Microbiome studies in veterinary field: communities' diversity measurements pitfalls

Ugo Ala[1], Angela Del Carro[1], Mario Giacobini[1], Barbara Colitti[1], Ada Rota[1], Luigi Bertolotti[1]

[1] *Department of Veterinary Sciences, Università degli Studi di Torino, Largo Paolo Braccini 2, 10095 Grugliasco, Torino, Italy*

ABSTRACT
In recent years, the role of the microbiota has proved to be extremely important in medicine as one of the most important aspects for the characterization of living beings in both healthy and pathological conditions. Moreover, the development of shotgun technology, and in particular the cheaper 16S ribosomal RNA (rRNA) gene sequencing, made possible its wide diffusion. In veterinary sciences, microbiome studies have seen applications not only in medicine in the strict sense (e.g diagnosis) but also, for example, in food inspection (quality, fraud, etc.) and in animal feed preparation itself. However, focusing on microbial profiling by 16S rRNA sequencing, there are several crucial aspects to be considered: from the experimental design definition and the sample size problem to the data analysis steps. This latter involves several layers, e.g. which 16S rRNA databases to use, which metrics for alpha and beta diversity, etc. In this work, we want to present, as a case study, a critical discussion about the large number of alpha and beta diversity metrics and their impact in the statistical comparisons among groups.

Corresponding author: Luigi Bertolotti, e-mail: luigi.bertolotti@unito.it

1. INTRODUCTION

The wide technological availability in the analysis of the microbiome (16S rRNA sequencing, shotgun metagenomics and metatranscriptomics) has allowed a very versatile and widespread application in many fields: from ecology to medicine, both human and veterinary [1]–[3]. In particular, the study and characterization of the microbiome have revealed its crucial role with metabolism and its interconnection with host physiology [4], [5].

Many studies have been proposed, several aimed at the characterization of the microbial environment through the description of which microorganisms are found, several others provided a more complex experimental design: case-control studies, clinical trials or, following the temporary evolution of the data themselves, time-series experiments. Within these inferential studies some specific problems emerge, both in the definition of the experimental design and in the subsequent analysis of the data obtained from the sequencing [6], [7]. The preventive assessment of the sample size is not straightforward: investigation and comparison of microbial composition requires the appropriate definition of the null hypothesis for the statistical test. Indeed, due to the multivariate nature of the microbial data, there are many hypotheses that can be investigated, reflecting the specific scientific question and the chosen experimental design. They can start with a generic request of equality of community composition [8]–[10] and can continue to focus on differential abundance for specific Operational Taxonomic Units (OTU), thus being able to evaluate which OTUs characterize a condition and, more generally, if there are patterns of OTUs that characterize specific conditions.

Moving at the data analysis, several choices have to be made:

I. the sequenced reads association strategy through clustering methods, implemented for example in Mothur with the definition of OTUs [11] compared to the denoising methods implemented in DADA2 with the definition of Amplicon Sequence Variants (ASV) [12],

II. the 16S rRNA databases you can rely on for taxonomic association (Silva [13], Greengenes [14], RDP [15] NCBI [16]),

III. the methods used to identify differentially abundant (DA) microorganisms [17] and

IV. the algorithms used to calculate and to compare alpha and beta diversities [18], [19].

These diversity metrics differ in the basic [20], [21]: some focus more on the number of features, some others more on the evenness with few that considers the phylogenetic relationship among features (Faith's phylogenetic diversity [22] and UniFrac distance [23]) too.

Specifically, in this work we aim to discuss the diversity metrics problem by showing the strong variability of results and the consequent challenge in choosing between the 24 different alpha and 19 beta diversity metrics, as implemented on Qiime2 [24].

2. METHODS

In order to test and to evaluate the possible metrics in the Qiime2 pipeline, bacterial population structure was investigated in dog samples collected during parturitions (see the experimental design scheme at Figure 1).

Briefly, 4 pregnant bitches were included in the study and samples were collected. After birth, the puppies and the mothers were followed and sampled at day 0 (the birth), day 2, and day 30. Samples from vagina and rectum were sampled from the mothers and meconium was collected from the puppies. Environmental controls were sampled during the collection time as well as a laboratory negative control during the molecular protocols. DNA was extracted using the RNeasy Power Microbiome KIT (Qiagen, Hilden, Germany) following the manufacturer's instructions. One microliter of RNaseA (Thermo Fisher Scientific, Waltham, MA, USA) was added to digest RNA, with an incubation of 1 h at 37 °C.DNA was quantified with fluorimetric method Qubit High Sensitive dsDNA kit (Life Technologies, Carlsbad, CA, USA) and standardized at 5 ng/µl. The 16S rRNA gene was amplified following the Illumina 16S Metagenomic Sequencing Library Preparation Protocol (Illumina Inc. San Diego, CA, USA), with minor modifications. Briefly, the V3-V4 region of the 16S gene was amplified with unique barcoded PCR primers containing the Illumina adapter overhang nucleotide sequences. PCR amplicons were cleaned up and size selected using NucleoMag® NGS Clean-up and Size Select (Macherey- Nagel, Allentown, PA, USA). The resulting products were tagged by using the Nextera XT Index Kit (Illumina Inc., San Diego, CA, USA). After the second purification step, amplicon products were quantified using Qubit High Sensitive dsDNA kit (Life Technologies, Carlsbad, CA, USA). Purified and normalized libraries were then pooled and diluted to a 4 nM concentration. The pooled library was then denatured with 0.2 N NaOH, diluted to 10 pM, and combined with 20% (vol/vol) denatured 10 pM PhiX and sequenced with the MiSeq Illumina platform (Illumina Inc., San Diego, CA, USA) with V3-600 cycles chemistry. Sequencing results were analyzed using Trimmomatic

Figure 1. Experimental design scheme.

and Qiime 2-2019.10 standard pipelines. General Qiime2 pipeline was applied, including all the possible metric evaluations for Alpha and Beta diversities. Alpha diversity measures were subdivided into three classes according to their specific focus on abundance or uniformity: Diversity (Richness), Evenness and Phylogenesis.

Alpha diversity metrics, as identified in Qiime2, belonging to Diversity class are thirteen: *ace*, *brillouin_d*, *chao1*, *doubles*, *enspie*, *fisher_alpha*, *gini_index*, *margalef*, *menhinick*, *michaelis_mentin_fit*, *observed_otus*, *shannon* and *singles*. Those belonging to Evenness class are ten: *berger_parker_d*, *dominance*, *heip_e*, *lladser_pe*, *mcintosh_e*, *pielou_e*, *robbins*, *simpson*, *simpson_e* and *strong*. The one that considers the phylogenesis is the *Faith_pd*. Beta diversity metrics were subdivided into two classes according to their use of phylogenesis. Beta diversity metrics, as identified in Qiime2, belonging to non-phylogenetic class are sixteen: *braycurtis*, *cityblock*, *correlation*, *cosine*, *dice*, *euclidean*, *hamming*, *jaccard*, *kulsinski*, *matching*, *rogerstanimoto*, *russelrao*, *seuclidean*, *sokalmichener*, *sqeuclidean* and *yule*. Those belonging to the phylogenetic class are three: *unweighted_unifrac*, *weighted_unifrac* and *generalized_unifrac*. In addition to the beta diversity found in Qiime2, we also tested another beta diversity metric (the Robust Aitchison PCA capable of linking specific features to beta diversity sorting [25], available as a Qiime2 plugin at https://github.com/biocore/DEICODE).

Pearson correlation coefficient has been calculated for all possible pairs of alpha diversity metric. Samples were grouped according to

i) the sample features (biological samples vs environmental control (white) samples),
ii) the collection time (day 0, day 2 and day 30) and
iii) the tissue origin (vaginal, rectal, meconium, and control).

The Wilcoxon test and Kruskal-Wallis test were used to assess possible statistical significant differences among groups according to the alpha metrics. Bonferroni correction was applied to nominal p-values. Beta diversities significances were evaluated using the Permanova test, followed by False Discovery Rate estimation as implemented in Qiime2. All the statistical analyses were conducted with R statistical software [26].

3. RESULTS

The alpha diversity values have been calculated for all samples according to the 24 different algorithms present in Qiime2 suite and the correlations among the different algorithms have been calculated, as reported in Figure 2.

Several clusters of highly correlated (both positively or negatively) metrics are highlighted together with some metrics that seem to be poorly associated with the others, like *Lladser* point estimate of the unsampled taxa or *Robbins* estimator of the probability of unobserved outcomes.

Successively, these alpha diversity measures have been used to compare the different experimental groups as defined in the Methods section. In particular, three main comparisons have been analyzed by grouping samples according to: environmental control versus biological samples; the collection time or the tissue of origin. These three comparisons were used to evaluate the ability to highlight statistical significance of different metrics, as reported in Table 1.

In particular, by analysing the nominal and the corrected p-values of metrics subdivided into the three classes, it emerges that the class of metrics based on Diversity is able to highlight more statistical differences among the experimental groups. In the first comparison, between white and biological samples, the
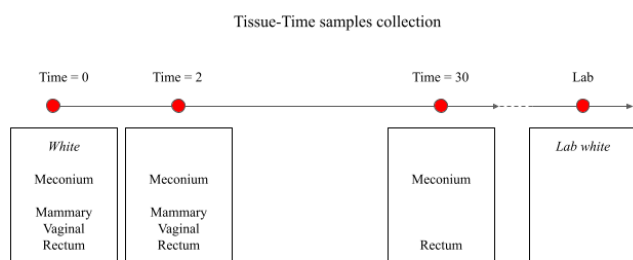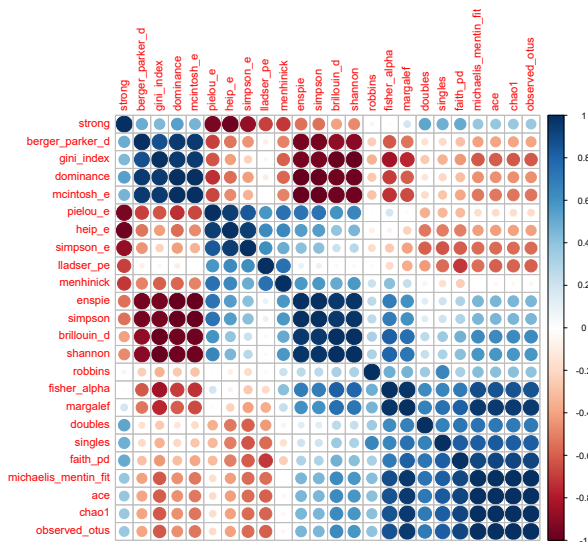
Figure 2. Correlation plot based on the 24 analyzed alpha diversity metrics. For each pairwise correlation, the Pearson coefficient is based on the alpha diversity metrics values of all samples.

difference among samples is hardly detectable (46% with the Diversity metrics versus 10% with the Evenness metrics, but 0% for both when Bonferroni correction is applied), highlighting the presence of possible false positives in the comparisons if nominal p-values alone were considered. Focusing on Time and Tissue based groups, the Diversity metrics show a range between 69% to 100% of statistical comparisons whereas the Evenness metrics range from 0% to 90%, generally showing a lower percentage of significant tests and a much greater range of variability. The Diversity metrics confirm in 69% of cases the significance both with the nominal and the corrected p-value for both Time and Tissue comparisons, the Evenness metrics instead show a 50% of cases that lose significance when p-value correction is applied. According to our data, the Faith Phylogenetic Distance is robust as respect to the p-value adjustment.

Time and Tissue groups' comparisons based on Beta diversity metrics instead show a more convergent behaviour, as reported in Table 2: the significant PERMANOVA test percentages, both according to the empirical nominal p-value and the False Discovery Rate (FDR) q-value, show a high value of significant

Table 1. Number of alpha diversity metrics significant comparisons according to environmental control (White), Time and Tissue analyses. In brackets, the percentages of significant metrics are reported. The metrics are considered all together and subdivided into classes based on their specific aim of measurement and their distinct focus on abundance or uniformity: Diversity (Richness), Evenness and Phylogenesis. Wilc: Wilcoxon rank-sum test; Bonf: Bonferroni correction; K-W: Kruskal-Wallis test.

Alpha Diversity Metrics Classes	White comparison		Tissue comparison		Time comparison	
	Wilc	Bonf	K-W	Bonf	K-W	Bonf
All metrics	7 (29.2%)	0 (0%)	17 (70.8%)	11 (45.8%)	23 (95.8%)	15 (62.5%)
Diversity metrics	6 (46.2%)	0 (0%)	12 (92,3%)	10 (76.9%)	13 (100%)	9 (69,2%)
Evenness metrics	1 (10%)	0 (0%)	4 (40%)	0 (0%)	9 (90%)	5 (50%)
Phylogenetic metrics	0 (0%)	0 (0%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)

Table 2. Number of beta diversity metrics significant comparisons according to Time and Tissue analyses. In brackets the percentages of significant metrics are reported. The metrics are considered all together and subdivided into classes based on their specific strategy of measurement based, or not based, on phylogenetic evolution.

Beta Diversity Metrics Classes	Time comparison		Tissue comparison	
	PERMANOVA	Bonferroni	PERMANOVA	Bonferroni
All metrics	19 (100%)	19 (100%)	17 (89.47%)	16 (84.21%)
Non-phylogenetic metrics	16 (100%)	16 (100%)	14 (87,5%)	13 (81.25%)
Phylogenetic metrics	3 (100%)	3 (100%)	3 (100%)	3 (100%)

tests and are well comparable between the two phylogenetically and non-phylogenetically based metrics. By analysing the pseudo-F statistics of the pairwise comparisons, as in Figure 3, we observed that when the differences between the experimental groups are small, the beta diversity metrics that are able to highlight these differences can vary between 16% of those analysed (three out of nineteen) to 84% (sixteen out of nineteen), but without a real preference for measures based on phylogeny.

The added Robust Aitchison PCA beta diversity metric is in line with other measures of beta diversity. Specifically, it highlights significant differences for both the Time and the Tissue comparisons ($pValue < 0.001$ with $pseudo\text{-}F = 12.49$ and $pValue < 0.001$ with $pseudo\text{-}F = 7.66$, respetively). The specific Time pairwise comparisons put in evidence an overall significance for all comparisons except for samples collected at time=0 and the white samples ($qValue = 0.13$ with $pseudo\text{-}F = 1.81$). Tissue pairwise comparisons show instead non significances in the following cases: mammary vs meconium ($qValue = 0.39$ with $pseudo\text{-}F = 1.07$), mammary vs vaginal ($qValue = 0.87$ with $pseudo\text{-}F = 0.15$), meconium vs vaginal ($qValue = 0.33$ with $pseudo\text{-}F = 1.27$) and vaginal vs blank ($qValue = 0.15$ with $pseudo\text{-}F = 2.27$).
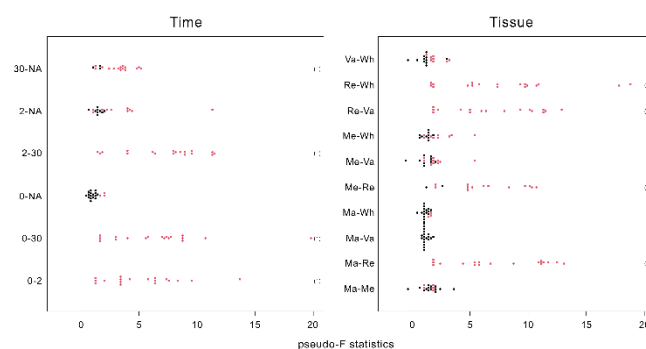

Figure 3. The PERMANOVA based pseudo-F statistics are plotted. The study of pairwise comparisons deriving from the 19 beta diversity metrics analysis is considered: pairwise comparisons according to time evolution are reported in left panel, whereas for tissues comparison in right panel. On the x-axis the pseudo-F values, as from the PERMANOVA analysis, are considered; on the y-axis the pairwise comparison they are related to , in the left panel: 0, 2 and 30 are the days of collection and NA are the white samples, whereas in the right panel: Me - Meconium, Ma - Mammary, Re - Rectal, Va - Vaginal, and Wh - White. When pseudo-F values exceed the value of 20, the number of beta diversity metrics is reported. Black dots are not significant according to the q- value cut-off of 0.05, red dots are those significant.

4. DISCUSSION

The microbiome studies, in all their forms, have reached a huge number of items in public reference databases. Even if the largest part of them is very well structured and developed, in several cases the choice of metrics remains a crucial point. Especially in the case of complex studies, the real biological differences (if any) are hard to be highlighted. Our case study shows how when biological differences are not so strong, metrics can provide different statistical outcomes. This could be due by different reasons both biological (i.e. kind of samples, lab procedures...) and statistical (i.e. experimental design, sample size, statistical power...). In order to explore this scenario, we applied 24 alpha and 19 beta diversity metrics on the same sample set. The final results provide a quite large variability in the statistical significance, making the results interpretation strictly dependent from the specific metric used. Our study generally suggests, as expected, that when the biological differences among the investigated groups are strong and clear, each single diversity metric is able to statistically confirm this difference. On the other hand, the use of a large number of metrics, both correlated or not (as in the alpha diversity metrics case), could be a key to avoid misinterpretation of data. In particular, by choosing to consider together those metrics that show correlation (critically discussing them without selecting only those with a significant result), evidences of similarities of differences among conditions would be strengthen (by avoiding the p-hacking problem [20]). Whereas, by considering no correlated metrics, it is possible to explore different aspects of richness and evenness that otherwise would be neglected.

REFERENCES

[1] F. H. Coutinho, G. B. Gregoracci, J. M. Walter, C. C. Thompson, F. L. Thompson, Metagenomics sheds light on the ecology of marine microbes and their viruses, Trends Microbiol. 26 (2018) pp. 955–965.
DOI: 10.1016/j.tim.2018.05.015

[2] L. Curini, B. Alushi, M. R. Christopher, S. Baldi, L. Di Gloria, P. Stefano, A. Laganà, L. Iannone, (+ 6 more authors), The first taxonomic and functional characterization of human CAVD-associated microbiota, Microb. Cell Graz Austria 10 (2023), pp. 36–48.
DOI: 10.15698/mic2023.02.791

[3] A. Talat, K. S. Blake, G. Dantas, A. U. Khan, Metagenomic insight into microbiome and antibiotic resistance genes of high clinical concern in urban and rural hospital wastewater of Northern India Origin: a Major Reservoir of Antimicrobial Resistance, Microbiol. Spectr. 11 (2023), art no. e0410222.
DOI: 10.1128/spectrum.04102-22

[4] M. H. Mohajeri, R. J M Brummer, R. A Rastall, R. K Weersma, H. J. M. Harmsen, M. Faas, M. Eggersdorfer, The role of the microbiome for human health: from basic science to clinical applications, Eur. J. Nutr. 57 (2018), pp. 1–14.
DOI: 10.1007/s00394-018-1703-4

[5] H. Tilg, N. Zmora, T. E. Adolph, E. Elinav, The intestinal microbiota fuelling metabolic inflammation, Nat. Rev. Immunol. 20 (2020), pp. 40–54.
DOI: 10.1038/s41577-019-0198-4

[6] B. J. Kelly, R. Gross, K. Bittinger, S. Sherrill-Mix, J. Lewis, R. Collman, F. Bushman, Hongzhe Li, Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA, Bioinforma. Oxf. Engl. 31 (2015), pp. 2461–2468.
DOI: 10.1093/bioinformatics/btv183

[7] F. Mattiello, B. Verbist, K. Faust, J. Raes, W. D. Shannon, L. Bijnens, O. Thas, A web application for sample size and power calculation in case-control microbiome studies, Bioinforma. Oxf. Engl. 32 (2016), pp. 2038–2040.
DOI: 10.1093/bioinformatics/btw099

[8] K. J. Koehler, J. R. Wilson, Chi–square tests for comparing vectors of proportions for several cluster samples, Commun. Stat. - Theory Methods 15 (1986), pp. 2977–2990.
DOI: 10.1080/03610928608829290

[9] By IMPACTT investigators, Beta-diversity distance matrices for microbiome sample size and power calculations - How to obtain good estimates, Comput. Struct. Biotechnol. J. 20 (2022), pp. 2259–2267.
DOI: 10.1016/j.csbj.2022.04.032

[10] T. Ferdous, L. Jiang, I. Dinu, J. Groizeleau, A. L Kozyrskyj, C. M. T. Greenwood, M.-C. Arrieta, The rise to power of the microbiome: power and sample size calculation for microbiome studies, Mucosal Immunol. 15 (2022), pp. 1060–1070.
DOI: 10.1038/s41385-022-00548-1

[11] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A Lesniewski, B. B Oakley, (+7 more authors), Introducing mothur: Open- source, platform-independent, community-supported software for describing and comparing microbial communities, Appl. Environ. Microbiol. 75 (2009), pp. 7537–7541.
DOI: 10.1128/AEM.01541-09

[12] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, S. P. Holmes, DADA2: High- resolution sample inference from Illumina amplicon data, Nat. Methods 13 (2016) pp. 581–583.
DOI: 10.1038/nmeth.3869

[13] P. Yilmaz, L. Wegener Parfrey, P. Yarza, J. Gerken, E. Pruesse, C. Quast, T. Schweer, J. Peplieset, (+ 2 more authors), The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks', Nucleic Acids Res. 42(D1) (2014), pp. D643–D648.
DOI: 10.1093/nar/gkt1209

[14] D. McDonald, M. N. Price, J. Goodrich, E. P. Nawrocki, T. Z. DeSantis, A. Probst, G. L. Andersen, R. Knight, (+ 1 more author), An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea, ISME J. 6 (2012), pp. 610–618.
DOI: 10.1038/ismej.2011.139

[15] Q. Wang, G. M. Garrity, J. M. Tiedje, J. R. Cole, Naïve bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy, Appl. Environ. Microbiol. 73 (2007), pp. 5261–5267.
DOI: 10.1128/AEM.00062-07

[16] S. Federhen, The NCBI taxonomy database, Nucleic Acids Res. 40 (2012), pp. D136–D143.
DOI: 10.1093/nar/gkr1178

[17] J. T. Nearing, G. M. Douglas, M. G. Hayes, J. MacDonald, D. K. Desai, N. Allward, C. M. A. Jones, R. J. Wright, A. S. Dhanani, A. M. Comeau, M. G. I. Langille, Microbiome differential abundance methods produce different results across 38 datasets, Nat. Commun. 13 (2022) art. no. 342.
DOI: 10.1038/s41467-022-28034-z

[18] J. Fukuyama, P. J. McMurdie, L. Dethlefsen, D. A. Relman, S. Holmes, Comparisons of distance methods for combining covariates and abundances in microbiome studies, Pac. Symp. Biocomput. Pac. Symp. Biocomput. (2012), pp. 213–224.
DOI: 10.1142/9789814366496_0021

[19] A. D. Willis, Rarefaction, alpha diversity, and statistics, Front. Microbiol. 10 (2019), pp 2407.
DOI: 10.3389/fmicb.2019.02407

[20] J. G. Kers, E. Saccenti, The power of microbiome studies: Some considerations on which alpha and beta metrics to use and how to report results, Front. Microbiol. 12 (2021), art. no. 796025.
DOI: 10.3389/fmicb.2021.796025

[21] S. L. Hagerty, K. E. Hutchison, C. A. Lowry, A. D. Bryan, An empirically derived method for measuring human gut microbiome alpha diversity: Demonstrated utility in predicting health-related

outcomes among a human clinical sample, PloS One 15 (2020) art. no. e0229204.
DOI: 10.1371/journal.pone.0229204

[22] D. P. Faith, The role of the phylogenetic diversity measure, PD, in bio-informatics: getting the definition right, Evol. Bioinforma. Online 2 (2007) pp. 277–283.
DOI: 10.1177/117693430600200008

[23] C. Lozupone, M. E. Lladser, D. Knights, J. Stombaugh, R. Knight, UniFrac: an effective distance metric for microbial community comparison, ISME J. 5 (2011), pp. 169–172.
DOI: 10.1038/ismej.2010.133

[24] E. Bolyen, J. Ram Rideout, M. R. Dillon, N. A. Bokulich, C. C. Abnet, G. A. Al-Ghalith, H. Alexander, E. J. Alm, (+ 104 more authors), Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2, Nat. Biotechnol. 37 (2019), pp. 852–857.
DOI: 10.1038/s41587-019-0209-9

[25] C. Martino, J. T. Morton, C. A. Marotz, L. R. Thompson, A. Tripathi, R. Knight, K. Zengler, A novel sparse compositional technique reveals microbial perturbations, mSystems 4 (2019), art. no. e00016-19.
DOI: 10.1128/mSystems.00016-19

[26] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2020. Online [Accessed 09 March 2024].
https://www.R-project.org/