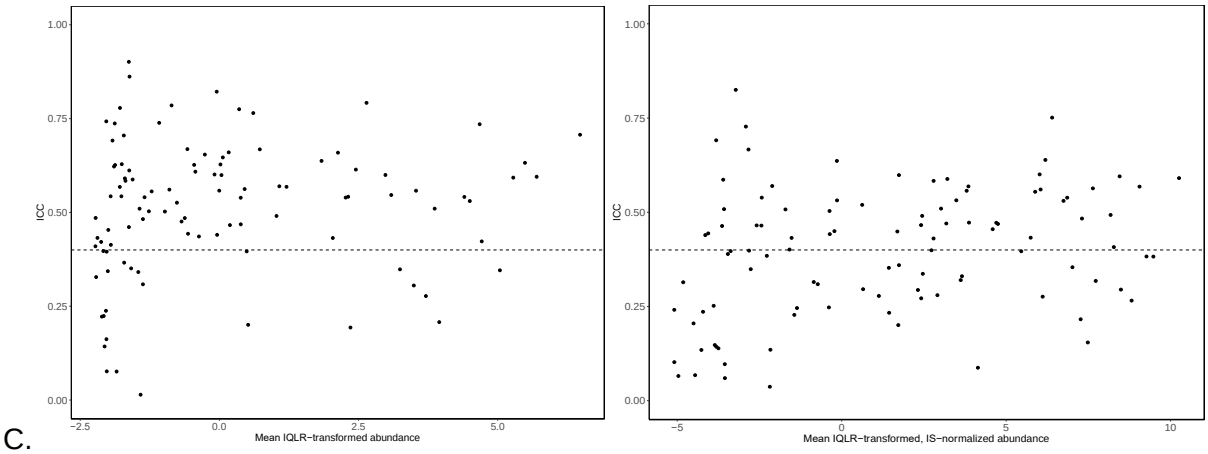
Figure 1. ICCs plotted against mean abundances of all genera for the MEC participants. All plots have 103 genera, except for plot B in the right column which has 102 genera because the ICCs could not be calculated for one genus due to some variances equaling zero, which indicates that there might be singularity. A Plots: All participants, B Plots: No antibiotics participants, C Plots: Antibiotics participants. All figures in the left column have IQLR-transformed abundances and all figures in the right column have IQLR-transformed IS-normalized abundances. Dotted line is an ICC of 0.40.



A.



B.



C.

Table 3. ICCs calculated with mean IQLR-transformed abundances and ICCs calculated with mean IQLR-transformed IS-normalized abundances of 103 genera for all MEC participants.

Genus	ICC	Mean IQLR-transformed abundance	ICC (IS-normalized	Mean IQLR-transformed, IS-normalized abundance
Bacteria.Other.Other.Other.Other.Other	0.15	-2.10	0.05	-4.72
Bacteria.__Actinobacteria.__Bifidobacteriales.__Bifidobacteriaceae.__Bifidobacterium.__uncultured_bacterium	0.59	0.45	0.43	3.79
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Adlercreutzia	0.55	-1.39	0.44	-0.60
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Collinsella	0.60	1.69	0.52	5.38
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Eggerthella	0.63	-1.90	0.59	-3.09
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Enterorhabdus	0.61	-1.50	0.39	-1.23
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Slackia	0.74	-1.81	0.65	-3.14
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__uncultured	0.64	-0.58	0.60	1.27
Bacteria.__Actinobacteria.__Corynebacteriales.__Nocardiaceae.__Rhodococcus.__unidentified	0.49	-0.54	0.24	1.72
Bacteria.__Bacteroidetes.Other.Other.Other.Other	0.52	-0.14	0.52	1.84
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.Other.Other	0.64	0.52	0.49	3.91
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Bacteroidaceae.__Bacteroides	0.82	6.53	0.70	10.38
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Porphyromonadaceae.__Barnesiella	0.64	4.58	0.61	8.43
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Porphyromonadaceae.__Butyricimonas	0.79	0.17	0.56	3.10
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Porphyromonadaceae.__Odoribacter	0.56	-2.16	0.29	-4.90
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Porphyromonadaceae.__Parabacteroides	0.69	4.58	0.57	8.42
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Prevotellaceae.Other	0.59	-1.75	0.50	-2.72
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Prevotellaceae.__Paraprevotella	0.62	-0.51	0.37	2.19
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Prevotellaceae.__Prevotella	0.81	2.71	0.75	6.53
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Prevotellaceae.__uncultured	0.80	-1.78	0.32	-2.81
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__RF16.__uncultured_bacterium	0.74	-0.03	0.41	2.92
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Rikenellaceae.__Alistipes	0.64	3.76	0.60	7.60
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Rikenellaceae.__RC9_gut_group	0.87	-1.79	0.58	-3.22
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__S24.7.__uncultured_bacterium	0.48	2.32	0.50	6.14
Bacteria.__Bacteroidetes.__VC2.1_Bac22.__uncultured_bacterium.Other.Other	0.75	0.73	0.61	4.08
Bacteria.__Firmicutes.__Bacilli.__Lactobacillales.__Lactobacillaceae.__Lactobacillus	0.52	-1.90	0.48	-3.67
Bacteria.__Firmicutes.__Bacilli.__Lactobacillales.__Leuconostocaceae.__Weissella	0.43	-2.12	0.16	-4.59
Bacteria.__Firmicutes.__Bacilli.__Lactobacillales.__Streptococcaceae.__Streptococcus	0.52	1.04	0.45	4.70
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.Other.Other	0.55	0.46	0.54	3.82
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Christensenellaceae.__uncultured	0.65	2.90	0.55	6.74
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Clostridiaceae.__Clostridium	0.59	0.74	0.39	4.08
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Clostridiaceae.__Sarcina	0.68	-1.69	0.51	-2.66
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Family_XIII_Incertae_Sedis.__uncultured	0.49	-1.88	0.50	-3.08
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.Other	0.64	1.86	0.54	5.67
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Anaerostipes	0.55	3.65	0.55	7.50
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Blautia	0.70	4.56	0.63	8.41
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Butyrivibrio	0.81	-1.89	0.45	-3.63
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Coprococcus	0.50	3.12	0.44	6.96
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Dorea	0.31	2.24	0.33	6.07
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Howardella	0.74	-1.05	0.60	-0.76
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Lachnospira	0.61	2.39	0.62	6.21
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Moryella	0.48	0.48	0.48	3.95
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Pseudobutyrivibrio	0.42	4.98	0.40	8.82
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Roseburia	0.45	3.43	0.41	7.28
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Shuttleworthia	0.68	-1.68	0.46	-2.44
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__uncultured	0.66	5.44	0.53	9.28
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae_Incertae_Sedis.__Clostridia	0.68	-0.79	0.49	0.58
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae_Incertae_Sedis.__Clostridium	0.72	-0.99	0.50	0.43
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae_Incertae_Sedis.__human_gut_metagenome	0.58	-0.03	0.53	2.50
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae_Incertae_Sedis.__uncultured_bacterium	0.71	5.62	0.52	9.46
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Peptostreptococcaceae.__uncultured	0.67	0.22	0.48	3.38
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Peptostreptococcaceae_Incertae_Sedis.__uncultured_bacterium	0.53	-0.06	0.51	3.02
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.Other	0.59	0.13	0.45	3.27
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Anaerotruncus	0.52	-0.37	0.36	2.18
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Faecalibacterium	0.51	5.17	0.51	9.02
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Oscillibacter	0.61	-1.36	0.52	-0.81
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Oscillospira	0.55	-0.45	0.43	2.15
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Ruminococcus	0.51	3.77	0.49	7.61
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Subdoligranulum	0.76	2.83	0.62	6.66
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__uncultured	0.61	4.68	0.41	8.52
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.Other	0.54	-1.22	0.36	-0.32
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.__Clostridium	0.67	-1.89	0.55	-3.53
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.__Ruminococcus	0.68	-1.84	0.65	-3.31
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.__human_gut_metagenome	0.38	-1.67	0.22	-2.29
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.__uncultured_bacterium	0.45	3.66	0.40	7.50
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.Other	0.63	-0.13	0.52	2.07
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Acidaminococcus	0.65	-1.36	0.48	-1.51
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Allisonella	0.57	-1.53	0.44	-2.36
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Dialister	0.70	-1.48	0.33	-1.48
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Megamonas	0.46	-1.96	0.33	-3.85
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Megasphaera	0.55	-1.75	0.43	-2.92
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Phascolarctobacterium	0.69	1.99	0.57	5.72
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Veillonella	0.52	-1.10	0.37	0.00
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__uncultured	0.45	-1.79	0.28	-3.01
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__uncultured.Other	0.52	-1.94	0.31	-3.57
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__uncultured.__uncultured_bacterium	0.47	-2.02	0.37	-4.12
Bacteria.__Firmicutes.__Clostridia.__Halanaerobiales.__Halanaerobiaceae.__Halocella	0.41	-1.93	0.33	-3.39
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae.Other	0.47	-2.19	0.28	-5.03
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae.__Catenibacterium	0.62	-0.56	0.40	1.52
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae.__Coprobacillus	0.38	-1.97	0.24	-3.72
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae.__uncultured	0.62	2.28	0.59	6.08
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae_Incertae_Sedis.__Clostridium	0.58	-1.25	0.56	-0.33
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae_Incertae_Sedis.__Eubacteriaceae_bacterium_DJF_VR85	0.67	-1.73	0.67	-2.90
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae_Incertae_Sedis.__uncultured_Firmicutes_bacterium	0.64	-0.87	0.55	0.71
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae_Incertae_Sedis.__uncultured_bacterium	0.72	-0.12	0.42	2.20
Bacteria.__Fusobacteria.__Fusobacteriales.__CFT112H7.__uncultured_bacterium.Other	0.83	-1.52	0.33	-1.32
Bacteria.__Lentisphaerae.__Lentisphaeria.__Victivallales.__Victivallaceae.__Victivallis	0.63	-1.59	0.52	-2.41
Bacteria.__Proteobacteria.__Alphaproteobacteria.__Rhodospirillales.__Rhodospirillaceae.__Thalassospira	0.43	-0.33	0.23	2.34
Bacteria.__Proteobacteria.__Betaproteobacteria.__Burkholderiales.__Alcaligenaceae.__Parasutterella	0.70	0.76	0.50	3.99
Bacteria.__Proteobacteria.__Betaproteobacteria.__Burkholderiales.__Alcaligenaceae.__Sutterella	0.65	0.02	0.40	2.87
Bacteria.__Proteobacteria.__Deltaproteobacteria.__Desulfovibrionales.__Desulfovibrionaceae.__Desulfovibrio	0.76	-1.94	0.54	-4.01
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Aeromonadales.__Succinivibrionaceae.__Succinivibrio	0.75	-1.37	0.31	-0.88
Bacteria.__Proteobacteria.__Gammaproteobacteria.__B38.__uncultured_bacterium.Other	0.56	-1.99	0.33	-3.93
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Enterobacteriales.__Enterobacteriaceae.__Enterobacter	0.53	-0.23	0.31	2.84
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Enterobacteriales.__Enterobacteriaceae.__Escherichia.Shigella	0.50	-0.66	0.31	1.17
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Enterobacteriales.__Enterobacteriaceae.__Kluyvera	0.30	-1.93	0.19	-3.54
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Enterobacteriales.__Enterobacteriaceae.__Raoultella	0.56	-0.51	0.26	1.63
Bacteria.__Synergistetes.__Synergistia.__Synergistales.__Synergistaceae.__Cloacibacillus	0.54	-2.19	0.32	-4.98
Bacteria.__Tenericutes.__Mollicutes.__RF9.__uncultured_bacterium.Other	0.17	-1.94	0.14	-3.57
Bacteria.__Verrucomicrobia.__Opitutae.Other.Other.Other	0.67	-1.91	0.46	-4.02
Bacteria.__Verrucomicrobia.__Opitutae.__vadinHA64.__uncultured_bacterium.Other	0.63	-2.18	0.65	-4.99
Bacteria.__Verrucomicrobia.__Verrucomicrobiae.__Verrucomicrobiales.__Verrucomicrobiaceae.__Akkermansia	0.40	1.69	0.39	5.45
Eukaryota.Other.Other.Other.Other.Other	0.54	-1.74	0.51	-2.08

Table 4. ICCs calculated with mean IQLR-transformed abundances and ICCs calculated with mean IQLR-transformed IS-normalized abundances of 103 genera for the No Antibiotics MEC participants. There are NAs for the one of the IS-normalized genera because the ICCs could not be calculated for the first genus due to some variances equaling zero, which indicates that there might be singularity.

Genus	ICC	Mean IQLR-transformed abundance	ICC (IS-normalized	Mean IQLR-transformed IS-normalized abundance
Bacteria.Other.Other.Other.Other.Other	0.17	-2.13	NA	NA
Bacteria.__Actinobacteria.__Bifidobacteriales.__Bifidobacteriaceae.__Bifidobacterium.__uncultured_bacterium	0.70	0.51	0.52	3.90
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Adlercreutzia	0.56	-1.43	0.39	-0.80
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Collinsella	0.57	2.12	0.50	5.94
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Eggerthella	0.65	-1.92	0.52	-3.32
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Enterorhabdus	0.71	-1.55	0.46	-1.57
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Slackia	0.73	-1.62	0.70	-2.33
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__uncultured	0.68	-0.31	0.67	1.84
Bacteria.__Actinobacteria.__Corynebacteriales.__Nocardiaceae.__Rhodococcus.__unidentified	0.50	-0.47	0.27	1.71
Bacteria.__Bacteroidetes.Other.Other.Other.Other	0.49	-0.26	0.43	1.92
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.Other.Other	0.67	0.04	0.46	3.31
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Bacteroidaceae.__Bacteroides	0.89	6.58	0.77	10.49
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Porphyromonadaceae.__Barnesiella	0.72	4.75	0.71	8.65
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Porphyromonadaceae.__Butyricimonas	0.82	0.00	0.54	3.01
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Porphyromonadaceae.__Odoribacter	0.60	-2.11	0.40	-4.84
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Porphyromonadaceae.__Parabacteroides	0.65	4.49	0.55	8.40
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Prevotellaceae.Other	0.54	-1.87	0.38	-3.25
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Prevotellaceae.__Paraprevotella	0.65	-0.88	0.31	1.72
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Prevotellaceae.__Prevotella	0.83	2.76	0.75	6.64
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Prevotellaceae.__uncultured	0.38	-1.92	0.12	-3.00
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__RF16.__uncultured_bacterium	0.81	0.17	0.55	3.46
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Rikenellaceae.__Alistipes	0.74	3.67	0.65	7.56
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Rikenellaceae.__RC9_gut_group	0.80	-1.93	0.60	-3.92
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__S24.7.__uncultured_bacterium	0.41	2.36	0.39	6.25
Bacteria.__Bacteroidetes.__VC2.1_Bac22.__uncultured_bacterium.Other.Other	0.74	0.83	0.66	4.27
Bacteria.__Firmicutes.__Bacilli.__Lactobacillales.__Lactobacillaceae.__Lactobacillus	0.54	-1.75	0.49	-3.25
Bacteria.__Firmicutes.__Bacilli.__Lactobacillales.__Leuconostocaceae.__Weissella	0.45	-2.13	0.13	-4.67
Bacteria.__Firmicutes.__Bacilli.__Lactobacillales.__Streptococcaceae.__Streptococcus	0.57	1.05	0.45	4.69
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.Other.Other	0.69	0.43	0.54	3.84
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Christensenellaceae.__uncultured	0.72	2.74	0.55	6.63
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Clostridiaceae.__Clostridium	0.60	0.99	0.46	4.48
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Clostridiaceae.__Sarcina	0.77	-1.60	0.60	-2.52
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Family_XIII_Incertae_Sedis.__uncultured	0.49	-1.78	0.44	-2.61
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.Other	0.59	1.62	0.52	5.48
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Anaerostipes	0.55	3.76	0.60	7.66
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Blautia	0.82	4.62	0.73	8.52
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Butyrivibrio	0.92	-1.78	0.64	-3.53
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Coprococcus	0.58	3.02	0.49	6.91
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Dorea	0.44	2.15	0.41	6.03
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Howardella	0.70	-1.22	0.56	-1.29
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Lachnospira	0.62	2.34	0.62	6.22
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Moryella	0.66	0.45	0.58	3.78
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Pseudobutyrivibrio	0.51	4.93	0.51	8.83
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Roseburia	0.59	3.38	0.55	7.28
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Shuttleworthia	0.77	-1.67	0.55	-2.58
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__uncultured	0.67	5.39	0.61	9.30
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae_Incertae_Sedis.__Clostridia	0.84	-0.98	0.55	-0.17
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae_Incertae_Sedis.__Clostridium	0.71	-0.92	0.47	0.93
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae_Incertae_Sedis.__human_gut_metagenome	0.47	-0.21	0.49	2.25
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae_Incertae_Sedis.__uncultured_bacterium	0.81	5.54	0.61	9.45
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Peptostreptococcaceae.__uncultured	0.69	0.35	0.45	3.69
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Peptostreptococcaceae_Incertae_Sedis.__uncultured_bacterium	0.63	-0.07	0.54	2.89
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.Other	0.64	-0.09	0.37	3.08
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Anaerotruncus	0.58	-0.37	0.42	1.98
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Faecalibacterium	0.44	5.08	0.47	8.98
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Oscillibacter	0.67	-1.49	0.61	-1.20
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Oscillospira	0.56	-0.25	0.41	2.56
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Ruminococcus	0.64	3.60	0.59	7.50
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Subdoligranulum	0.85	2.70	0.70	6.58
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__uncultured	0.71	4.65	0.49	8.56
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.Other	0.63	-1.09	0.41	0.04
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.__Clostridium	0.66	-1.87	0.62	-3.43
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.__Ruminococcus	0.50	-1.90	0.45	-3.39
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.__human_gut_metagenome	0.40	-1.86	0.14	-3.04
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.__uncultured_bacterium	0.62	3.61	0.59	7.51
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.Other	0.64	-0.25	0.57	1.76
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Acidaminococcus	0.71	-1.06	0.65	-0.94
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Allisonella	0.65	-1.67	0.47	-3.04
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Dialister	0.76	-1.42	0.39	-1.59
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Megamonas	0.50	-1.83	0.41	-3.49
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Megasphaera	0.50	-1.62	0.37	-2.34
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Phascolarctobacterium	0.75	2.12	0.73	5.95
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Veillonella	0.54	-1.20	0.40	-0.58
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__uncultured	0.49	-1.58	0.29	-2.30
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__uncultured.Other	0.28	-2.09	0.12	-4.28
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__uncultured.__uncultured_bacterium	0.58	-1.97	0.47	-4.03
Bacteria.__Firmicutes.__Clostridia.__Halanaerobiales.__Halanaerobiaceae.__Halocella	0.41	-1.92	0.28	-3.40
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae.Other	0.46	-2.16	0.30	-4.99
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae.__Catenibacterium	0.62	-0.65	0.44	1.32
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae.__Coprobacillus	0.43	-1.94	0.31	-3.60
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae.__uncultured	0.69	2.25	0.63	6.10
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae_Incertae_Sedis.__Clostridium	0.63	-0.96	0.61	0.72
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae_Incertae_Sedis.__Eubacteriaceae_bacterium_DJF_VR85	0.64	-1.75	0.62	-2.88
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae_Incertae_Sedis.__uncultured_Firmicutes_bacterium	0.67	-0.52	0.61	1.52
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae_Incertae_Sedis.__uncultured_bacterium	0.63	-0.19	0.37	1.98
Bacteria.__Fusobacteria.__Fusobacteriales.__CFT112H7.__uncultured_bacterium.Other	0.90	-1.23	0.45	-0.55
Bacteria.__Lentisphaerae.__Lentisphaeria.__Victivallales.__Victivallaceae.__Victivallis	0.76	-1.56	0.59	-2.39
Bacteria.__Proteobacteria.__Alphaproteobacteria.__Rhodospirillales.__Rhodospirillaceae.__Thalassospira	0.22	-0.78	0.15	1.83
Bacteria.__Proteobacteria.__Betaproteobacteria.__Burkholderiales.__Alcaligenaceae.__Parasutterella	0.74	0.80	0.53	4.09
Bacteria.__Proteobacteria.__Betaproteobacteria.__Burkholderiales.__Alcaligenaceae.__Sutterella	0.73	0.01	0.39	2.94
Bacteria.__Proteobacteria.__Deltaproteobacteria.__Desulfovibrionales.__Desulfovibrionaceae.__Desulfovibrio	0.81	-2.00	0.35	-4.18
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Aeromonadales.__Succinivibrionaceae.__Succinivibrio	0.86	-1.33	0.35	-1.32
Bacteria.__Proteobacteria.__Gammaproteobacteria.__B38.__uncultured_bacterium.Other	0.63	-2.02	0.42	-3.97
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Enterobacteriales.__Enterobacteriaceae.__Enterobacter	0.49	-0.05	0.28	3.17
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Enterobacteriales.__Enterobacteriaceae.__Escherichia.Shigella	0.49	-0.57	0.34	1.21
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Enterobacteriales.__Enterobacteriaceae.__Kluyvera	0.30	-1.84	0.29	-3.54
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Enterobacteriales.__Enterobacteriaceae.__Raoultella	0.48	-0.46	0.29	1.79
Bacteria.__Synergistetes.__Synergistia.__Synergistales.__Synergistaceae.__Cloacibacillus	0.64	-2.19	0.32	-5.13
Bacteria.__Tenericutes.__Mollicutes.__RF9.__uncultured_bacterium.Other	0.19	-1.88	0.19	-3.58
Bacteria.__Verrucomicrobia.__Opitutae.Other.Other.Other	0.84	-2.04	0.58	-4.51
Bacteria.__Verrucomicrobia.__Opitutae.__vadinHA64.__uncultured_bacterium.Other	0.74	-2.14	0.85	-4.91
Bacteria.__Verrucomicrobia.__Verrucomicrobiae.__Verrucomicrobiales.__Verrucomicrobiaceae.__Akkermansia	0.35	1.39	0.33	5.19
Eukaryota.Other.Other.Other.Other.Other	0.49	-1.78	0.52	-2.41

Table 5. ICCs calculated with mean IQLR-transformed abundances and ICCs calculated with mean IQLR-transformed IS-normalized abundances of 103 genera for the Antibiotics MEC participants.

Genus	ICC	Mean IQLR-transformed abundance	ICC (IS-normalized)	Mean IQLR-transformed IS-normalized abundance
Bacteria.Other.Other.Other.Other.Other	0.14	-2.06	0.07	-4.45
Bacteria.__Actinobacteria.__Bifidobacteriales.__Bifidobacteriaceae.__Bifidobacterium.__uncultured_bacterium	0.47	0.38	0.33	3.66
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Adlercreutzia	0.54	-1.34	0.50	-0.37
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Collinsella	0.57	1.21	0.47	4.75
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Eggerthella	0.63	-1.87	0.67	-2.83
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Enterorhabdus	0.51	-1.43	0.31	-0.85
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__Slackia	0.74	-2.03	0.44	-4.06
Bacteria.__Actinobacteria.__Coriobacteriia.__Coriobacteriales.__Coriobacteriaceae.__uncultured	0.56	-0.90	0.52	0.62
Bacteria.__Actinobacteria.__Corynebacteriales.__Nocardiaceae.__Rhodococcus.__unidentified	0.48	-0.62	0.20	1.72
Bacteria.__Bacteroidetes.Other.Other.Other.Other	0.56	0.00	0.60	1.74
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.Other.Other	0.57	1.08	0.46	4.59
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Bacteroidaceae.__Bacteroides	0.71	6.48	0.59	10.26
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Porphyromonadaceae.__Barnesiella	0.54	4.40	0.49	8.17
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Porphyromonadaceae.__Butyricimonas	0.77	0.36	0.59	3.21
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Porphyromonadaceae.__Odoribacter	0.33	-2.21	0.07	-4.96
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Porphyromonadaceae.__Parabacteroides	0.73	4.68	0.60	8.45
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Prevotellaceae.Other	0.61	-1.62	0.57	-2.11
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Prevotellaceae.__Paraprevotella	0.60	-0.09	0.40	2.73
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Prevotellaceae.__Prevotella	0.79	2.64	0.75	6.40
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Prevotellaceae.__uncultured	0.86	-1.61	0.47	-2.58
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__RF16.__uncultured_bacterium	0.65	-0.26	0.29	2.32
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Rikenellaceae.__Alistipes	0.51	3.87	0.56	7.64
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__Rikenellaceae.__RC9_gut_group	0.90	-1.63	0.54	-2.43
Bacteria.__Bacteroidetes.__Bacteroidia.__Bacteroidales.__S24.7.__uncultured_bacterium	0.54	2.27	0.60	6.02
Bacteria.__Bacteroidetes.__VC2.1_Bac22.__uncultured_bacterium.Other.Other	0.76	0.61	0.57	3.85
Bacteria.__Firmicutes.__Bacilli.__Lactobacillales.__Lactobacillaceae.__Lactobacillus	0.40	-2.08	0.44	-4.15
Bacteria.__Firmicutes.__Bacilli.__Lactobacillales.__Leuconostocaceae.__Weissella	0.42	-2.12	0.21	-4.50
Bacteria.__Firmicutes.__Bacilli.__Lactobacillales.__Streptococcaceae.__Streptococcus	0.49	1.03	0.47	4.70
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.Other.Other	0.40	0.49	0.56	3.80
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Christensenellaceae.__uncultured	0.55	3.09	0.54	6.86
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Clostridiaceae.__Clostridium	0.56	0.45	0.32	3.62
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Clostridiaceae.__Sarcina	0.57	-1.79	0.40	-2.82
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Family_XIII_Incertae_Sedis.__uncultured	0.45	-1.99	0.59	-3.60
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.Other	0.66	2.13	0.55	5.88
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Anaerostipes	0.56	3.53	0.48	7.30
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Blautia	0.53	4.50	0.41	8.27
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Butyrivibrio	0.40	-2.02	0.14	-3.75
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Coprococcus	0.35	3.24	0.35	7.01
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Dorea	0.19	2.35	0.28	6.11
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Howardella	0.79	-0.86	0.64	-0.15
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Lachnospira	0.61	2.45	0.64	6.19
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Moryella	0.20	0.52	0.09	4.14
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Pseudobutyrivibrio	0.35	5.04	0.27	8.81
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Roseburia	0.31	3.49	0.22	7.26
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__Shuttleworthia	0.59	-1.70	0.38	-2.28
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae.__uncultured	0.63	5.49	0.38	9.27
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae_Incertae_Sedis.__Clostridia	0.44	-0.56	0.35	1.43
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae_Incertae_Sedis.__Clostridium	0.74	-1.08	0.53	-0.14
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae_Incertae_Sedis.__human_gut_metagenome	0.66	0.17	0.58	2.79
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Lachnospiraceae_Incertae_Sedis.__uncultured_bacterium	0.60	5.70	0.38	9.47
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Peptostreptococcaceae.__uncultured	0.65	0.06	0.51	3.02
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Peptostreptococcaceae_Incertae_Sedis.__uncultured_bacterium	0.44	-0.04	0.47	3.18
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.Other	0.54	0.38	0.53	3.49
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Anaerotruncus	0.44	-0.37	0.27	2.42
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Faecalibacterium	0.59	5.28	0.57	9.05
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Oscillibacter	0.56	-1.22	0.44	-0.36
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Oscillospira	0.48	-0.68	0.45	1.69
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Ruminococcus	0.21	3.95	0.32	7.72
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__Subdoligranulum	0.60	2.98	0.53	6.74
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae.__uncultured	0.42	4.71	0.29	8.49
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.Other	0.31	-1.37	0.31	-0.73
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.__Clostridium	0.69	-1.92	0.46	-3.64
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.__Ruminococcus	0.78	-1.79	0.82	-3.22
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.__human_gut_metagenome	0.34	-1.46	0.23	-1.44
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Ruminococcaceae_Incertae_Sedis.__uncultured_bacterium	0.28	3.71	0.15	7.48
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.Other	0.63	0.02	0.47	2.41
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Acidaminococcus	0.37	-1.71	0.14	-2.17
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Allisonella	0.48	-1.37	0.40	-1.59
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Dialister	0.59	-1.56	0.25	-1.36
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Megamonas	0.22	-2.11	0.13	-4.26
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Megasphaera	0.62	-1.89	0.51	-3.58
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Phascolarctobacterium	0.64	1.83	0.40	5.46
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__Veillonella	0.50	-0.98	0.30	0.65
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__Veillonellaceae.__uncultured	0.16	-2.03	0.14	-3.81
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__uncultured.Other	0.54	-1.76	0.35	-2.76
Bacteria.__Firmicutes.__Clostridia.__Clostridiales.__uncultured.__uncultured_bacterium	0.22	-2.08	0.24	-4.21
Bacteria.__Firmicutes.__Clostridia.__Halanaerobiales.__Halanaerobiaceae.__Halocella	0.41	-1.95	0.40	-3.38
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae.Other	0.49	-2.22	0.24	-5.09
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae.__Catenibacterium	0.63	-0.45	0.36	1.74
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae.__Coprobacillus	0.34	-2.00	0.15	-3.86
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae.__uncultured	0.54	2.31	0.56	6.05
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae_Incertae_Sedis.__Clostridium	0.35	-1.59	0.43	-1.52
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae_Incertae_Sedis.__Eubacteriaceae_bacterium_DJF_VR85	0.71	-1.72	0.73	-2.91
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae_Incertae_Sedis.__uncultured_Firmicutes_bacterium	0.50	-1.27	0.45	-0.22
Bacteria.__Firmicutes.__Erysipelotrichi.__Erysipelotrichales.__Erysipelotrichaceae_Incertae_Sedis.__uncultured_bacterium	0.82	-0.05	0.49	2.45
Bacteria.__Fusobacteria.__Fusobacteriales.__CFT112H7.__uncultured_bacterium.Other	0.08	-1.85	0.04	-2.18
Bacteria.__Lentisphaerae.__Lentisphaeria.__Victivallales.__Victivallaceae.__Victivallis	0.46	-1.62	0.46	-2.44
Bacteria.__Proteobacteria.__Alphaproteobacteria.__Rhodospirillales.__Rhodospirillaceae.__Thalassospira	0.47	0.19	0.28	2.91
Bacteria.__Proteobacteria.__Betaproteobacteria.__Burkholderiales.__Alcaligenaceae.__Parasutterella	0.67	0.73	0.47	3.87
Bacteria.__Proteobacteria.__Betaproteobacteria.__Burkholderiales.__Alcaligenaceae.__Sutterella	0.60	0.04	0.43	2.79
Bacteria.__Proteobacteria.__Deltaproteobacteria.__Desulfovibrionales.__Desulfovibrionaceae.__Desulfovibrio	0.74	-1.88	0.69	-3.81
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Aeromonadales.__Succinivibrionaceae.__Succinivibrio	0.01	-1.41	0.25	-0.39
Bacteria.__Proteobacteria.__Gammaproteobacteria.__B38.__uncultured_bacterium.Other	0.54	-1.95	0.25	-3.89
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Enterobacteriales.__Enterobacteriaceae.__Enterobacter	0.61	-0.43	0.34	2.47
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Enterobacteriales.__Enterobacteriaceae.__Escherichia.Shigella	0.53	-0.76	0.28	1.13
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Enterobacteriales.__Enterobacteriaceae.__Kluyvera	0.24	-2.04	0.06	-3.55
Bacteria.__Proteobacteria.__Gammaproteobacteria.__Enterobacteriales.__Enterobacteriaceae.__Raoultella	0.67	-0.57	0.23	1.44
Bacteria.__Synergistetes.__Synergistia.__Synergistales.__Synergistaceae.__Cloacibacillus	0.43	-2.19	0.31	-4.82
Bacteria.__Tenericutes.__Mollicutes.__RF9.__uncultured_bacterium.Other	0.08	-2.02	0.10	-3.55
Bacteria.__Verrucomicrobia.__Opitutae.Other.Other.Other	0.63	-1.75	0.39	-3.46
Bacteria.__Verrucomicrobia.__Opitutae.__vadinHA64.__uncultured_bacterium.Other	0.41	-2.23	0.10	-5.09
Bacteria.__Verrucomicrobia.__Verrucomicrobiae.__Verrucomicrobiales.__Verrucomicrobiaceae.__Akkermansia	0.43	2.04	0.43	5.74
Eukaryota.Other.Other.Other.Other.Other	0.58	-1.69	0.51	-1.71

Table 6. ICCs of all 103 genera from the MEC samples using Bray-Curtis PC1.

Genera	No antibiotic use (n=27)	Antibiotic use (n=23)	Total (n=50)
Counts	0.92	0.82	0.87
IS-Normalized Counts	0.92	0.82	0.87

Table 7. ICCs of all 1,272 OTUs from the MEC samples using Bray-Curtis PC1.

OTUs	No antibiotic use (n=27)	Antibiotic use	Total (n=50)
Counts	0.96	0.88	0.90
IS-Normalized Counts	0.96	0.88	0.90

Table 8. Phylum level mean, Standard Deviation, and Coefficient of Variation for all 11 FHC samples and samples by plate.

	Firmicutes	Bacteroidetes	Actinobacteria
All 11 samples			
Abundances	16145 ± 7142 (44)*	12515 ± 6981 (56)	345 ± 208 (60)
IS-Normalized abundances	43265375 ± 19319456 (45)	33571913 ± 18063175 (54)	903585 ± 531460 (59)
R01 plate (n=3)			
Abundances	13637 ± 3699 (27)	9929 ± 3352 (34)	272 ± 104 (38)
IS-Normalized abundances	41448692 ± 20501818 (49)	29207870 ± 13163449 (45)	781045 ± 332705 (43)
Plate 1 (n=3)			
Abundances	16258 ± 6952 (43)	11475 ± 6604 (58)	305 ± 240 (79)
IS-Normalized abundances	41302890 ± 18777740 (45)	29170691 ± 17514894 (60)	780615 ± 631139 (81)
Plate 2 (n=2)			
Abundances	19282 ± 8381 (43)	13947 ± 7864 (56)	471 ± 250 (53)
IS-Normalized abundances	43339409 ± 30920734 (71)	31971887 ± 26146001 (82)	1074429 ± 850940 (79)

Plate 3 (n=3)			
Abundances	16449 ± 11767 (72)	15187 ± 11555 (76)	375 ± 298 (80)
IS-Normalized abundances	46995188 ± 24158492 (51)	43403863 ± 24583581 (57)	1035199 ± 663922 (64)

*Mean ± SD (%CV)

Table 9. Order level mean, Standard Deviation, and Coefficient of Variation for all 11 FHC samples and samples by plate.

	Clostridiales	Bacteroidales	Erysipelotrichales	Verrucomicrobiales	Rhodospirillales	Coriobacteriales	Bifidobacteriaceae	VC2.1Bac22 uncultured bacterium	Burkholderiales	Lactobacillales
All 11 samples										
Abundances	15520 ± 6687 (43)*	12400 ± 6856 (55)	585 ± 440 (75)	483 ± 488 (101)	239 ± 243 (102)	229 ± 137 (60)	115 ± 75.41 (66)	78.09 ± 72.21 (92)	71.18 ± 59.10 (83)	39.45 ± 32.16 (82)
IS-Normalized abundances	41637974 ± 18257471 (44)	33266799 ± 17755019 (53)	1520873 ± 1098872 (72)	1226967 ± 1238035 (101)	604216 ± 609964 (101)	599256 ± 346805 (58)	302572 ± 202944 (67)	206972 ± 183926 (89)	183214 ± 146991 (80)	106529 ± 82401 (77)
R01 plate (n=3)										
Abundances	13183 ± 3418 (26)	9879 ± 3324 (34)	426 ± 270 (63)	362 ± 265 (73)	172 ± 146 (85)	180 ± 82.94 (46)	91.00 ± 39.85 (44)	45.33 ± 24.09 (53)	53.00 ± 40.15 (76)	27.67 ± 11.93 (43)
IS-Normalized abundances	40281199 ± 20141926 (50)	29076430 ± 13123687 (45)	1089733 ± 364118 (33)	865725 ± 303316 (35)	385960 ± 35821 (9)	503154 ± 213847 (43)	275897 ± 191917 (70)	121044 ± 39983 (33)	127101 ± 40268 (32)	77760 ± 31572 (41)
Plate 1 (n=3)										
Abundances	15696 ± 6452 (41)	11399 ± 6541 (57)	530 ± 478 (90)	435 ± 544 (125)	219 ± 264 (121)	206 ± 146 (71)	98.67 ± 98.74 (100)	67.00 ± 55.07 (82)	66.00 ± 71.36 (108)	31.33 ± 22.23 (71)
IS-Normalized abundances	39868608 ± 17476815 (44)	28976003 ± 17350387 (60)	1354418 ± 1250922 (92)	1118927 ± 1412488 (126)	561768 ± 687490 (122)	524172 ± 384073 (73)	256444 ± 256342 (100)	170123 ± 144326 (85)	168239 ± 186238 (111)	79863 ± 58579 (73)
Plate 2 (n=2)										
Abundances	18535 ± 7766 (42)	13869 ± 7800 (56)	707 ± 580 (82)	886 ± 937 (106)	401 ± 467 (116)	296 ± 163 (55)	174 ± 87.68 (50)	70.00 ± 57.98 (83)	96.50 ± 89.80 (93)	40.50 ± 34.65 (86)
IS-Normalized abundances	41559077 ± 29114700 (70)	31784986 ± 25956722 (82)	1683404 ± 1705339 (101)	2181330 ± 2577782 (118)	1002492 ± 1256290 (125)	676263 ± 547075 (81)	395264 ± 304310 (77)	166872 ± 170050 (102)	233460 ± 255124 (109)	96929 ± 100695 (104)
Plate 3 (n=3)										
Abundances	15672 ± 11079 (71)	14942 ± 11308 (76)	718 ± 643 (90)	385 ± 429 (111)	218 ± 248 (114)	258 ± 206 (80)	116 ± 92.22 (80)	127 ± 124 (97)	77.67 ± 74.10 (95)	58.67 ± 55.43 (94)
IS-Normalized abundances	44816713 ± 22562959 (50)	42735838 ± 23987826 (56)	2010112 ± 1476580 (73)	1060007 ± 1043154 (98)	599402 ± 603662 (101)	719105 ± 456832 (64)	313579 ± 209887 (67)	356481 ± 289842 (81)	220804 ± 171693 (78)	168363 ± 127735 (76)

*Mean ± SD (%CV)

Table 10. Genus level mean, Standard Deviation, and Coefficient of Variation for all 11 FHC samples and samples by plate.

	Prevotella	Bacteroides	Faecalibacterium	Lachnospiraceae uncultured	Lachnospiraceae Incertae Sedis uncultured bacterium	Ruminococcaceae uncultured	Coprococcus	Parabacteroides	Alistipes	Ruminococcus
All 11 samples										
Abundances	4969 ± 2797 (56)*	3775 ± 2086 (55)	3165 ± 1463 (46)	2533 ± 987 (39)	2157 ± 913 (42)	1647 ± 849 (52)	1183 ± 407 (34)	1056 ± 521 (49)	992 ± 706 (71)	765 ± 279 (36)
IS-Normalized abundances	13485984 ± 7339301 (54)	10058880 ± 5369864 (53)	8373835 ± 3839540 (46)	6817540 ± 2895009 (42)	5821060 ± 2525459 (43)	4376731 ± 2197656 (50)	3220880 ± 1345320 (42)	2817321 ± 1378830 (49)	2609167 ± 1779234 (68)	2075366 ± 814444 (39)
R01 plate (n=3)										
Abundances	3893 ± 1147 (29)	2971 ± 1051 (35)	2677 ± 904 (34)	2272 ± 484 (21)	1875 ± 474 (25)	1301 ± 437 (34)	1118 ± 156 (14)	919 ± 297 (32)	785 ± 389 (50)	658 ± 116 (18)
IS-Normalized abundances	11681513 ± 5525641 (47)	8680705 ± 3864345 (45)	7906278 ± 3707601 (47)	7099361 ± 3751893 (53)	5734471 ± 2844395 (50)	3841672 ± 1789325 (47)	3599836 ± 1999658 (56)	2721134 ± 1250051 (46)	2135051 ± 756720 (35)	2082030 ± 1110535 (53)
Plate 1 (n=3)										
Abundances	4259 ± 2174 (51)	3663 ± 2165 (59)	3335 ± 1532 (46)	2655 ± 934 (35)	2112 ± 806 (38)	1677 ± 917 (55)	1263 ± 407 (32)	1041 ± 561 (54)	921 ± 807 (88)	758 ± 213 (28)
IS-Normalized abundances	10815127 ± 5791805 (54)	9319861 ± 5740943 (62)	8472168 ± 4111531 (49)	6743584 ± 2569550 (38)	5363303 ± 2197826 (41)	4265018 ± 2438289 (57)	3211601 ± 1142478 (36)	2644979 ± 1491639 (56)	2343465 ± 2112186 (90)	1917748 ± 580415 (30)
Plate 2 (n=2)										
Abundances	4989 ± 2384 (48)	4530 ± 2626 (58)	4046 ± 1826 (45)	3056 ± 1139 (37)	2433 ± 1040 (43)	2144 ± 1141 (53)	1386 ± 434 (31)	1286 ± 701 (55)	1263 ± 1051 (83)	864 ± 252 (29)
IS-Normalized abundances	11288316 ± 8453339 (75)	10409352 ± 8642865 (83)	9117298 ± 6629162 (73)	6802372 ± 4503692 (66)	5461521 ± 3864919 (71)	4891450 ± 3877326 (79)	3057042 ± 1869754 (61)	2938512 ± 2359289 (80)	3012441 ± 3077797 (102)	1899143 ± 1125886 (59)
Plate 3 (n=3)										
Abundances	6742 ± 4798 (71)	4187 ± 3272 (78)	2895 ± 2141 (74)	2322 ± 1629 (70)	2302 ± 1589 (69)	1633 ± 1222 (75)	1031 ± 673 (65)	1054 ± 792 (75)	1091 ± 988 (90)	812 ± 518 (64)
IS-Normalized abundances	19426424 ± 9806740 (50)	11942427 ± 7058563 (59)	8247418 ± 4490805 (54)	6619787 ± 3301247 (50)	6605097 ± 3181667 (48)	4680356 ± 2575522 (55)	2960429 ± 1285852 (43)	3005055 ± 1676049 (56)	3080136 ± 2262648 (73)	2343803 ± 967546 (41)

*Mean ± SD (%CV)