

# Periodontitis

Jaewoong Lee

Seunghoon Kim

Semin Lee

2020-12-29

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Microbiome . . . . .	5
1.2	Ribosomal RNA . . . . .	5
1.3	16S rRNA Gene Sequencing . . . . .	5
1.4	Periodontitis . . . . .	5
<b>2</b>	<b>Materials</b>	<b>5</b>
2.1	16S rRNA Gene Sequencing . . . . .	5
<b>3</b>	<b>Methods</b>	<b>5</b>
3.1	QIIME2 Workflow . . . . .	5
3.1.1	Denoising techniques . . . . .	5
3.1.2	Taxonomy Classification . . . . .	5
3.1.3	Merging Denoising and Taxonomy Classification . . . . .	5
3.1.4	Rarefaction . . . . .	5
3.1.5	Alpha-diversity . . . . .	8
3.1.6	Beta-diversity . . . . .	8
3.1.7	ANCOM . . . . .	8
3.2	Python Packages . . . . .	8
3.2.1	Pandas . . . . .	8
3.2.2	Scikit-learn . . . . .	8
3.2.3	Matplotlib . . . . .	8
3.2.4	Seaborn . . . . .	8
3.2.5	statannot . . . . .	9
3.3	t-SNE . . . . .	9
3.4	Classification . . . . .	9
3.4.1	Random Forest Classification . . . . .	9
<b>4</b>	<b>Results</b>	<b>9</b>
4.1	Quality Filter . . . . .	9
4.2	Rarefaction . . . . .	9
4.3	Alpha-diversity . . . . .	9
4.4	Beta-diversity . . . . .	17
4.5	ANCOM . . . . .	17
4.6	t-SNE Plot with Whole Microbiome . . . . .	17
4.7	t-SNE Plot with ANCOM Selected Microbiome Data . . . . .	17
4.8	Random Forest Classifier with Every Class . . . . .	17
4.9	Random Forest Classifier with Merging (Healthy+Early) Classes . . . . .	17
4.10	Random Forest Classifier with Merging (Moderate+Severe) Classes . . . . .	17
4.11	Random Forest Classifier with Merging (Healthy+Early) & (Moderate+Severe) Classes . . . . .	46
4.12	Random Forest Classifier with Healthy Class and Early Class Only . . . . .	46
<b>5</b>	<b>Discussion</b>	<b>46</b>
5.1	Alpha-diversity . . . . .	46
5.2	Beta-diversity . . . . .	46
5.3	t-SNE Plot . . . . .	50
5.4	Random Forest Classifier . . . . .	50
<b>6</b>	<b>References</b>	<b>50</b>

## List of Tables

1	Confusion Matrix . . . . .	10
2	Kruskal-Wallis Tests among All Group with DADA2 . . . . .	11
3	Kruskal-Wallis Tests from Evenness Index with DADA2 . . . . .	12
4	Kruskal-Wallis Tests from Faith PD Index with DADA2 . . . . .	13
5	Kruskal-Wallis Tests from Observed Features Index with DADA2 . . . . .	13
6	Kruskal-Wallis Tests from Shannon's Diversity Index with DADA2 . . . . .	13
7	Kruskal-Wallis Tests among All Group with Deblur . . . . .	13

8	Kruskal-Wallis Tests from Evenness Index with Deblur . . . . .	15
9	Kruskal-Wallis Tests from Faith PD Index with Deblur . . . . .	15
10	Kruskal-Wallis Tests from Observed Features Index with Deblur . . . . .	15
11	Kruskal-Wallis Tests from Shannon's Diversity Index with Deblur . . . . .	15
12	Bray-Curtis Distance Index with DADA2 . . . . .	18
13	Jaccard Distance Index with DADA2 . . . . .	18
14	Unweighted UniFrac Distance Index with DADA2 . . . . .	18
15	Weighted UniFrac Distance Index with DADA2 . . . . .	18
16	Bray-Curtis Distance Index with Deblur . . . . .	18
17	Jaccard Distance Index with Deblur . . . . .	18
18	Unweighted UniFrac Distance Index with Deblur . . . . .	23
19	Weighted UniFrac Distance Index with Deblur . . . . .	23
20	ANCOM Significant Taxa with DADA2 and GG . . . . .	27
21	ANCOM Significant Taxa with DADA2 and SILVA . . . . .	29
22	ANCOM Significant Taxa with DADA2 and HOMD . . . . .	30
23	ANCOM Significant Taxa with Deblur and GG . . . . .	31
24	ANCOM Significant Taxa with Deblur and SILVA . . . . .	33
25	ANCOM Significant Taxa with Deblur and HOMD . . . . .	34
26	Taxa with Deblur and HOMD Ordered by Random Forest . . . . .	41
27	Taxa with DADA2 and HOMD Ordered by Random Forest for Merging (Healthy+Early) Classes . . . . .	43
28	Taxa with Deblur and HOMD Ordered by Random Forest for Merging (Moderate+Severe) Classes . . . . .	45
29	Taxa with DADA2 and SILVA Ordered by Random Forest for Merging (Healthy+Early) & (Moderate+Severe) Classes . . . . .	47
30	Taxa with DADA2 and HOMD Ordered by Random Forest for Healthy Class and Early Class Only . . . . .	49

## List of Figures

1	Concept of a Core Human Microbiome (Turnbaugh et al., 2007) . . . . .	6
2	A Theoretic Overview of QIIME2 Workflow (Bolyen et al., 2019, 2018) . . . . .	6
3	Denoising Techniques which provided by QIIME2 . . . . .	7
4	Taxonomy Classification which provided by QIIME2 . . . . .	7
5	Example Diagram for Merging Denoising and Taxonomy Classification . . . . .	7
6	Example ANCOM Volcano Plot which Provided by QIIME2 (Bolyen et al., 2019, 2018) . . . . .	10
7	Visualization by t-SNE (Maaten & Hinton, 2008) . . . . .	10
8	Workflow of Classification . . . . .	10
9	Deciding the Best Features . . . . .	11
10	Random Forest Classifier Workflow . . . . .	11
11	Random Forest Classifier Workflow with Merging . . . . .	11
12	Sequence Quality Plot . . . . .	11
13	Frequency and Number per Sample by DADA2 . . . . .	12
14	Frequency and Number per Sample by Deblur . . . . .	12
15	Evenness Index from DADA2 . . . . .	13
16	Faith PD Index from DADA2 . . . . .	14
17	Observed Features Index from DADA2 . . . . .	14
18	Shannon's Diversity Index from DADA2 . . . . .	14
19	Evenness Index from Deblur . . . . .	15
20	Faith PD Index from Deblur . . . . .	16
21	Observed Features Index from Deblur . . . . .	16
22	Shannon's Diversity Index from Deblur . . . . .	16
23	t-SNE Plot from Bray-Curtis Distance Index with DADA2 . . . . .	19
24	Bray-Curtis Distance Index with DADA2 . . . . .	19
25	t-SNE Plot from Jaccard Distance Index with DADA2 . . . . .	20
26	Jaccard Distance Index with DADA2 . . . . .	20
27	t-SNE Plot from Unweighted UniFrac Distance Index with DADA2 . . . . .	21
28	Unweighted UniFrac Distance Index with DADA2 . . . . .	21
29	t-SNE Plot from Weighted UniFrac Distance Index with DADA2 . . . . .	22
30	Weighted UniFrac Distance Index with DADA2 . . . . .	22
31	t-SNE Plot from Bray-Curtis Distance Index with Deblur . . . . .	23
32	Bray-Curtis Distance Index with Deblur . . . . .	24
33	t-SNE Plot from Jaccard Distance Index with Deblur . . . . .	24

34	Jaccard Distance Index with Deblur . . . . .	25
35	t-SNE Plot from Unweighted UniFrac Distance Index with Deblur . . . . .	25
36	Unweighted UniFrac Distance Index with Deblur . . . . .	26
37	t-SNE Plot from Weighted UniFrac Distance Index with Deblur . . . . .	26
38	Weighted UniFrac Distance Index with Deblur . . . . .	27
39	ANCOM Volcano Plot with DADA2 and GG . . . . .	28
40	ANCOM Volcano Plot with DADA2 and HOMD . . . . .	28
41	ANCOM Volcano Plot with DADA2 and SILVA . . . . .	28
42	ANCOM Volcano Plot with Deblur and GG . . . . .	32
43	ANCOM Volcano Plot with Deblur and SILVA . . . . .	32
44	ANCOM Volcano Plot with Deblur and HOMD . . . . .	32
45	t-SNE Plot with Whole Microbiome from DADA2 and GG (328 taxa) . . . . .	35
46	t-SNE Plot with Whole Microbiome from DADA2 and SILVA (633 taxa) . . . . .	35
47	t-SNE Plot with Whole Microbiome from DADA2 and HOMD (425 taxa) . . . . .	36
48	t-SNE Plot with Whole Microbiome from Deblur and GG (232 taxa) . . . . .	36
49	t-SNE Plot with Whole Microbiome from Deblur and SILVA (414 taxa) . . . . .	37
50	t-SNE Plot with Whole Microbiome from Deblur and HOMD (235 taxa) . . . . .	37
51	t-SNE Plot with ANCOM Selected Microbiome Data from DADA2 and GG (15 taxa) . . . . .	38
52	t-SNE Plot with ANCOM Selected Microbiome Data from DADA2 and SILVA (23 taxa) . . . . .	38
53	t-SNE Plot with ANCOM Selected Microbiome Data from DADA2 and HOMD (20 taxa) . . . . .	39
54	t-SNE Plot with ANCOM Selected Microbiome Data from Deblur and GG (27 taxa) . . . . .	39
55	t-SNE Plot with ANCOM Selected Microbiome Data from Deblur and SILVA (20 taxa) . . . . .	40
56	t-SNE Plot with ANCOM Selected Microbiome Data from Deblur and HOMD (28 taxa) . . . . .	40
57	Metrics by Feature Count with Deblur and HOMD . . . . .	42
58	Most and Second Most Important Features with Deblur and HOMD . . . . .	42
59	Metrics by Feature Count with DADA2 and HOMD for Merging (Healthy+Early) Classes . . . . .	42
60	Most and Second Most Important Features with DADA2 and HOMD for Merging (Healthy+Early) Classes . . . . .	44
61	Metrics by Feature Count with Deblur and HOMD for Merging (Moderate+Severe) Classes . . . . .	44
62	Most and Second Most Important Features with Deblur and HOMD for Merging (Moderate+Severe) Classes . . . . .	44
63	Metrics by Feature Count with DADA2 and SILVA for Merging (Healthy+Early) & (Moderate+Severe) Classes . . . . .	48
64	Most and Second Most Important Features with DADA2 and HOMD for Merging (Healthy+Early) & (Moderate+Severe) Classes . . . . .	48
65	Metrics by Feature Count with DADA2 and HOMD for Healthy Class and Early Class Only . . . . .	48
66	Most and Second Most Important Features with DADA2 and HOMD for Healthy Class and Early Class Only . . . . .	52

# 1 Introduction

## 1.1 Microbiome

Microbiome is consist of microbiota, the micro-organisms which live inside and on humans (Turnbaugh et al., 2007). Microbiome is also about  $10^{13}$  micro-organisms whose which 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 which effecting periodontium that surround and support teeth. Major components of periodontitis are clinical attachment loss and bone loss (Flemmig, 1999). Previous study 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 shown as 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). Major difference between DADA2 and Deblur, as shown as figure 3, is a strategy, the strategy used to divide as different species. DADA2 uses amplicon sequence variants (ASVs), strictly divides sequences even one-base mismatch. However, Deblur uses operational taxonomic units (OTUs), considers as same sequence when sequences are 97 % or more matched.

### 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). Major difference among these databases is resolution. Resolution of GG and HOMD is from kingdom to species; however, resolution of SILVA is from domain to genus. Previous research have found that a higher accuracy at taxonomic levels above genus level; but accuracy drops at species level (Gihawi et al., 2019).

### 3.1.3 Merging Denoising and Taxonomy Classification

After denoising and taxonomy classification steps, some different IDs (ASVs or OTUs) have been identified as same taxonomy. In that case, the different IDs will be merged into one taxonomy (Figure 5).

### 3.1.4 Rarefaction

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

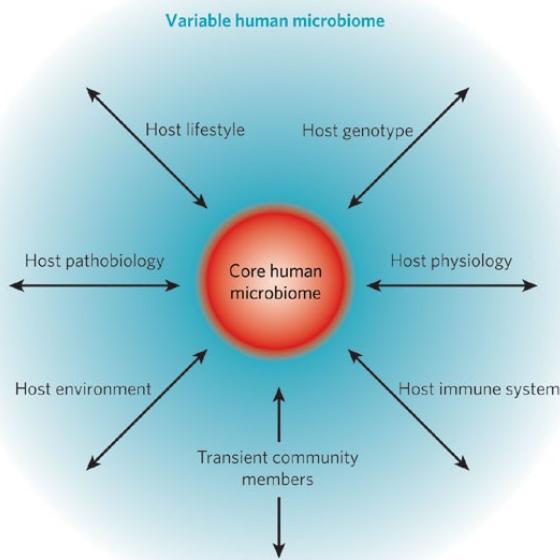


Figure 1: Concept of a Core Human Microbiome (Turnbaugh et al., 2007)

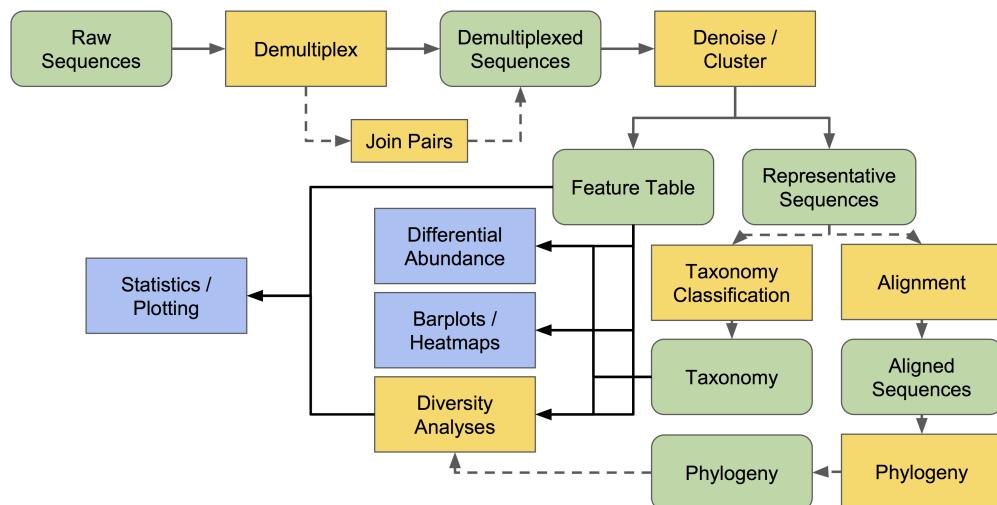


Figure 2: A Theoretic Overview of QIIME2 Workflow (Bolyen et al., 2019, 2018)

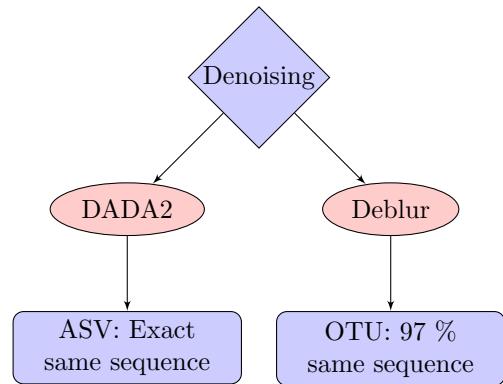


Figure 3: Denoising Techniques which provided by QIIME2

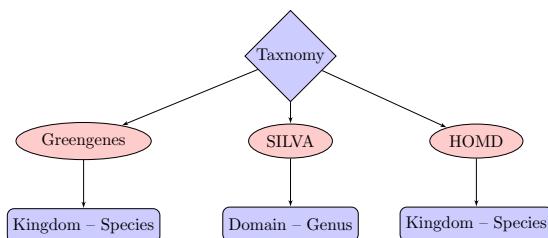


Figure 4: Taxonomy Classification which provided by QIIME2

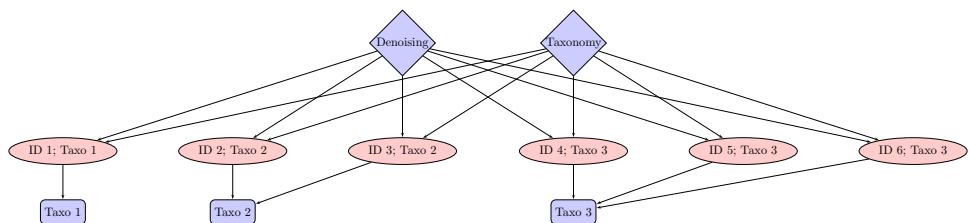


Figure 5: Example Diagram for Merging Denoising and Taxonomy Classification

### **3.1.5 Alpha-diversity**

Alpha-diversity is a metric which shows the richness of taxa at a single community. There are four alpha-diversity indices which provided from QIIME2:

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

Evenness index shows a measurement of diversity in different type at community (Pielou, 1966); Faith's phylogenetic diversity, however, indicates a qualitative measurement of community richness which priorities for species conservation which incorporates with taxic diversity (Faith, 1992). Observed features, as its name, is a number of observed features in microbiome. Moreover, Shannon's diversity index means a significant aspect of community richness (Shannon, 1948).

### **3.1.6 Beta-diversity**

Beta-diversity is a metric which indicates the taxonomic differentiation between multiple communities. There are four beta-diversity indices which provided from QIIME2:

- 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 a measurement of local distribution among communities. UniFrac distance indices reveal measurements of phylogenetic distances (McDonald et al., 2018). Difference between unweighted UniFrac distance index and weighted UniFrac distance index is a qualitative and a quantitative, respectively.

### **3.1.7 ANCOM**

ANCOM (Analysis of composition of microbiomes) can be used for analyzing the composition of microbiome in multiple populations (Mandal et al., 2015). Example ANCOM volcano plot is shows as 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 a location in 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 supervised learning which identifies a class of a new observation, depends on given information which consist of training observations and their classes.

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).

### 4.3 Alpha-diversity

Alpha-diversity analysis with DADA2 was done: Evenness index (Table 3 and Figure 15), Faith PD (Table 4 and Figure 16), observed feature index (Table 5 and Figure 17) and Shannon's diversity index (Table 6 and Figure 18). Also, alpha-diversity analysis with DADA2 was done: Evenness index (Table 8 and Figure 19), Faith PD (Table 9 and Figure 20), observed feature index (Table 10 and Figure 21) and Shannon's diversity index (Table 11 and Figure 22). Moreover, Kruskal-Wallis tests among all groups are shown as table 2 (with DADA2) and table 7 (with Deblur).

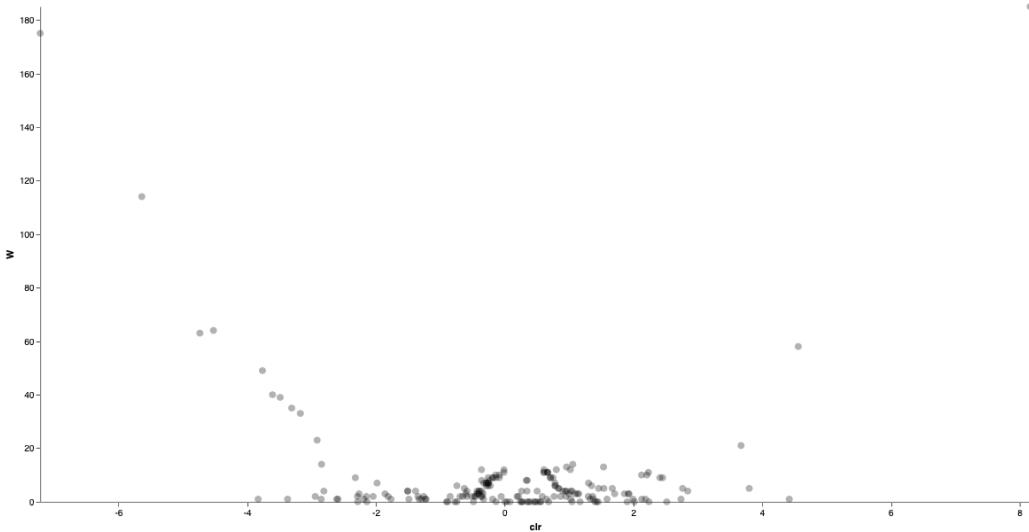


Figure 6: Example ANCOM Volcano Plot which Provided by QIIME2 (Bolyen et al., 2019, 2018)

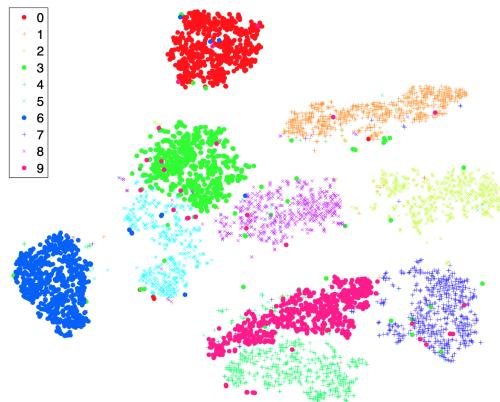


Figure 7: Visualization by t-SNE (Maaten & Hinton, 2008)

Table 1: Confusion Matrix

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

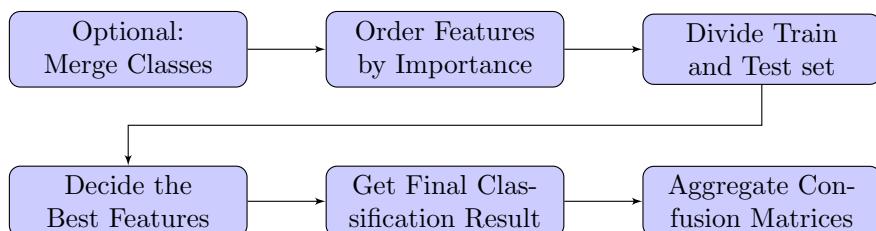


Figure 8: Workflow of Classification

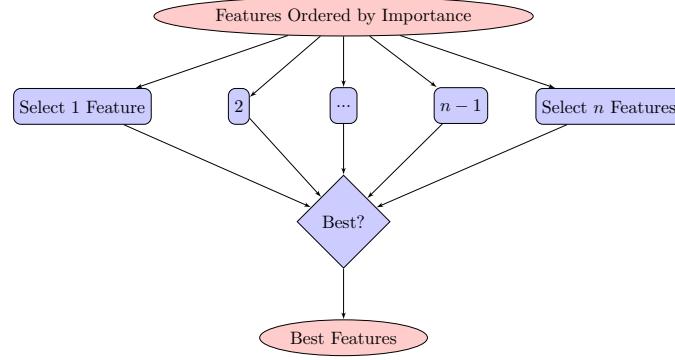


Figure 9: Deciding the Best Features

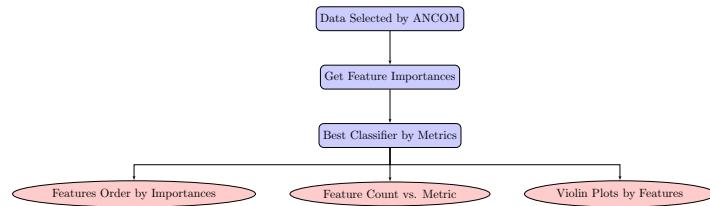


Figure 10: Random Forest Classifier Workflow

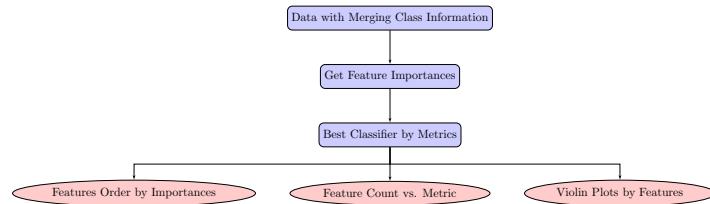


Figure 11: Random Forest Classifier Workflow with Merging

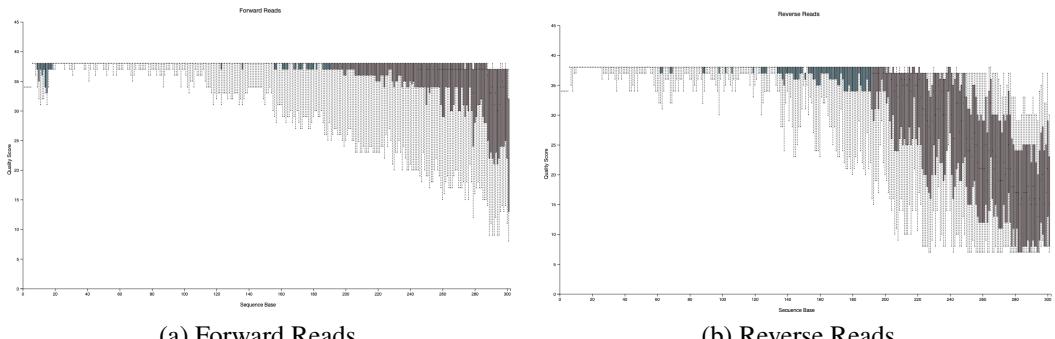


Figure 12: Sequence Quality Plot

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
Shannon's Diversity	7.311350438247132	0.06260902704190516

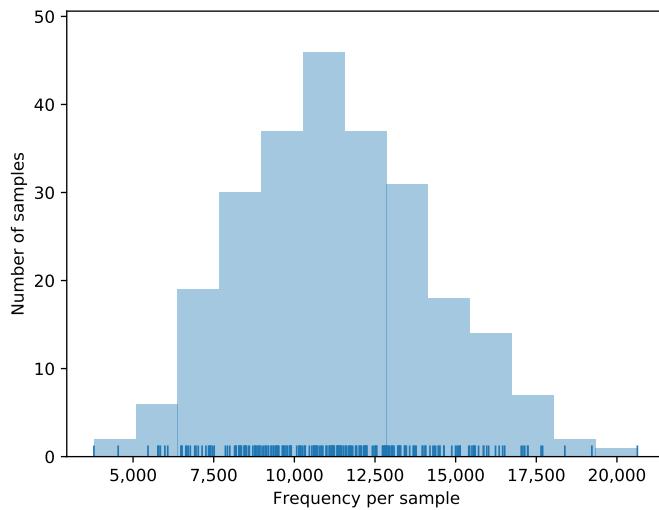


Figure 13: Frequency and Number per Sample by DADA2

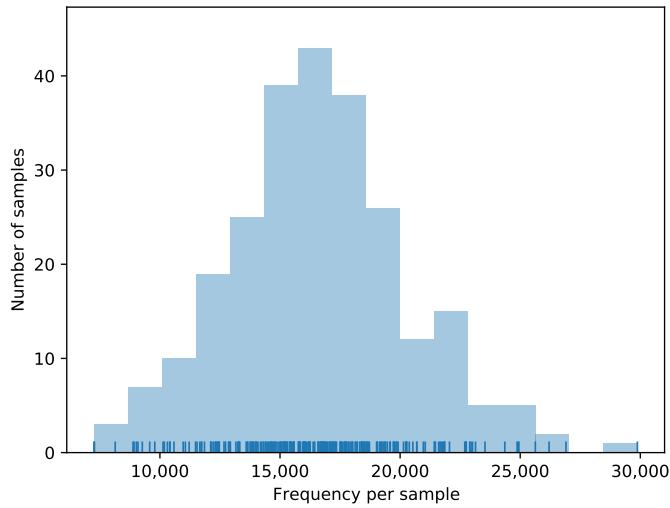


Figure 14: Frequency and Number per Sample by Deblur

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

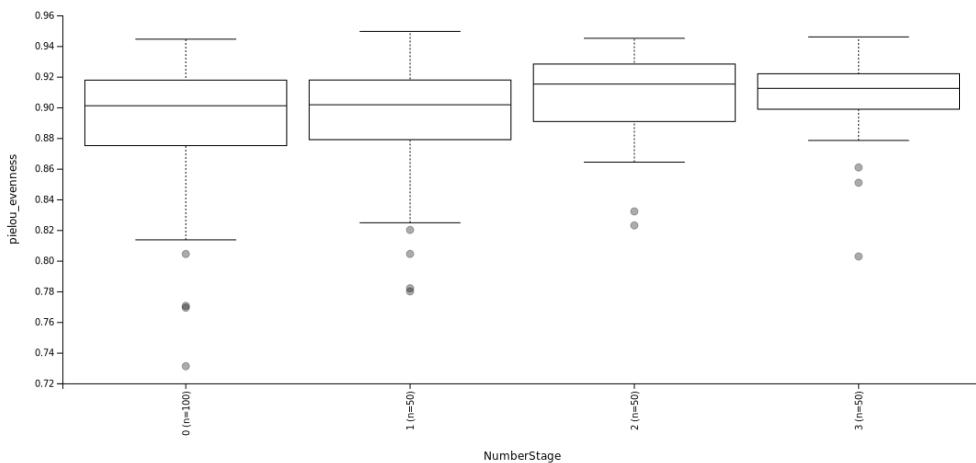


Figure 15: Evenness Index from DADA2

Table 7: Kruskal-Wallis Tests among All Group with Deblur

Alpha-Diversity	H	p-value
Evenness	9.242885737051779	0.026229960554059864
Faith PD	87.83605864541846	6.386769940789011e-19
Observed Features	59.59138364929631	7.186872791755095e-13
Shnnon's Diversity	24.823351075697246	0.000016810908296023026

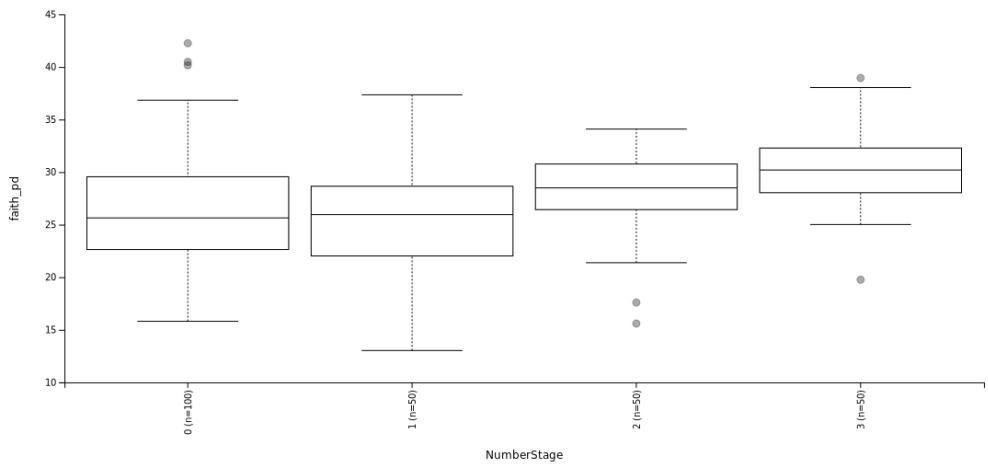


Figure 16: Faith PD Index from DADA2

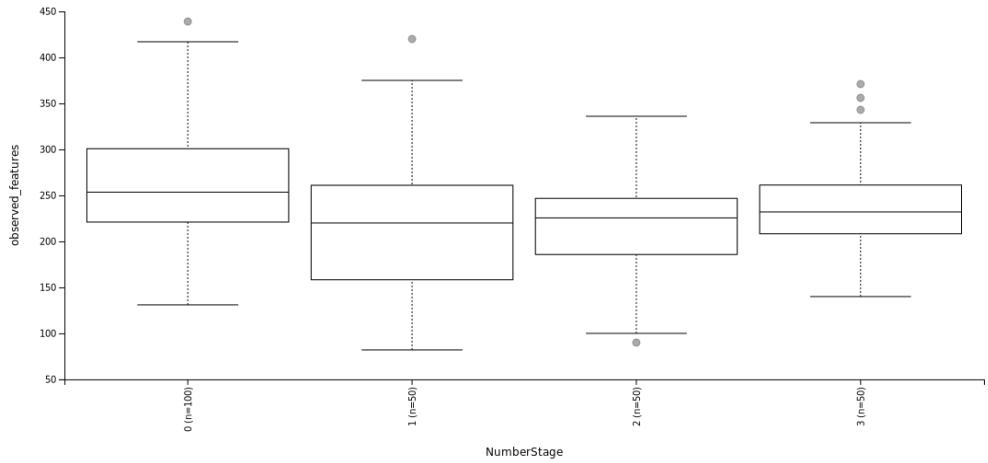


Figure 17: Observed Features Index from DADA2

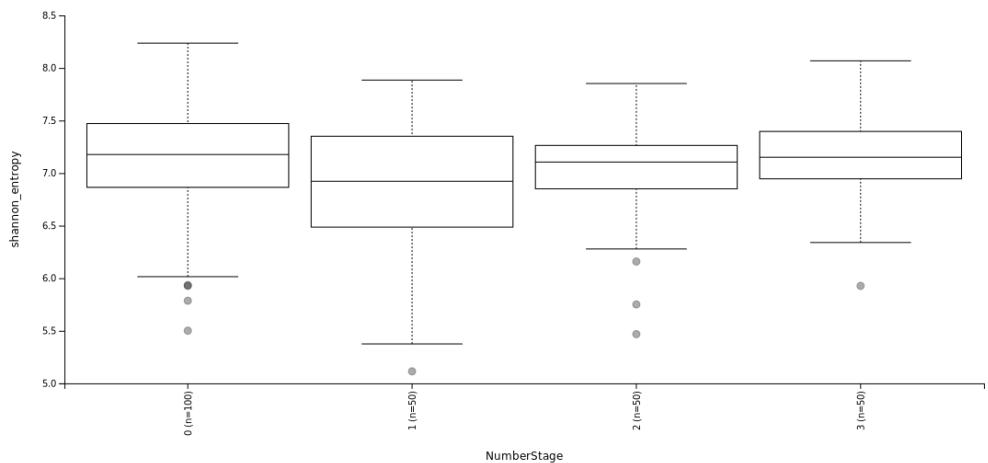


Figure 18: Shannon's Diversity Index from DADA2

Table 8: Kruskal-Wallis Tests from Evenness Index with Deblur

Group 1	Group 2	H	p-value	q-value
Slight (n=50)	Healthy (n=100)	2.884386754966897	0.0894420544121846	0.15829564582637523
Slight (n=50)	Moderate (n=50)	4.392047524752456	0.03610692636685824	0.10832077910057474
Slight (n=50)	Severe (n=50)	8.828245544554477	0.002966034055389358	0.017796204332336148
Healthy (n=100)	Moderate (n=50)	0.6168317880794802	0.43222705558822094	0.43597874518665736
Healthy (n=100)	Severe (n=50)	2.6199099337748066	0.1055304305509168	0.15829564582637523
Moderate (n=50)	Severe (n=50)	0.6068435643564385	0.43597874518665736	0.43597874518665736

Table 9: Kruskal-Wallis Tests from Faith PD Index with Deblur

Group 1	Group 2	H	p-value	q-value
Slight (n=50)	Healthy (n=100)	2.7110304635762077	0.09965659889456922	0.11958791867348306
Slight (n=50)	Moderate (n=50)	26.80400792079206	2.251698564500841e-07	3.3775478467512613e-07
Slight (n=50)	Severe (n=50)	29.06252673267329	7.007948881210323e-08	1.4015897762420645e-07
Healthy (n=100)	Moderate (n=50)	51.153949668874134	8.539868055189094e-13	2.5619604165567283e-12
Healthy (n=100)	Severe (n=50)	54.86883178807949	1.288482355374052e-13	7.730894132244311e-13
Moderate (n=50)	Severe (n=50)	0.005750495049483106	0.9395527422741722	0.9395527422741722

Table 10: Kruskal-Wallis Tests from Observed Features Index with Deblur

Group 1	Group 2	H	p-value	q-value
Slight (n=50)	Healthy (n=100)	0.4675226919952207	0.49412905906624816	0.5929548708794977
Slight (n=50)	Moderate (n=50)	18.684815977243918	1.542055834477253e-05	2.31308375171588e-05
Slight (n=50)	Severe (n=50)	20.703272962949605	5.362426456004328e-06	1.0724852912008657e-05
Healthy (n=100)	Moderate (n=50)	35.26606516292951	2.875998708064018e-09	8.627996124192055e-09
Healthy (n=100)	Severe (n=50)	37.015293460828644	1.1720632904898772e-09	7.032379742939263e-09
Moderate (n=50)	Severe (n=50)	0.003849966992737873	0.9505245257136643	0.9505245257136643

Table 11: Kruskal-Wallis Tests from Shannon's Diversity Index with Deblur

Group 1	Group 2	H	p-value	q-value
Slight (n=50)	Healthy (n=100)	0.3867973509933357	0.5339876723058008	0.6407852067669609
Slight (n=50)	Moderate (n=50)	10.767968316831627	0.0010327180791227218	0.0020654361582454436
Slight (n=50)	Severe (n=50)	14.428562376237608	0.00014557751137778065	0.000627545643904027
Healthy (n=100)	Moderate (n=50)	10.172185430463571	0.0014257517732722547	0.002138627659908382
Healthy (n=100)	Severe (n=50)	13.746754966887409	0.0002091818813013423	0.000627545643904027
Moderate (n=50)	Severe (n=50)	0.15987326732670226	0.6892732232396639	0.6892732232396639

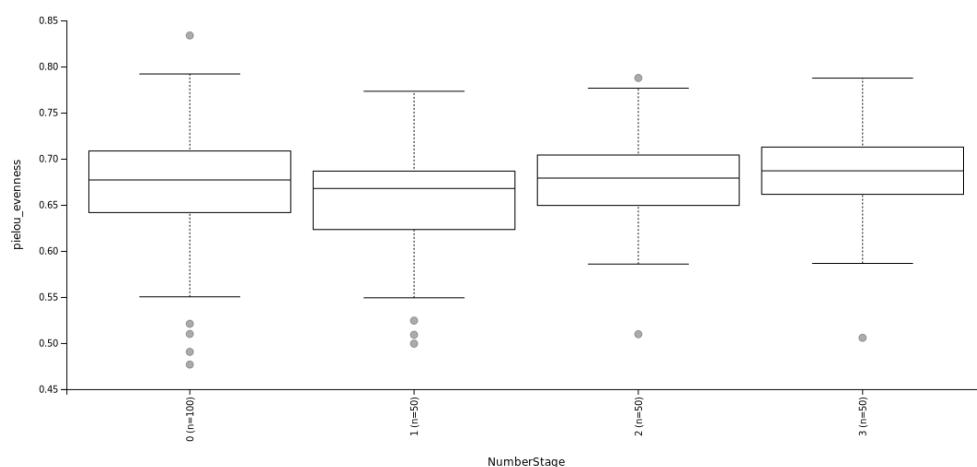


Figure 19: Evenness Index from Deblur

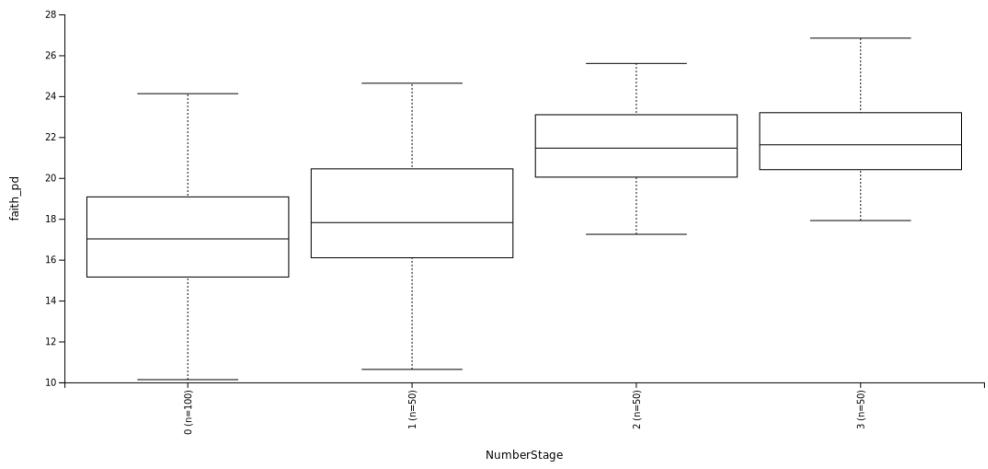


Figure 20: Faith PD Index from Deblur

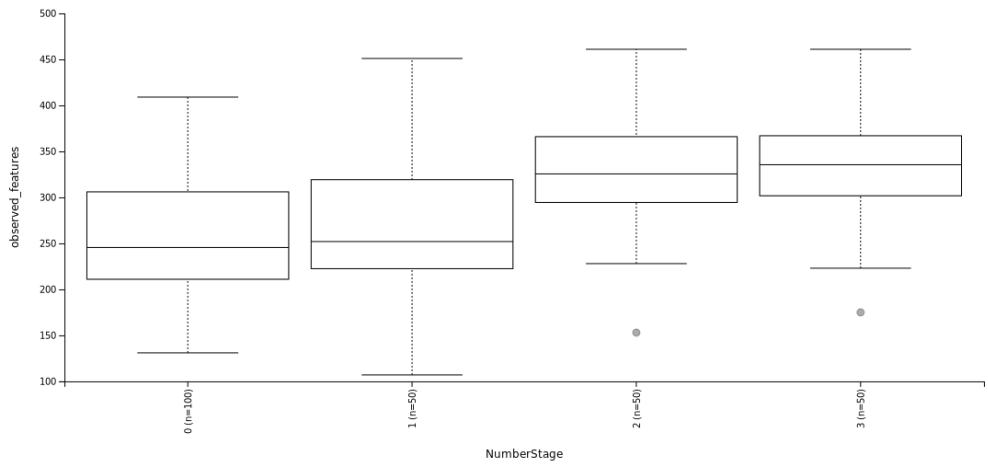


Figure 21: Observed Features Index from Deblur

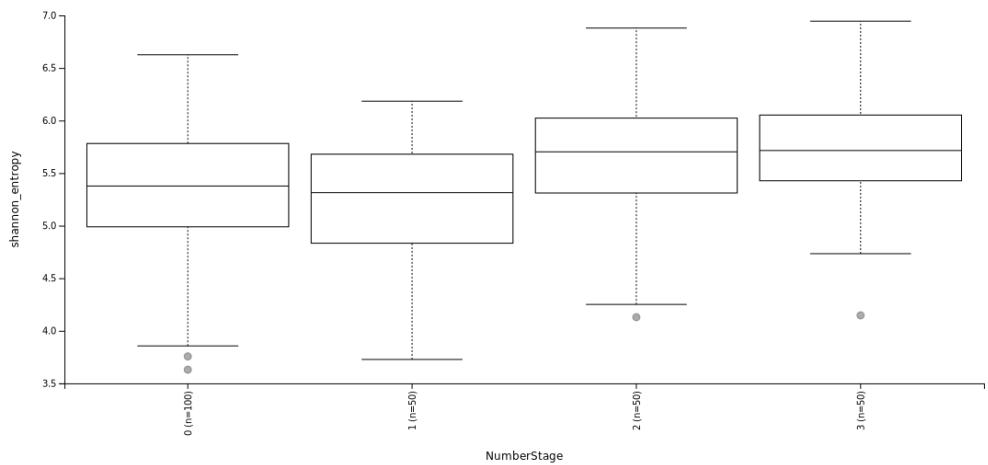


Figure 22: Shannon's Diversity Index from Deblur

#### 4.4 Beta-diversity

Beta-diversity analysis with DADA2 was done: Bray-Curtis distance (Table 12 and Figure 24), Jaccard distance (Table 13 and Figure 26), unweighted UniFrac distance (Table 14 and Figure 28) and weighted UniFrac distance (Table 15 and Figure 28). Also, beta-diversity analysis with Deblur was done: Bray-Curtis distance (Table 16 and Figure 32), Jaccard distance (Table 17 and Figure 34), unweighted UniFrac distance (Table 18 and Figure 36) and weighted UniFrac distance (Table 19 and Figure 36).

#### 4.5 ANCOM

Statistically significant different taxa and volcano plots by ANCOM were derived: DADA2 and GG (Table 20 and Figure 39), DADA2 and SILVA (Table 21 and Figure 40), DADA2 and HOMD (Table 22 and Figure 41), Deblur and GG (Table 23 and Figure 42), Deblur and SILVA (Table 24 and Figure 43) and Deblur and HOMD (Table 25 and Figure 44).

#### 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 45), 633 taxa from DADA2 and SILVA (Figure 46), 425 taxa from DADA2 and HOMD (Figure 47), 232 taxa from Deblur and GG (Figure 48), 414 taxa from Deblur and SILVA (Figure 49) and 235 taxa from Deblur and HOMD (Figure 50).

#### 4.7 t-SNE Plot with ANCOM Selected Microbiome Data

As whole microbiome data, ANCOM selected microbiome data are also multi-dimensional data, even though their columns are selected by ANCOM. Hence, with t-SNE, ANCOM selected microbiome data have also been deflated their dimension: 15 taxa (as Table 20) from DADA2 and GG (Figure 51), 23 taxa (as Table 21) from DADA2 and SILVA (Figure 52), 20 taxa (as Table 22) from DADA2 and HOMD (Figure 53), 27 taxa (as Table 23) from Deblur and GG (Figure 48), 20 taxa (as Table 24) from Deblur and SILVA (Figure 55) and 28 taxa (as Table 25) from Deblur and HOMD (Figure 56).

#### 4.8 Random Forest Classifier with Every Class

As figure 3 and figure 4, there are six combinations. Thus, classification algorithm is carried out on these six combinations (Figure 10). Among these six combinations, Deblur and HOMD pair gives the best result in terms of balanced accuracy. Features which are ordered by importance are in table 26. Also, five metrics by feature count are shown as figure 57; then, the highest value of balanced accuracy is 0.778 with using 13 taxa. Moreover, violin plots from the two most important taxa, *Porphyromonas gingivalis* and *Actinomyces*, are in figure 58.

#### 4.9 Random Forest Classifier with Merging (Healthy+Early) Classes

As figure 3 and figure 4, there are six combinations. However, there is no statistically significant difference between Healthy and Early classes. Thus, classification algorithm is carried out on these six combinations with merging Healthy and Early classes. Among these six combinations, DADA2 and HOMD pair gives the best result in terms of balanced accuracy. Features which are ordered by importance are in table 27. Also, five metrics by feature count are shown as figure 59; then, the highest value of balanced accuracy is 0.938 with using 11 taxa. Moreover, violin plots from the two most important taxa, *Porphyromonas gingivalis* and *Actinomyces*, are in figure 60.

#### 4.10 Random Forest Classifier with Merging (Moderate+Severe) Classes

As figure 3 and figure 4, there are six combinations. However, there is no statistically significant difference between Moderate and Severe classes. Thus, classification algorithm is carried out on these six combinations with merging Moderate and Severe classes. Among these six combinations, Deblur and HOMD pair gives the best result in terms of balanced accuracy. Features which are ordered by importance are in table 28. Also, five metrics by feature count are shown as figure 61; then, the highest value of balanced accuracy is 0.777 with using 28 taxa. Moreover, violin plots from the two most important taxa, *Porphyromonas gingivalis* and *Actinomyces*, are in figure 62.

Table 12: 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

Table 13: 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 14: 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 15: 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

Table 16: Bray-Curtis Distance Index with Deblur

Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
Slight	Healthy	150	999	1.7634974220433302	0.019	0.022799999999999997
Slight	Moderate	100	999	3.203442604434298	0.001	0.0015
Slight	Severe	100	999	4.192790849454974	0.001	0.0015
Healthy	Moderate	150	999	6.953487468508356	0.001	0.0015
Healthy	Severe	150	999	7.5433379986347155	0.001	0.0015
Moderate	Severe	100	999	1.0959020597220626	0.313	0.313

Table 17: Jaccard Distance Index with Deblur

Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
Slight	Healthy	150	999	1.3701039884255466	0.001	0.0012
Slight	Moderate	100	999	2.198029993855521	0.001	0.0012
Slight	Severe	100	999	2.237738583770674	0.001	0.0012
Healthy	Moderate	150	999	4.528432929980079	0.001	0.0012
Healthy	Severe	150	999	4.374635292015638	0.001	0.0012
Moderate	Severe	100	999	1.0036296853126103	0.429	0.429

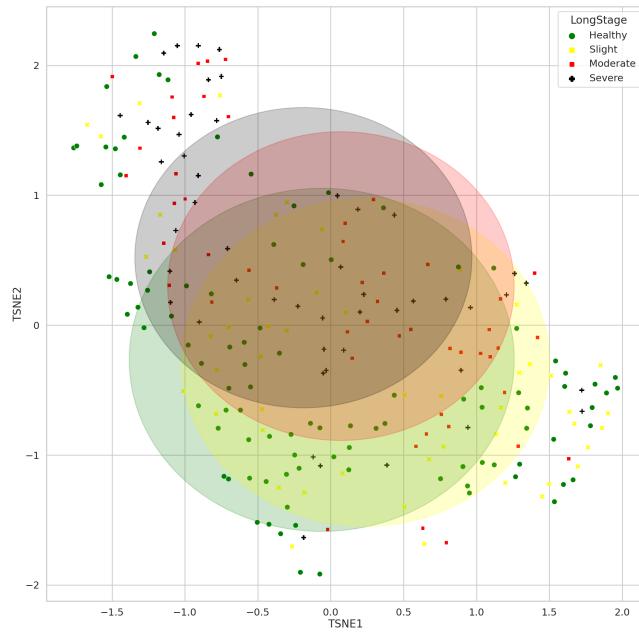


Figure 23: t-SNE Plot from Bray-Curtis Distance Index with DADA2

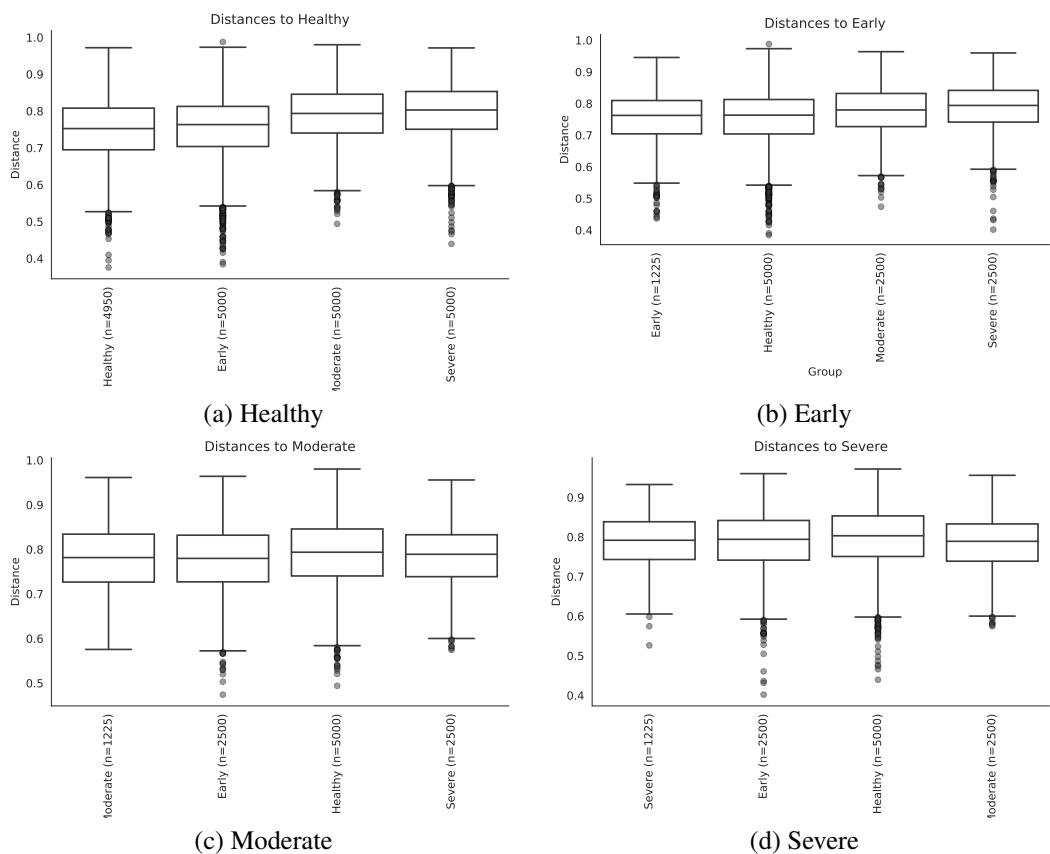


Figure 24: Bray-Curtis Distance Index with DADA2

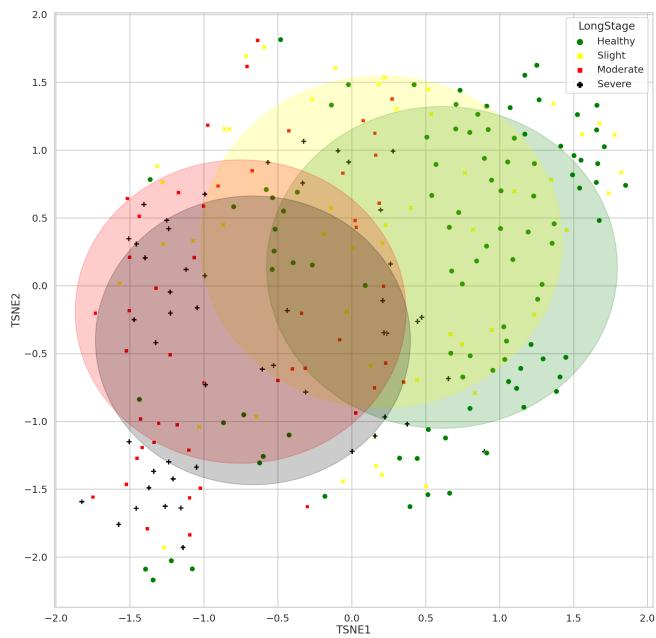


Figure 25: t-SNE Plot from Jaccard Distance Index with DADA2

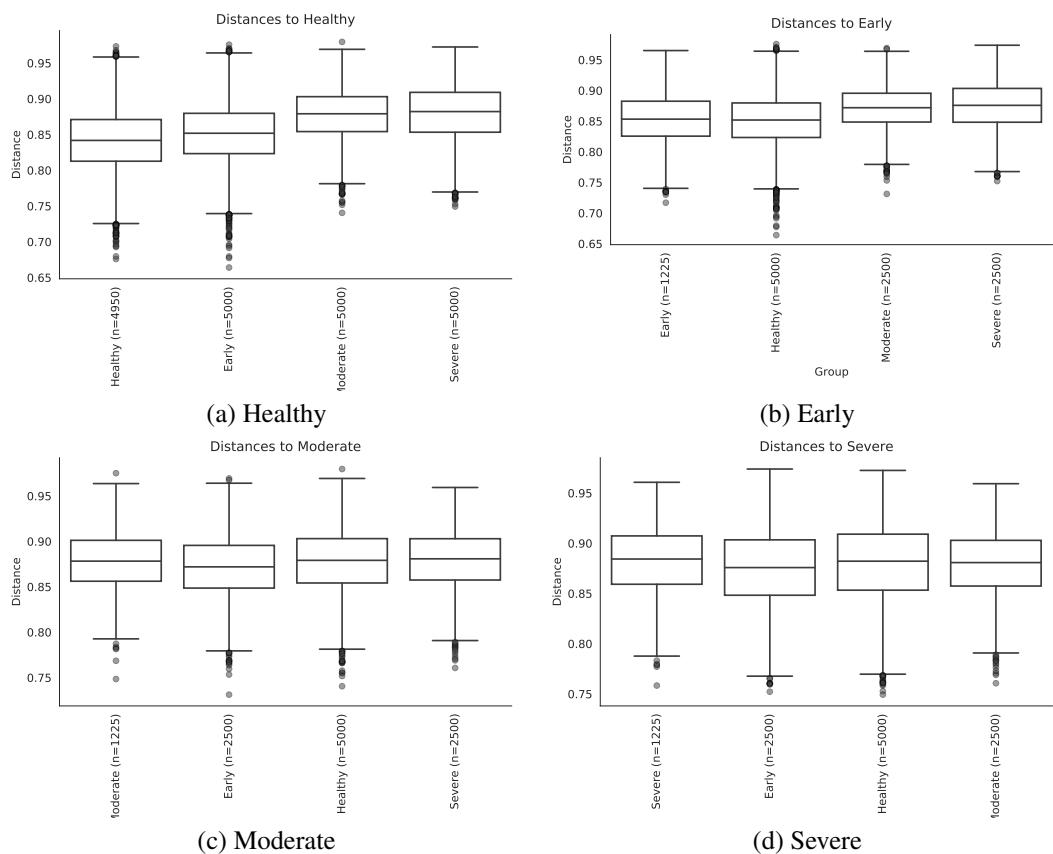


Figure 26: Jaccard Distance Index with DADA2

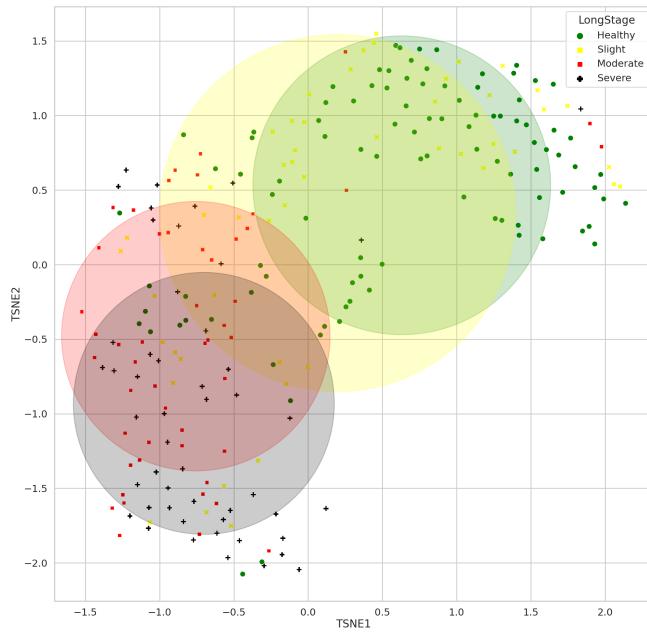


Figure 27: t-SNE Plot from Unweighted UniFrac Distance Index with DADA2

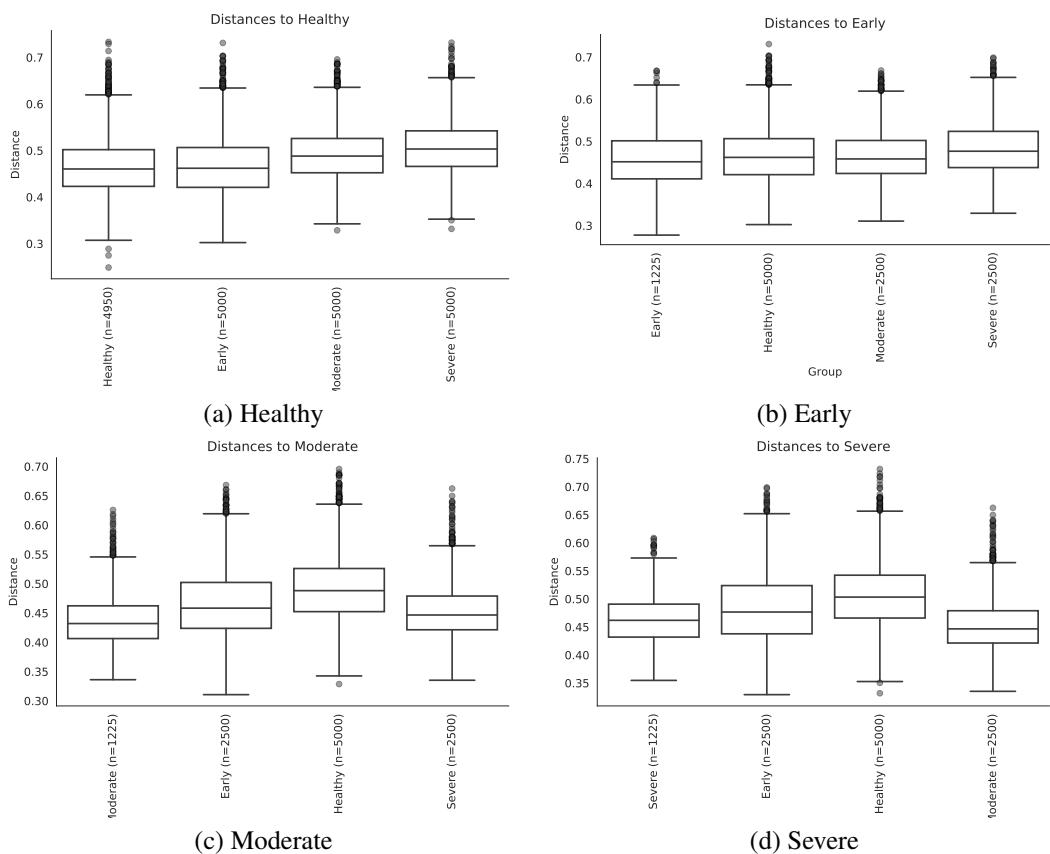


Figure 28: Unweighted UniFrac Distance Index with DADA2

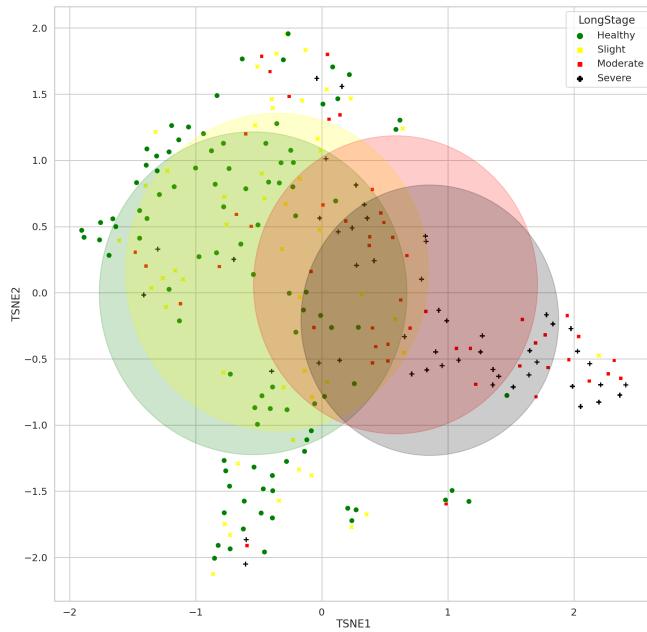


Figure 29: t-SNE Plot from Weighted UniFrac Distance Index with DADA2

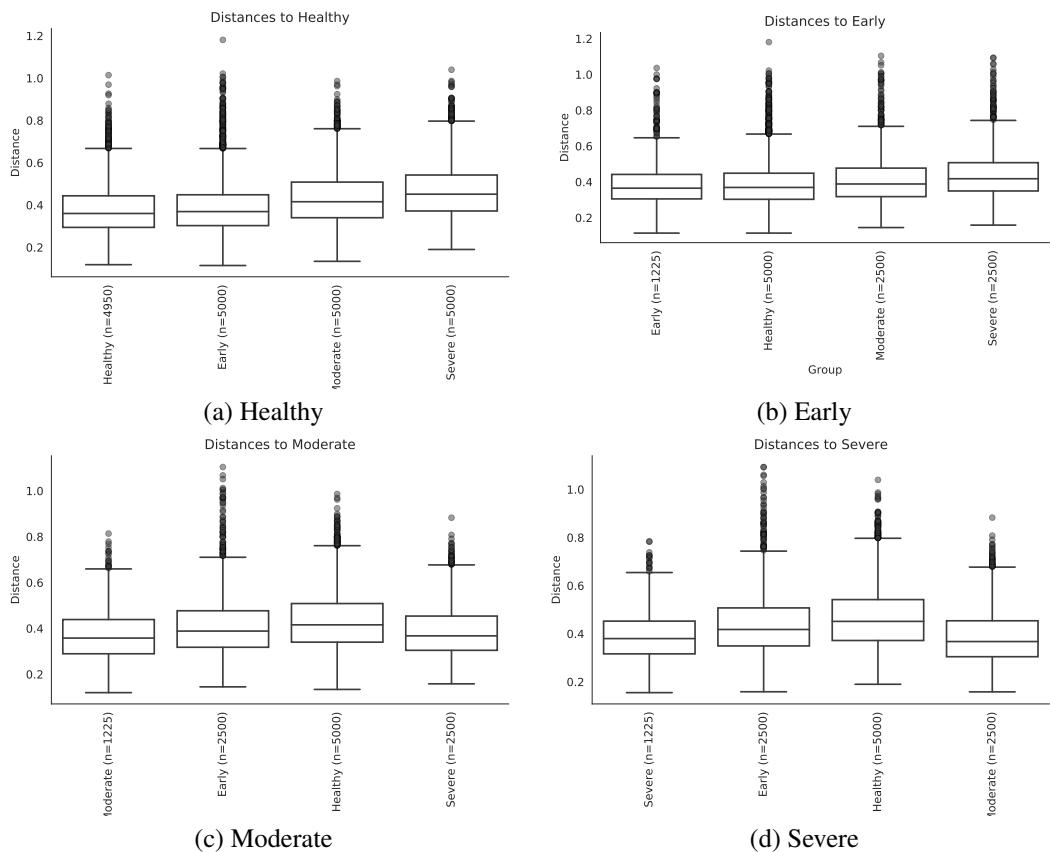


Figure 30: Weighted UniFrac Distance Index with DADA2

Table 18: Unweighted UniFrac Distance Index with Deblur

Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
Slight	Healthy	150	999	2.709074154153053	0.003	0.0036
Slight	Moderate	100	999	7.547240014264336	0.001	0.0015
Slight	Severe	100	999	7.772239667697252	0.001	0.0015
Healthy	Moderate	150	999	19.48285778321118	0.001	0.0015
Healthy	Severe	150	999	20.254907535032658	0.001	0.0015
Moderate	Severe	100	999	1.061788954262309	0.34	0.34

Table 19: Weighted UniFrac Distance Index with Deblur

Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
Slight	Healthy	150	999	2.0087857905677193	0.088	0.088
Slight	Moderate	100	999	5.981646579135783	0.002	0.003
Slight	Severe	100	999	16.572566883582837	0.001	0.002
Healthy	Moderate	150	999	9.494764618252377	0.001	0.002
Healthy	Severe	150	999	20.338834647304648	0.001	0.002
Moderate	Severe	100	999	5.026218407543304	0.003	0.0036

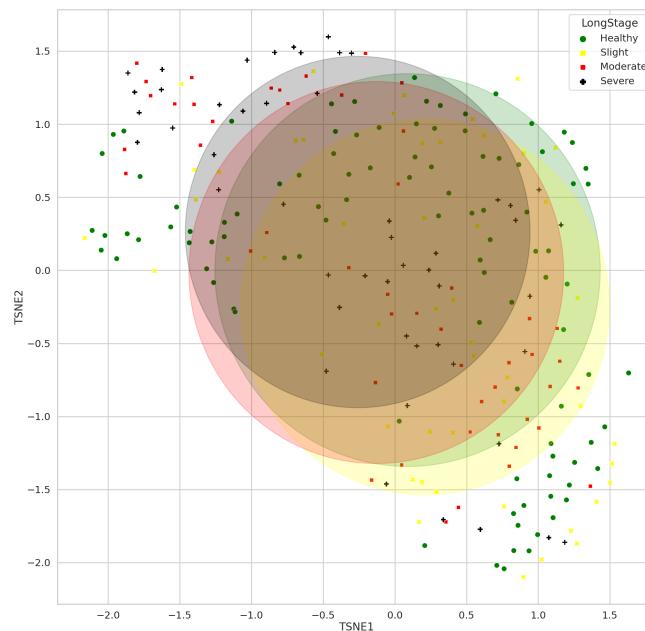


Figure 31: t-SNE Plot from Bray-Curtis Distance Index with Deblur

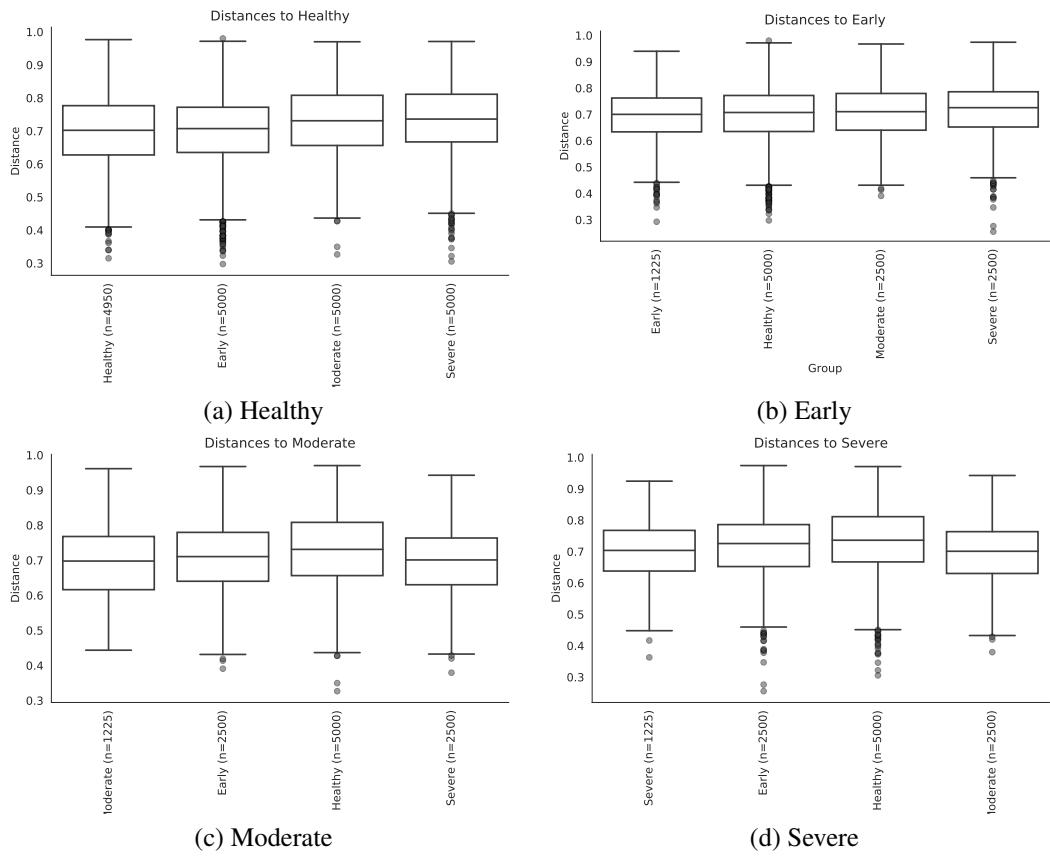


Figure 32: Bray-Curtis Distance Index with Deblur

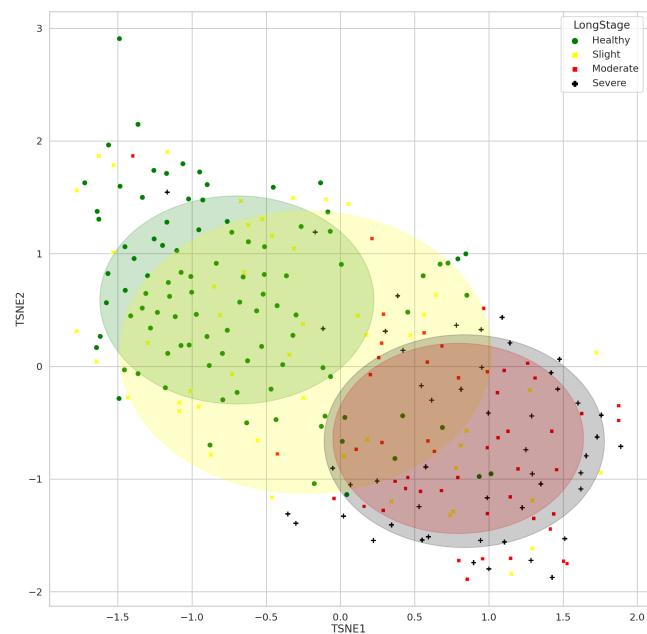


Figure 33: t-SNE Plot from Jaccard Distance Index with Deblur

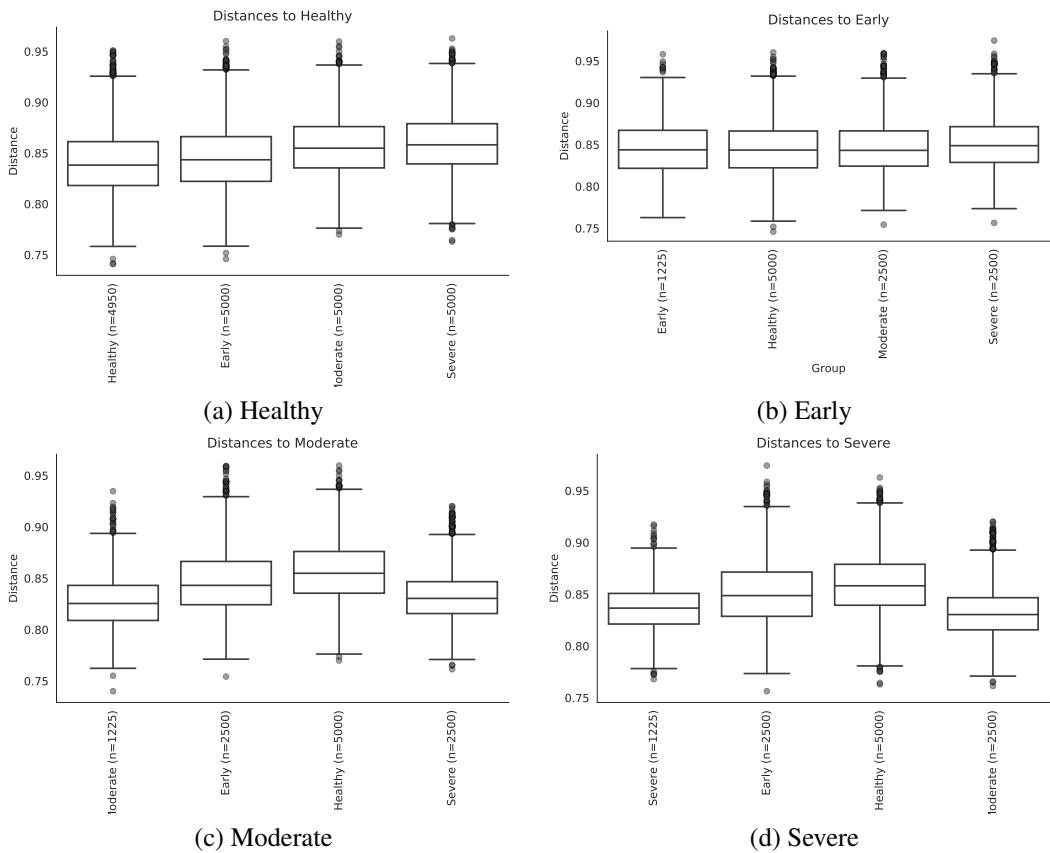


Figure 34: Jaccard Distance Index with Deblur

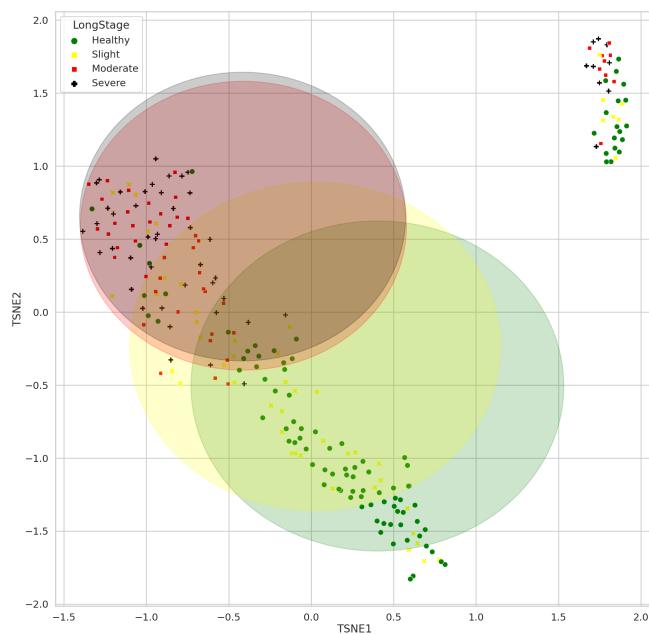


Figure 35: t-SNE Plot from Unweighted UniFrac Distance Index with Deblur

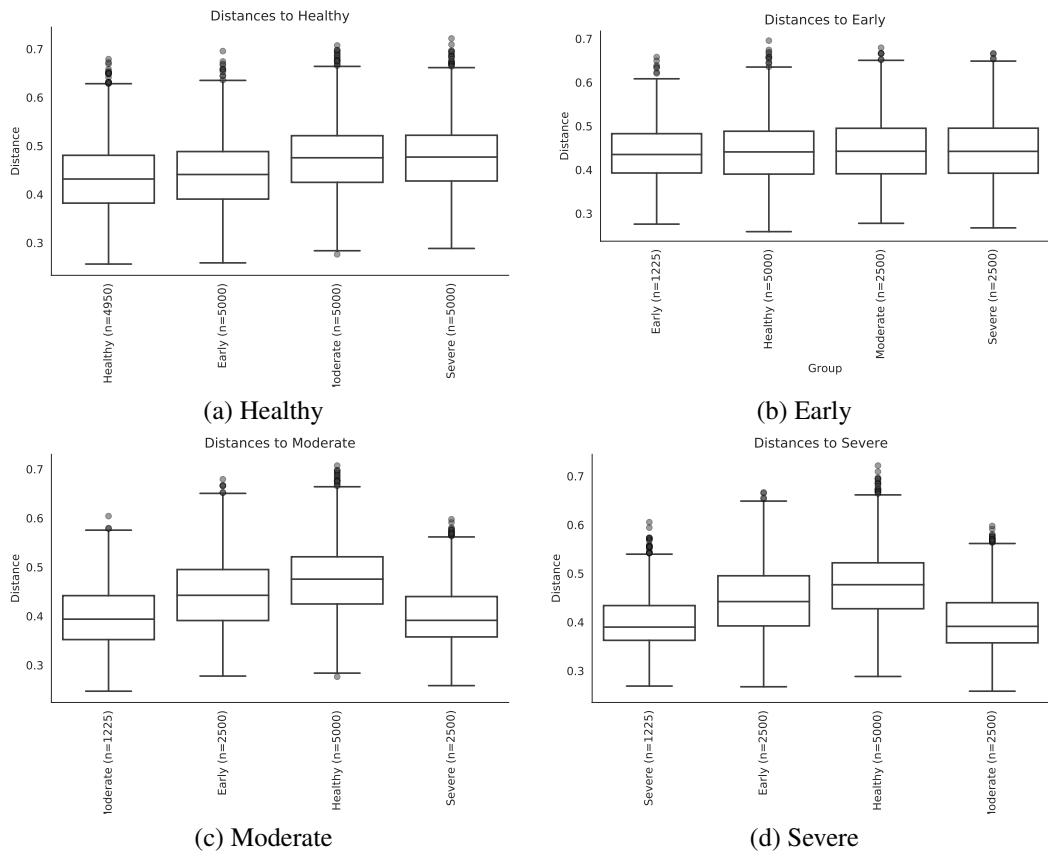


Figure 36: Unweighted UniFrac Distance Index with Deblur

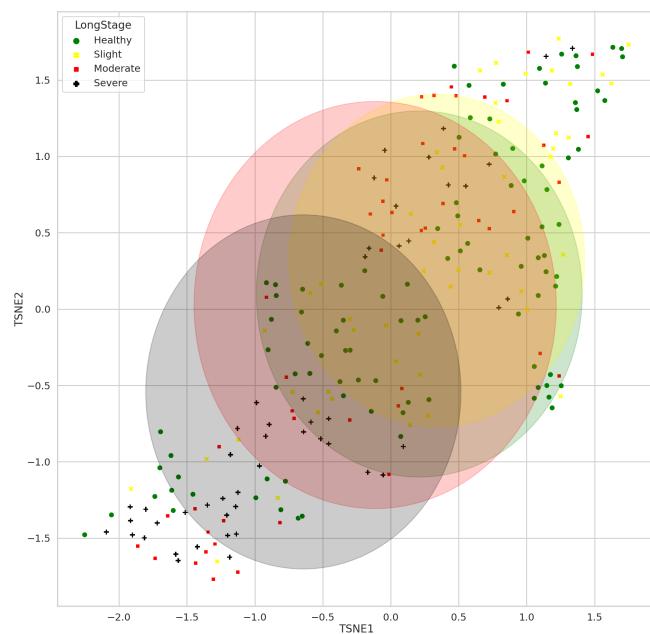


Figure 37: t-SNE Plot from Weighted UniFrac Distance Index with Deblur

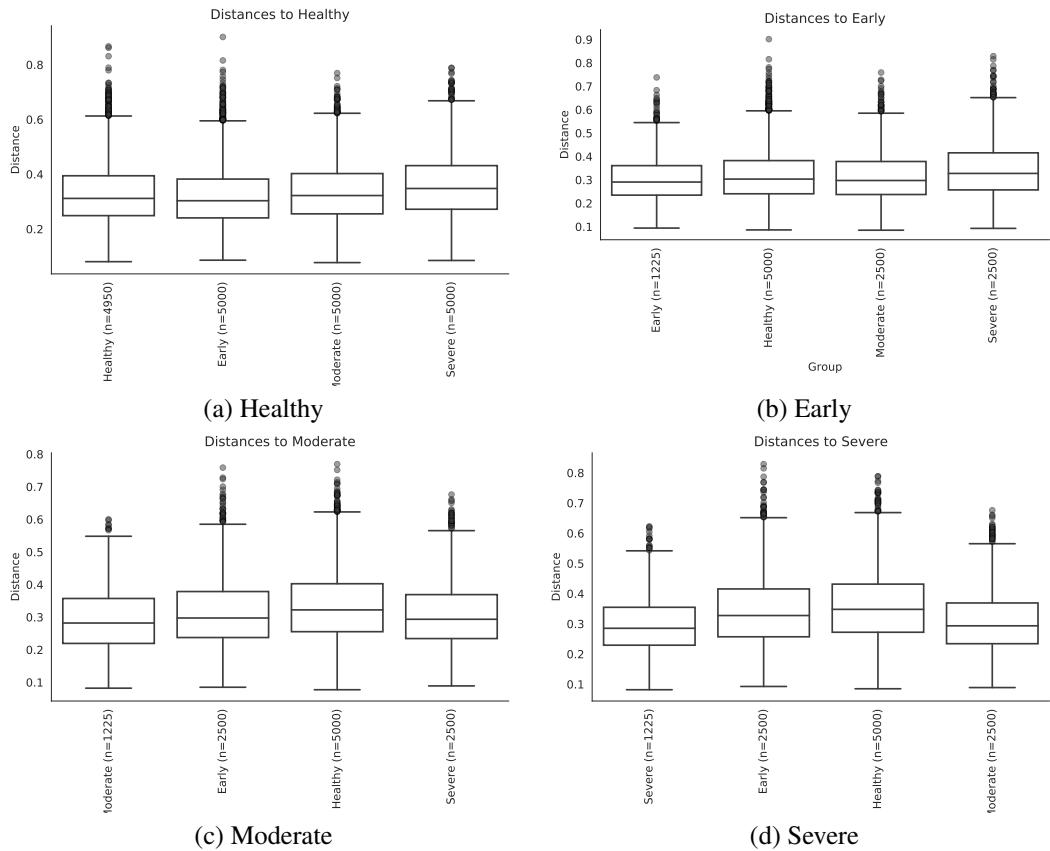


Figure 38: Weighted UniFrac Distance Index with Deblur

Table 20: ANCOM Significant Taxa with DADA2 and GG

		W	Reject null hypothesis			
Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	326	True
Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	325	True
Bacteria	Spirochaetes	Spirochaetes	Spirochaetales	Spirochaetaceae	325	True
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	323	True
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	321	True
Bacteria	Synergistetes	Synergistia	Synergistales	Dethiosulfovibrionaceae	319	True
Bacteria	Tenericutes	Mollicutes	Mycoplasmatales	Mycoplasmataceae	318	True
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	315	True
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	313	True
Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	309	True
Bacteria	Firmicutes	Clostridia	Clostridiales	[Mogibacteriaceae]	306	True
Bacteria	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	305	True
Bacteria	Firmicutes	Clostridia	Clostridiales	[Acidaminobacteraceae]	304	True

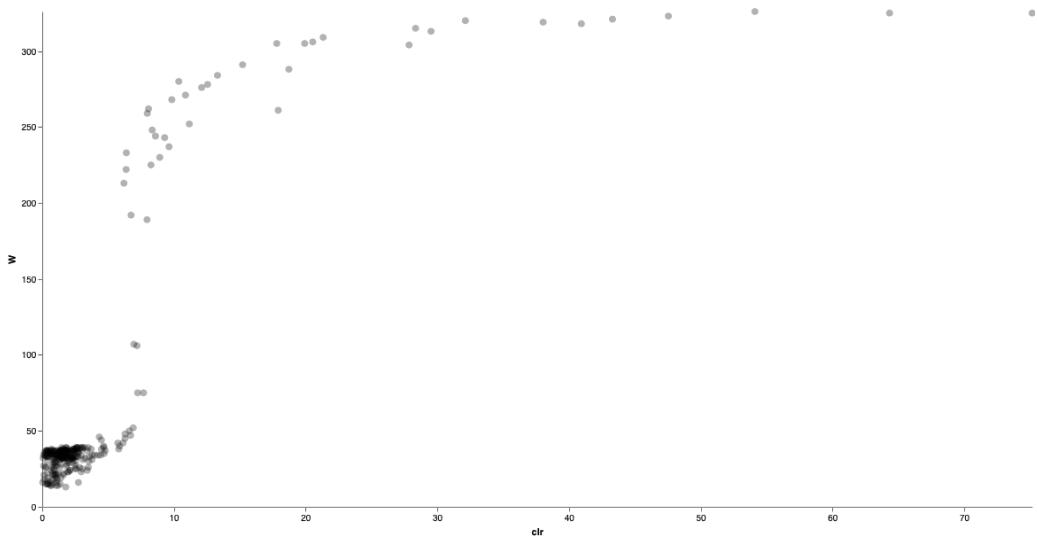


Figure 39: ANCOM Volcano Plot with DADA2 and GG

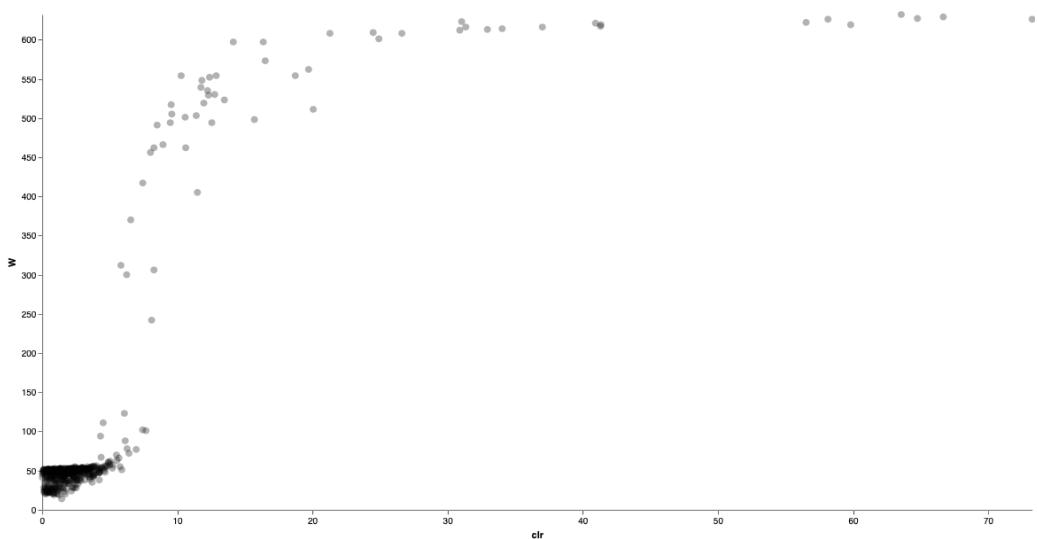


Figure 40: ANCOM Volcano Plot with DADA2 and HOMD

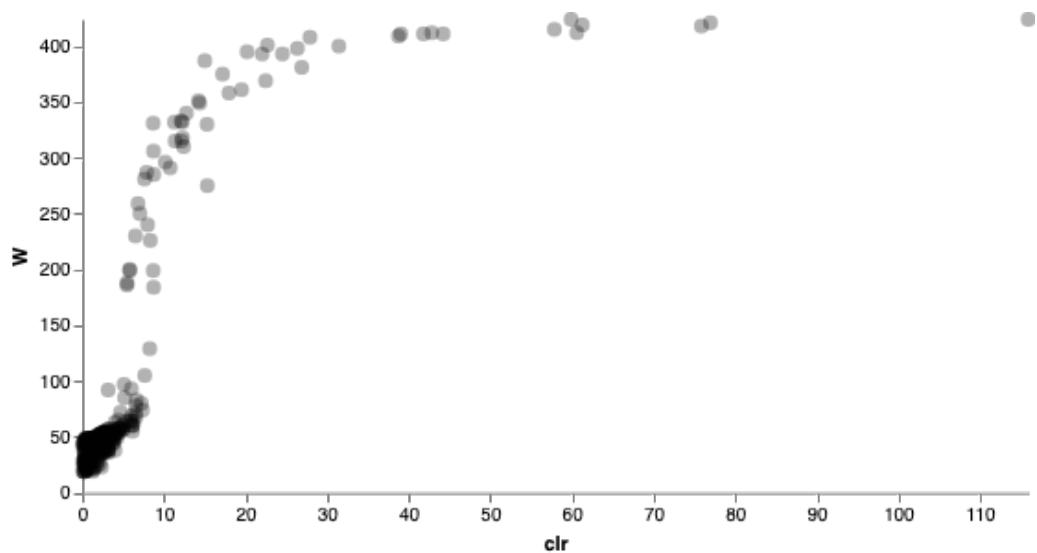


Figure 41: ANCOM Volcano Plot with DADA2 and SILVA

Table 21: ANCOM Significant Taxa with DADA2 and SILVA

	W	Reject null hypothesis
Bacteria Actinobacteriota Actinobacteria Actinomycetales Actinomycetaceae Actinomyces	632	True
Bacteria Bacteroidota Bacteroidia Bacteroidales Porphyromonadaceae Porphyromonas Porphyromonas gingivalis	629	True
Bacteria Firmicutes Clostridia Peptostreptococcales-Tissierellales Peptostreptococcaceae Filifactor Filifactor alocis	627	True
Bacteria Bacteroidota Bacteroidia Bacteroidales Prevotellaceae Prevotella Prevotella intermedia	626	True
Bacteria Spirochaetota Spirochaetia Spirochaetales Spirochaetaceae Treponema Treponema denticola	626	True
Bacteria Actinobacteriota Actinobacteria Actinomycetales Actinomycetaceae Actinomyces Schaalialia odontolytica	623	True
Bacteria Bacteroidota Bacteroidia Bacteroidales Tannerellaceae Tannerella Tannerella forsythia	622	True
Bacteria Spirochaetota Spirochaetia Spirochaetales Spirochaetaceae Treponema Treponema medium	621	True
Bacteria Firmicutes Clostridia Peptostreptococcales-Tissierellales Anaerovoracaceae [Eubacterium] nodatum group [Eubacterium] nodatum	619	True
Bacteria Spirochaetota Spirochaetia Spirochaetales Spirochaetaceae Treponema uncultured bacterium	619	True
Bacteria Firmicutes Bacilli Mycoplasmatales Mycoplasmataceae Mycoplasma Metamyoplasma faecium	617	True
Bacteria Synergistota Synergistia Synergistales Synergistaceae Fretibacterium	616	True
Bacteria Spirochaetota Spirochaetia Spirochaetales Spirochaetaceae Treponema	616	True
Bacteria Firmicutes Clostridia Lachnospirales Defluvitaleaceae Defluvitaleaceae UCG-011 Lachnospiraceae bacterium	614	True
Bacteria Bacteroidota Bacteroidia Bacteroidales Porphyromonadaceae Porphyromonas	613	True
Bacteria Firmicutes Clostridia Peptostreptococcales-Tissierellales Anaerovoracaceae [Eubacterium] brachy group [Eubacterium] brachy	612	True
Bacteria Bacteroidota Bacteroidia Bacteroidales Prevotellaceae	609	True
Bacteria Actinobacteriota Actinobacteria Corynebacteriales Corynebacteriaceae Corynebacterium Corynebacterium durum	608	True
Bacteria Firmicutes Clostridia Peptostreptococcales-Tissierellales Anaerovoracaceae [Eubacterium] saphenum group Eubacterium saphenum	608	True
Bacteria Spirochaetota Spirochaetia Spirochaetales Spirochaetaceae Treponema Treponema maltophilum	601	True
Bacteria Campylobacterota Campylobacteria Campylobacterales Campylobacteraceae Campylobacter Campylobacter showae	597	True
Bacteria Actinobacteriota Actinobacteria Actinomycetales Actinomycetaceae Actinomyces Actinomyces graevenitzii	597	True
Bacteria Firmicutes Clostridia Lachnospirales Lachnospiraceae Oribacterium	573	True

Table 22: ANCOM Significant Taxa with DADA2 and HOMD

		W	Reject null hypothesis
Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae	424	True	
Porphyromonas gingivalis			
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 faicum	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	

Table 23: ANCOM Significant Taxa with Deblur and GG

	W	Reject null hypothesis
Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae Filifactor	229	True
Bacteria Spirochaetes Spirochaetes Spirochaetales Spirochaetaceae Treponema	229	True
Bacteria Spirochaetes Spirochaetes Spirochaetales Spirochaetaceae Treponema amylovorum	227	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella intermedia	226	True
Bacteria Tenericutes Mollicutes Mycoplasmatales Mycoplasmataceae Mycoplasma	226	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Porphyromonas endodontalis	226	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Tannerella	226	True
Bacteria Synergistetes Synergistia Synergistales Dethiosulfovibrionaceae TG5	226	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae Porphyromonas	224	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales	220	True
Bacteria Firmicutes Clostridia Clostridiales [Mogibacteriaceae]	211	True
Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae Peptostreptococcus	209	True
Bacteria Proteobacteria Deltaproteobacteria Desulfobacterales Desulfovulbaceae Desulfobulbus	207	True
Bacteria Spirochaetes Spirochaetes Spirochaetales Spirochaetaceae Treponema socranskii	201	True
Bacteria Proteobacteria Epsilonproteobacteria Campylobacterales Campylobacteraceae Campylobacter	199	True
Bacteria Firmicutes Clostridia Clostridiales [Acidaminobacteraceae]	199	True
Bacteria Proteobacteria Gammaproteobacteria Pasteurellales Pasteurellaceae Haemophilus parainfluenzae	196	True
Bacteria Firmicutes Clostridia Clostridiales [Tissierellaceae] Parvimonas	195	True
Bacteria Proteobacteria Betaproteobacteria Neisseriales Neisseriaceae Neisseria subflava	194	True
Bacteria Firmicutes Clostridia Clostridiales [Mogibacteriaceae] Mogibacterium	187	True
Bacteria Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Actinomyces	187	True
Bacteria Firmicutes Clostridia Clostridiales [Tissierellaceae]	186	True
Bacteria Actinobacteria Actinobacteria Actinomycetales	185	True
Bacteria Firmicutes Clostridia Clostridiales	184	True
Bacteria Firmicutes Clostridia Clostridiales Lachnospiraceae Oribacterium	180	True
Bacteria Firmicutes Clostridia Clostridiales Lachnospiraceae	179	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella nanceiensis	173	True

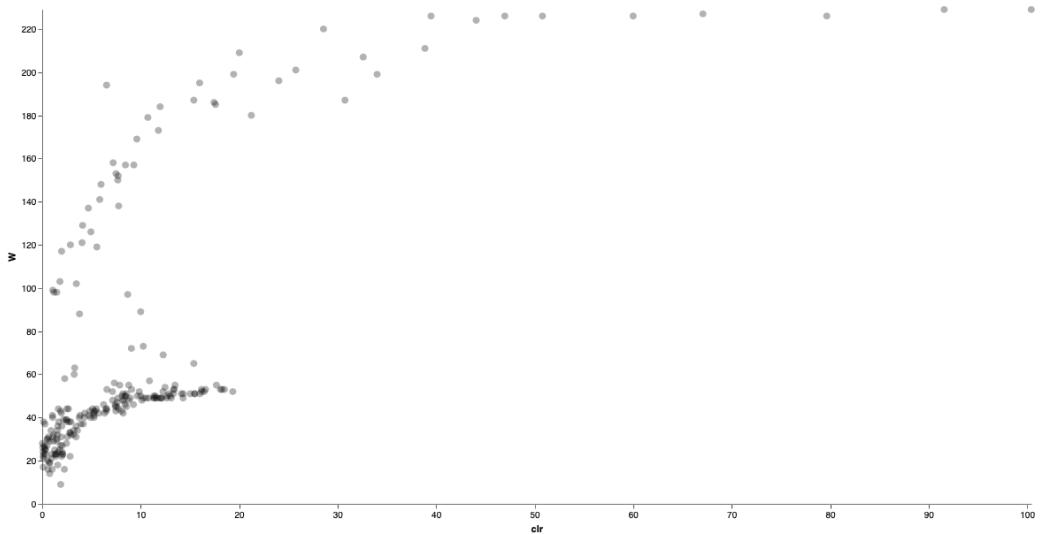


Figure 42: ANCOM Volcano Plot with Deblur and GG

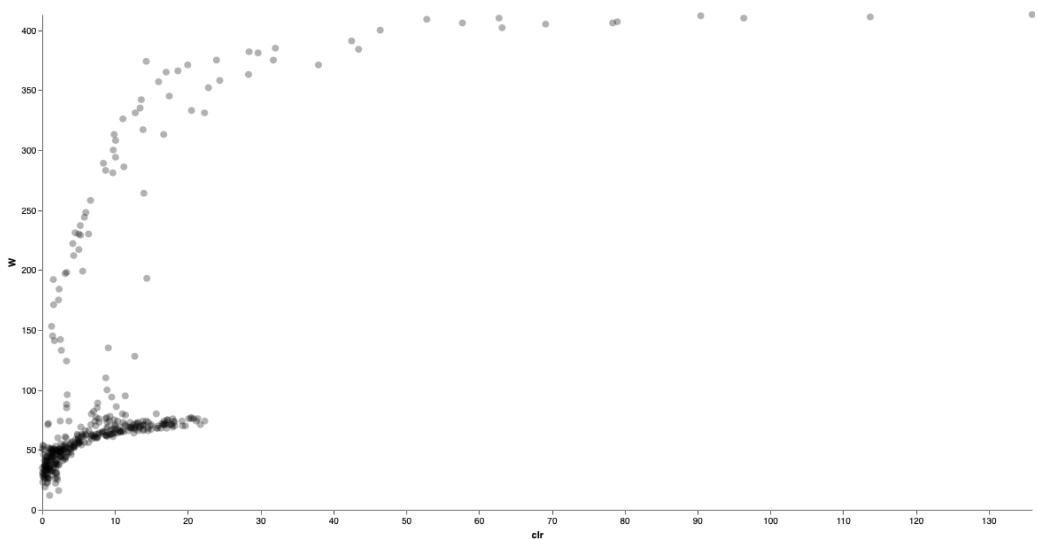


Figure 43: ANCOM Volcano Plot with Deblur and SILVA

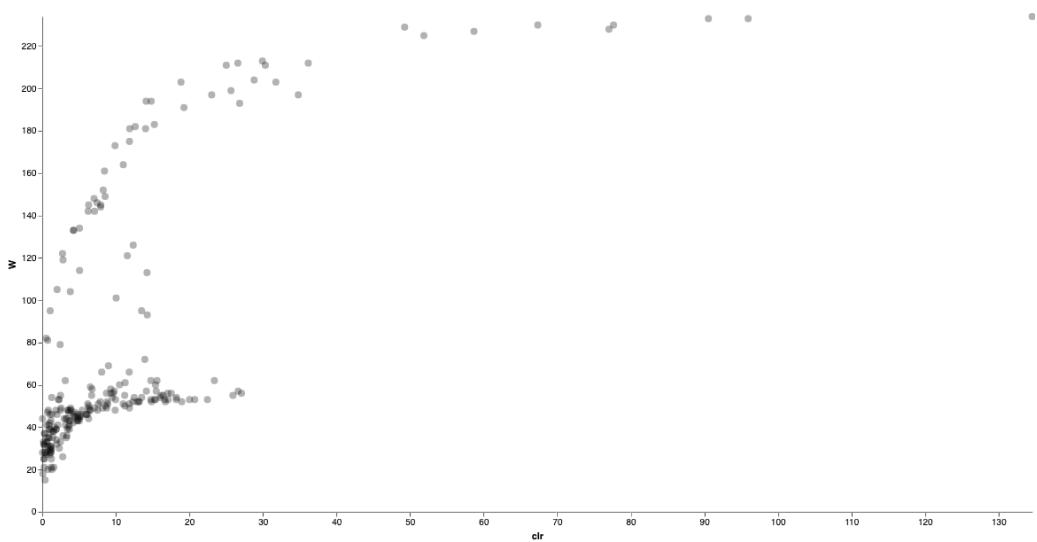


Figure 44: ANCOM Volcano Plot with Deblur and HOMD

Table 24: ANCOM Significant Taxa with Deblur and SILVA

	W	Reject null hypothesis
Bacteria Actinobacteriota Actinobacteria Actinomycetales Actinomycetaceae Actinomyces	632	True
Bacteria Bacteroidota Bacteroidia Bacteroidales Porphyromonadaceae Porphyromonas Porphyromonas gingivalis	629	True
Bacteria Firmicutes Clostridia Peptostreptococcales-Tissierellales Peptostreptococcaceae Filifactor Filifactor alocis	627	True
Bacteria Bacteroidota Bacteroidia Bacteroidales Prevotellaceae Prevotella Prevotella intermedia	626	True
Bacteria Spirochaetota Spirochaetia Spirochaetales Spirochaetaceae Treponema Treponema denticola	626	True
Bacteria Actinobacteriota Actinobacteria Actinomycetales Actinomycetaceae Actinomyces Schaalialia odontolytica	623	True
Bacteria Bacteroidota Bacteroidia Bacteroidales Tannerellaceae Tannerella Tannerella forsythia	622	True
Bacteria Spirochaetota Spirochaetia Spirochaetales Spirochaetaceae Treponema Treponema medium	621	True
Bacteria Firmicutes Clostridia Peptostreptococcales-Tissierellales Anaerovoracaceae [Eubacterium] nodatum group [Eubacterium] nodatum	619	True
Bacteria Spirochaetota Spirochaetia Spirochaetales Spirochaetaceae Treponema uncultured bacterium	619	True
Bacteria Firmicutes Bacilli Mycoplasmatales Mycoplasmataceae Mycoplasma Metamyoplasma faecium	617	True
Bacteria Synergistota Synergistia Synergistales Synergistaceae Fretibacterium	616	True
Bacteria Spirochaetota Spirochaetia Spirochaetales Spirochaetaceae Treponema	616	True
Bacteria Firmicutes Clostridia Lachnospirales Defluvitaleaceae Defluvitaleaceae UCG-011 Lachnospiraceae bacterium	614	True
Bacteria Bacteroidota Bacteroidia Bacteroidales Porphyromonadaceae Porphyromonas	613	True
Bacteria Firmicutes Clostridia Peptostreptococcales-Tissierellales Anaerovoracaceae [Eubacterium] brachy group [Eubacterium] brachy	612	True
Bacteria Bacteroidota Bacteroidia Bacteroidales Prevotellaceae	609	True
Bacteria Actinobacteriota Actinobacteria Corynebacteriales Corynebacteriaceae Corynebacterium Corynebacterium durum	608	True
Bacteria Firmicutes Clostridia Peptostreptococcales-Tissierellales Anaerovoracaceae [Eubacterium] saphenum group Eubacterium saphenum	608	True
Bacteria Spirochaetota Spirochaetia Spirochaetales Spirochaetaceae Treponema Treponema maltophilum	601	True
Bacteria Campylobacterota Campylobacteria Campylobacteriales Campylobacteraceae Campylobacter Campylobacter showae	597	True
Bacteria Actinobacteriota Actinobacteria Actinomycetales Actinomycetaceae Actinomyces Actinomyces graevenitzii	597	True
Bacteria Firmicutes Clostridia Lachnospirales Lachnospiraceae Oribacterium	573	True

Table 25: ANCOM Significant Taxa with Deblur and HOMD

	W	Reject null hypothesis
Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae	234	True
Porphyromonas gingivalis		
Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae [XI]	233	True
Filifactor alocis		
Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema	233	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae	230	True
Tannerella forsythia		
Bacteria Synergistetes Synergistia Synergistales Synergistaceae Fretibacterium	230	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales Porphyromonadaceae	229	True
Porphyromonas sp. HMT 285		
Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae [XI] Peptostreptococcaceae [XI][G-6] nodatum	228	True
Bacteria Firmicutes Mollicutes Mycoplasmatales Mycoplasmataceae Mycoplasma	227	True
Bacteria Firmicutes Clostridia Clostridiales Lachnospiraceae [XIV] Lachnospiraceae [G-8] bacterium HMT 500	225	True
Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae [XI] Peptostreptococcaceae [XI][G-5] saphenum	213	True
Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae [XI] Peptostreptococcaceae [XI][G-9] brachy	212	True
Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema sp. HMT 260	212	True
Bacteria Proteobacteria Deltaproteobacteria Desulfobacterales Desulfovulbaceae Desulfobulbus sp. HMT 041	211	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella dentalis	211	True
Bacteria Bacteroidetes Bacteroidetes [C-1] Bacteroidetes [O-1] Bacteroidetes [F-1] Bacteroidetes [G-3]	204	True
Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae [XI] Peptostreptococcus anaerobius	203	True
Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae [XI] Peptostreptococcaceae [XI][G-4] bacterium HMT 369	203	True
Bacteria Firmicutes Clostridia Clostridiales Clostridiales [F-1] Clostridiales [F-1][G-1] bacterium HMT 093	199	True
Bacteria Firmicutes Clostridia Clostridiales Lachnospiraceae [XIV] Oribacterium sinus	197	True
Bacteria Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Actinomyces	197	True
Bacteria Firmicutes Clostridia Clostridiales Peptoniphilaceae Parvimonas sp. HMT 393	194	True
Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella nanceiensis	194	True
Bacteria Proteobacteria Gammaproteobacteria Pasteurellales Pasteurellaceae	193	True
Bacteria Firmicutes Clostridia Clostridiales Lachnospiraceae [XIV] Johnsonella sp. HMT 166	191	True
Bacteria Spirochaetes Spirochaetia Spirochaetales Spirochaetaceae Treponema sp. HMT 258	183	True
Bacteria Firmicutes Negativicutes Veillonellales Veillonellaceae Veilonella denticariosi	182	True
Bacteria Firmicutes Clostridia Clostridiales Lachnospiraceae [XIV] Catonella sp. HMT 164	181	True
Bacteria Firmicutes Clostridia Clostridiales Peptostreptococcaceae [XI] Mogibacterium	181	True

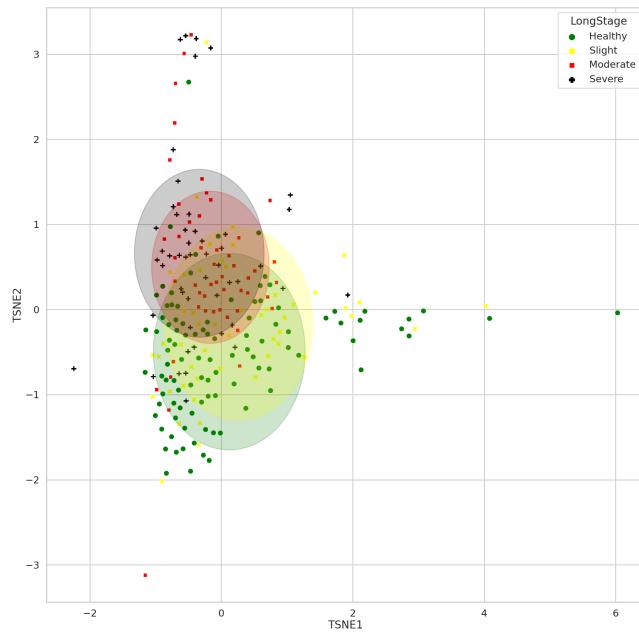


Figure 45: t-SNE Plot with Whole Microbiome from DADA2 and GG (328 taxa)

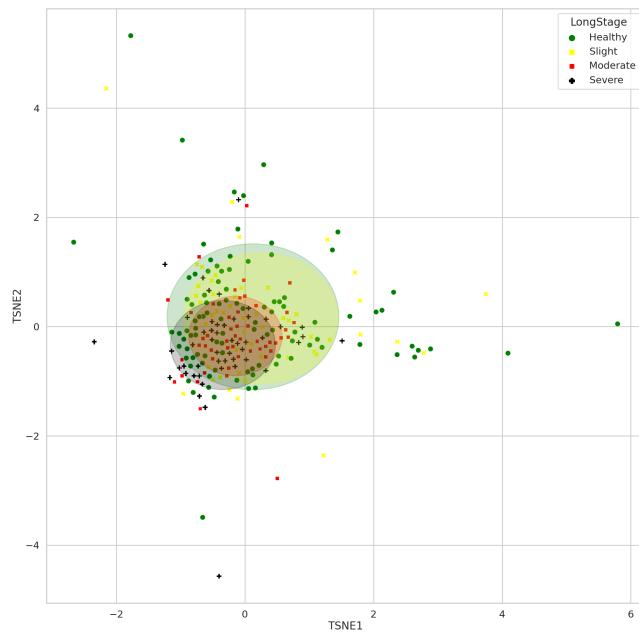


Figure 46: t-SNE Plot with Whole Microbiome from DADA2 and SILVA (633 taxa)

Table 26: Taxa with Deblur and HOMD Ordered by Random Forest

---

Table 27: Taxa with DADA2 and HOMD Ordered by Random Forest for Merging (Healthy+Early) Classes

---

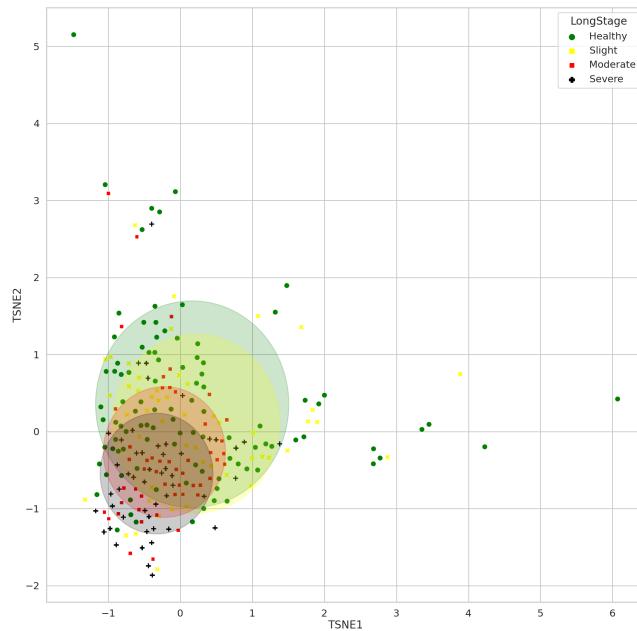


Figure 47: t-SNE Plot with Whole Microbiome from DADA2 and HOMD (425 taxa)

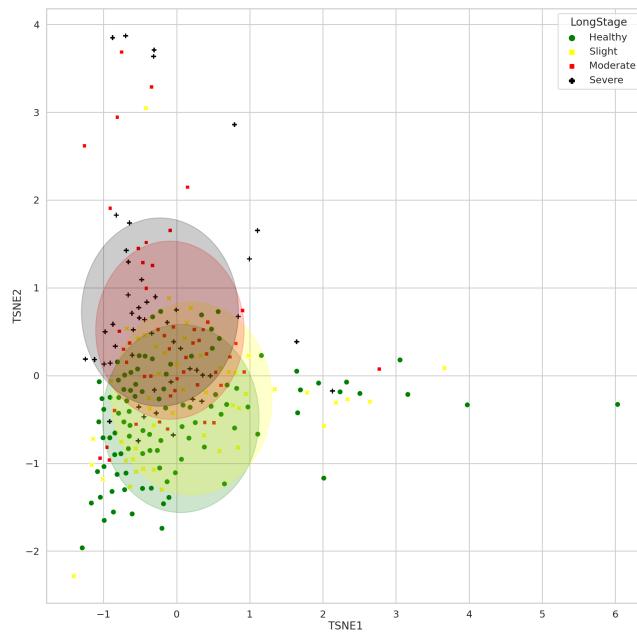


Figure 48: t-SNE Plot with Whole Microbiome from Deblur and GG (232 taxa)

Table 28: Taxa with Deblur and HOMD Ordered by Random Forest for Merging (Moderate+Severe) Classes

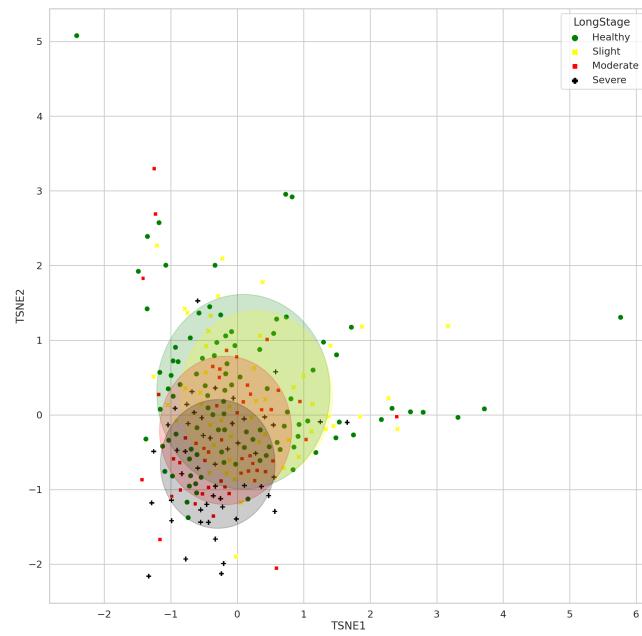


Figure 49: t-SNE Plot with Whole Microbiome from Deblur and SILVA (414 taxa)

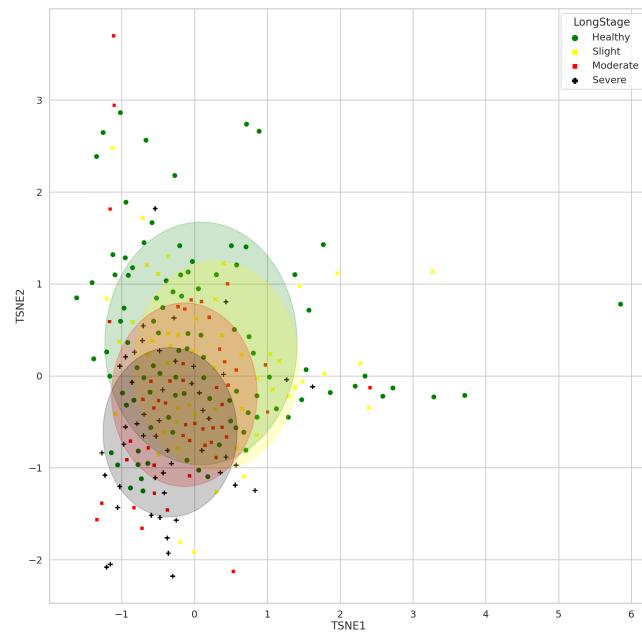


Figure 50: t-SNE Plot with Whole Microbiome from Deblur and HOMD (235 taxa)

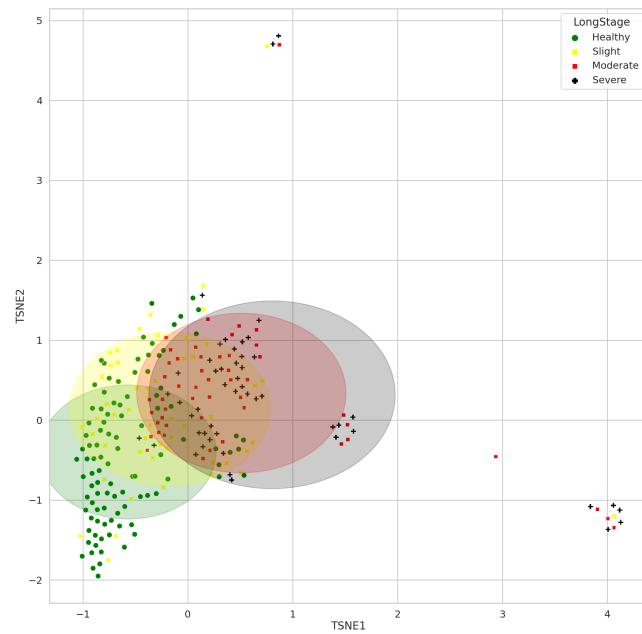


Figure 51: t-SNE Plot with ANCOM Selected Microbiome Data from DADA2 and GG (15 taxa)

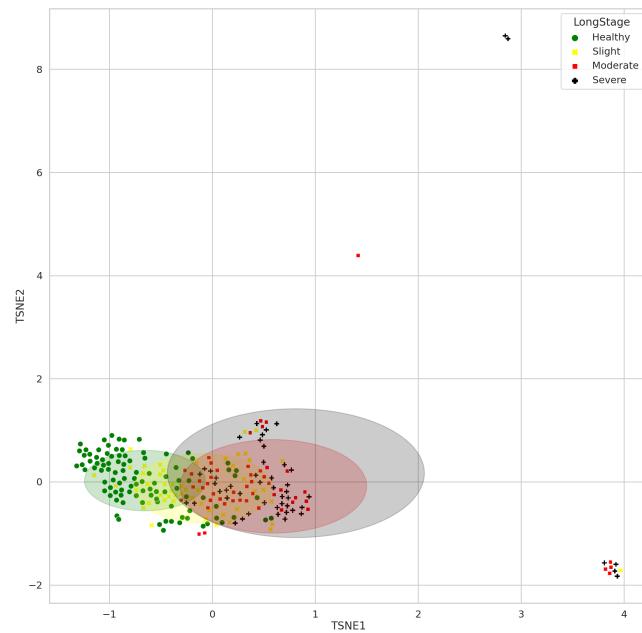


Figure 52: t-SNE Plot with ANCOM Selected Microbiome Data from DADA2 and SILVA (23 taxa)

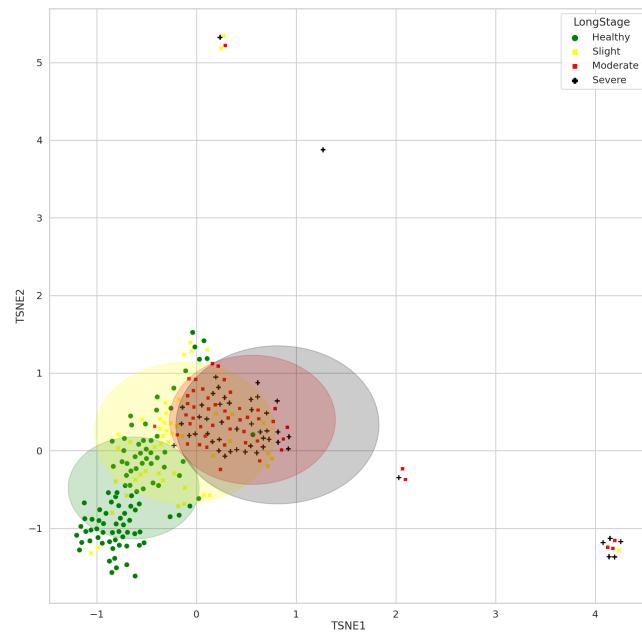


Figure 53: t-SNE Plot with ANCOM Selected Microbiome Data from DADA2 and HOMD (20 taxa)

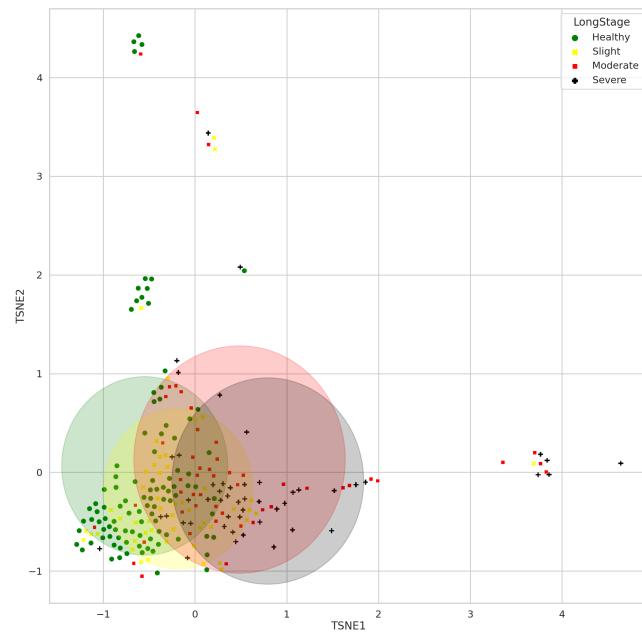


Figure 54: t-SNE Plot with ANCOM Selected Microbiome Data from Deblur and GG (27 taxa)

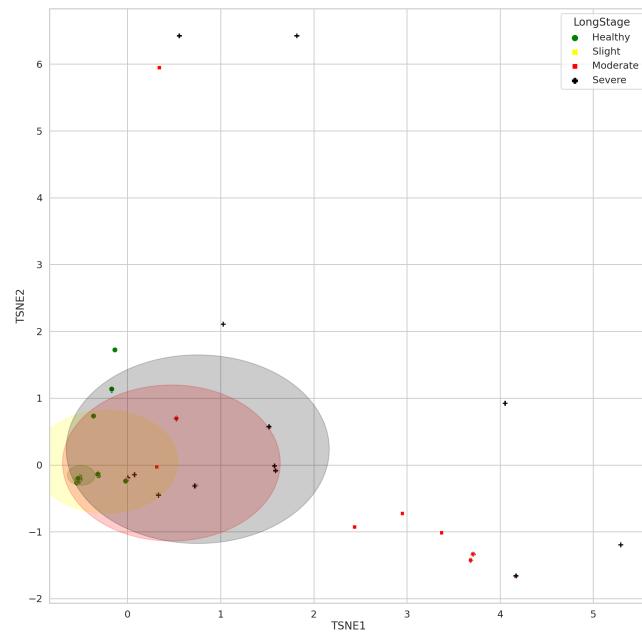


Figure 55: t-SNE Plot with ANCOM Selected Microbiome Data from Deblur and SILVA (20 taxa)

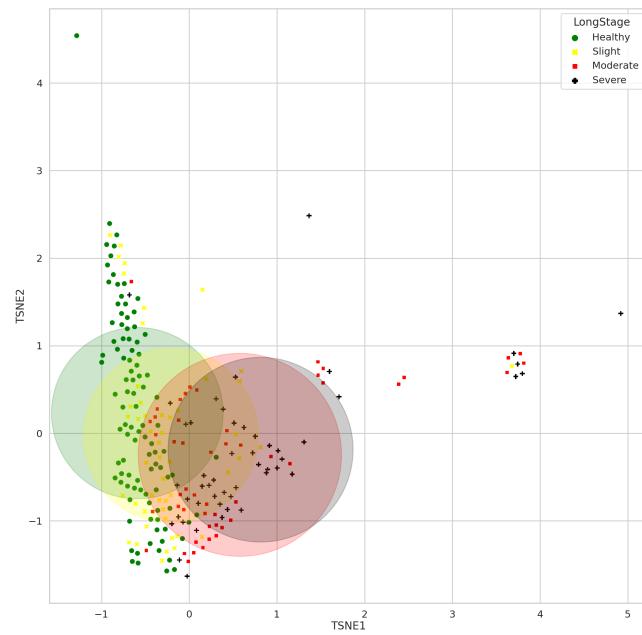


Figure 56: t-SNE Plot with ANCOM Selected Microbiome Data from Deblur and HOMD (28 taxa)



figures/RandomForest/ANCOM.Deblur.homd/metrics.png

Figure 57: Metrics by Feature Count with Deblur and HOMD



(a) *Porphyromonas gingivalis*

(b) *Actinomyces*

Figure 58: Most and Second Most Important Features with Deblur and HOMD

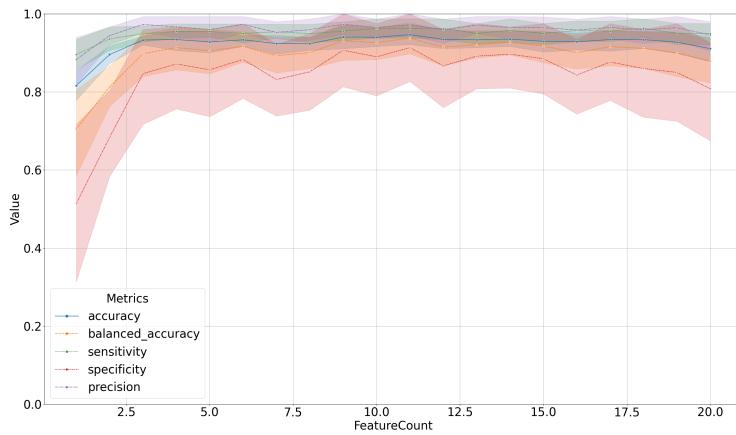


Figure 59: Metrics by Feature Count with DADA2 and HOMD for Merging (Healthy+Early) Classes

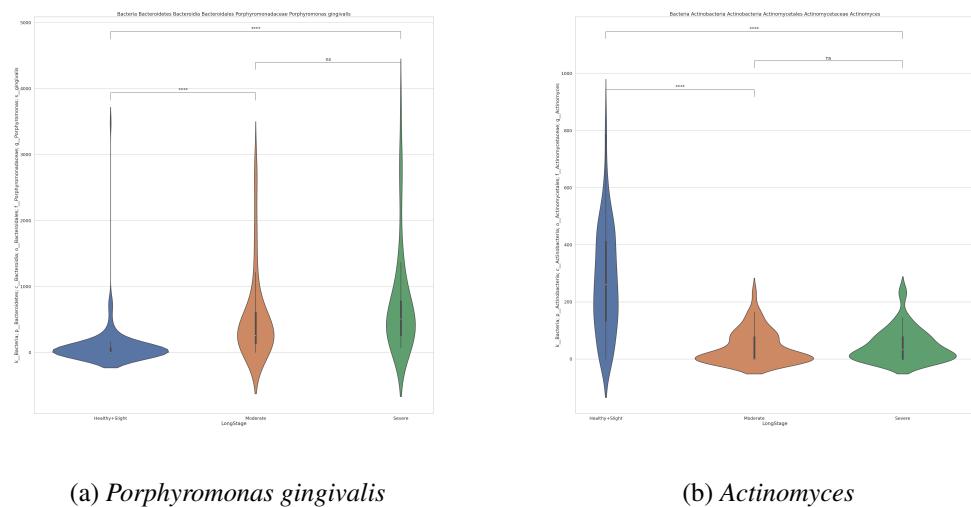


Figure 60: Most and Second Most Important Features with DADA2 and HOMD for Merging (Healthy+Early) Classes

figures/RandomForest/two.Deblur.homd/metrics.png

Figure 61: Metrics by Feature Count with Deblur and HOMD for Merging (Moderate+Severe) Classes

figures/RandomForest/two.Deblur.homed/Feature\_1.png

(a) *Porphyromonas gingivalis*

(b) *Actinomyces*

Figure 62: Most and Second Most Important Features with Deblur and HOMD for Merging (Moderate+Severe) Classes

## 4.11 Random Forest Classifier with Merging (Healthy+Early) & (Moderate+Severe) Classes

As figure 3 and figure 4, there are six combinations. However, there are not statistically significant difference between (Healthy and Early) classes and between (Moderate and Severe) classes. Thus, classification algorithm is carried out on these six combinations with merging (Healthy and Early) classes and (Moderate and Severe) classes. Among these six combinations, DADA2 and SILVA pair gives the best result in terms of balanced accuracy. Features which are ordered by importance are in talbe 29. Also, five metrics by feature count are shown as figure 63; then, the highest value of balanced accuracy is 0.934 with using 15 taxa. Moreover, violin plots from the two most important taxa, *Porphyromonas gingivalis* and *Treponema denticola*, are in figure 64.

## 4.12 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 7. 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 8 and Figure 19). 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 9 and Figure 20). Moreover, there are no statistically significant differences between (Healthy and Early) classes and (Moderate and Severe) classes with observed features index from Deblur (Table 10 and Figure 21). 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 11 and Figure 22).

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 12, Figure 23 and Figure 24), Jaccard distance index (Table 13, Figure 25 and Figure 26), unweighted UniFrac distance index (Table 14, Figure 27 and Figure 28) and weighted UniFrac distance index (Table 15, Figure 29 and Figure 30), show statistically significant differences in every pair of classes, except (Moderate and Severe) classes.

Bray-Curtis distance index with Deblur has no statistically significant differences between (Healthy and Early) classes and (Moderate and Severe) classes (Table 16, Figure 31 and Figure 32). Moreover, Jaccard distance index with Deblur has no statistically significant difference between (Moderate and Severe) classes (Table 17, Figure 33 and Figure 34). Additionally, unweighted UniFrac distance index with Deblur has no statistically significant difference between (Moderate and Severe) classes (Table 18, Figure 35 and Figure 36). Furthermore, weighted UniFrac distance index with Deblur has no statistically significant difference between (Healthy and Early) classes (Table 38, Figure 37 and Figure 38).

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.

Table 29: Taxa with DADA2 and SILVA Ordered by Random Forest for Merging (Healthy+Early) & (Moderate+Severe) Classes

figures/RandomForest/one-two.DADA2.silva/metrics.png

Figure 63: Metrics by Feature Count with DADA2 and SILVA for Merging (Healthy+Early) & (Moderate+Severe) Classes

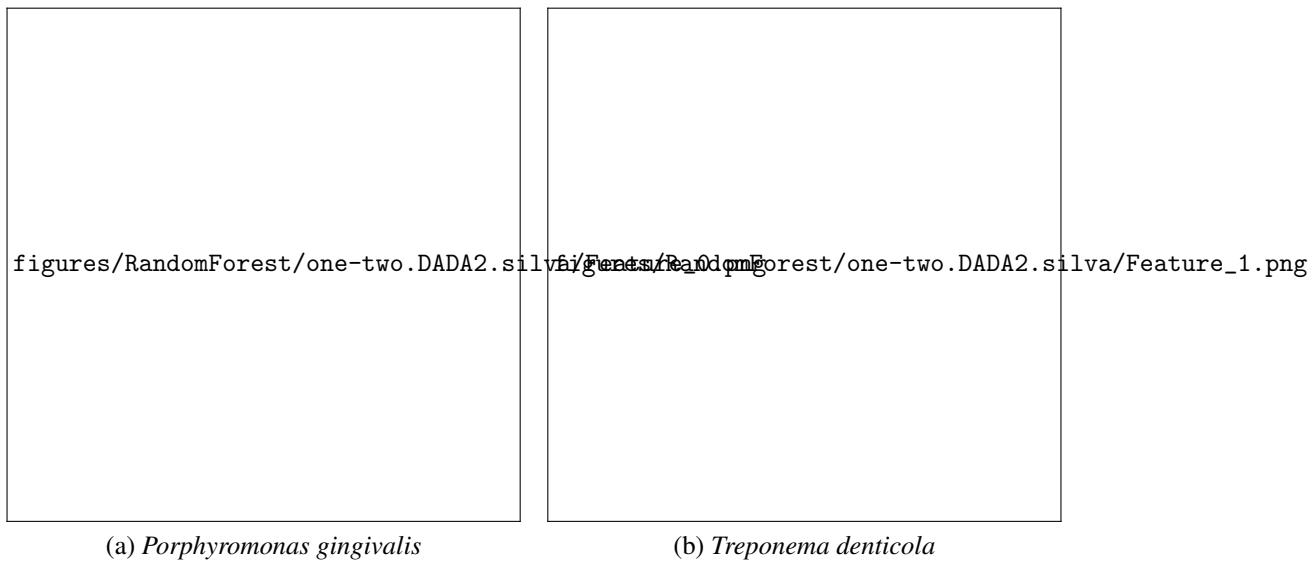


Figure 64: Most and Second Most Important Features with DADA2 and HOMD for Merging (Healthy+Early) & (Moderate+Severe) Classes

Table 30: Taxa with DADA2 and HOMD Ordered by Random Forest for Healthy Class and Early Class Only

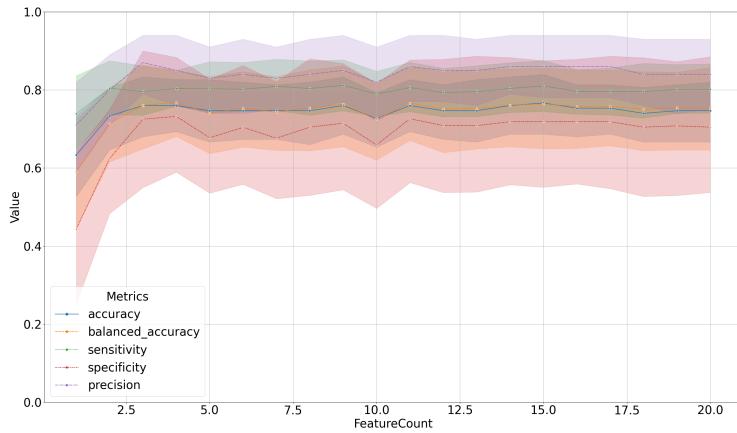


Figure 65: Metrics by Feature Count with DADA2 and HOMD for Healthy Class and Early Class Only

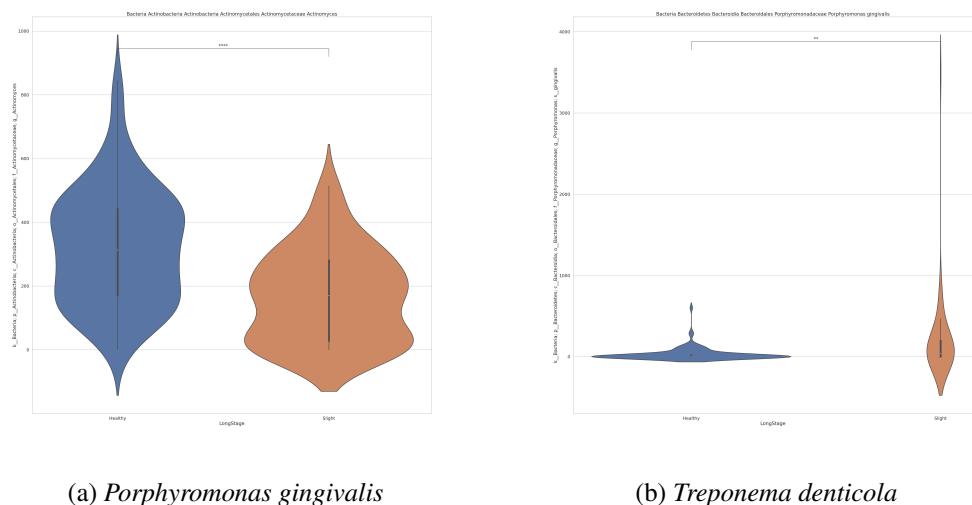


Figure 66: Most and Second Most Important Features with DADA2 and HOMD for Healthy Class and Early Class Only

### 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 45, Figure 46, Figure 47, Figure 48, Figure 49 and Figure 50) are more evenly distributed, whereas t-SNE plots with ANCOM selected microbiome data (Figure 51, Figure 52, Figure 53, Figure 54, Figure 55 and Figure 56) 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 26, Table 27, Table 28, Table 29 and Table 30). Thus, the two most important features are displayed for results of Random Forest classifier (Figure 58, Figure 60, Figure 62, Figure 64 and Figure 66).

## 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.