Periodontitis

Jaewoong Lee Seunghoon Kim Semin Lee

2021-01-08

Contents

List of Tables

List of Figures

1 Introduction

1.1 Microbiome

The microbiome is consists of microbiota, the micro-organisms that live inside and on humans (Turnbaugh et al., 2007). The microbiome is also about 10^{13} micro-organisms whose that collective genome (Gill et al., 2006).

1.2 Ribosomal RNA

Ribosomal RNA (rRNA) is well-known as a key to phylogeny (Olsen & Woese, 1993).

1.3 16S rRNA Gene Sequencing

1.4 Periodontitis

Periodontitis is an inflammatory condition that affects the periodontium that surrounds and supports teeth. Major syndromes of periodontitis are clinical attachment loss and bone loss (Flemmig, 1999). Previous studies found risk factors of periodontitis, such as smoking, diabetes, genetic factors, and host response (Van Dyke & Dave, 2005).

2 Materials

2.1 16S rRNA Gene Sequencing

- 100 Healthy samples
- 50 Chronic Early Periodontitis Sample
- 50 Chronic Moderate Periodontitis Sample
- 50 Chronic Severe Periodontitis Sample

3 Methods

3.1 QIIME2 Workflow

QIIME2 is a capable, expandable and distributed microbiome analysis package with transparent analysis (Bolyen et al., 2019, 2018). A theoretic overview of QIIME2 workflow is figure 2.

3.1.1 Denoising techniques

There are two denoising techniques provided by QIIME2: DADA2 (Callahan et al., 2016) and Deblur (Amir et al., 2017). The most meaningful difference between DADA2 and Deblur is the strategy that divides them into different variants (Figure 3). DADA2 uses amplicon sequence variants (ASVs) that strictly divide sequences even one-base mismatch. However, Deblur uses operational taxonomic units (OTUs), considers the same taxonomies when they are 97 % or more matched. We chose DADA2 rather than Deblur by the result of two reasons. First, DADA2 has internal filtering methods that cut the sequences with low-quality out. Second, DADA2 can be designated trimmed length both in forward and reverse.

3.1.2 Taxonomy Classification

There are three taxonomy classification databases: Greengenes (GG) (DeSantis et al., 2006), SILVA (Pruesse et al., 2007) and Human Oral Microbiome Database (HOMD) (Chen et al., 2010). The essential difference is its resolution. Previous researches have found that a higher accuracy at taxonomic levels above the genus level. However, accuracy drops at the species level (Gihawi et al., 2019).

3.1.3 Merging Denoising and Taxonomy Classification

After denoising and taxonomy classification steps, different IDs, such as ASVs or OTUs, have been identified as the single taxonomy. We considered those IDs as a single taxonomy (Figure 5).

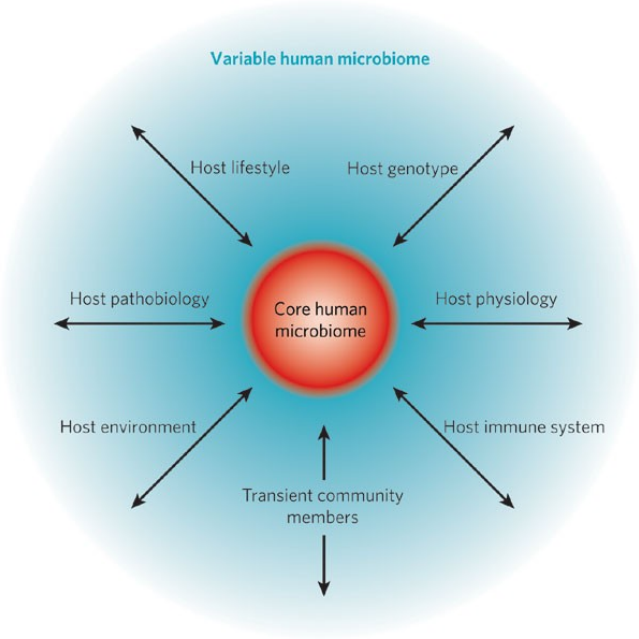


Figure 1: Concept of a Core Human Microbiome (Turnbaugh et al., 2007)

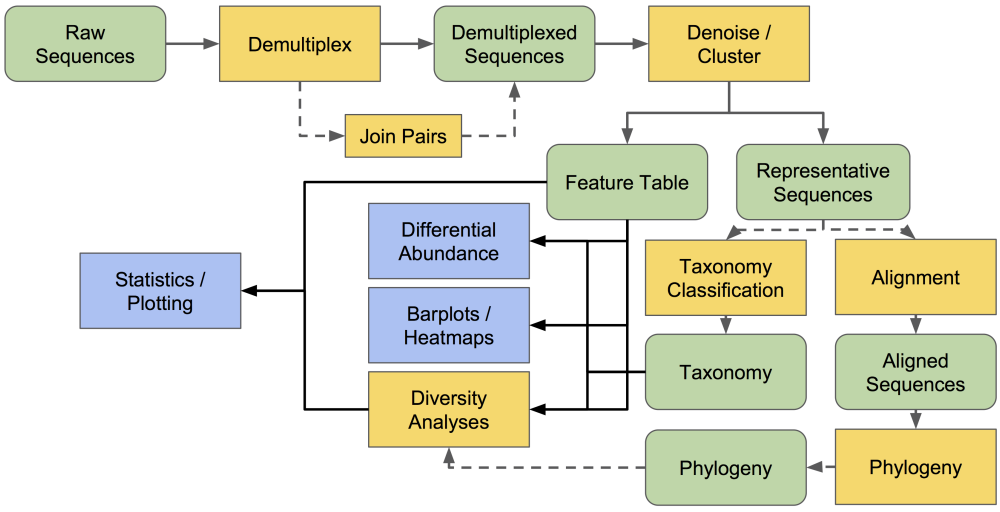


Figure 2: A Theoretic Overview of QIIME2 Workflow (Bolyen et al., 2019, 2018)

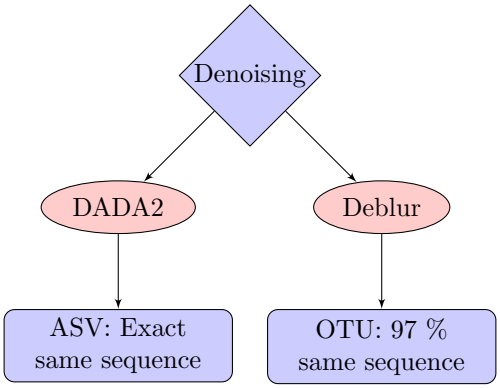


Figure 3: Denoising Techniques which provided by QIIME2

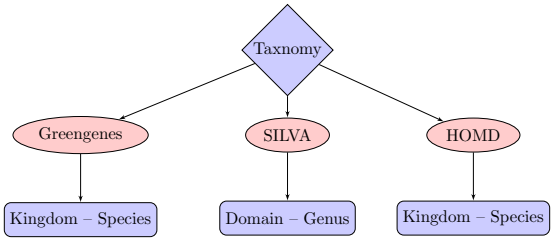


Figure 4: Taxonomy Classification which provided by QIIME2

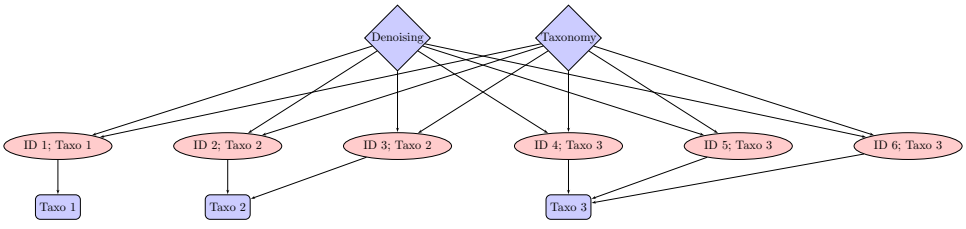


Figure 5: Example Diagram for Merging Denoising and Taxonomy Classification

3.1.4 Rarefaction

Rarefaction is a statistical method of estimating the number of species expected in a random sample taken from a collection (James & Rathbun, 1981). Moreover, rarefaction allows comparisons of the species richness among communities. Thus, rarefaction is one of the best choices for normalization (Weiss et al., 2017).

3.1.5 Alpha-diversity

Alpha-diversity is a metric that shows the richness of taxa in a single community. QIIME2 provides four alpha-diversity indices:

- Evenness index (Pielou, 1966).
- Faith's phylogenetic diversity (Faith PD) (Faith, 1992).
- Observed features.
- Shannon's diversity index (Shannon, 1948).

The evenness index shows a measurement of diversity in a different type of community (Pielou, 1966). Faith's phylogenetic diversity index, however, indicates a qualitative evaluation of community richness. The index prefers species conservation that incorporates taxic diversity (Faith, 1992). Observed features index, as its name, is the number of detected features in the microbiome. Furthermore, Shannon's diversity index means a significant aspect of community richness (Shannon, 1948).

3.1.6 Beta-diversity

Beta-diversity is a metric that indicates the taxonomic differentiation among multiple communities. QIIME2 provides four beta-diversity indices:

- Bray-Curtis distance index (Sørensen, 1948).
- Jaccard distance index (Jaccard, 1912).
- Unweighted UniFrac distance index (McDonald et al., 2018).
- Weighted UniFrac distance index (McDonald et al., 2018).

Bray-Curtis distance index shows a quantitative measurement of community dissimilarity (Sørensen, 1948). Jaccard distance index, however, indicates an evaluation of local distribution among communities (Jaccard, 1912). UniFrac distance indices reveal measures of phylogenetic distances (McDonald et al., 2018). The unweighted UniFrac distance index and the weighted UniFrac distance index promote qualitative and quantitative, respectively.

3.1.7 ANCOM

ANCOM (analysis for the composition of microbiomes) analyzes the architecture of the microbiome in multiple populations (Mandal et al., 2015). An example ANCOM volcano plot is figure 6. In figure 6, two metrics are clearly shown: clr and W. clr stands for centered log ratio, and W is a count of the number of sub-hypothesis which have passed for given species.

3.2 Python Packages

3.2.1 Pandas

Pandas is a Python package of rich data structures and tools for analyzing with structured data sets (McKinney et al., 2011).

3.2.2 Scikit-learn

Scikit-learn grants state-of-the-art implementation of many machine learning algorithms, while controlling an easy-to-use interface tightly integrated the Python code (Pedregosa et al., 2011).

3.2.3 Matplotlib

Matplotlib is a Python graphics package which used for application development, interactive scripting and publication quality image generation (Barrett, Hunter, Miller, Hsu, & Greenfield, 2005). Matplotlib, also, is designed to create simple plots with a few commands (Hunter, 2007).

3.2.4 Seaborn

Seaborn is a Python data visualization package which based on matplotlib, allows a high-level interface for displaying engaging and descriptive statistical graphics (Waskom & the seaborn development team, 2020).

3.2.5 statannot

Statannot is a python package which computes statistical test and adds statistical annotations on violin plot generated by seaborn package.

3.3 t-SNE

t-SNE (t-distributed stochastic neighbor embedding) reveals high-dimensional data in a location in a two-dimensional map (Maaten & Hinton, 2008). Figure 7 is example of t-SNE with hand-writing digits (Maaten & Hinton, 2008). In figure 7, all 10 digits are grouped into 10 groups clearly; some hand-writings, however, are classified into wrong groups due to their similar shapes, such as 0 and 6.

3.4 Classification

In machine learning, classification is one of the supervised problems. Classifier identifies a class of new observations, depends on training observations.

In this study, classification will be carried out as figure 8; and the third step in figure 8 is demonstrated in minute detail as figure 9. Note that the first step in figure 8 is optional: due to tables herein-after, such as table3, show that no statistically significant differences between healthy samples and early periodontitis samples and between moderate periodontitis samples and severe periodontitis samples.

Moreover, evaluations of classification algorithm are carried out with derivations from confusion matrix (table 1):

- Accuracy (ACC) $= \frac{TP+TN}{TP+TN+FP+FN}$
- Balanced Accuracy (BA) $= \frac{TP}{2\times(TP+FN)} + \frac{TN}{2\times(TN+FP)}$
- Sensitivity (SEN) $= \frac{TP}{TP+FN}$
- Specificity (SPE) $= \frac{TN}{TN+FP}$
- Precision (PRE) $= \frac{TP}{TP+FP}$

3.4.1 Random Forest Classification

As figure 8, importance of features have to be derived by classifier. Random Forest classifier (Breiman, 2001) can get this information, and is used frequently by researchers. Hence, Random Forest classifier will be carried out with every class (Figure 10) or with merged classes (Figure 11).

4 Results

4.1 Quality Filter

Longer sequences have more fallen sequence quality than shorter. Thus, sequences which longer than threshold should be trimmed out due to their low quality. However, gold-standard strategy for deciding the threshold does not exist; the threshold is set as longest sequence length which have half of sequences have greater than 30 quality score. Hence, sequence quality plot is shown as figure 12; trimmed length in forward reads is 300, and trimmed length in reverse reads is 265.

4.2 Rarefaction

Sampling depth should be decided for rarefaction. Gold-standard method for determining sampling depth is minimum frequency in the samples. Hence, sampling depth with DADA2 is 3,786 (Figure 13), and sampling depth with Deblur is 7,253 (Figure 14).

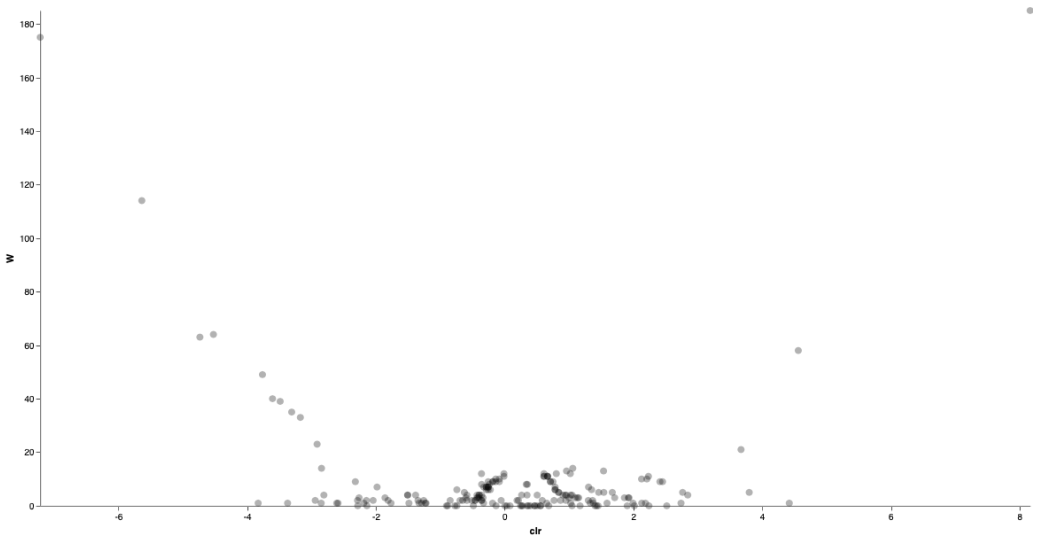


Figure 6: Example ANCOM Volcano Plot which Provided by QIIME2 (Bolyen et al., 2019, 2018)

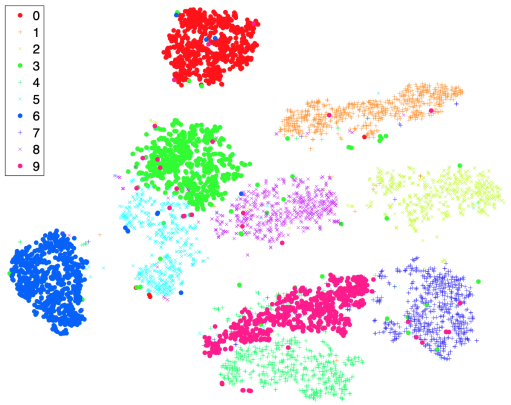


Figure 7: Visualization by t-SNE (Maaten & Hinton, 2008)

Table 1: Confusion Matrix

		Actual Class	
		Positive	Negative
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

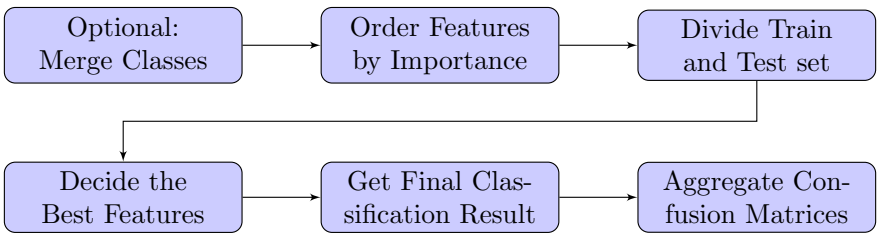


Figure 8: Workflow of Classification

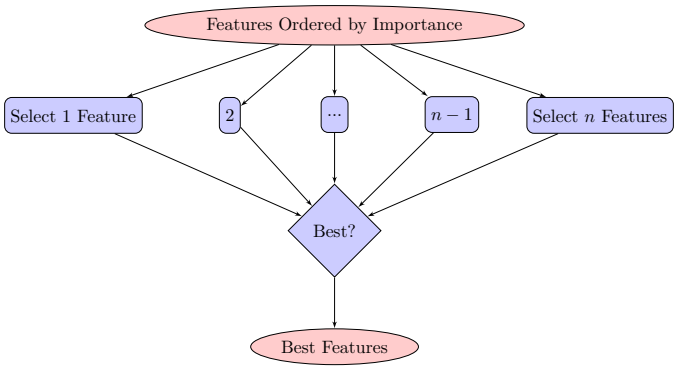


Figure 9: Deciding the Best Features

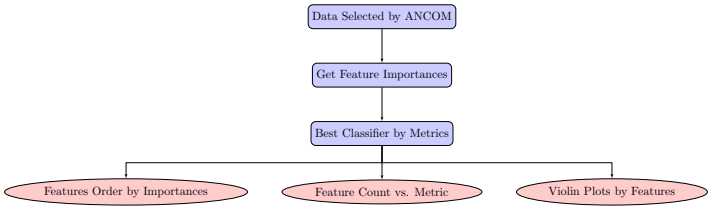


Figure 10: Random Forest Classifier Workflow

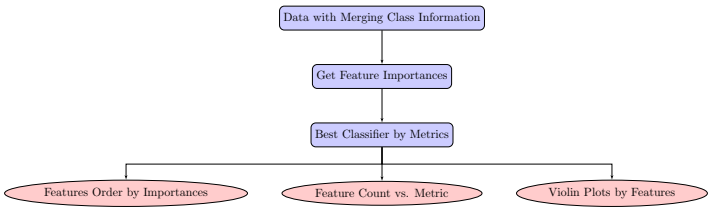
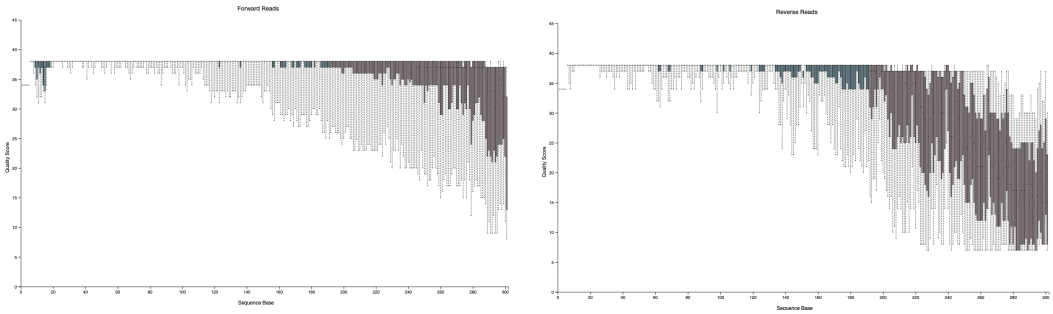


Figure 11: Random Forest Classifier Workflow with Merging



(a) Forward Reads (b) Reverse Reads

Figure 12: Sequence Quality Plot

Table 2: Kruskal-Wallis Tests among All Group with DADA2

Alpha-Diversity	H	p-value
Evenness	12.185457848605665	0.006774123738087294
Faith PD	33.42272318725111	2.6227945981005624e-7
Observed Features	21.019370066584198	0.0001043055436502384
Shnnon's Diversity	7.311350438247132	0.06260902704190516

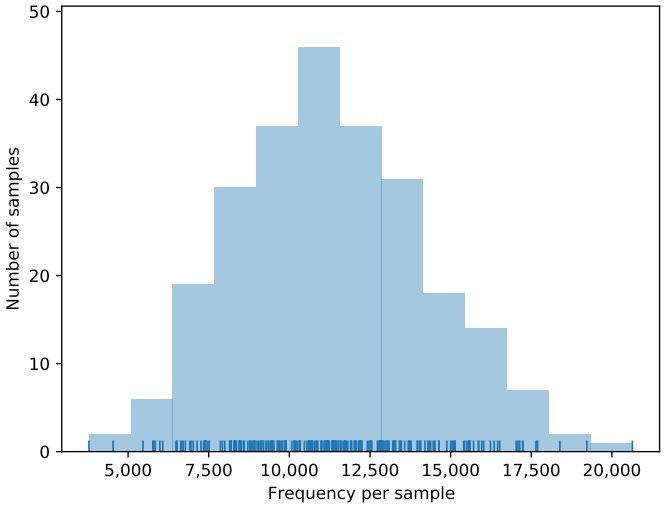


Figure 13: Frequency and Number per Sample by DADA2

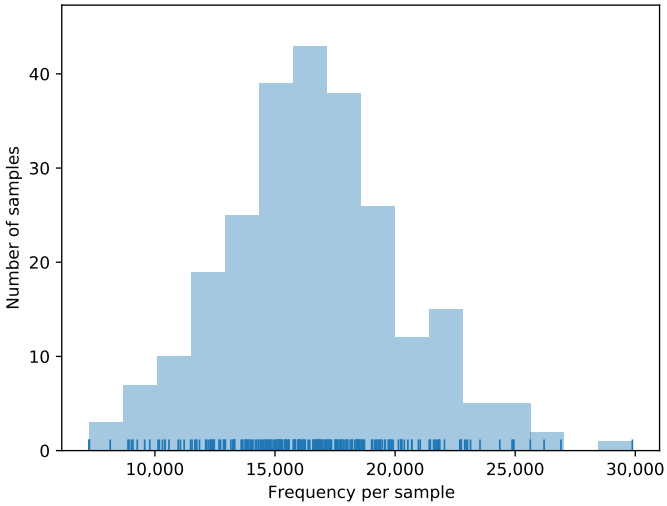


Figure 14: Frequency and Number per Sample by Deblur

Table 3: Kruskal-Wallis Tests from Evenness Index with DADA2

Group 1	Group 2	H	p-value	q-value
Slight (n=50)	Healthy (n=100)	0.003576158940404639	0.9523141335184352	0.9523141335184352
Slight (n=50)	Moderate (n=50)	5.112902970297	0.02374855135702787	0.03562282703554181
Slight (n=50)	Severe (n=50)	5.206859405940577	0.022497939047433364	0.03562282703554181
Healthy (n=100)	Moderate (n=50)	6.591830463576116	0.01024477815032801	0.03073433445098403
Healthy (n=100)	Severe (n=50)	6.756619867549659	0.0093400517403089	0.03073433445098403
Moderate (n=50)	Severe (n=50)	0.01216633663364064	0.9121705706341857	0.9523141335184352

Table 4: Kruskal-Wallis Tests from Faith PD Index with DADA2

Group 1	Group 2	H	p-value	q-value
Slight (n=50)	Healthy (n=100)	0.3434543046357703	0.557842085850555	0.557842085850555
Slight (n=50)	Moderate (n=50)	7.833790099009889	0.005127846488653557	0.0076917697329803355
Slight (n=50)	Severe (n=50)	19.832839603960394	8.451807369366e-06	2.5355422108098e-05
Healthy (n=100)	Moderate (n=50)	8.964254304635801	0.0027531304578610103	0.005506260915722021
Healthy (n=100)	Severe (n=50)	24.32056688741727	8.156352492752821e-07	4.893811495651693e-06
Moderate (n=50)	Severe (n=50)	5.461592079207946	0.019438927334967618	0.02332671280196114

Table 5: Kruskal-Wallis Tests from Observed Features Index with DADA2

Group 1	Group 2	H	p-value	q-value
Slight (n=50)	Healthy (n=100)	9.559750209810552	0.001988901703187571	0.005966705109562713
Slight (n=50)	Moderate (n=50)	0.01069480203811357	0.9176330712208788	0.9176330712208788
Slight (n=50)	Severe (n=50)	1.8918489487993617	0.1689935259025544	0.20279223108306527
Healthy (n=100)	Moderate (n=50)	16.280824652808626	5.461383546704547e-05	0.0003276830128022728
Healthy (n=100)	Severe (n=50)	6.9139163882453465	0.008552745576573654	0.017105491153147308
Moderate (n=50)	Severe (n=50)	2.1161415616917054	0.145753334857958	0.20279223108306527

Table 6: Kruskal-Wallis Tests from Shannon's Diversity Index with DADA2

Group 1	Group 2	H	p-value	q-value
Slight (n=50)	Healthy (n=100)	5.291586754966886	0.021428686619934936	0.11394854365524665
Slight (n=50)	Moderate (n=50)	1.3095920792079028	0.2524685249140654	0.3029622298968785
Slight (n=50)	Severe (n=50)	4.305790099009869	0.037982847885082216	0.11394854365524665
Healthy (n=100)	Moderate (n=50)	2.223194701986756	0.13595148461788642	0.27190296923577284
Healthy (n=100)	Severe (n=50)	0.06109668874171348	0.8047709009969876	0.8047709009969876
Moderate (n=50)	Severe (n=50)	1.3573544554455452	0.2439965042398798	0.3029622298968785

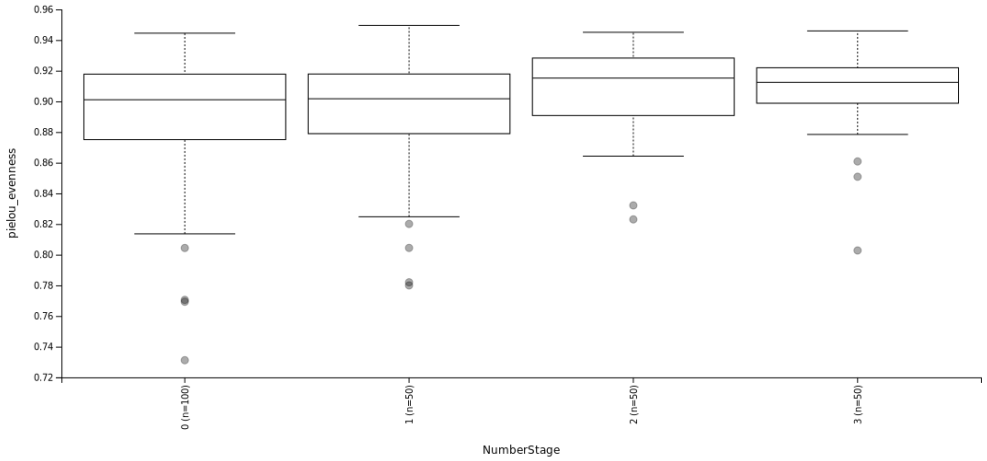


Figure 15: Evenness Index from DADA2

Table 7: Bray-Curtis Distance Index with DADA2

Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
Slight	Healthy	150	999	1.8288671026193992	0.004	0.0048
Slight	Moderate	100	999	2.4738348324475568	0.001	0.0015
Slight	Severe	100	999	3.3691960533567005	0.001	0.0015
Healthy	Moderate	150	999	5.602936565444328	0.001	0.0015
Healthy	Severe	150	999	6.325447306476738	0.001	0.0015
Moderate	Severe	100	999	1.1018815494184453	0.219	0.219

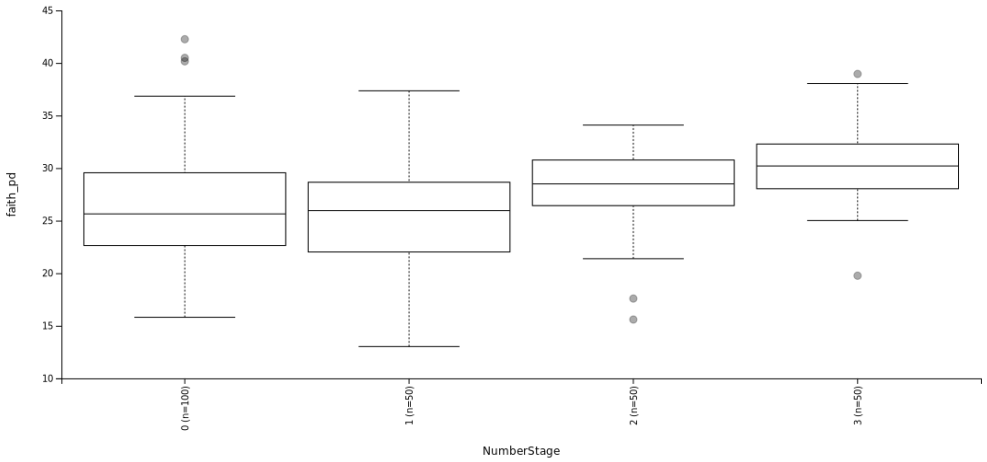


Figure 16: Faith PD Index from DADA2

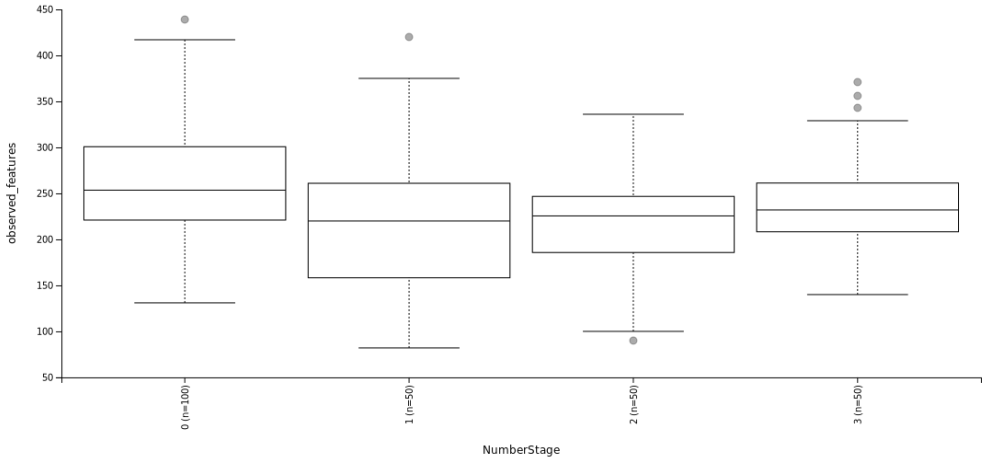


Figure 17: Observed Features Index from DADA2

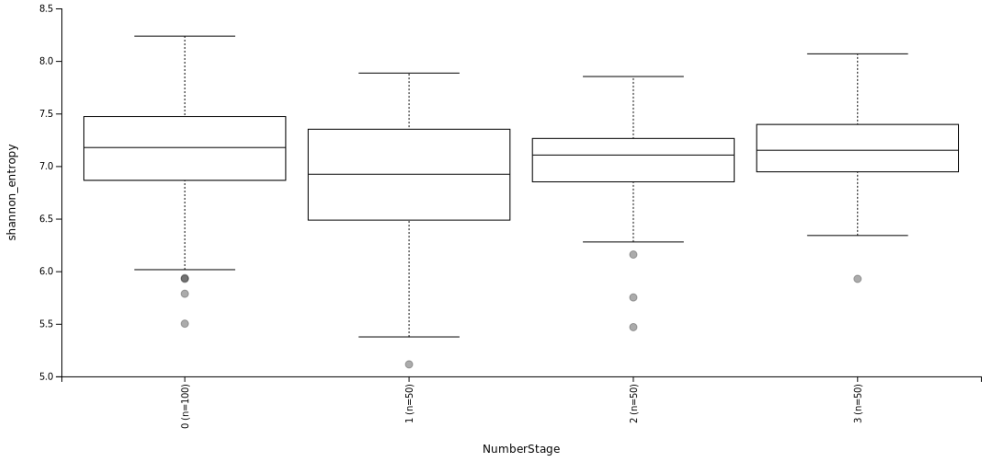


Figure 18: Shannon's Diversity Index from DADA2

Table 8: Jaccard Distance Index with DADA2

Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
Slight	Healthy	150	999	1.5875955458962276	0.001	0.0012
Slight	Moderate	100	999	1.7486415070626309	0.001	0.0012
Slight	Severe	100	999	1.8371794988000507	0.001	0.0012
Healthy	Moderate	150	999	3.9547515710373635	0.001	0.0012
Healthy	Severe	150	999	3.8380356039546784	0.001	0.0012
Moderate	Severe	100	999	0.9700395015774723	0.62	0.62

Table 9: Unweighted UniFrac Distance Index with DADA2

Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
Slight	Healthy	150	999	2.414078271406213	0.002	0.0024
Slight	Moderate	100	999	4.941256726696032	0.001	0.0015
Slight	Severe	100	999	6.184322196061149	0.001	0.0015
Healthy	Moderate	150	999	12.484494695636283	0.001	0.0015
Healthy	Severe	150	999	13.432593034368626	0.001	0.0015
Moderate	Severe	100	999	1.2428267228930112	0.084	0.084

Table 10: Weighted UniFrac Distance Index with DADA2

Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
Slight	Healthy	150	999	2.6584441800971716	0.019	0.022799999999999997
Slight	Moderate	100	999	8.702906307484113	0.001	0.0015
Slight	Severe	100	999	14.068214366598513	0.001	0.0015
Healthy	Moderate	150	999	22.059259782524673	0.001	0.0015
Healthy	Severe	150	999	31.310013450629775	0.001	0.0015
Moderate	Severe	100	999	1.7543213081828324	0.115	0.115

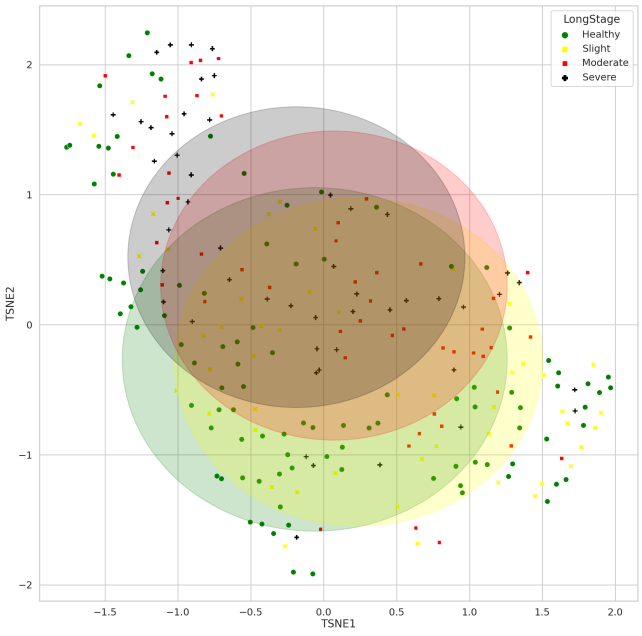
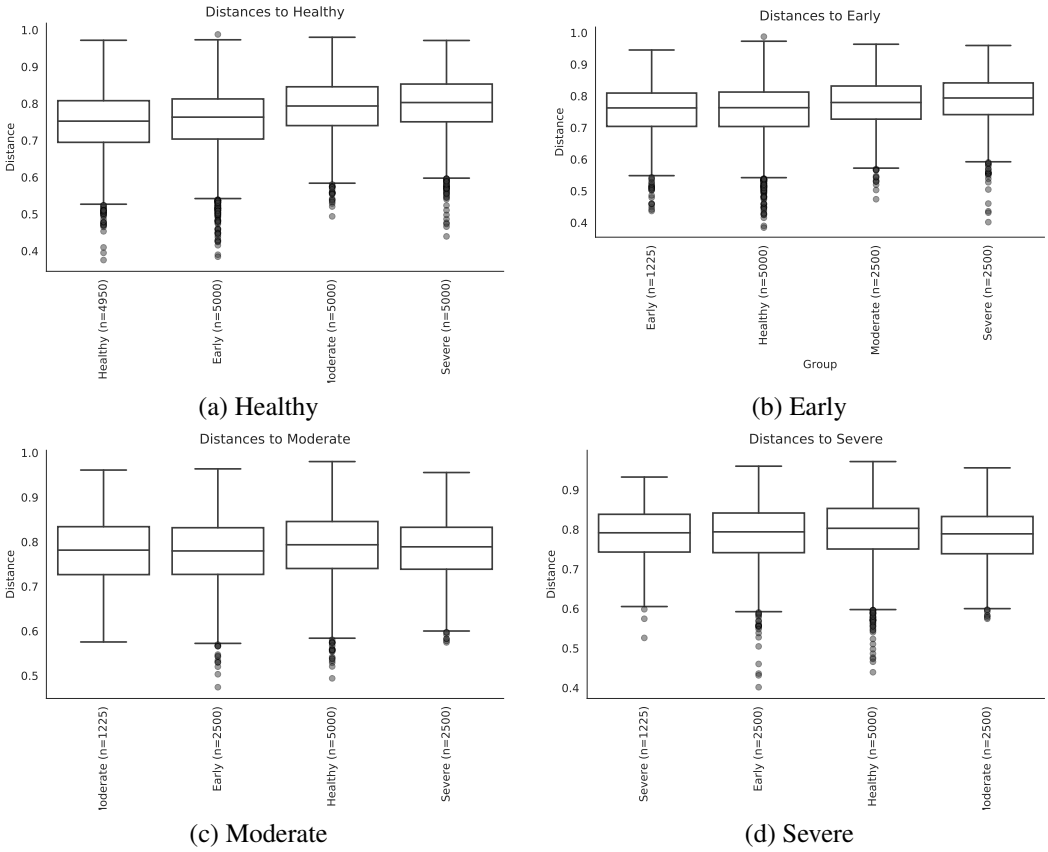


Figure 19: t-SNE Plot from Bray-Curtis Distance Index with DADA2



(a) Healthy (b) Early

(c) Moderate (d) Severe

Figure 20: Bray-Curtis Distance Index with DADA2

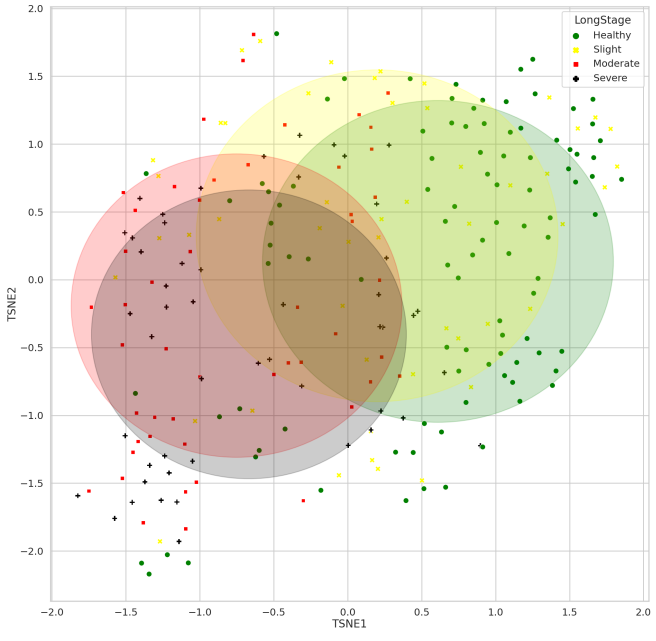
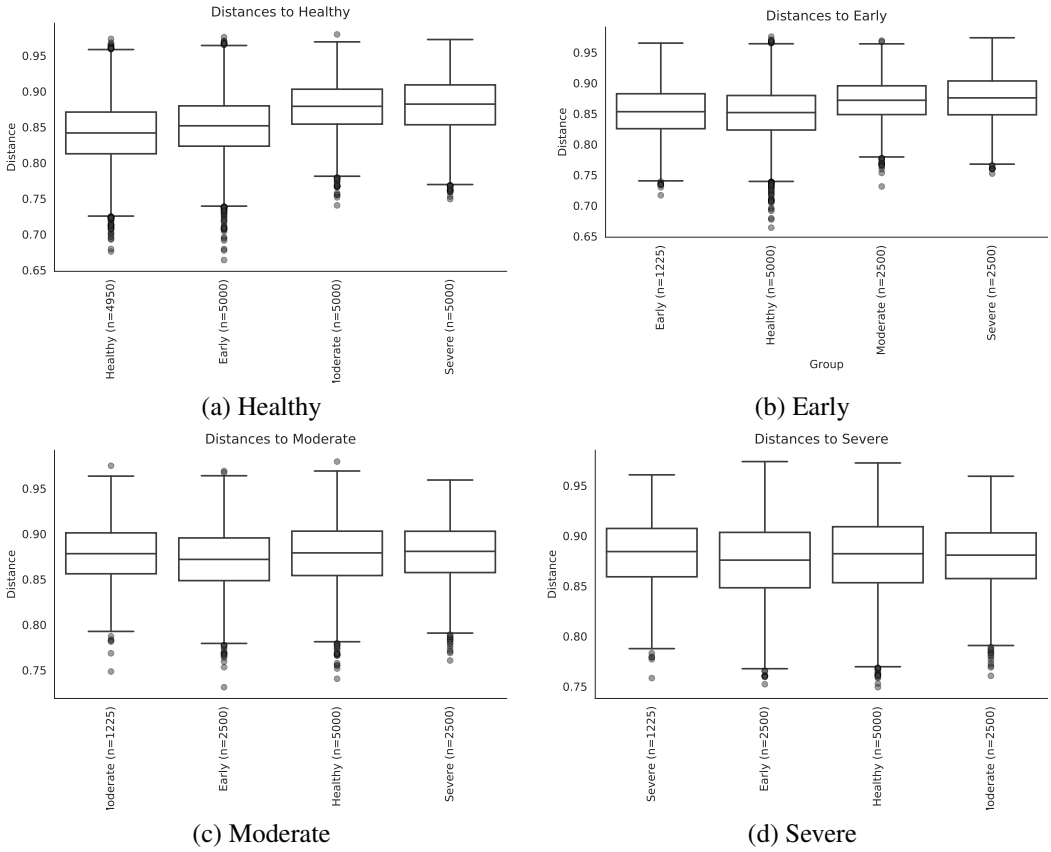


Figure 21: t-SNE Plot from Jaccard Distance Index with DADA2



(a) Healthy (b) Early

(c) Moderate (d) Severe

Figure 22: Jaccard Distance Index with DADA2

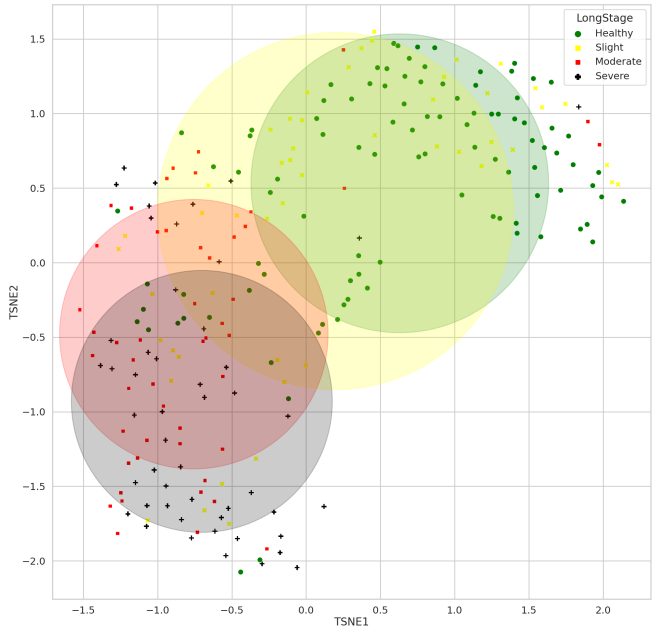
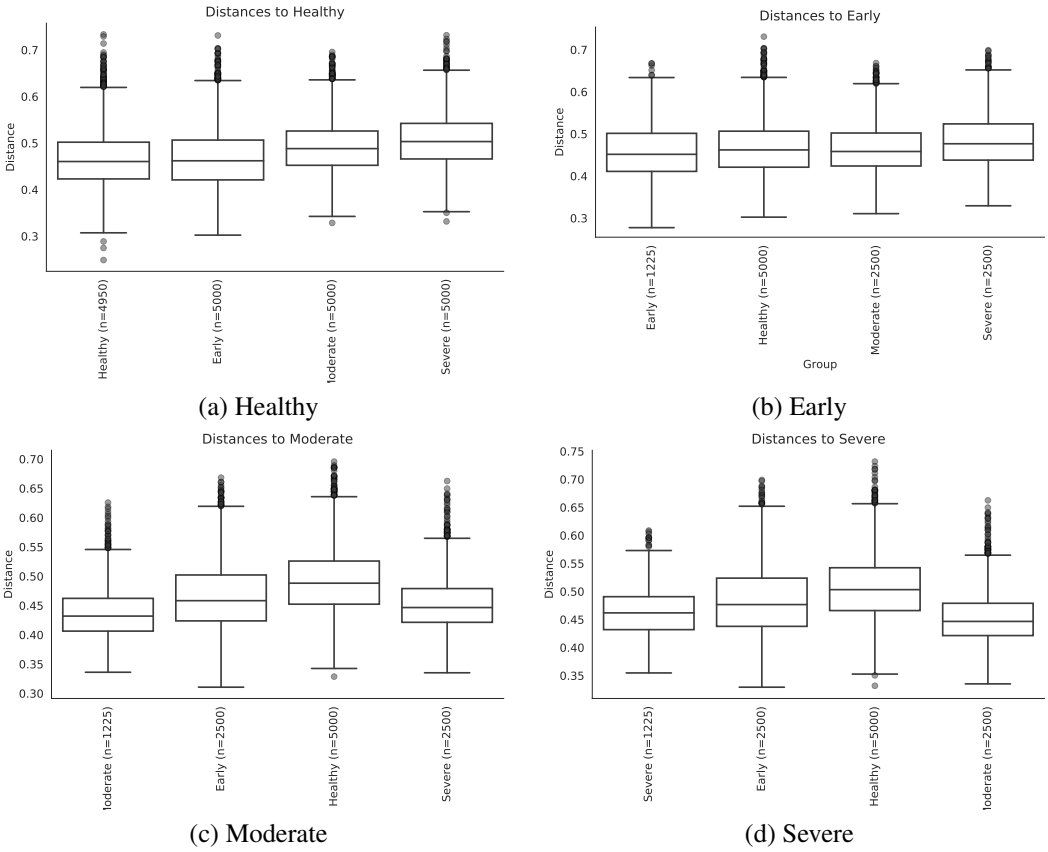


Figure 23: t-SNE Plot from Unweighted UniFrac Distance Index with DADA2



(a) Healthy (b) Early

(c) Moderate (d) Severe

Figure 24: Unweighted UniFrac Distance Index with DADA2

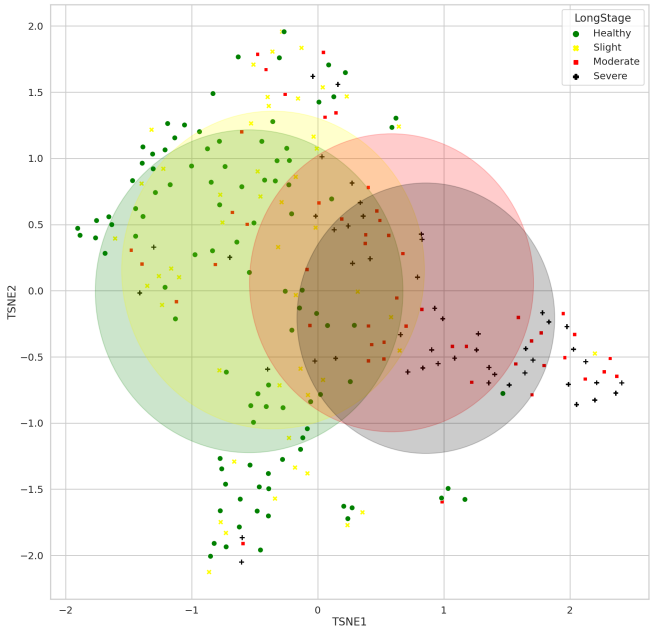


Figure 25: t-SNE Plot from Weighted UniFrac Distance Index with DADA2

4.3 Alpha-diversity

4.4 Beta-diversity

4.5 ANCOM

4.6 t-SNE Plot with Whole Microbiome

As mentioned herein-before, t-SNE is a technique which reduce multi-dimensional data into two-dimension. Whole microbiome data are multi-dimensional data, which have *circa* 600 columns, so the data should be reduced their dimension for readability. Hence, by the grace of t-SNE, the microbiome data have been deflated their dimension: 328 taxa from DADA2 and GG (Figure 28), 633 taxa from DADA2 and SILVA (Figure 29), 425 taxa from DADA2 and HOMD (Figure 30), 232 taxa from Deblur and GG (Figure 31), 414 taxa from Deblur and SILVA (Figure 32) and 235 taxa from Deblur and HOMD (Figure 33).

4.7 t-SNE Plot with ANCOM Selected Microbiome Data

4.8 Random Forest Classifier with Every Class

4.9 Random Forest Classifier with Merging (Moderate+Severe) Classes

4.10 Random Forest Classifier with Healthy Class and Early Class Only

5 Discussion

5.1 Alpha-diversity

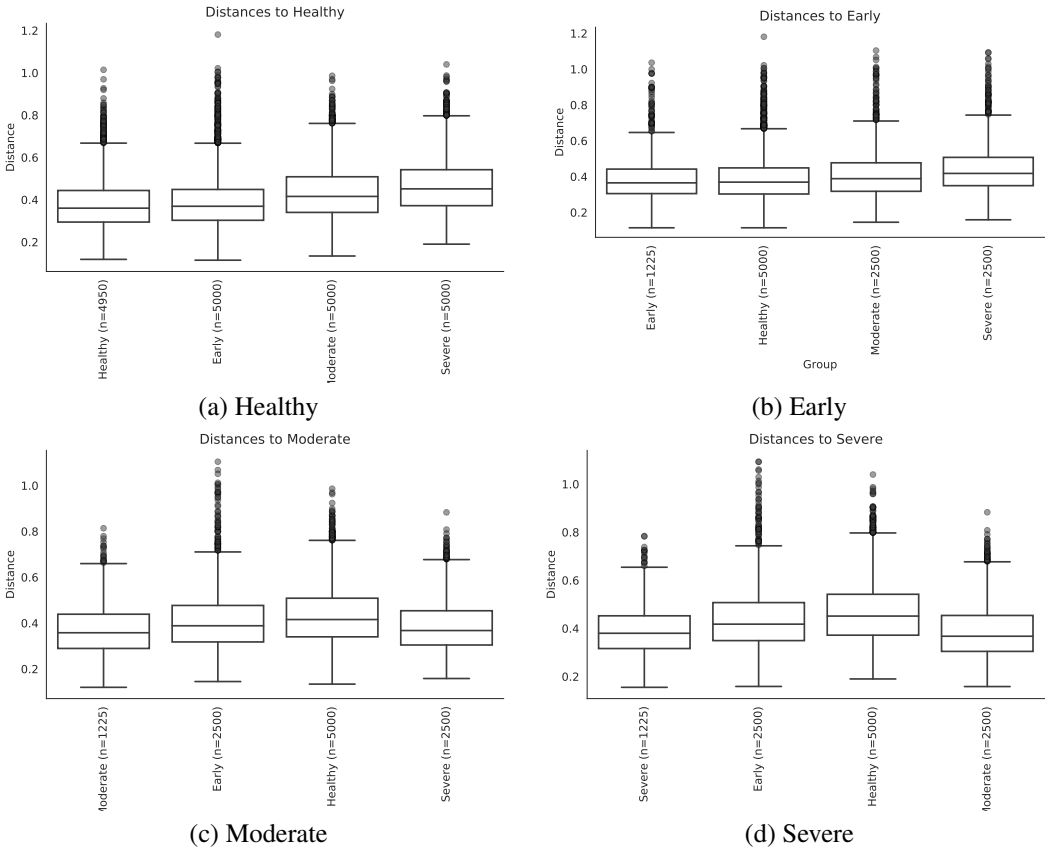
Alpha-diversity indices among all groups from DADA2 are in table 2. Shannon's diversity index in DADA2, though, has marginally significant p-value; the other indices have strongly significant p-values. Additionally, there are no statistically significant differences between (Healthy and Early) classes and (Moderate and Severe) classes with evenness index from DADA2 (Table 3 and Figure 15). Also, there is no statistically significant difference between (Healthy and Early) classes with Faith's phylogenetic diversity index from DADA2 (Table 4 and Figure 16). Moreover, there are no statistically significant differences between (Early and Moderate) classes, (Early and Severe) classes and (Moderate and Severe) classes with observed feature index from DADA2 (Table 5 and Figure 17). Furthermore, there are no statistically significant differences between (Healthy and Moderate) classes, (Healthy and Severe) classes, (Early and Moderate) classes and (Moderate and Severe) classes from Shannon's diversity index from DADA2 (Table 6 and Figure 18).

Alpha-diversity indices among all groups from Deblur are in table **??**. Every index have strongly significant p-values. Additionally, there are no statistically significant differences between (Healthy and Early) classes, (Healthy and Moderate) classes, (Healthy and Severe) classes and (Moderate and Severe) classes with evenness index from Deblur (Table **??** and Figure **??**). Also, there are no statistically significant differences between (Healthy and Early) classes and (Moderate and Severe) classes with Faith's phylogenetic diversity index from Deblur (Table **??** and Figure **??**). Moreover, there are no statistically significant differences between (Healthy and Early) classes and (Moderate and Severe) classes with observed features index from Deblur (Table **??** and Figure **??**). Furthermore, there are no statistically significant differences between (Healthy and Early) classes and (Moderate and Severe) classes with Shannon's diversity index from Deblur (Table **??** and Figure **??**).

Merging similar classes could elevate classification metrics; while merging must result mere primitive classification than without merging classes. Accordingly, deciding merged classes should be rigorous and resolutely reasoned. In result, many pairs of classes should be merged as value of alpha-diversity indices, for instance (Healthy and Early) classes, (Healthy and Severe) classes and (Moderate and Severe) classes. Despite alpha-diversity indices show no significant differences, some pairs of classes have to refuse to be merged in two major reasons. First, merging those classes is fallacious. For example, (Healthy and Severe) classes, without loss of generality. Healthy class and Severe class does not adjoin each other, in terms of clinical stage. Second, even those classes are adjacent each other, some pairs of classes have not enough results to merge. For instance, null hypothesis from (Early and Moderate) classes is only sustained by Shannon's diversity index (Table 6), so merging Early class and Moderate class cannot be reasoned. Hence, two pairs of classes will be merged in classification: (Healthy and Early) classes and (Moderate and Severe) classes.

5.2 Beta-diversity

From data with DADA2, all beta-diversity distance index, includes Bray-Curtis distance index (Table 7, Figure 19 and Figure 20), Jaccard distance index (Table 8, Figure 21 and Figure 22), unweighted UniFrac distance index (Table 9, Figure 23 and Figure 24) and weighted UniFrac distance index (Table 10, Figure 25 and Figure 26), show statistically significant differences in every pair of classes, except (Moderate and Severe) classes.



(a) Healthy (b) Early

(c) Moderate (d) Severe

Figure 26: Weighted UniFrac Distance Index with DADA2

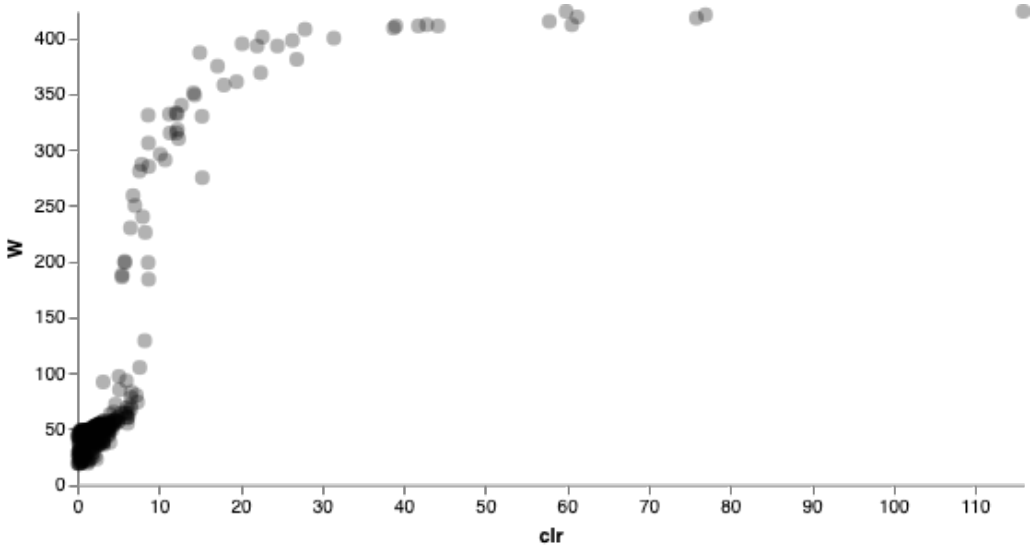


Figure 27: ANCOM Volcano Plot with DADA2 and GG

Table 11: ANCOM Significant Taxa with DADA2 and HOMD

	W	Reject null hypothesis
Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Porphyromonas gingivalis	424	True
Bacteria Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Actinomyces	424	True
Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae [XI] Filifactor alocis	421	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella intermedia	419	True
Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema putidum	418	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Tannerella forsythia	415	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Porphyromonas sp. HMT 285	412	True
Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae [XI] Peptostreptococcaceae [XI][G-6] nodatum	412	True
Bacteria Synergistetes Synergistia Synergistales Synergistaceae Fretibacterium	411	True
Bacteria Firmicutes Mollicutes Mycoplasmatales Mycoplasmataceae Mycoplasma faucium	411	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella sp. HMT 304	411	True
Bacteria Firmicutes Clostridia Clostridiales Lachnospiraceae [XIV] Lachnospiraceae [G-8] bacterium HMT 500	409	True
Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema	408	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella sp. HMT 526	401	True
Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae [XI] Peptostreptococcaceae [XI][G-9] brachy	400	True
Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae [XI] Peptostreptococcaceae [XI][G-5] saphenum	398	True
Bacteria Proteobacteria Epsilonproteobacteria Campylobacterales Campylobacteraceae Campylobacter showae	395	True
Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema sp. HMT 260	393	True
Bacteria Actinobacteria Actinobacteria Corynebacteriales Corynebacteriaceae Corynebacterium durum	393	True
Bacteria Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Actinomyces graevenitzii	387	True

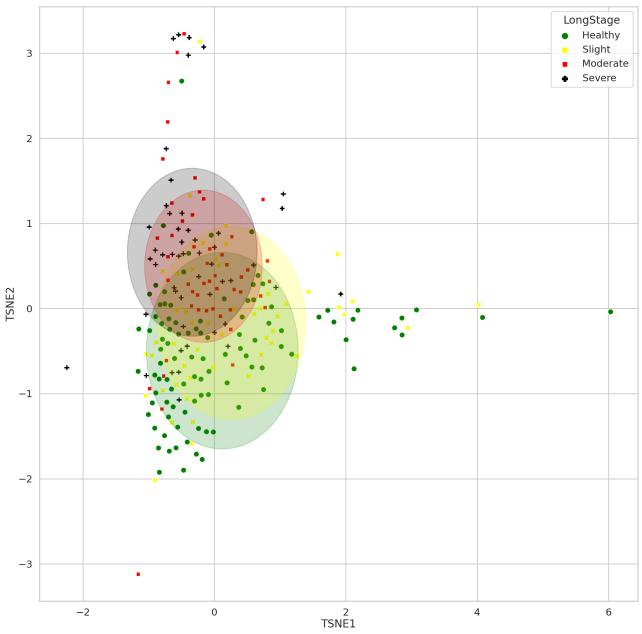


Figure 28: t-SNE Plot with Whole Microbiome from DADA2 and GG (328 taxa)

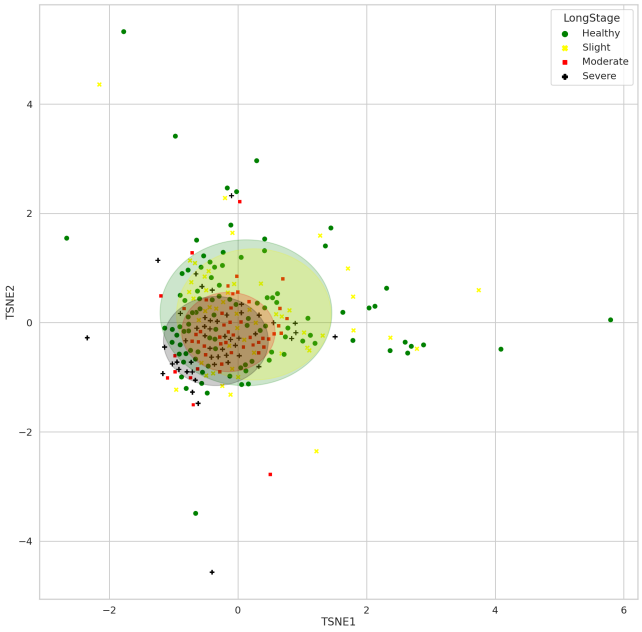


Figure 29: t-SNE Plot with Whole Microbiome from DADA2 and SILVA (633 taxa)

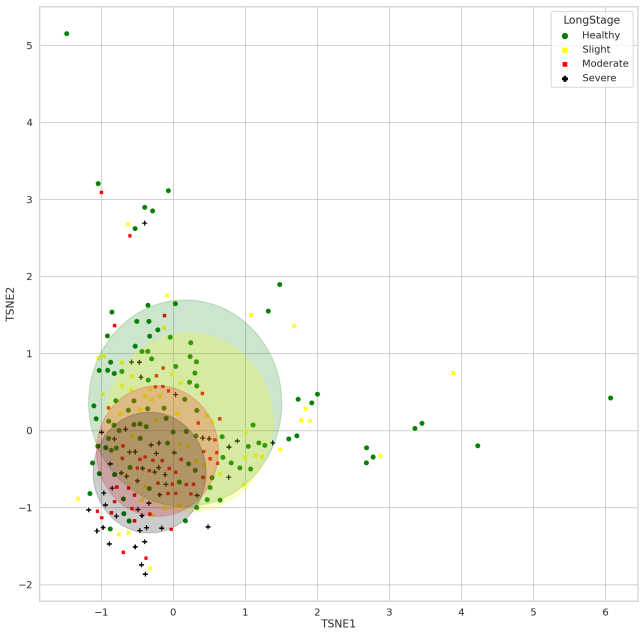


Figure 30: t-SNE Plot with Whole Microbiome from DADA2 and HOMD (425 taxa)

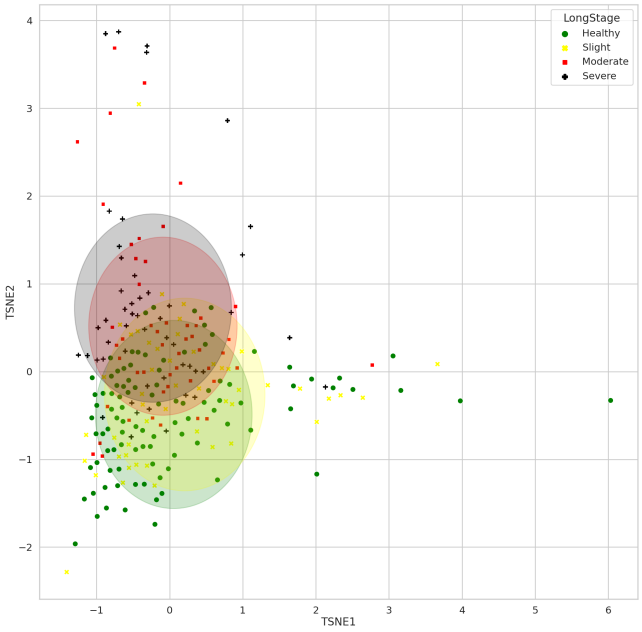


Figure 31: t-SNE Plot with Whole Microbiome from Deblur and GG (232 taxa)

Figure 32: t-SNE Plot with Whole Microbiome from Deblur and SILVA (414 taxa)

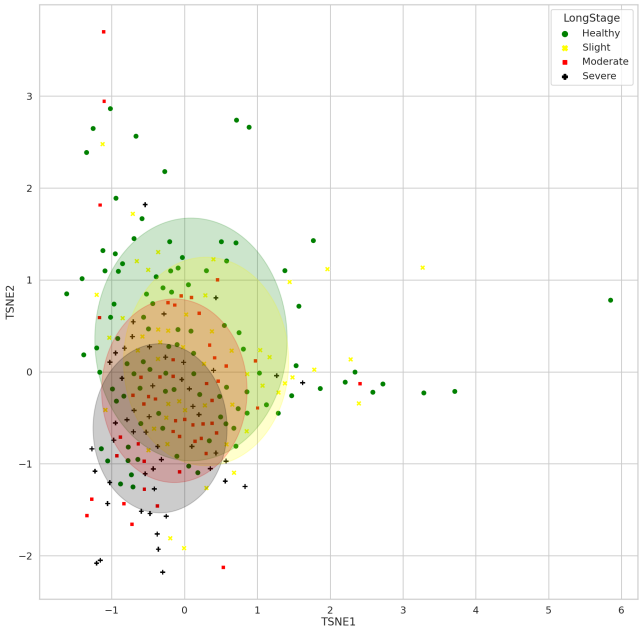


Figure 33: t-SNE Plot with Whole Microbiome from Deblur and HOMD (235 taxa)

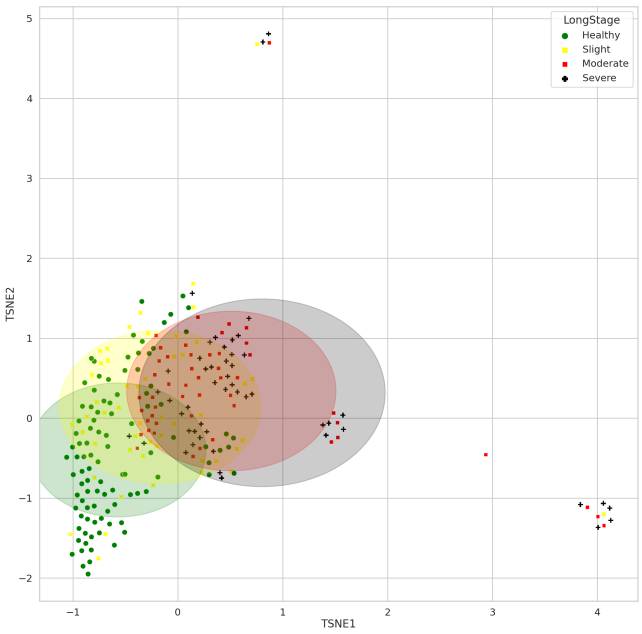


Figure 34: t-SNE Plot with ANCOM Selected Microbiome Data from DADA2 and GG (15 taxa)

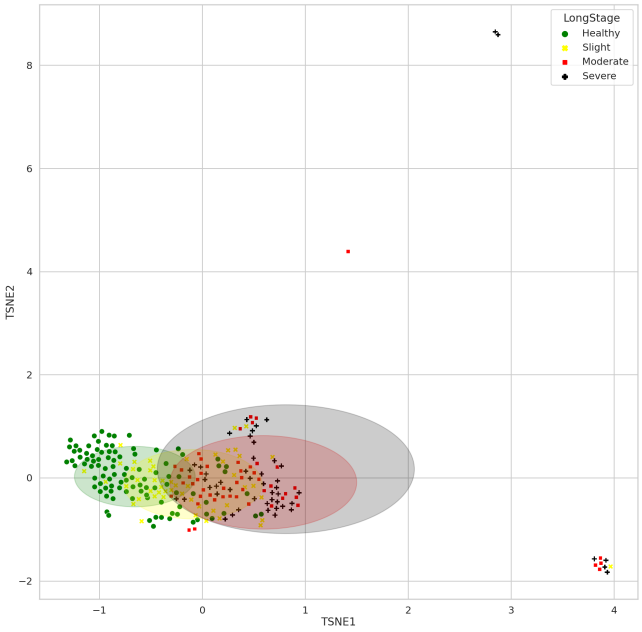


Figure 35: t-SNE Plot with ANCOM Selected Microbiome Data from DADA2 and SILVA (23 taxa)

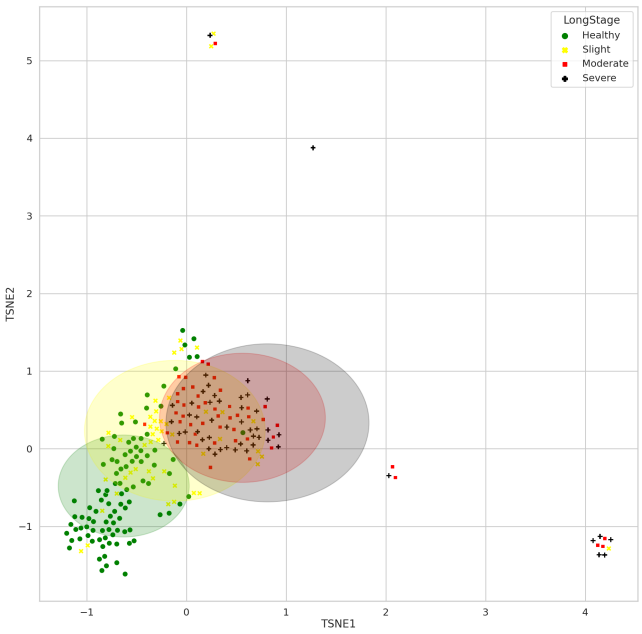


Figure 36: t-SNE Plot with ANCOM Selected Microbiome Data from DADA2 and HOMD (20 taxa)

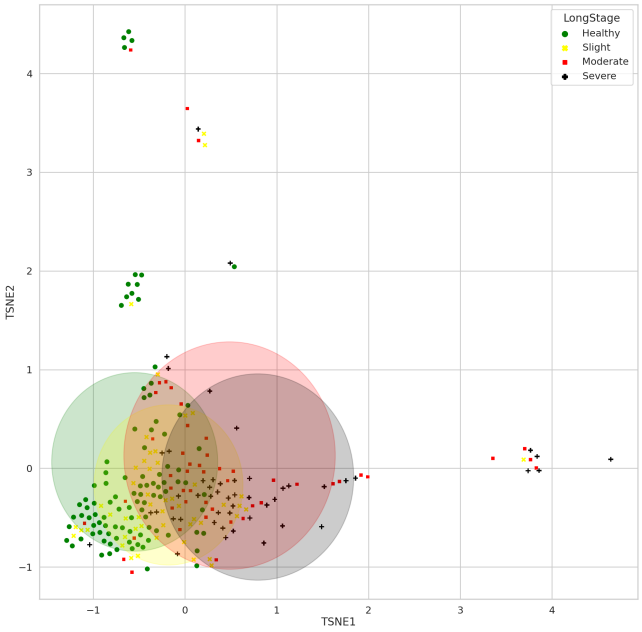


Figure 37: t-SNE Plot with ANCOM Selected Microbiome Data from Deblur and GG (27 taxa)

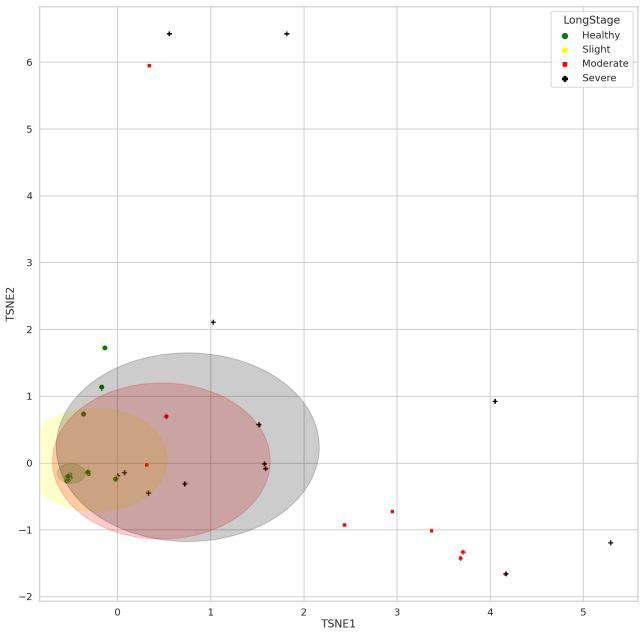


Figure 38: t-SNE Plot with ANCOM Selected Microbiome Data from Deblur and SILVA (20 taxa)

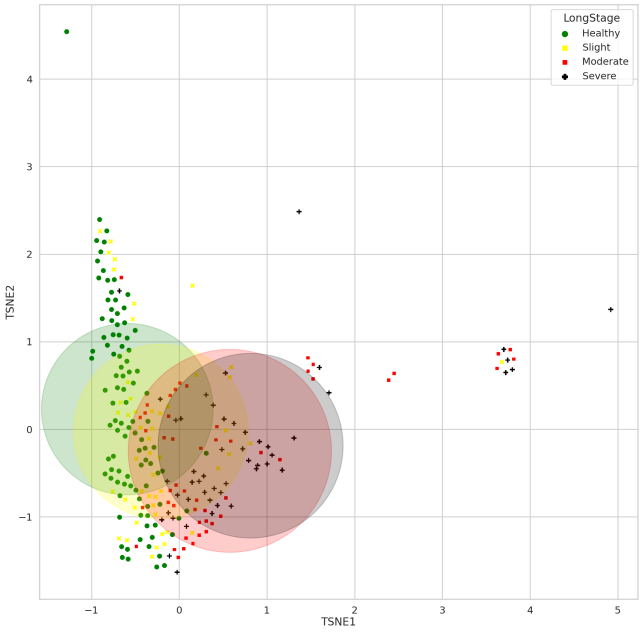


Figure 39: t-SNE Plot with ANCOM Selected Microbiome Data from Deblur and HOMD (28 taxa)

Table 12: Taxa with DADA2 and HOMD Ordered by Random Forest

Order	Taxonomy Classification	Importances
0	Bacteria Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Actinomyces	0.2563358219539378
1	Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Porphyromonas gingivalis	0.23196557322229505
2	Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae XI Filifactor alocis	0.05939593656609779
3	Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella intermedia	0.047788324527495964
4	Bacteria Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Actinomyces graevenitzii	0.0436241876822214
5	Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Porphyromonas sp. HMT 285	0.04330230979636331
6	Bacteria Proteobacteria Epsilonproteobacteria Campylobacterales Campylobacteraceae Campylobacter showae	0.041507035664934466
7	Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema sp. HMT 260	0.035914066247796333
8	Bacteria Actinobacteria Actinobacteria Corynebacteriales Corynebacteriaceae Corynebacterium durum	0.028488659124825007
9	Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Tannerella forsythia	0.02755242782877868
10	Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema putidum	0.02407253635910223
11	Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema	0.023393398798141077
12	Bacteria Firmicutes Clostridia Clostridiales Lachnospiraceae XIV Lachnospiraceae G-8 bacterium HMT 500	0.021534750022711564
13	Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae XI Peptostreptococcaceae XIG-9 brachy	0.021026493991759487
14	Bacteria Synergistetes Synergistia Synergistales Synergistaceae Fretibacterium	0.017748489962311726
15	Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella sp. HMT 526	0.01743345729576584
16	Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae XI Peptostreptococcaceae XIG-6 nodatum	0.016063177113771455
17	Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella sp. HMT 304	0.01592053147101768
18	Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae XI Peptostreptococcaceae XIG-5 saphenum	0.01446869437546418
19	Bacteria Firmicutes Mollicutes Mycoplasmatales Mycoplasmataceae Mycoplasma faucium	0.012464127995209072

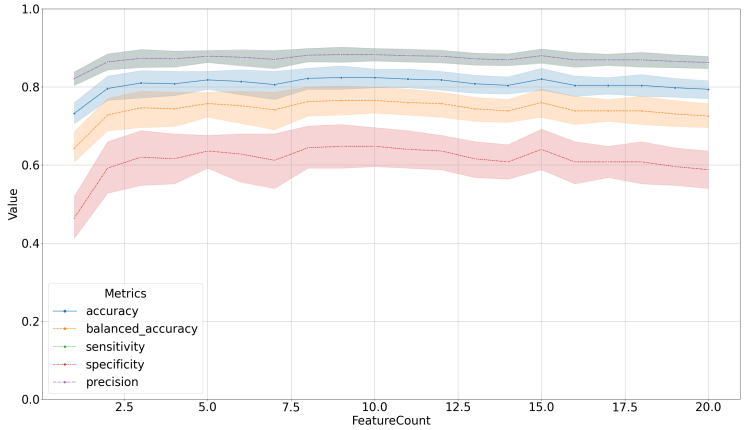
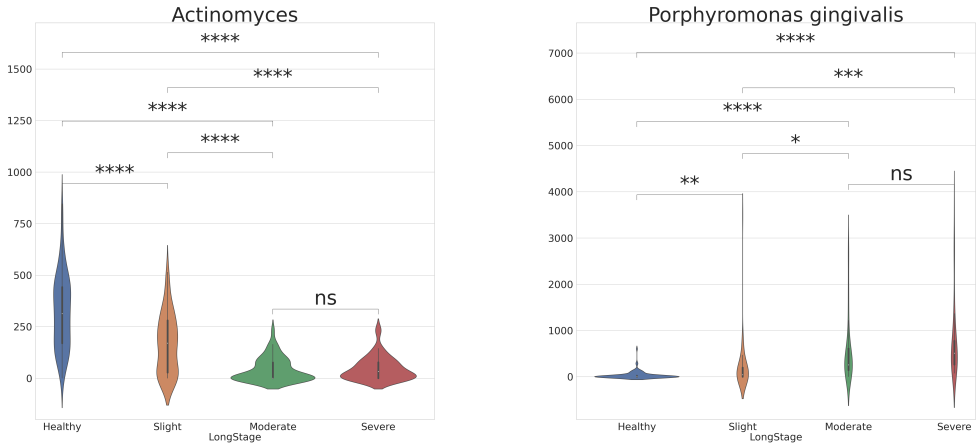


Figure 40: Metrics by Feature Count with DADA2 and HOMD



(a) *Porphyromonas gingivalis* (b) *Actinomyces*

Figure 41: Most and Second Most Important Features with DADA2 and HOMD

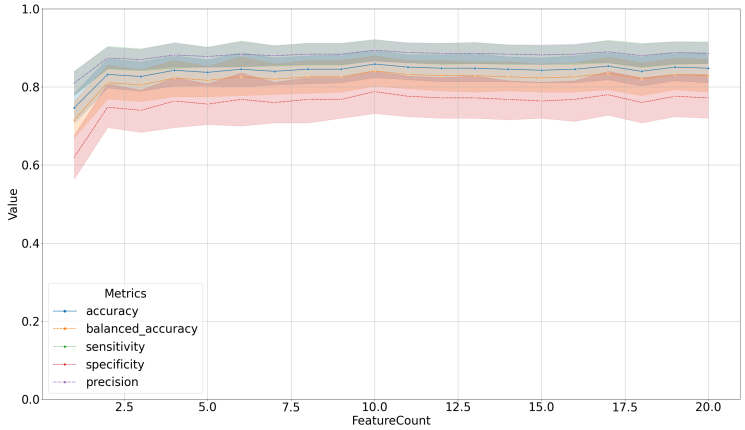
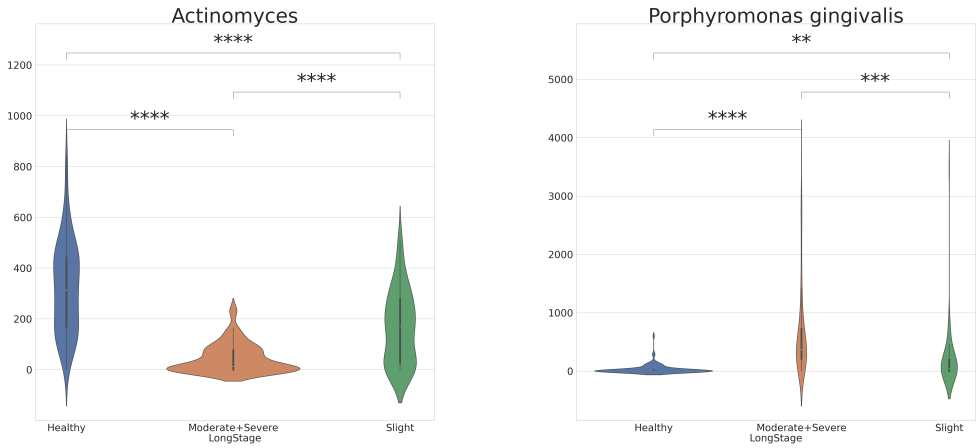


Figure 42: Metrics by Feature Count with Deblur and HOMD for Merging (Moderate+Severe) Classes

Table 13: Taxa with DADA2 and HOMD Ordered by Random Forest for Merging (Moderate+Severe) Classes

Order	Taxonomy Classification	Importances
0	Bacteria Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Actinomyces	0.33184235755114905
1	Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Porphyromonas gingivalis	0.32946935972610425
2	Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae XI Filifactor alocis	0.049427879731767196
3	Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella intermedia	0.04090774831842739
4	Bacteria Actinobacteria Actinobacteria Corynebacteriales Corynebacteriaceae Corynebacterium durum	0.0355991505558754
5	Bacteria Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Actinomyces graevenitzii	0.03406921215901854
6	Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Porphyromonas sp. HMT 285	0.026925634219470605
7	Bacteria Proteobacteria Epsilonproteobacteria Campylobacterales Campylobacteraceae Campylobacter showae	0.018680829306872235
8	Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema	0.01652121052090326
9	Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae XI Peptostreptococcaceae XIG-9 brachy	0.015815368803538402
10	Bacteria Firmicutes Clostridia Clostridiales Lachnospiraceae XIV Lachnospiraceae G-8 bacterium HMT 500	0.015255014883278186
11	Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema putidum	0.013803438286006666
12	Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella sp. HMT 304	0.012318276768942822
13	Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae XI Peptostreptococcaceae XIG-5 saphenum	0.011900138407501468
14	Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Tannerella forsythia	0.010750708299765504
15	Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae XI Peptostreptococcaceae XIG-6 nodatum	0.009403597999485315
16	Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella sp. HMT 526	0.007951045940208033
17	Bacteria Synergistetes Synergistia Synergistales Synergistaceae Fretibacterium	0.006881435611400768
18	Bacteria Firmicutes Mollicutes Mycoplasmatales Mycoplasmataceae Mycoplasma faucium	0.006351360361540656
19	Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema sp. HMT 260	0.006126232548744201



(a) *Porphyromonas gingivalis* (b) *Actinomyces*

Figure 43: Most and Second Most Important Features with Deblur and HOMD for Merging (Moderate+Severe) Classes

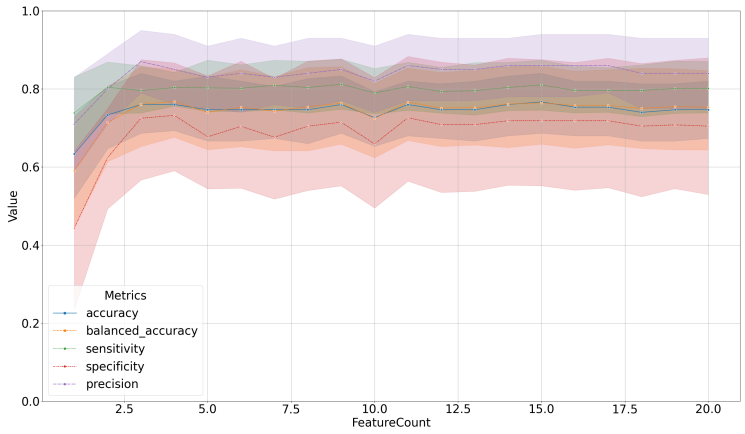
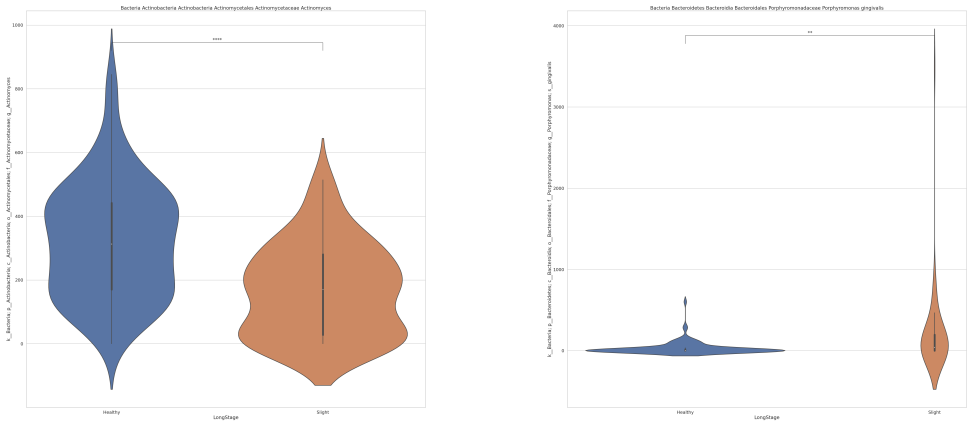


Figure 44: Metrics by Feature Count with DADA2 and HOMD for Healthy Class and Early Class Only



(a) *Porphyromonas gingivalis* (b) *Treponema denticola*

Figure 45: Most and Second Most Important Features with DADA2 and HOMD for Healthy Class and Early Class Only

Table 14: Taxa with DADA2 and HOMD Ordered by Random Forest for Healthy Class and Early Class Only

Order	Taxonomy Classification	Importances
0	Bacteria Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Actinomyces	0.3800329614200073
1	Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Porphyromonas gingivalis	0.11599155442053782
2	Bacteria Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Actinomyces graevenitzii	0.08377015225545748
3	Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae XI Filifactor alocis	0.07355851820527473
4	Bacteria Firmicutes Clostridia Clostridiales Lachnospiraceae XIV Lachnospiraceae G-8 bacterium HMT 500	0.055862102028897874
5	Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Porphyromonas sp. HMT 285	0.051266737996019496
6	Bacteria Proteobacteria Epsilonproteobacteria Campylobacterales Campylobacteraceae Campylobacter showae	0.03768573586643332
7	Bacteria Actinobacteria Actinobacteria Corynebacteriales Corynebacteriaceae Corynebacterium durum	0.032556906788312044
8	Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella intermedia	0.02510972430269449
9	Bacteria Synergistetes Synergistia Synergistales Synergistaceae Fretibacterium	0.020454891486853158
10	Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Tannerella forsythia	0.02023382279668738
11	Bacteria Firmicutes Mollicutes Mycoplasmatales Mycoplasmataceae Mycoplasma faucium	0.019175163953451355
12	Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema	0.01699482819483286
13	Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae XI Peptostreptococcaceae XIG-9 brachy	0.01474114962345347
14	Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema putidum	0.012552567408042328
15	Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella sp. HMT 526	0.009883103591414436
16	Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae XI Peptostreptococcaceae XIG-6 nodatum	0.009487308947517453
17	Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema sp. HMT 260	0.008833841116352032
18	Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae XI Peptostreptococcaceae XIG-5 saphenum	0.008254892599607861
19	Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella sp. HMT 304	0.0035540369981530828

Bray-Curtis distance index with Deblur has no statistically significant differences between (Healthy and Early) classes and (Moderate and Severe) classes (Table **??**, Figure **??** and Figure **??**). Moreover, Jaccard distance index with Deblur has no statistically significant difference between (Moderate and Severe) classes (Table **??**, Figure **??** and Figure **??**). Additionally, unweighted UniFrac distance index with Deblur has no statistically significant difference between (Moderate and Severe) classes (Table **??**, Figure **??** and Figure **??**). Furthermore, weighted UniFrac distance index with Deblur has no statistically significant difference between (Healthy and Early) classes (Table **??**, Figure **??** and Figure **??**).

As results of alpha-diversity indices, (Healthy and Early) classes and (Moderate and Severe) classes will be merged. Mercifully, alternative fact does totally not sustained by beta-divesity indices. Hence, (Healthy and Early) classes and (Moderate and Severe) classes, as mentioned herein-before, will be merged in classification.

5.3 t-SNE Plot

Overall distribution of taxa from each sample can be realized by t-SNE plot. If each class is evenly distributed on t-SNE plot, then the data might be difficult for classifying. In this manner, t-SNE plots with whole microbiome (Figure 28, Figure 29, Figure 30, Figure 31, Figure 32 and Figure 33) are more evenly distributed, whereas t-SNE plots with ANCOM selected microbiome data (Figure 34, Figure 35, Figure 36, Figure 37, Figure 38 and Figure 39) are biased by classes. *Id est*, data with ANCOM selected microbiome could result better in classifying. Hence, ANCOM selected taxa will be used in classification.

5.4 Random Forest Classifier

As results of Random Forest classifier, a feature or two features have dominant importance than others (Table **??**, Table **??**, Table **??**, Table **??** and Table 14). Thus, the two most important features are displayed for results of Random Forest classifier (Figure **??**, Figure **??**, Figure **??**, Figure **??** and Figure 45).

6 References

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., ... others (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems*, *2*(2).

Barrett, P., Hunter, J., Miller, J. T., Hsu, J.-C., & Greenfield, P. (2005). matplotlib–a portable python plotting package. In *Astronomical data analysis software and systems xiv* (Vol. 347, p. 91).

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C., Al-Ghalith, G. A., ... others (2018). *Qiime 2: Reproducible, interactive, scalable, and extensible microbiome data science* (Tech. Rep.). PeerJ Preprints.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... others (2019). Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature biotechnology*, *37*(8), 852–857.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). Dada2: high-resolution sample inference from illumina amplicon data. *Nature methods*, *13*(7), 581–583.

Chen, T., Yu, W.-H., Izard, J., Baranova, O. V., Lakshmanan, A., & Dewhirst, F. E. (2010). The human oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database*, *2010*.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... Andersen, G. L. (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, *72*(7), 5069–5072.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological conservation*, *61*(1), 1–10.

Flemmig, T. F. (1999). Periodontitis. *Annals of Periodontology*, *4*(1), 32–37.

Gihawi, A., Rallapalli, G., Hurst, R., Cooper, C. S., Leggett, R. M., & Brewer, D. S. (2019). Sepath: benchmarking the search for pathogens in human tissue whole genome sequence data leads to template pipelines. *Genome biology*, *20*(1), 1–15.

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., ... Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *science*, *312*(5778), 1355–1359.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, *9*(3), 90–95.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist*, *11*(2), 37–50.

James, F. C., & Rathbun, S. (1981). Rarefaction, relative abundance, and diversity of avian communities. *The Auk*, *98*(4), 785–800.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., & Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease*, *26*(1), 27663.

McDonald, D., Vázquez-Baeza, Y., Koslicki, D., McClelland, J., Reeve, N., Xu, Z., ... Knight, R. (2018). Striped unifrac: enabling microbiome analysis at unprecedented scale. *Nature methods*, *15*(11), 847–848.

McKinney, W., et al. (2011). pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, *14*(9).

Olsen, G. J., & Woese, C. R. (1993). Ribosomal rna: a key to phylogeny. *The FASEB journal*, *7*(1), 113–123.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Pielou, E. C. (1966). The measurement of diversity in different types of biological collections. *Journal of theoretical biology*, *13*, 131–144.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. *Nucleic acids research*, *35*(21), 7188–7196.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. Skr.*, *5*, 1–34.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, *449*(7164), 804–810.

Van Dyke, T. E., & Dave, S. (2005). Risk factors for periodontitis. *Journal of the International Academy of Periodontology*, *7*(1), 3.

Waskom, M., & the seaborn development team. (2020, September). *mwaskom/seaborn.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.592845` doi: 10.5281/zenodo.592845

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., ... others (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1), 27.