Periodontitis

Seunghoon Kim Jaewoong Lee Semin Lee

Ulsan National Institute of Science and Technology

jwlee230@unist.ac.kr

2020-12-07

Overview

1. Introduction
2. Materials
3. Methods
4. Results
5. Discussion

References

Introduction

Microbiome

- Microbiota: the micro-organisms which live inside & on humans (Turnbaugh et al., 2007)
- Microbiome: about 10^{13} micro-organisms whose which collective genome (Gill et al., 2006)

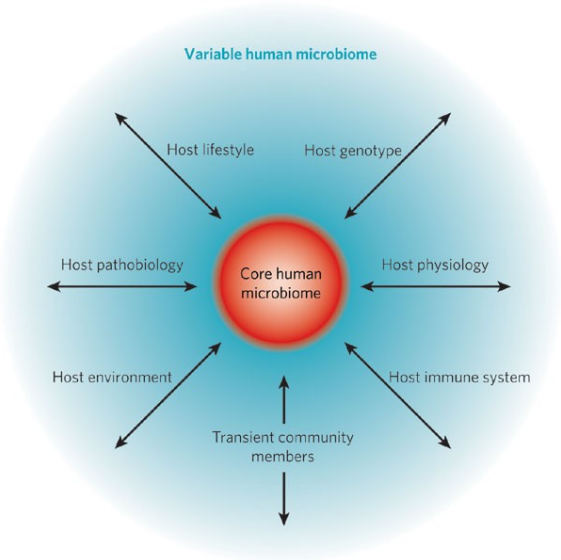


Figure: Concept of a core human microbiome (Turnbaugh et al., 2007)

rRNA

- Ribosomal RNA
- Well-known as a key to phylogeny (Olsen & Woese, 1993)

Periodontitis (Periodontal disease)

- CAL (Clinical Attachment Loss) & BL (Bone Loss) (Flemmig, 1999)
- Risk Factors (Van Dyke & Dave, 2005)
  1. Smoking
  2. Diabetes
  3. Genetic factor
  4. Host response

Materials

16S rRNA Sequencing

- 100 Healthy people
- 50 Chronic periodontitis – Early
- 50 Chronic periodontitis – Moderate
- 50 Chronic periodontitis – Severe

Methods

QIIME2 Workflow

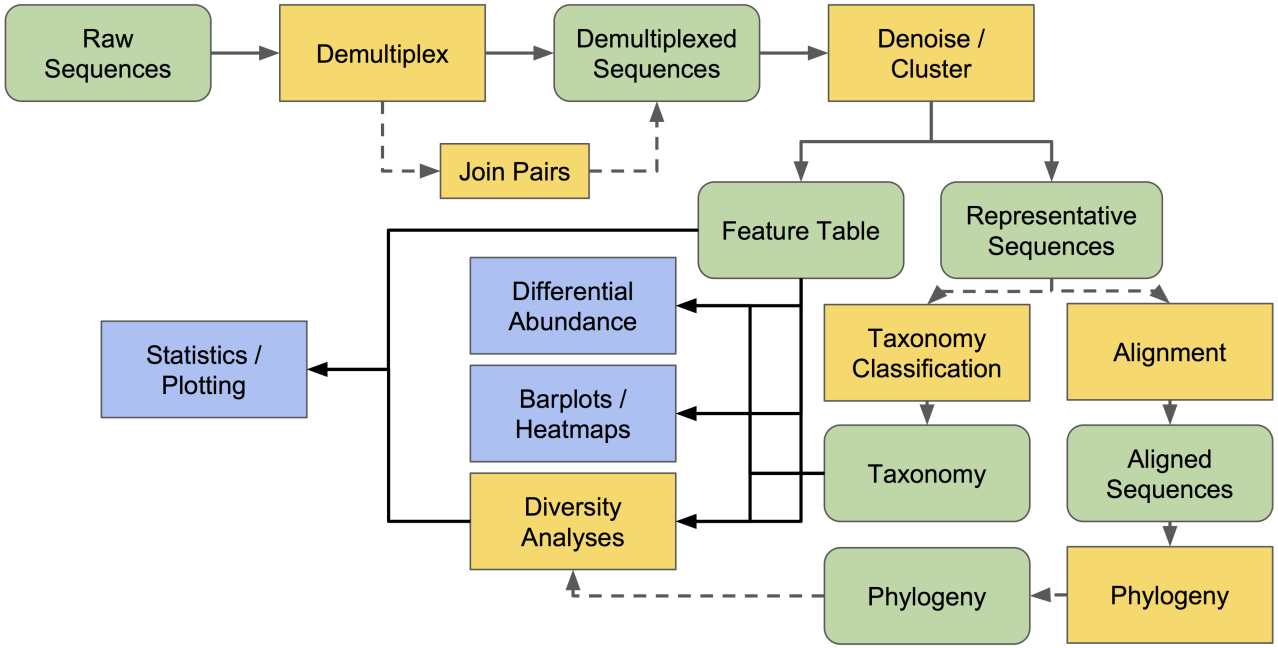


Figure: QIIME2 Workflow (Bolyen et al., 2019, 2018)

Denoising techniques

- DADA2: Amplicon Sequence Variants (ASVs) (Callahan et al., 2016)
- Deblur: Operational Taxonomic Units (OTUs) (Amir et al., 2017)

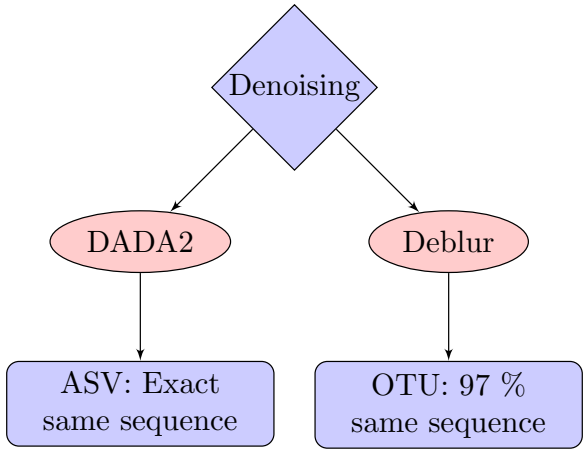


Figure: Denoising Techniques

Taxonomy Classification

- Greengenes (GG) (DeSantis et al., 2006)
- SILVA (Pruesse et al., 2007)

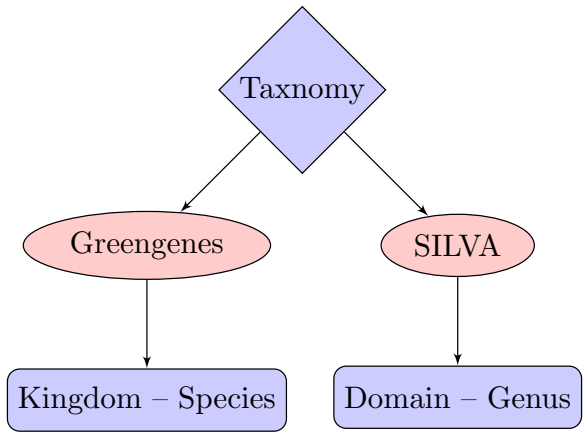


Figure: Taxonomy Classification

"A **higher** performance at taxonomic levels above *genus* level; but performance appears to drop at *species* level" (Gihawi et al., 2019)

Merging Denosing and Taxonomy Classification

Merging multiple IDs (ASVs and OTUs) into one, which have:

- Different IDs.
- Identified as same taxonomy.

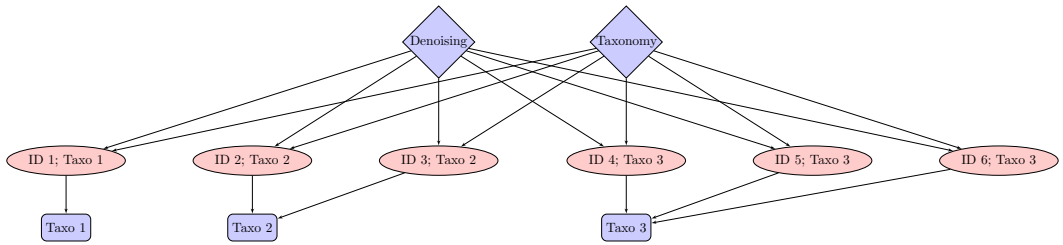


Figure: Example Diagram for Merging Denosing and Taxonomy Classification

Rarefaction

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

Alpha- & Beta-diversity

- alpha-diversity: the richness of taxa at a single community
- beta-diversity: the taxonomic differentiation between communities

Alpha-diversity

- Shannon's diversity index: a quantitative measure of community richness
- Observed Features: a qualitative measure of community richness
- Faith's Phylogenetic Diversity: a qualitative measure of community richness which incorporates phylogenetic relationship between the features
- Evenness: a measure of community evenness

(Bolyen et al., 2019, 2018)

Beta-diversity

- Bray-Curtis distance: a quantitative measure of community dissimilarity
- Jaccard distance: a qualitative measure of community dissimilarity
- Unweighted UniFrac distance: a qualitative measure of community dissimilarity which incorporates phylogenetic relationships between the features
- Weighted UniFrac distance: a quantitative measure of community dissimilarity which incorporates phylogenetic relationship between the features

(Bolyen et al., 2019, 2018)

ANCOM

- Analysis of composition of microbiomes
- ANCOM can be used for analyzing the composition of microbiomes in multiple populations (Mandal et al., 2015)
- Differential abundance testing

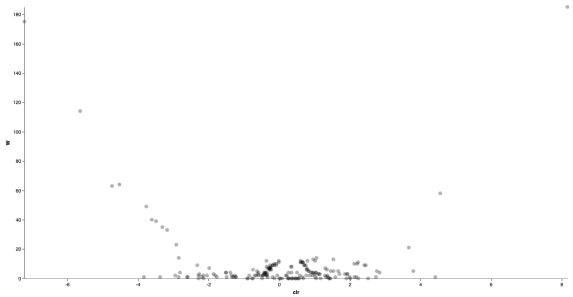


Figure: Example ANCOM Volcano Plot (Bolyen et al., 2019, 2018)

- clr: Centered log Ratio
- W: a count of the number of sub-hypothesis which have passed for given species

Python Packages

- Pandas (McKinney et al., 2011)
- Scikit-learn (Pedregosa et al., 2011)
- Matplotlib (Hunter, 2007; Barrett, Hunter, Miller, Hsu, & Greenfield, 2005)
- Seaborn (Waskom & the seaborn development team, 2020)

t-SNE

- t-distributed stochastic neighbor embedding
- reveals high-dimensional data a location in two-dimensional map (Maaten & Hinton, 2008)

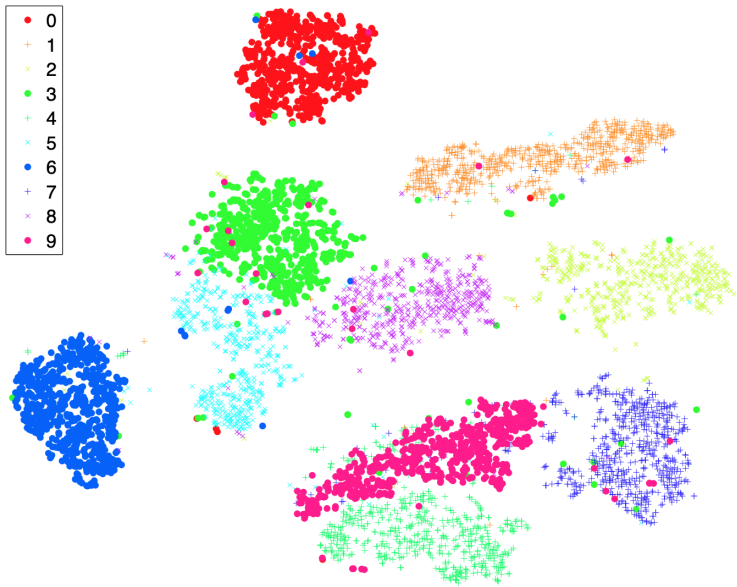


Figure: Visualization by t-SNE (Maaten & Hinton, 2008)

Classification I

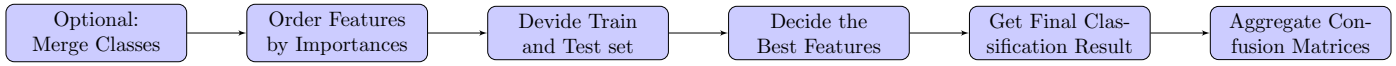


Figure: Workflow of Classification

Classification Metrics:

- Accuracy
- Balanced Accuracy
- Sensitivity
- Specificity
- Precision

Classification II

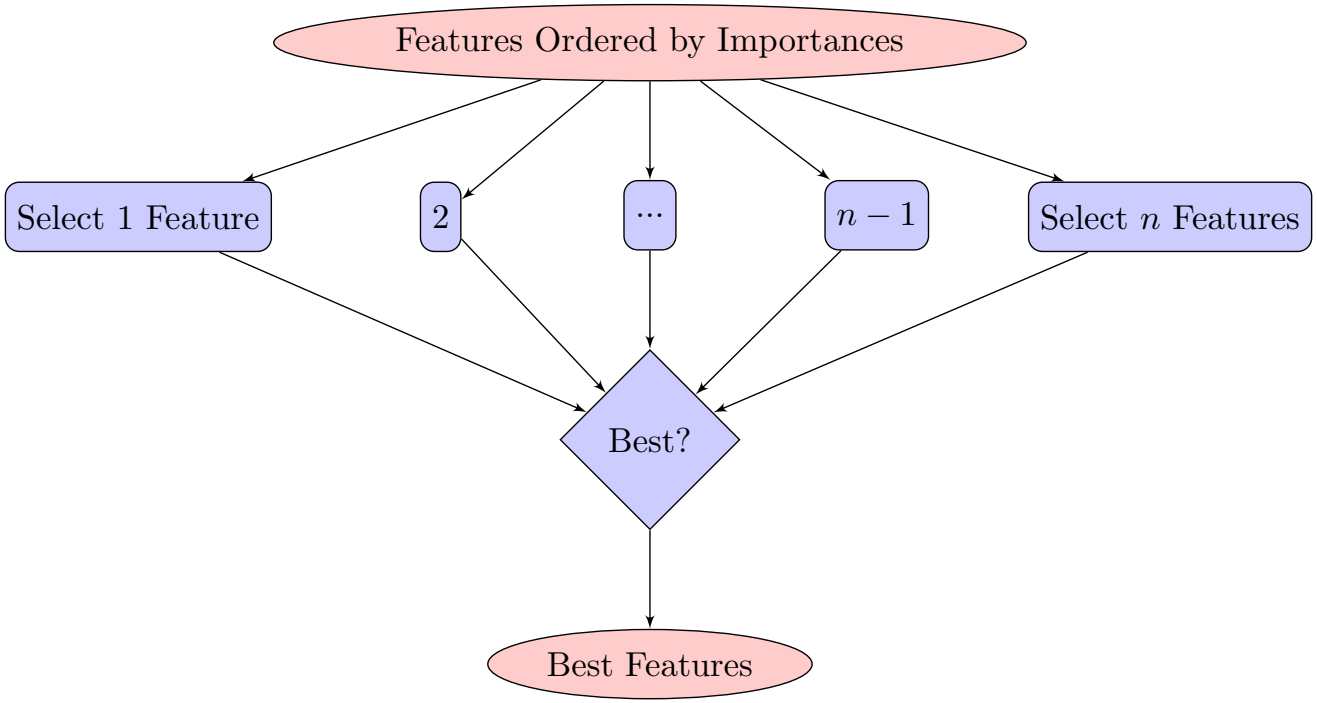
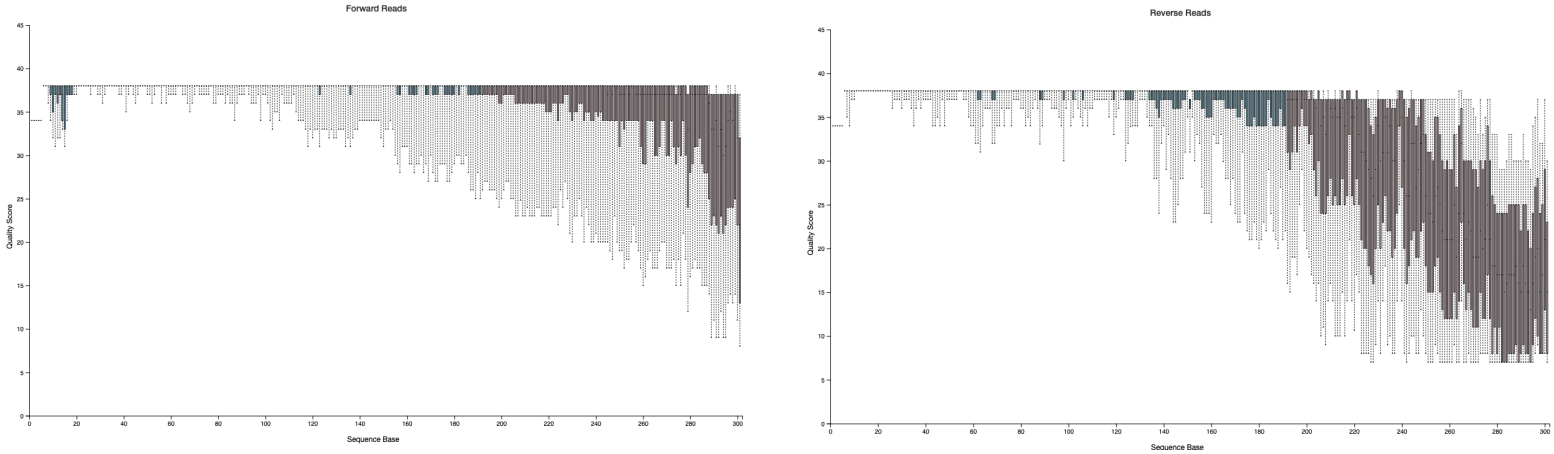


Figure: Deciding the Best Features

Results

Quality Filter



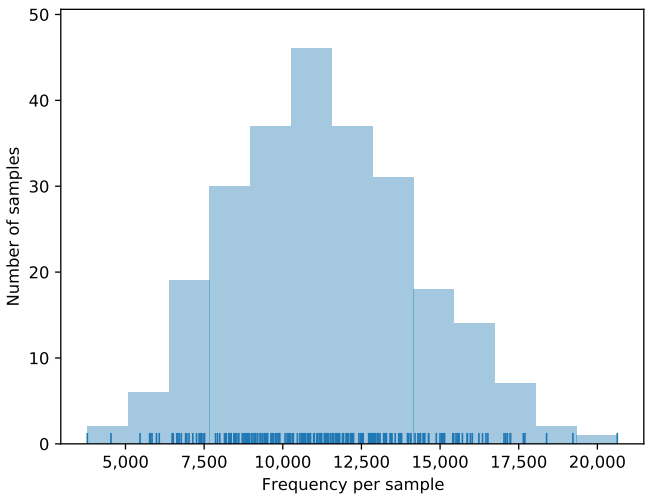
(a) Forward Reads (b) Reverse Reads

Figure: Sequence Quality Plot

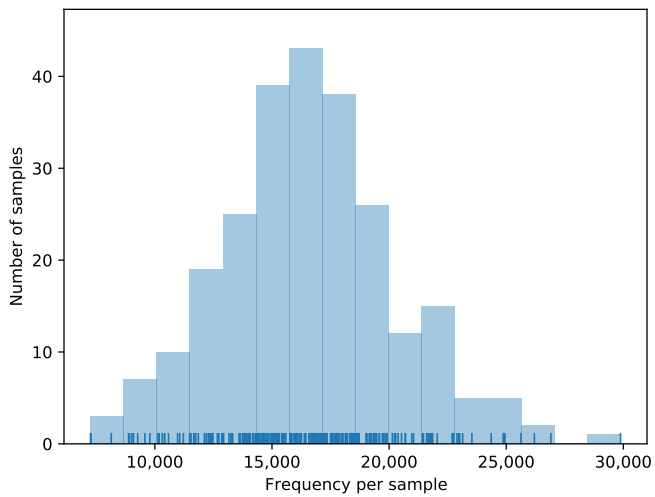
$\therefore$ Maximum Sequence Length $n_{forword} = 300$, $n_{reverse} = 265$

$\because$ The longest length which has sequence quality ≥ 30 at middle.

Rarefaction



(a) DADA2

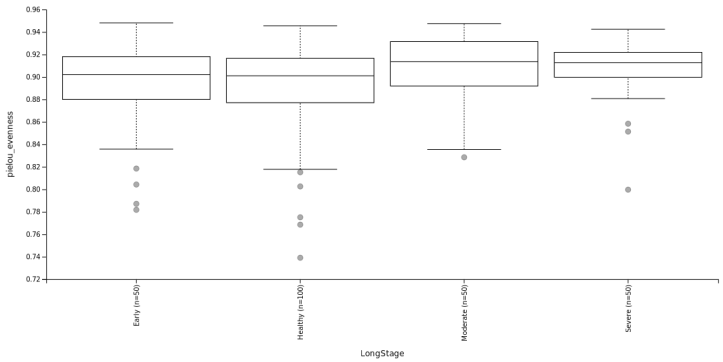


(b) Deblur

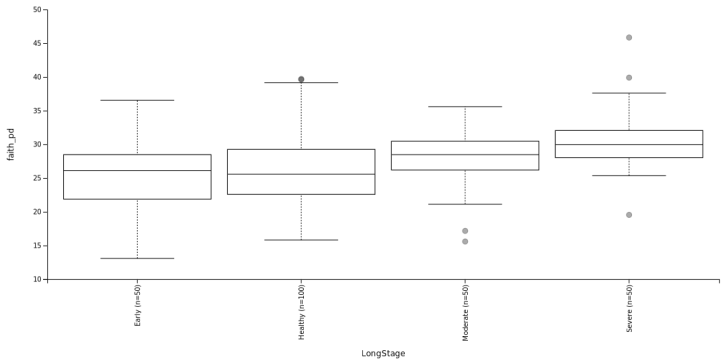
Figure: Frequency per sample

$\therefore$ p-sampling-depth $n_{DADA2} = 3786$ and $n_{Deblur} = 7253$

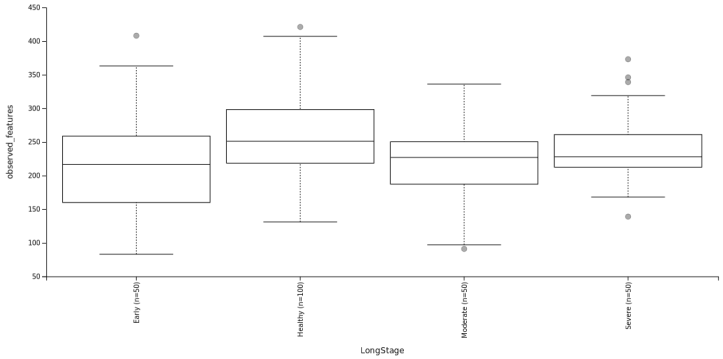
Alpha-diversity I



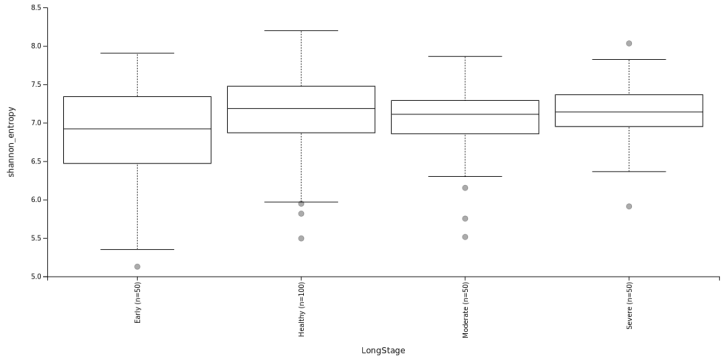
(a) Evenness ($p < 0.01$)



(b) Faith PD ($p < 10^{-6}$)



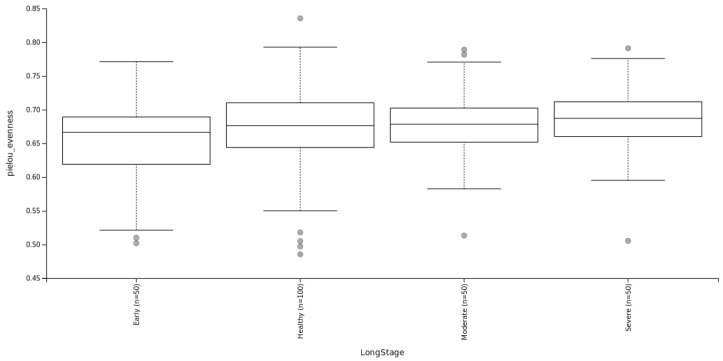
(c) Observed features ($p < 10^{-3}$)



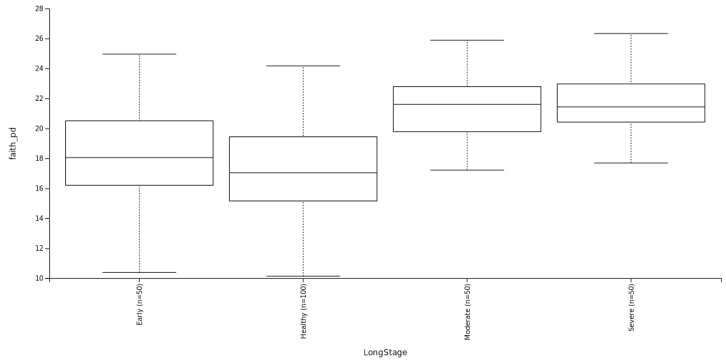
(d) Shannon's diversity ($p > 0.05$)

Figure: Alpha Diversity from DADA2 with Kruskal-Wallis among All Groups

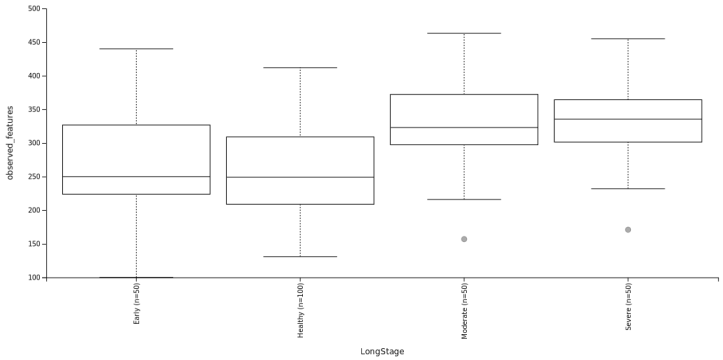
Alpha-diversity II



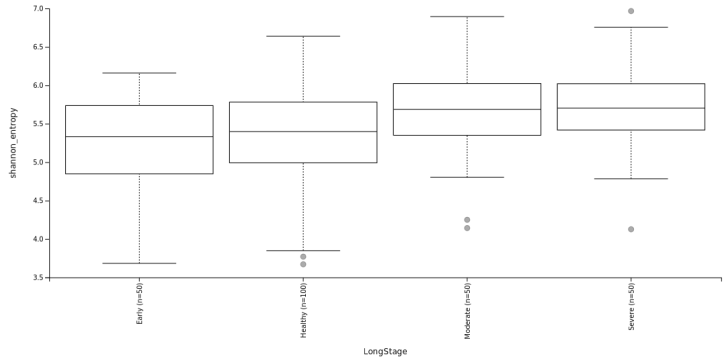
(a) Evenness ($p < 0.05$)



(b) Faith PD ($p < 10^{-18}$)



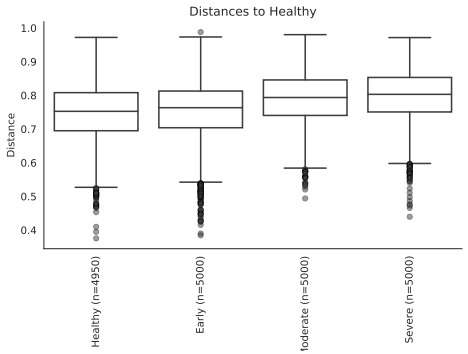
(c) Observed features ($p < 10^{-12}$)



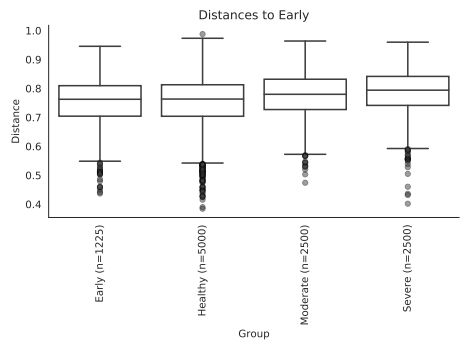
(d) Shannon's diversity ($p < 10^{-4}$)

Figure: Alpha Diversity from Deblur with Kruskal-Wallis among All Groups

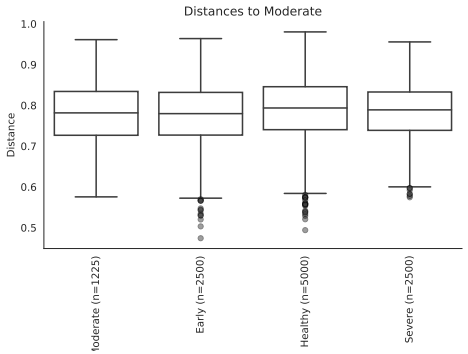
Beta-diversity I



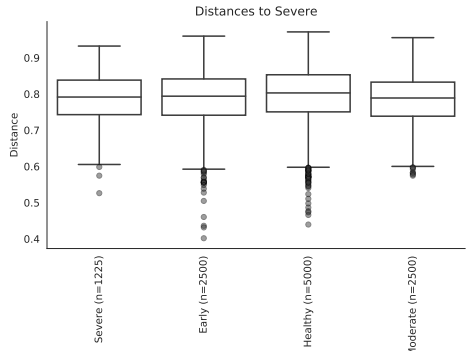
(a) Healthy



(b) Early



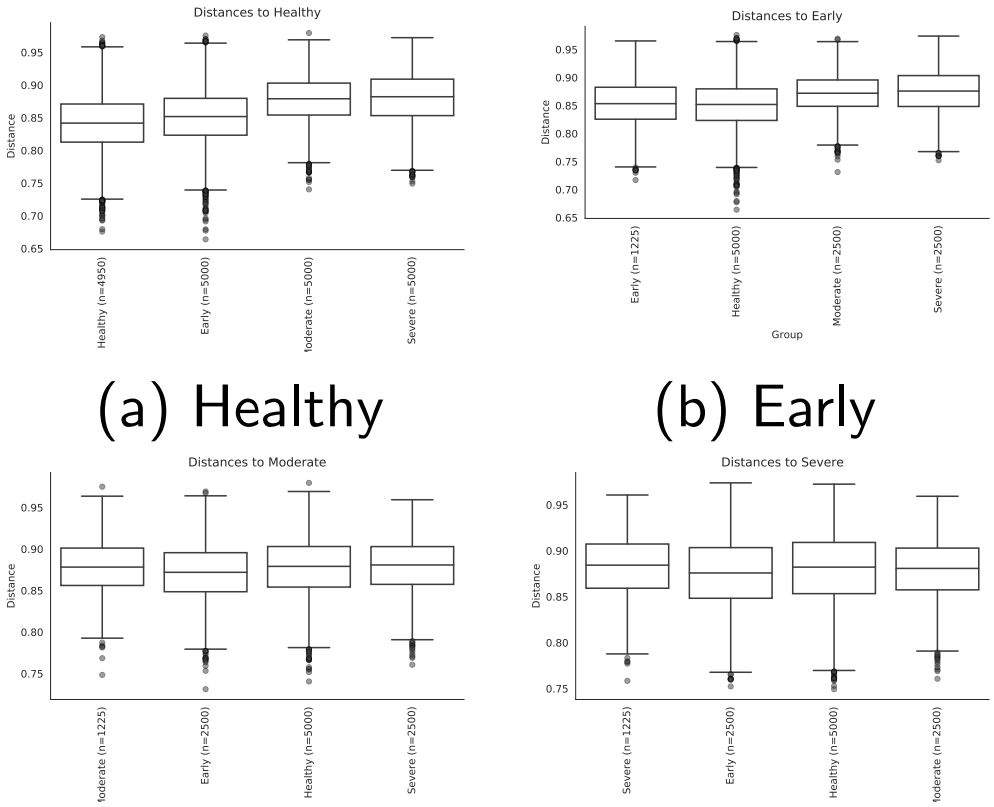
(c) Moderate



(d) Severe

Figure: Bray-Curtis Distance with DADA2

Beta-diversity II

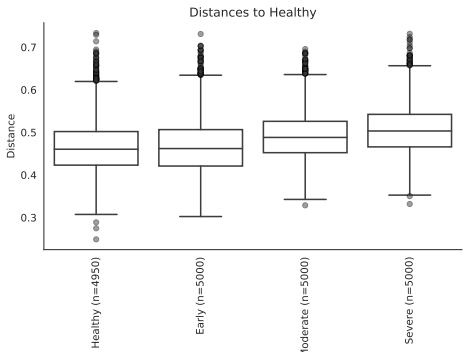


(a) Healthy (b) Early

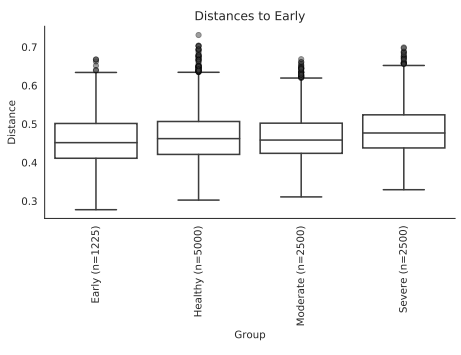
(c) Moderate (d) Severe

Figure: Jaccard Distance with DADA2

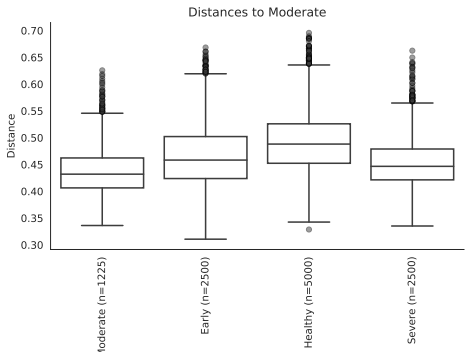
Beta-diversity III



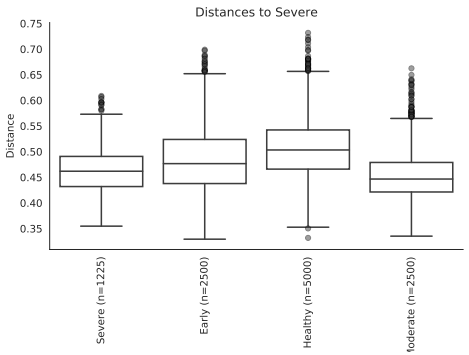
(a) Healthy



(b) Early



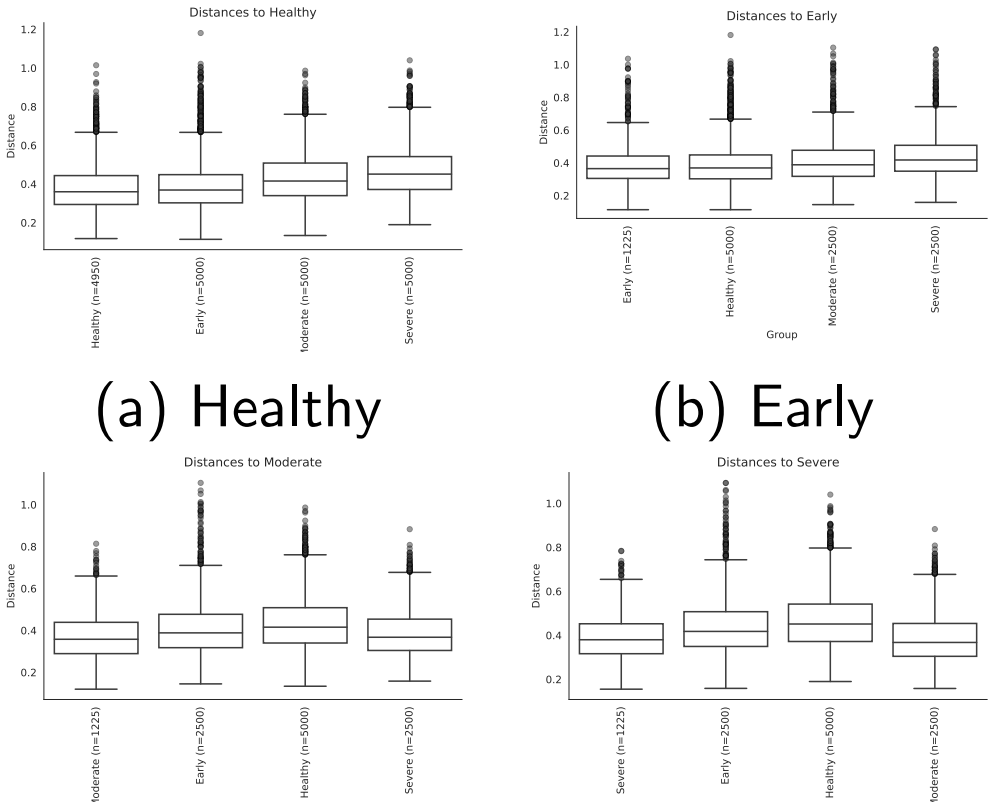
(c) Moderate



(d) Severe

Figure: Unweighted Unifrac Distance with DADA2

Beta-diversity IV

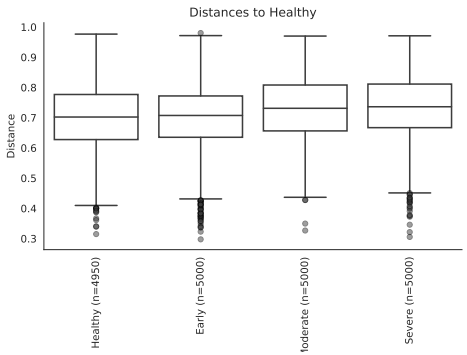


(a) Healthy (b) Early

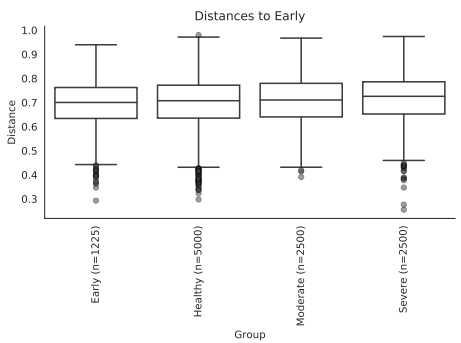
(c) Moderate (d) Severe

Figure: Weighted Unifrac Distance with DADA2

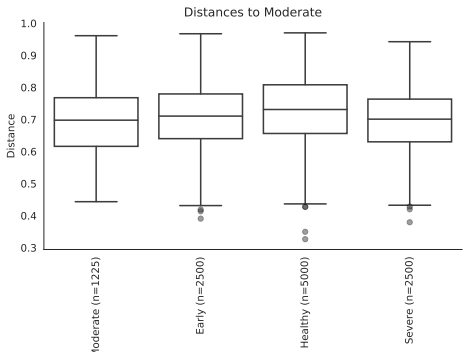
Beta-diversity V



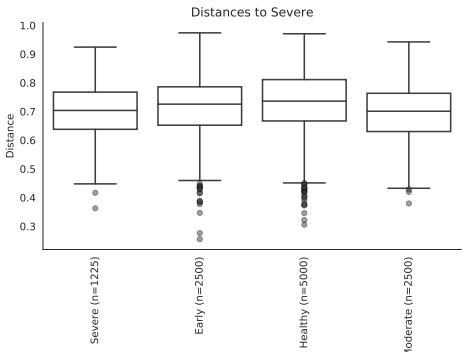
(a) Healthy



(b) Early



(c) Moderate

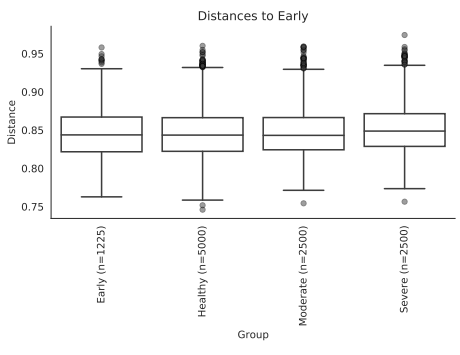


(d) Severe

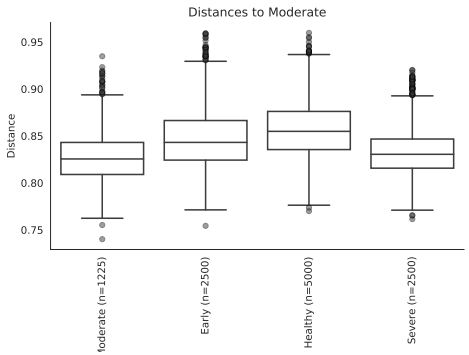
Figure: Bray-Curtis Distance with Deblur

Beta-diversity VI

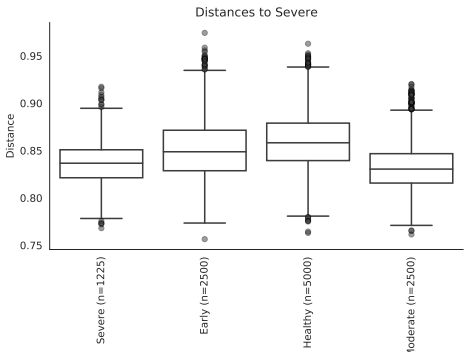
(a) Healthy



(b) Early



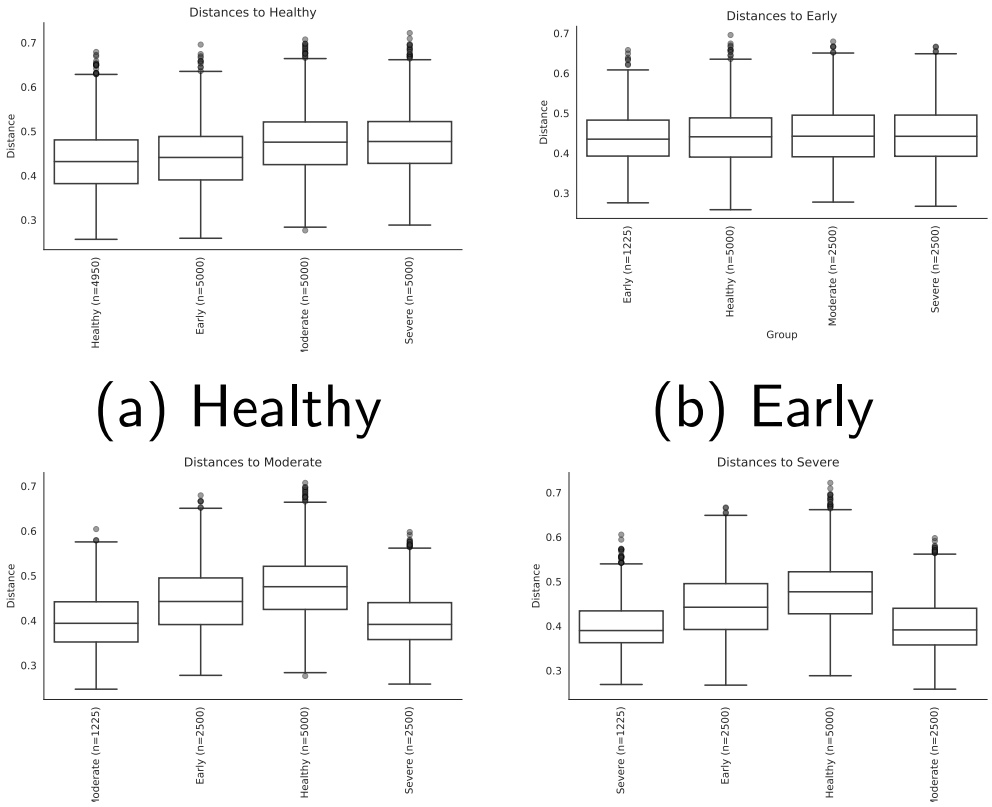
(c) Moderate



(d) Severe

Figure: Jaccard Distance with Deblur

Beta-diversity VII

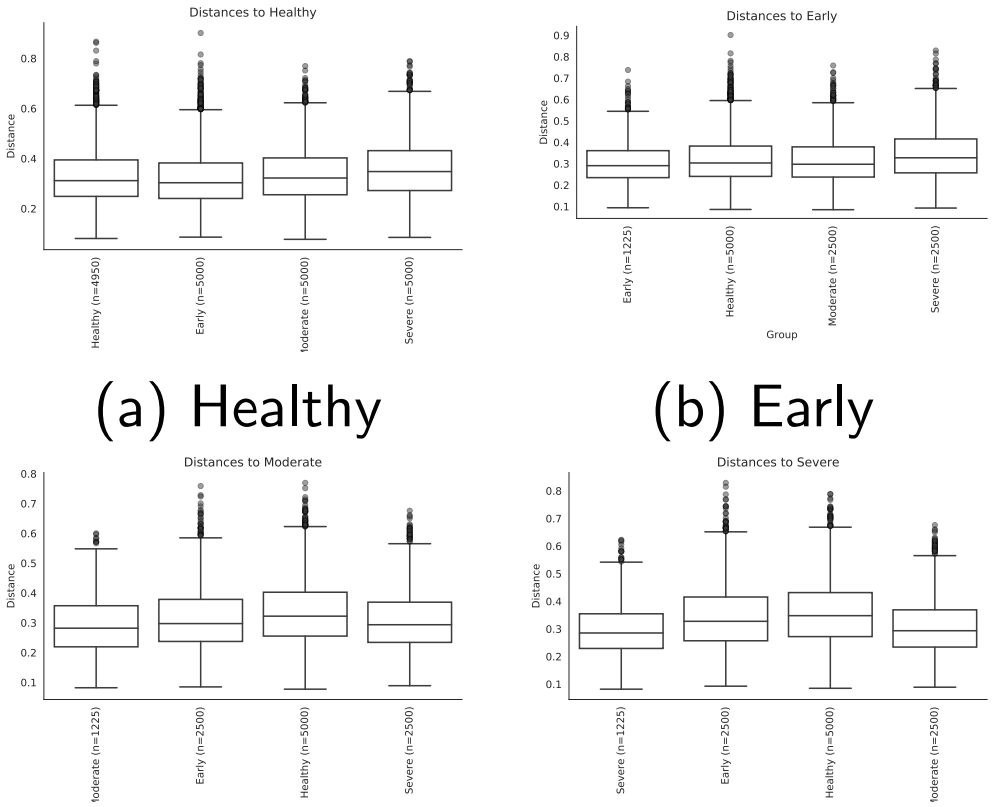


(a) Healthy (b) Early

(c) Moderate (d) Severe

Figure: Unweighted Unifrac Distance with Deblur

Beta-diversity VIII

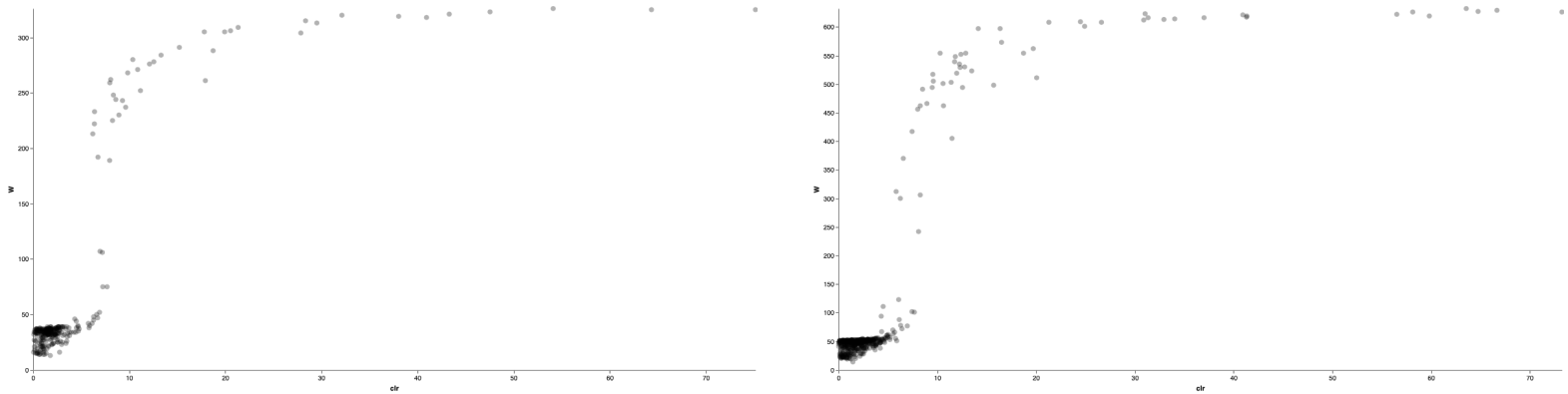


(a) Healthy (b) Early

(c) Moderate (d) Severe

Figure: Weighted Unifrac Distance with Deblur

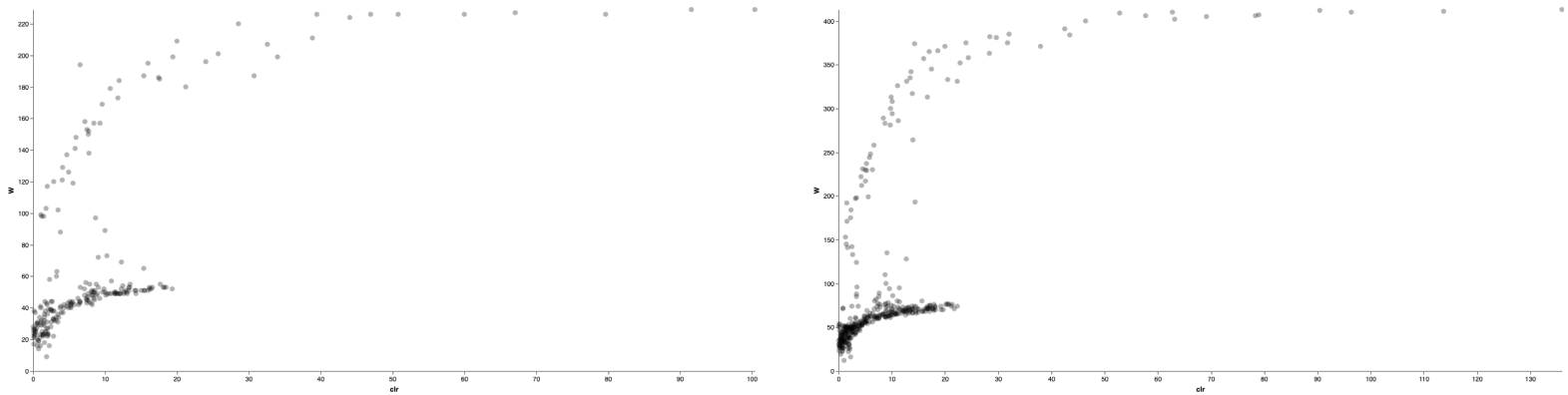
ANCOM I



(a) Greengenes (b) SILVA

Figure: ANCOM Volcano Plot with DADA2

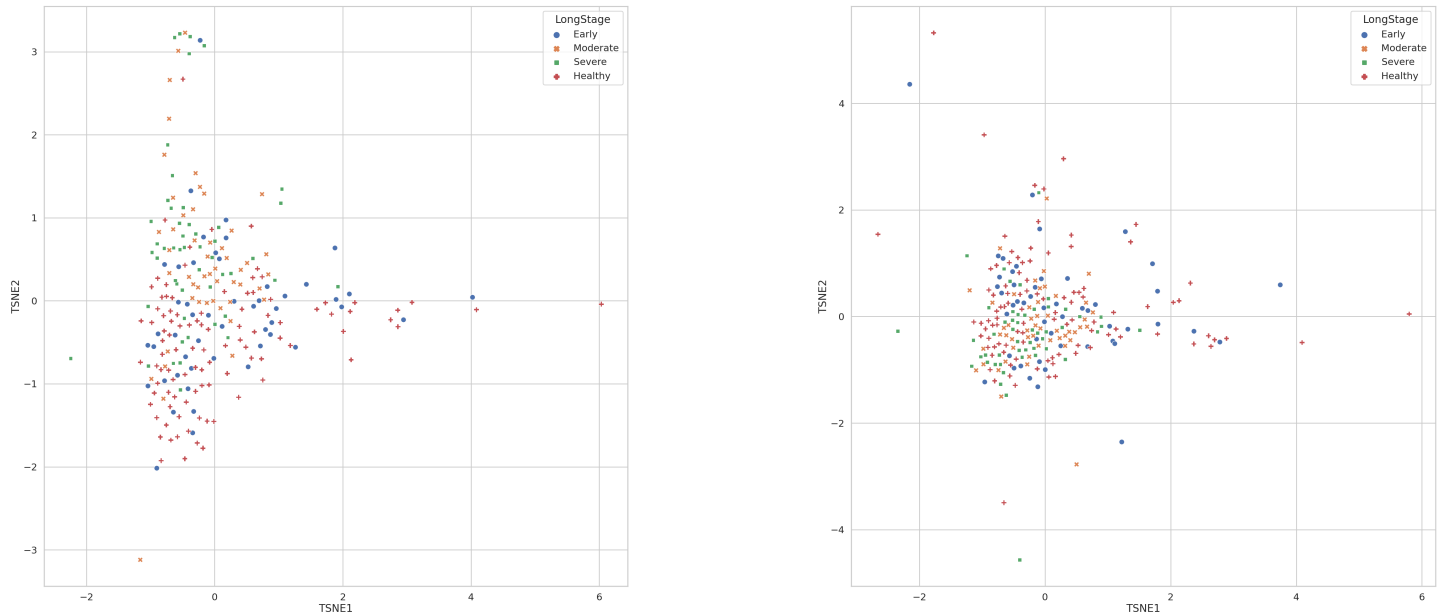
ANCOM II



(a) Greengenes (b) SILVA

Figure: ANCOM Volcano Plot with Deblur

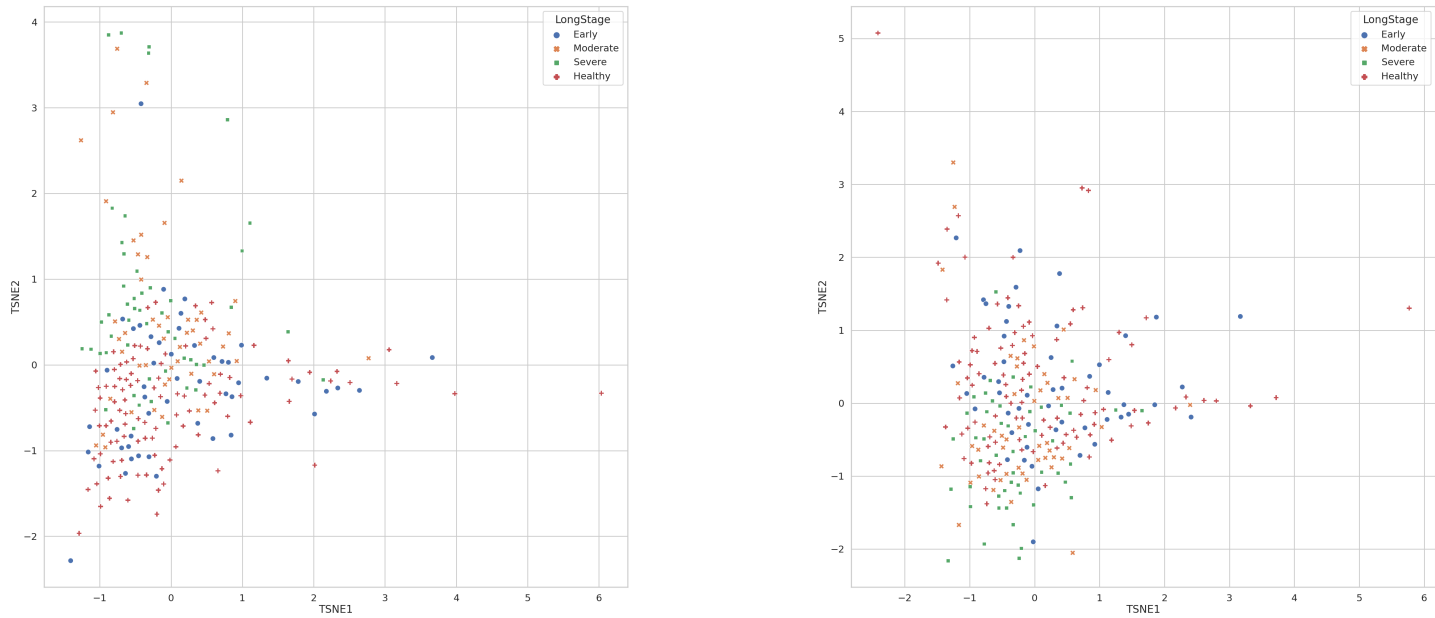
t-SNE with Whole Microbiome I



(a) Greengenes (328 Taxa) (b) SILVA (633 Taxa)

Figure: t-SNE Plot with Whole Microbiome from DADA2

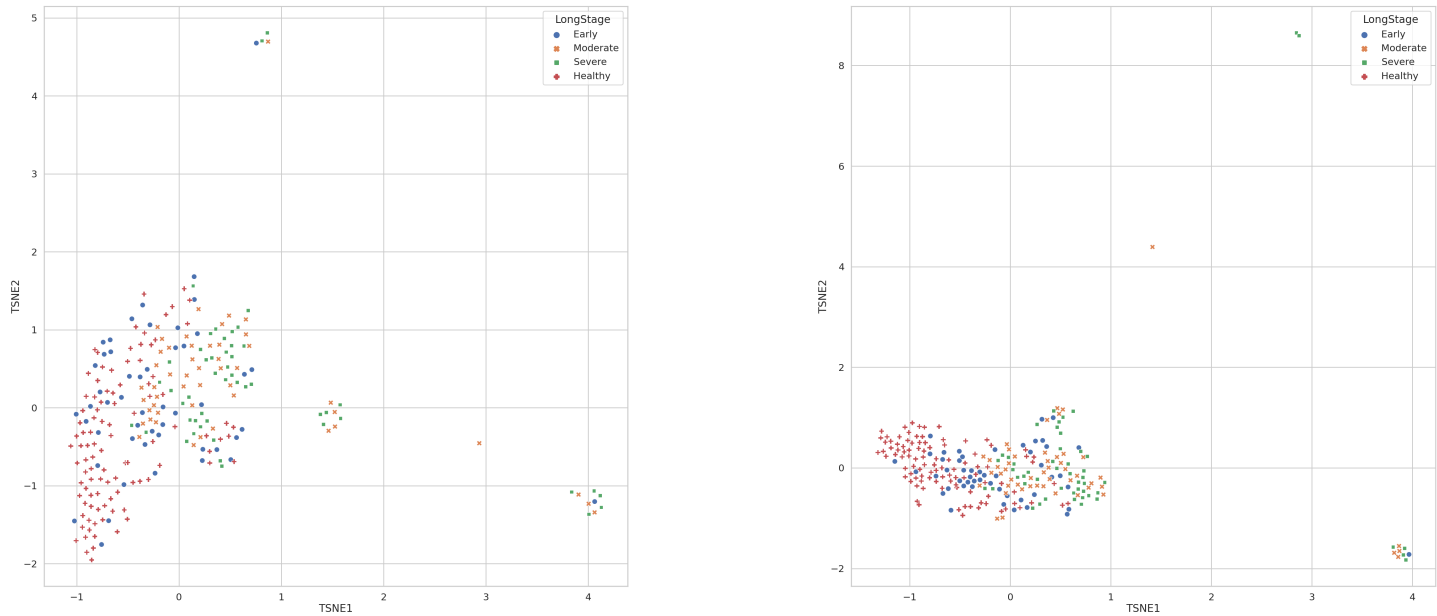
t-SNE with Whole Microbiome II



(a) Greengenes (232 Taxa) (b) SILVA (414 Taxa)

Figure: t-SNE Plot with Whole Microbiome from Deblur

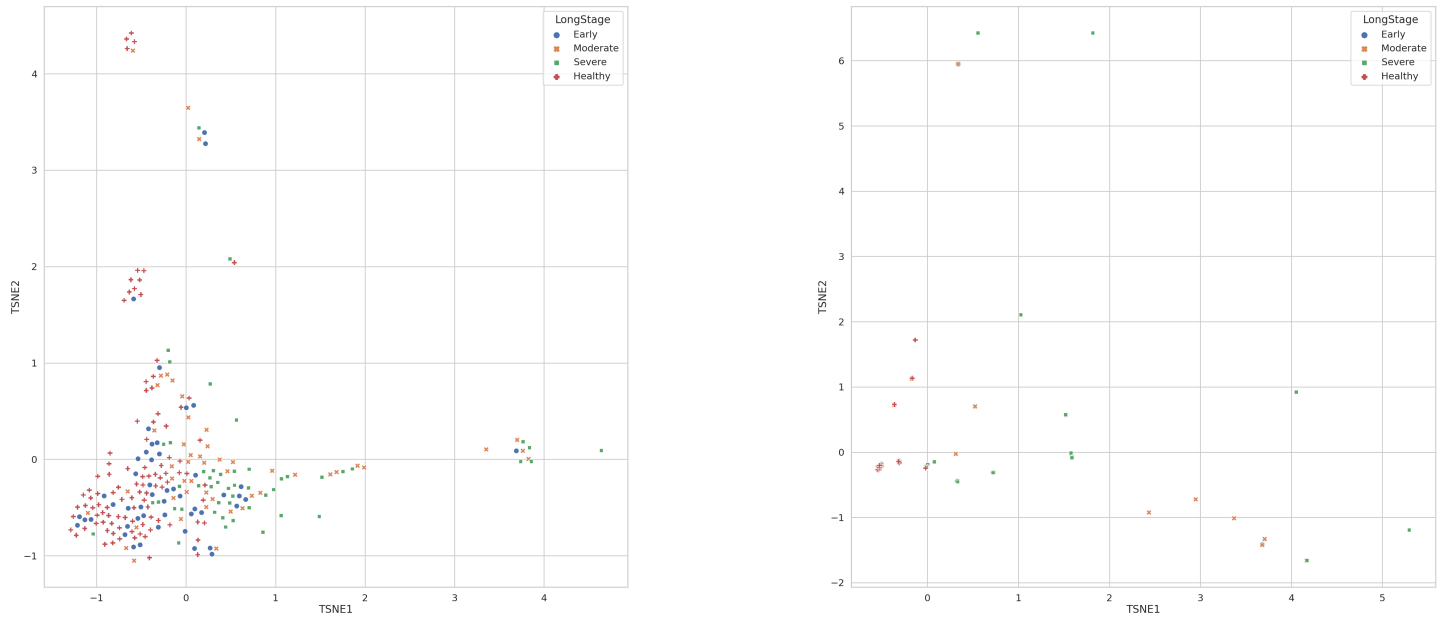
t-SNE with ANCOM Selected I



(a) Greengenes (15 Taxa) (b) SILVA (23 Taxa)

Figure: t-SNE Plot with ANCOM Selected from DADA2

t-SNE with ANCOM Selected II



(a) Greengenes (27 Taxa) (b) SILVA (20 Taxa)

Figure: t-SNE Plot with ANCOM Selected from Deblur

RandomForest Classifier I

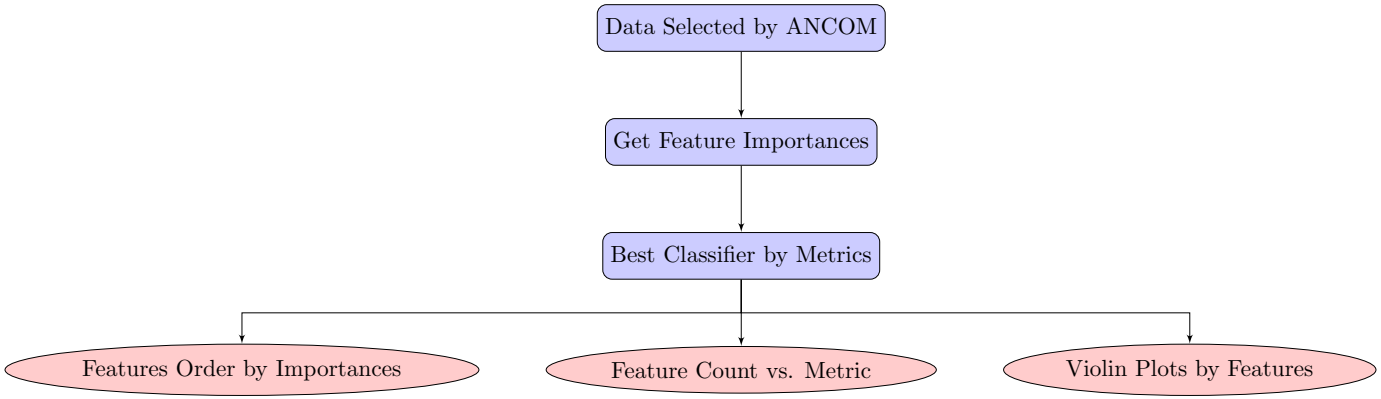


Figure: RandomForest Classifier Workflow

DADA2 + SILVA gives the best result with many metrics.

RandomForest Classifier II

Table: Features Order by Importances

Order	Taxonomy (Genus [Species])
1	*Actinomyces*
2	*Actinomyces Schaalia-odontolytica*
3	*Prevotella Prevotella-intermedia*
4	*Filifactor Filifactor-alocis*
5	*Oribacterium*
6	*Tannerella Tannerella-forsythia*

RandomForest Classifier III

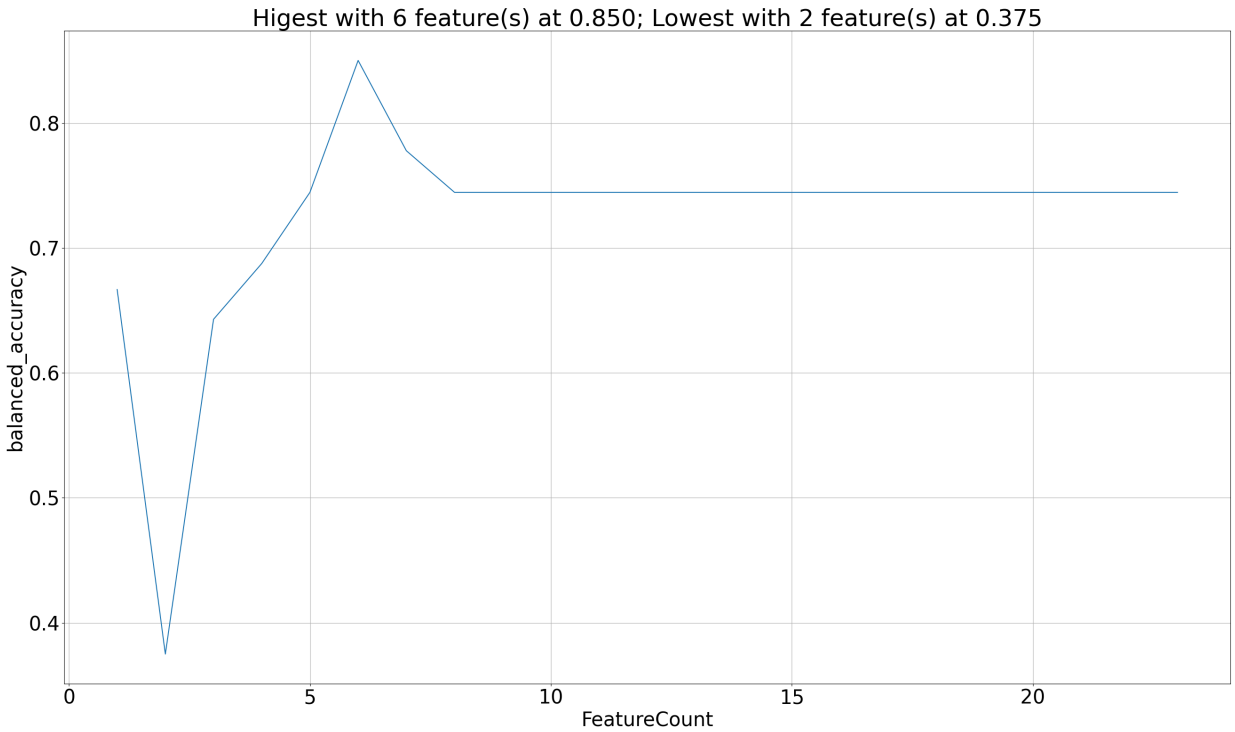
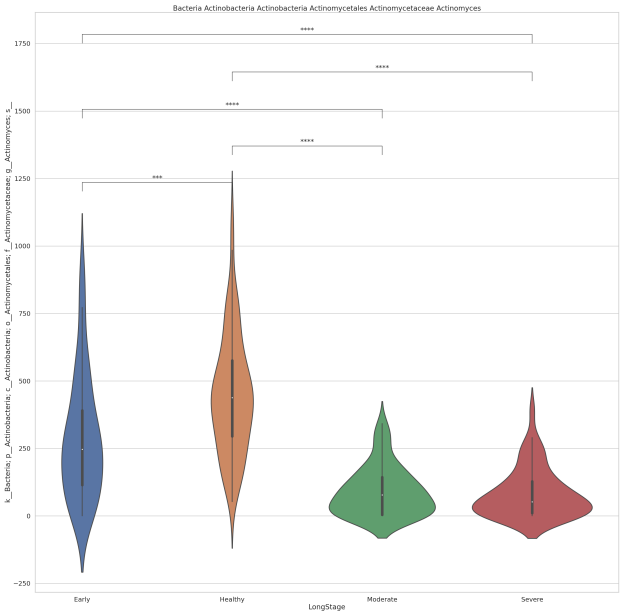
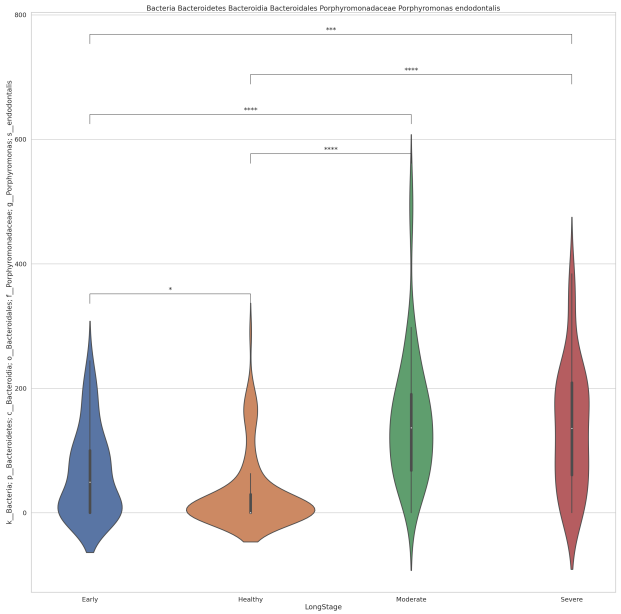


Figure: Balanced Accuracy by Feature Count

RandomForest Classifier IV



(a) *Actinomyces*



(b) *Actinomyces Schaalia-odontolytica*

Figure: Violin Plot by Features

RandomForest Classifier V

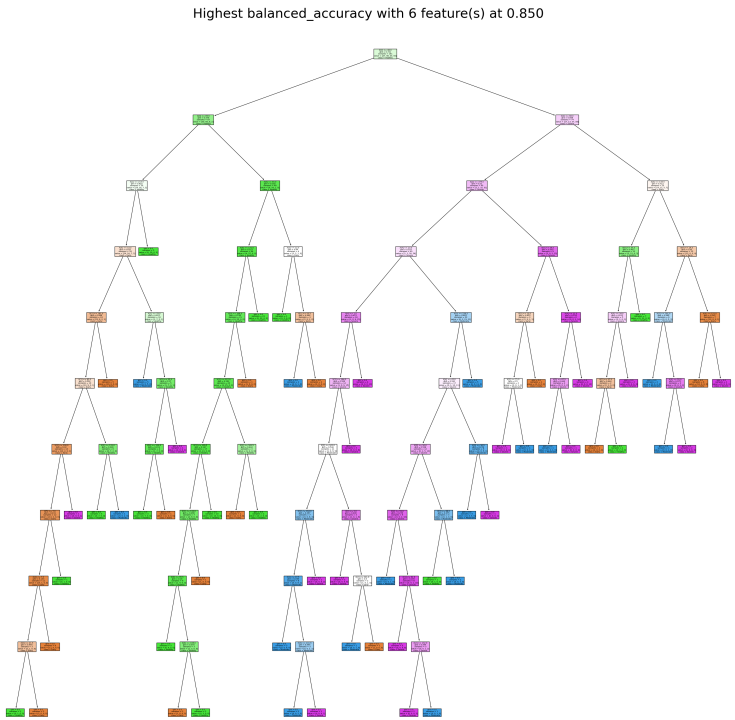


Figure: RandomForest Tree Plot

RandomForest Classifier – Merge Healthy+Early I

Discussion

References I

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.

References II

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.

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.

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.

References III

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.

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.

References IV

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.

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.

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.

References V

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.