Quantifying Stability of Common Distance Metrics and Similarity Scores

Meghana Tandon

Introduction

Nearly every metagenomic study involving samples of amplified reads has the same first step – sample reduction. Sample reduction is the process of using raw reads to produce operational taxonomic units (OTUs) and an associated table with abundance counts. This information is valuable to researchers because it allows them to deduce which species are present within each sample and assess community diversity across all the samples, also known as beta diversity.

It is difficult to identify species among prokaryotes because there is no means of differentiation by sexual reproduction and offspring viability. Thus, current approaches take the route of a phylo-phenotypic definition in which species are assigned based on phylogeny and primary function and characteristics. Phylogeny based species differentiation is done via sequencing specific regions of marker genes such as the V4 region of the 16S marker gene found in prokaryotic rRNA.

These reads offer important insight into how abundant a certain microbial species is in a particular location. However, they possess little meaning until organized into OTU tables. The process of selecting representative strings for each new species encountered and consolidating remaining reads into pre-existing clusters is dependent on a similarity score, a threshold which determines how similar two reads must be in order to be considered the same species and placed in the same cluster together. Similarity scores of 97% or 98.5% are very commonly seen and yet it is unclear why the default cut off is one of the aforementioned quantities. They have become the status quo among such studies but there lacks rigorous backing to justify why these thresholds are appropriate.

Aside from clustering methods, there are also denoising algorithms that produce ASVs or amplicon sequence variants. ASVs provide finer resolution and are deemed the equivalent of enforcing a 100% similarity score, although this is misleading since ASV tables will not be the same as an OTU table constructed from 100% similarity threshold. There is currently a shift towards using ASVs as they are more scalable but it is unclear how they perform in relation to OTUs.

In the current debate between ASVs and OTUs, there are many factors to consider. The biggest issue with lower similarity scores is that they may incorrectly cluster unique sequences and estimate lower diversity than intended. Meanwhile, although ASVs are produced by algorithms designed to identify differences between machine error and legitimate sequences that vary by one nucleotide or are in low abundance within the sample, they are not perfect and may overestimate diversity. This also threatens the reliability of higher similarity scores for OTU construction because more stringent requirements for clustering will lead to more clusters with low abundance counts. OTUs are dataset specific but easily present larger biological trends while ASVs allow for seamless cross-study analyses.

To perform analyses on microbiome data post clustering or denoising, beta diversity metrics are used to indicate biodiversity across samples, or within a community. The beta diversity metrics included in this project are abundance and set-based indices such as Bray-Curtis, Jaccard, and Morisita-Horn, as well as phylogenetic tree-based metrics such as weighted and unweighted UniFrac, and finally weighted Jaccard which does not fit into either category. Well-established metrics like Bray-Curtis and UniFrac are justifiably popular among researchers, it is unclear to that degree these measures are interchangeable.

Significance

Throughout the initial stages of microbiome analysis, every choice made has a chance of introducing error or bias. It is thus crucial that we do not compound upon this with arbitrarily chosen parameters if in fact these decisions have a significant downstream effect. With this project, the goal is to assess the robustness of commonly utilized similarity scores and distance metrics to identify the relatively least volatile parameters when constructing and comparing OTUs.

The significance of this project lies in the fact that the biosphere has become increasingly relevant in several domains, including agriculture, human health, and manufacturing. In order to accurately study the diversity of bacteria in the guts of aging humans, rivers across seasons, and in compost throughout decomposition, it is vital that the metrics used to assess said diversity are confirmed to be reliable. If similarity scores and distance metrics do indeed impact downstream results or restrict cross study analyses, it is vital that researchers take more care in selecting measures and become aware of assumptions that are and are not permitted

Hypothesis

Based on prior research conducted in this area, it is expected that the ASV tables will be more stable than the OTU tables constructed at 100% similarity (written as 100% OTUs from here on, for the sake of brevity). Stability will be determined by the Spearman correlation between pairwise distance matrices such that for the same distance metric, when both ASVs and 100% OTUs are plotted against lower similarity score tables, the ASVs will have a stronger correlation.

In terms of distance metrics, UniFrac is expected to be unstable compared to the abundance-based indices due to the fact that UniFrac only uses an evolutionary tree constructed from the OTU sequences themselves. Weighted UniFrac is expected to be stable and considered highly interchangeable because it takes both abundance counts and the phylogenetic tree into account. The definition of interchangeable here is not that one metric is a near identical replacement for another. The way it is used in this project is as following– if one diversity metric is replaced by another interchangeable metric, the relative diversity of the entire community will be the same, though often shifted significantly up or down in overall reporting of diversity.

Lastly, weighted Jaccard is expected to be incompatible with every other distance metric, particularly because it is computed by comparing every index of abundance vectors for a pair of samples and dividing the sum of the minimum abundance counts over the sum of the maximum abundance counts. It is rarely used in beta diversity analysis so it can serve as a bound on the worst correlation and treated as the most unstable.

Methodology

Though there are a wide variety of algorithms for both clustering and denoising, the two utilized in this project are VSEARCH, an open source equivalent to USEARCH, and DADA2. The data set analyzed consists of river samples collected through the Pre-College Computational Biology program at CMU. The reads were sequenced as single end reads using Illumina MiSeq technology.

The standard pre-processing steps of demultiplexing, quality filtering, chimera removal, and dereplicating were performed on the raw reads. Read merging was not needed since the data did not include paired end reads. Taxonomy assignment was also not implemented.

The pre-processing for both scripts was done exactly the same way. After taking a look at the quality plots of the reads, a truncated length of 240 base pairs was selected as the average location where a severe drop in read quality was observed across all samples. The maximum number of expected errors allowed in each read was set to 1 (this value greatly helped eliminate extraneous species counts in VSEARCH that produced very sparse OTU tables). The truncQ parameter was assigned a value of 11, meaning that reads would be truncated at the first instance of a quality score lower than 11. These parameters are stringent and may not be needed given a very high-quality data set.

For VSEARCH, there is an additional step of choosing which clustering method to employ. Prior research suggests that reference-based assignment underperforms and introduces bias depending on which database is utilized, with some common ones being Greengenes, RDP, and SILVA. Thus, this project applies de novo clustering, while acknowledging that this method does not necessarily extend well for cross study analyses and taxonomy assignment. However, since taxonomy assignment is not relevant to the goal of this project, this is not a concern for the time being.

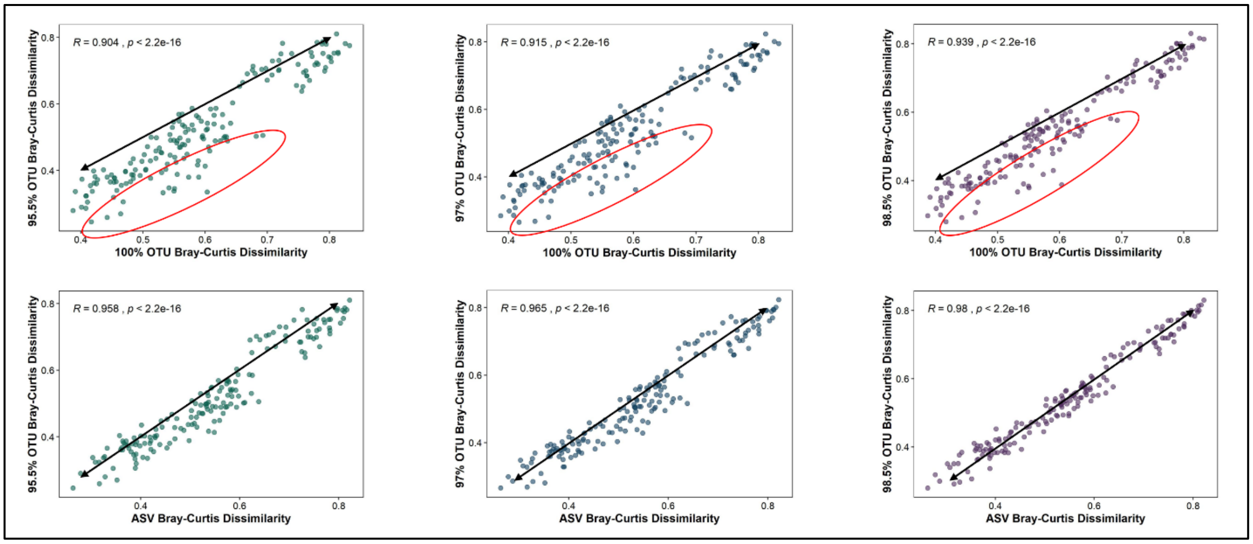
Once abundancy tables were successfully constructed for ASVs and OTUs at similarity scores of 95.5%, 97%, 98.5%, and 100%, R scripts with the vegdist, phangorn, DECIPHER, and phyloseq packages were used to compute pairwise distance matrices and further analysis regarding single-count variants (see Results section). For UniFrac and weighted Unifrac, multiple alignment of the OTUs as well as neighbor joining tree construction were used to construct the phylogenetic tree that is used in compute the distance matrices.

The Spearman correlation coefficient was utilized in place of Pearson's to prioritize general positive monotonic relationships regardless of whether they are linear relationships. This metric was used to assess stability as well, as seen in the next section.

Results and Discussion

Part I: Similarity Scores

The first part of the analysis involved comparing similarity scores while keeping the distance metric unchanged. To produce initial plots, the Bray-Curtis dissimilarity index was chosen due to its popularity but the drawbacks of the metric will be discussed and a more generalized takeaway will be offered.



The plots behaved as expected. As the similarity scores along the y-axis approach the similarity scores on the x-axis (assume that ASVs are produced at 100% similarity via an alternate algorithm), the spread reduces and Spearman's correlation coefficient increases. Note that the ASVs are more compatible with OTUs, as we see a stronger one-to-one correspondence. Meanwhile, the pairwise sample diversity reported from 100% OTUs is consistently higher than the diversity reported from any lower percentage OTUs, as indicated by a bulk of the points sitting below the respective lines in the first row of graphs. In fact, there are some outliers in which the pairwise sample diversity expressed through 100% OTUs is significantly higher (.15+ difference) than the diversity reported by the lower percentage OTUs. These are roughly found within the red ovals on the graphs.

To understand why this is the case, the distribution of frequencies in the OTU tables were analyzed, specifically between stable and unstable pairwise samples. Stable points are those which remain within a certain threshold of the direct correspondence line for all the similarity plots. The outliers and stable pairs of samples are listed in the table below. Note that assessing stability was tough and restricted this computation to three pairs each. See Challenges section for more information.

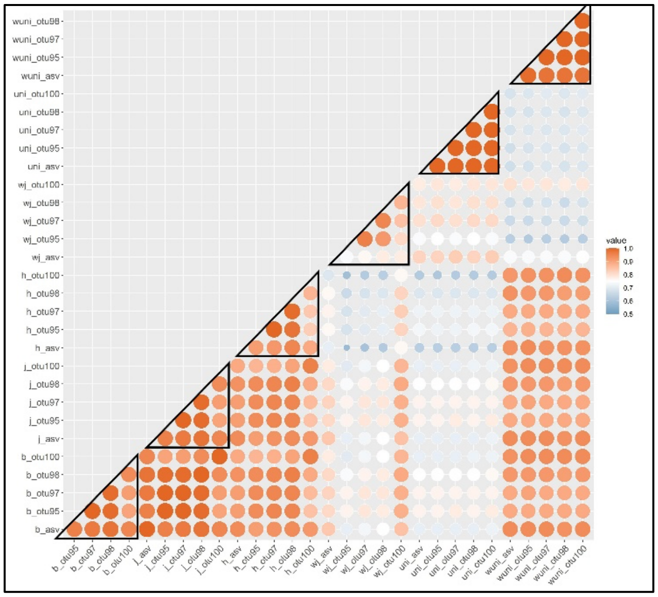
Stable (>.15 difference in diversity reporting across all graphs)	Unstable (<.02 difference in diversity reporting across all graphs)
S9/S12	C2/S12
S9/S14	S9/C2
S8/S14	C2/S13

Next, the number of rows in which one sample had only one count of the OTU while the other sample had zero counts were computed for each pair and averaged. The overall frequency of these "single-count variants" are shown below for each similarity score.

	100%	**98.5%**	**97%**	**95.5%**
Stable	318	190	154	136
Unstable	390	205	162	150

The huge difference in the number of single-count variants present in the OTU tables displays why diversity is overestimates at a 100% similarity score. As the similarity scores are reduced, allowing for more clustering and thus fewer overall OTUs, these extraneous reads do not contribute as much to the diversity computation. Thus, the count distribution in the OTU tables for the unstable pairs of samples is such that as the similarity score drops, the reads that were identified as unique species at 100% are repurposed so that they join pre-existing clusters and reduce net differences in the numerator of the Bray-Curtis computation. These results directly translate to the Jaccard dissimilarity index as well due to the similarity in how both measures are computed.

Thus, the 100% similarity score has a tendency to overreport diversity for pairs of samples that happen to have many reads that differ by just a couple bases (most likely due to machine error) *and* just so happen to align in a manner that exacerbates net differences in the numerator of abundance based diversity metrics. Do note that the isolated samples within unstable pairs did not necessarily have significantly more low abundance counts in comparison to stable samples. It was only when they were paired together that distance metrics such as Bray-Curtis, Jaccard, and Morisita-Horn fell short. Luckily, weighted UniFrac offsets this because it does not depend solely on the abundance counts. Rather, it also utilizes a phylogenetic tree to compute distance. The regions bound by the triangles in the Spearman Correlation Heat map indicate that the UniFrac metrics, weighted and unweighted, do not at all display variance based on similarity score. This is because they more heavily depend on presence/absence data will not change based on similarity score.



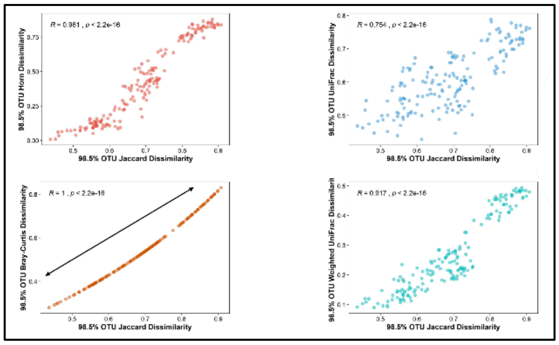
Spearman Correlation Heat Map

Not that, as hypothesized, ASVs are more stable than 100% OTUs in relation to lower similarity OTUs, with the highest correlation consistently going to the pairing of ASVs with 98.5% OTUs. The fact that ASVs and 100% OTUs experience a drop in correlation can also be attributed to the finding that 100% OTUs overestimate diversity. Thus, ASVs can be better viewed as equivalents to 98.5% OTUs, in terms of stability.

The exception to this trend is weighted Jaccard which does significantly better at lower similarity scores and performs terribly when ASVs are used. This is because the ASV table produces had fewer clusters than any of the OTU tables and was thus relatively much denser, which will make the diversity reported by weighted Jaccard more erratic since differences are not a matter of single count invariants.

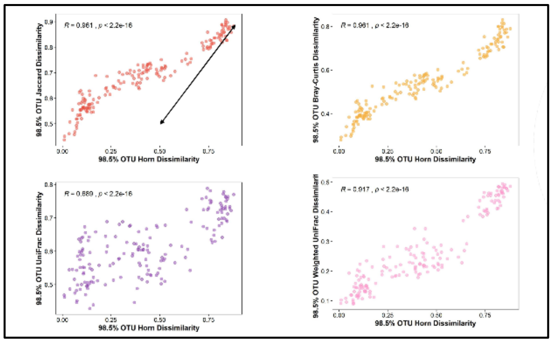
Part II: Distance Metrics

The second part of the analysis involves determining how "robust" or relatively "interchangeable" these distance metrics are among each other.



Correlations between Jaccard Dissimilarity Index and Horn, Bray-Curtis, UniFrac, and weighted UniFrac

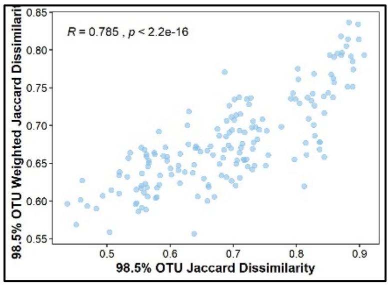
Jaccard and Bray-Curtis display a strong correlation due to how similarly the two indices are computed. Meanwhile, Morisita-Horn and Jaccard have a distribution that fits a sigmoid-function. Morisita-Horn is particularly sensitive at the tail ends of these distributions and tends to be more conservative in reporting diversity at the lower and upper bounds. This is seen in the bottom four plots on the right.

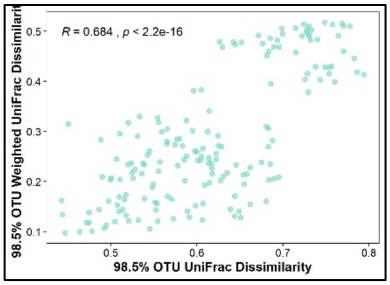


Correlations between Morisita-Horn Dissimilarity Index and Jaccard, Bray-Curtis, UniFrac, and weighted UniFrac

Though it is clear that these metrics are not interchangeable because they do not have a one-to-one correspondence, they can provide the same *relative* pairwise distances if they have a high Spearman correlation.

The heat map can be revisited for further analysis, this time with regard to the rectangles outside of the triangular regions. It is clear that weighted Jaccard and UniFrac are poor alternatives to set-based dissimilarity indices. Additionally, the weighted versions of Jaccard and UniFrac have very poor correlation with their unweighted dissimilarity indices.





Note that the correlation heat map cannot be taken at face value. When describing how relatively "interchangeable" these distance metrics are, merely taking the Spearman correlation into account is not sufficient, especially since this correlation coefficient is more forgiving with outliers in the tails of our data. Out of the distance metrics considered for this project, this consideration is necessitated by the Morisita-Horn index which consistently underreports at the tails, a fact that would be overlooked if just observing the heat map. Thus, if analyzing beta diversity in samples that will include samples that are either highly diverse or nearly identical, it would be more appropriate to select weighted UniFrac, Bray-Curtis, or Jaccard.

Challenges and Important Notes

Scope

It is important to note that the scope of this project did not involve much exploration into pre-processing parameters. However, it must be acknowledged that factors such as min length, max length, trimming options and other quality filtering metrics that decide which reads to throw out all have a significant role in the construction of OTUs and ASVs, and will need to be modified based on the data set. The effects of this were directly observed in this project when the suggested quality filtering parameters from one pipeline caused VSEARCH to produce extremely sparse abundance tables in which all pairwise sample distance measures indicated maximum diversity. Once these parameters were modified based on a separate pipeline, workable tables were attained, and then the pre-processing steps for the DADA2 script were updated to mitigate differences in pre-processing that could potentially introduce unnecessary variables.

Additionally, the algorithms chosen for the specific tasks of denoising and clustering can modify results. An example of this is the treatment of erroneous sequences between DADA2 and Deblur, in which the former alters these sequences to match an ASV they likely originated from, thus counting valid strings that would have been discarded by Deblur due to machine error.

Lastly, this project was conducted only on one data set (for reasons listed below) which is very incomplete. Thus, any observations and conclusions made from them in this analysis cannot be extrapolated to other data sets or processes until a more complete analysis is achieved.

Parallelization

Setting up a highly parallelizable project was not possible at this point in time. Because VSEARCH is currently not a package that is built into R, the VSEARCH algorithm was used in a Windows batch script and then the OTU tables were loaded into R for beta diversity analysis. In the future, a viable alternative would be to use QIIME2 which has both algorithms built in. However, because QIIME2 requires the input sample data to obey very specific naming conventions that require breaking apart merged FASTQ files into each lane and annotating each sample with sequence barcodes, lane numbers, set numbers, and more, an additional script would have to be written to ensure that this is done. Additionally, separate scripts would need to be written for single end and paired end reads and the data taken from studies would need to be categorized accordingly.

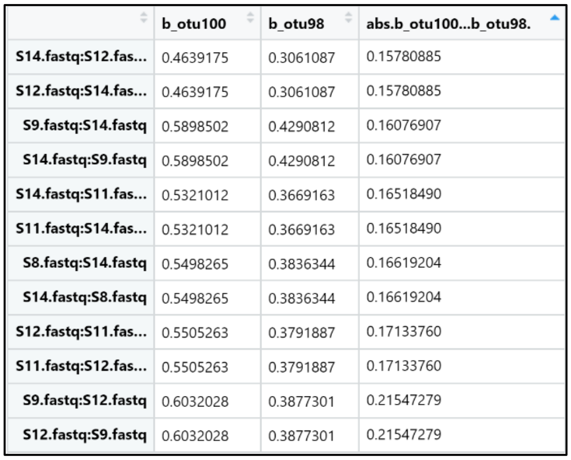
Availability of Data

In many public metagenomic databases, it is often the case that metadata or pre-constructed OTU tables with assigned taxonomy are readily available to perform downstream beta analysis. However, since the focus of this project was on the upstream construction of such tables, this information was not usable. In order to access the raw FASTQ files, further credentials were necessary. Other metagenomic databases contained fully sequenced genomes, which once again was not relevant to this project. There is one database called MG-RAST that is a crowd sourced repository that also contains Illumina MiSeq single-end reads, though only partial sets are available to download. Alternatively, mock data sets can also be generated using CAMISIM and specialized to mimic specific sequencing technologies as well. These resources are worth exploring further to expand the scope of this project.

Quantifying Stability

When assessing outliers in the similarity score plots of 98.5%, 97%, and 95.5% against 100%, the most disparate Bray-Curtis and Jaccard dissimilarity indices consistently belonged to the same six pairs of samples, always found in the same order as well.

Meanwhile, the "stable" samples, or those pairs that had as close to a one-to-one correspondence as possible did not have the same consistency across similarity scores. In fact, there was little to no pattern observed. Thus, selecting representative stable samples was tough. They were ultimately found by randomly choosing three pairs of samples that consistently fell beneath an arbitrary threshold for all plots.



	b_otu100	b_otu98	abs.b_otu100...b_otu98.
S14.fastq:S12.fas...	0.4639175	0.3061087	0.15780885
S12.fastq:S14.fas...	0.4639175	0.3061087	0.15780885
S9.fastq:S14.fastq	0.5898502	0.4290812	0.16076907
S14.fastq:S9.fastq	0.5898502	0.4290812	0.16076907
S14.fastq:S11.fas...	0.5321012	0.3669163	0.16518490
S11.fastq:S14.fas...	0.5321012	0.3669163	0.16518490
S8.fastq:S14.fastq	0.5498265	0.3836344	0.16619204
S14.fastq:S8.fastq	0.5498265	0.3836344	0.16619204
S12.fastq:S11.fas...	0.5505263	0.3791887	0.17133760
S11.fastq:S12.fas...	0.5505263	0.3791887	0.17133760
S9.fastq:S12.fastq	0.6032028	0.3877301	0.21547279
S12.fastq:S9.fastq	0.6032028	0.3877301	0.21547279

Example of outliers in Bray-Curtis plot of 98.5% OTUs plotted against 100% OTUs. The six outliers consistently occur in this order for all Bray-Curtis and Jaccard similarity score plots in which the x-axis is 100% similarity.

In the future, it is worth reassessing a more concrete definition of stability because this dataset heavily favored sample C2 paired with other samples, as these pairs were often found under the threshold selected for this project and deemed stable.

Conclusion

The most significant finding from this analysis is that utilizing a 100% similarity score for clustering algorithms is unadvisable due to overreporting of diversity for specific pairs of samples. Instead, for abundance and set-based distance metrics, if higher resolution is warranted, ASVs should be used and if OTUs are preferred, a 98.5% threshold is best as it provides enough resolution without overestimating diversity. Similarity scores play little to no role in UniFrac dissimilarity which is favorable to researchers who do not wish to deal with the uncertainly of choosing between ASVs or OTUs. However, this freedom comes at the loss of abundance counts. Weighted UniFrac achieves the advantages of both methods by being resistant to similarity scores while taking abundance counts into account. Thus, weighted UniFrac can be considered the most flexible or "interchangeable" metric except in relation to weighted Jaccard (which is acceptable as it has been established as a poor metric) and unweighted UniFrac (which loses too much relevant data).

Acknowledgements

I would like to acknowledge all the help I received from Professor Compeau in guiding me with this project. Additionally, I would like to thank Annie Nadkarni who taught us the distance metrics used in this project and Ayushi Sood for starting me off in the right direction.

Work Cited

Baselga, Andrés. Separating the Two Components of Abundance-based Dissimilarity: Balanced Changes in Abundance Vs. Abundance Gradients. https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.12029

Bloom, Stephen. Similarity Indices in Community Studies: Potential Pitfalls. 1981. https://www3.epa.gov/region1/npdes/merrimackstation/pdfs/ar/AR-1249.pdf

Callahan, Benjamin. DADA2 Pipeline Tutorial (1.12). GitHub repository. https://benjjneb.github.io/dada2/tutorial.html

Callahan, B., McMurdie, P. & Holmes, S. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11, 2639–2643 (2017). https://doi.org/10.1038/ismej.2017.119

Fritz, A., Hofmann, P., Majda, S. *et al.* CAMISIM: simulating metagenomes and microbial communities. *Microbiome* **7,** 17 (2019). https://doi.org/10.1186/s40168-019-0633-6

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.5. https://CRAN.R-project.org/package=dplyr

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry, H. Stevens, Eduard Szoecs and Helene Wagner (2019). vegan: Community Ecology Package. R package version 2.5-6. https://CRAN.R-project.org/package=vegan

Julien Tremblay, Etienne Yergeau, Systematic processing of ribosomal RNA gene amplicon sequencing data, GigaScience, Volume 8, Issue 12, December 2019, giz146, https://doi.org/10.1093/gigascience/giz146

McMurdie and Holmes (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE. 8(4):e61217

Rognes, Torbjørn. Alternative VSEARCH pipeline. 2019. GitHub repository. https://github.com/torognes/vsearch/wiki/Alternative-VSEARCH-pipeline

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Schliep K.P. 2011. phangorn: phylogenetic analysis in R. Bioinformatics, 27(4) 592-593

Wright ES (2016). "Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R." _The R Journal_, *8*(1), 352-359.