Periodontitis

Seunghoon Kim   Jaewoong Lee   Semin Lee

Ulsan National Institute of Science and Technology

jwlee230@unist.ac.kr

2020-12-28

Overview

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)

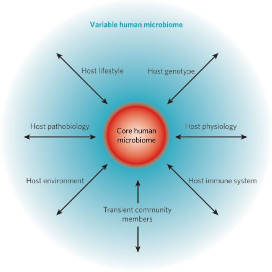


Figure: Concept of a core human microbiome (Turnbaugh et al., 2007)

rRNA

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

Periodontitis (Periodontal disease)

- Clinical Attachment Loss & 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

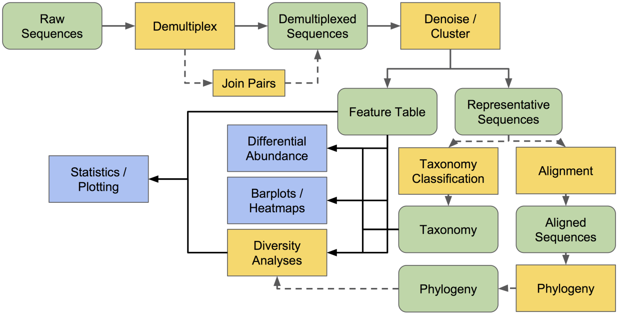


Figure: QIIME2 Workflow (Bolyen et al., 2019, 2018)

Denoising techniques

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

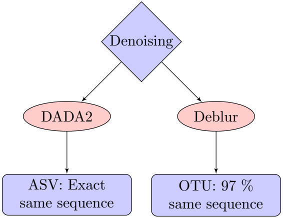


Figure: Denoising Techniques

Taxonomy Classification

- Greengenes (GG) (DeSantis et al., 2006)
- SILVA (Pruesse et al., 2007)
- Human Oral Microbiome Database (HOMD) (Chen et al., 2010)

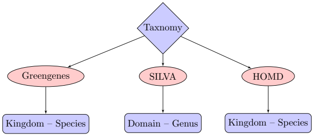


Figure: Taxonomy Classification

"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.

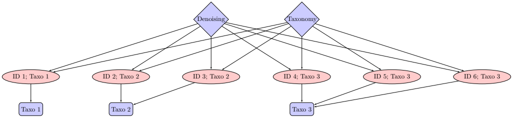


Figure: Example Diagram for Merging Denosing and Taxonomy Classification

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**

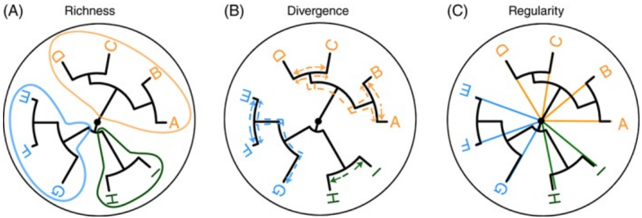


Figure: Three Dimensions of Phylogenetic Information (Tucker et al., 2017)

Alpha-diversity

- Evenness: a measurement of diversity in different type at community (Pielou, 1966)
- Faith's Phylogenetic Diversity: a qualitative measurement of community richness which priorities for species conservation, incorporates with taxic diversity (Faith, 1992)
- Observed Features: a number of observed taxa
- Shannon's diversity index: a significant aspect of community richness (Shannon, 1948)

Beta-diversity

- Bray-Curtis distance: a quantitative measurement of dissimilarity among communities (Sørensen, 1948)
- Jaccard distance: a measurement of local distribution among communities (Jaccard, 1912)
- Unweighted UniFrac distance: a qualitative measurement of phylogenetic distances (McDonald et al., 2018)
- Weighted UniFrac distance: a quantitative measurement of phylogenetic distances (McDonald et al., 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

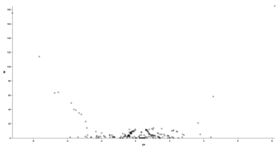


Figure: Example ANCOM Volcano Plot (Bolyen et al., 2019, 2018)

- 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)

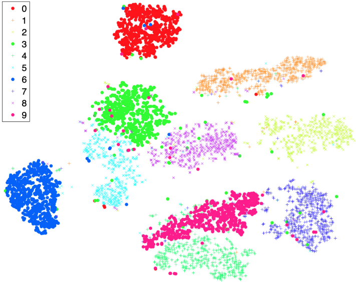


Figure: Visualization by t-SNE (Maaten & Hinton, 2008)

Classification I

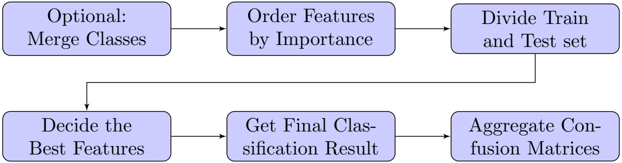


Figure: Workflow of Classification

Classification Metrics:

- Accuracy
- Balanced Accuracy
- Sensitivity

Classification II

- Specificity
- Precision

Classification III

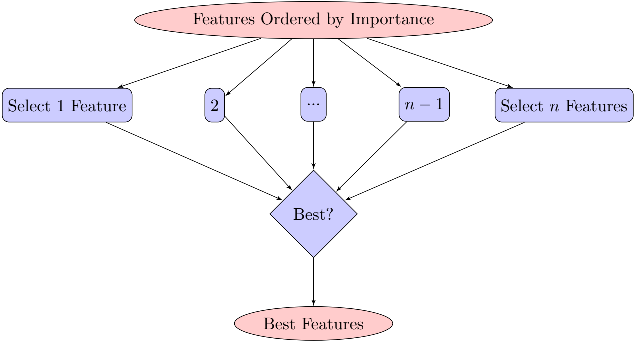
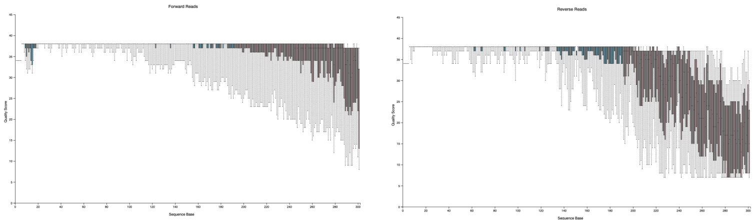


Figure: Deciding the Best Features

Results

Quality Filter



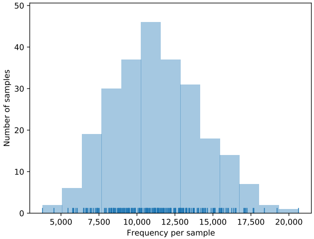
(a) Forward Reads (b) Reverse Reads

Figure: Sequence Quality Plot

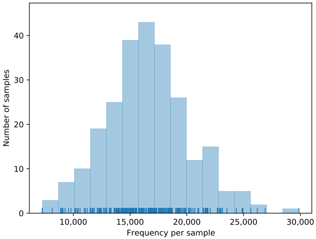
$\therefore$ Maximum Sequence Length $n_{forword} = 300$, $n_{reverse} = 265$

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

Rarefaction



(a) DADA2

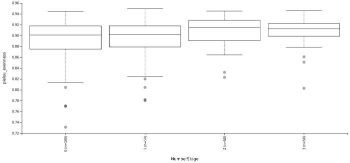


(b) Deblur

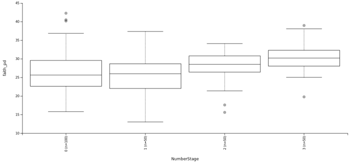
Figure: Frequency per sample

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

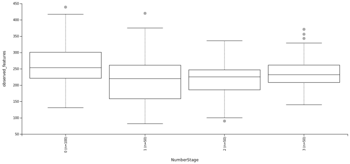
Alpha-diversity I



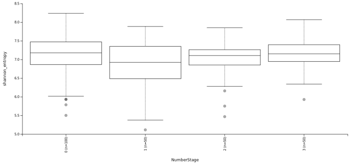
(a) Evenness ($p < 0.01$)



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



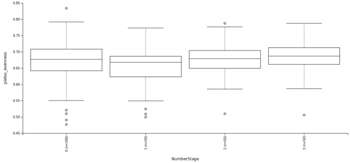
(c) Observed features ($p < 10^{-3}$)



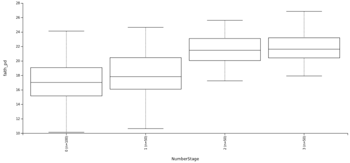
(d) Shannon's diversity ($p > 0.05$)

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

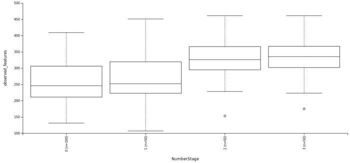
Alpha-diversity II



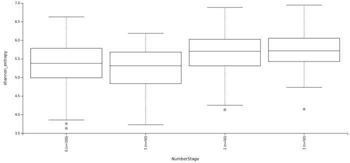
(a) Evenness ($p < 0.05$)



(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

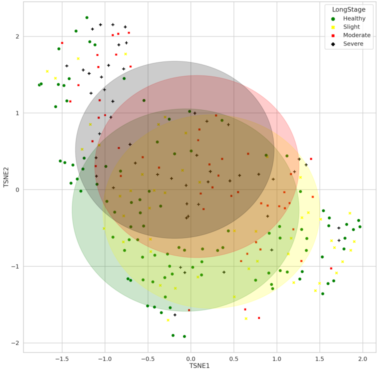


Figure: Bray-Curtis Distance with DADA2

Beta-diversity II

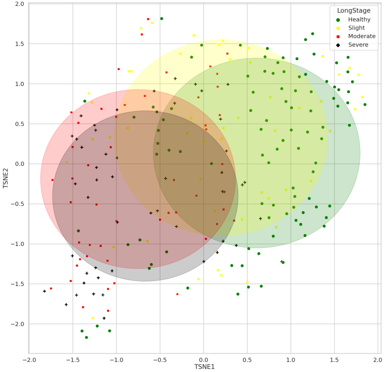


Figure: Jaccard Distance with DADA2

Beta-diversity III

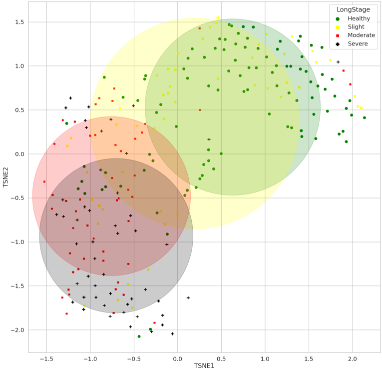


Figure: Unweighted Unifrac Distance with DADA2

Beta-diversity IV

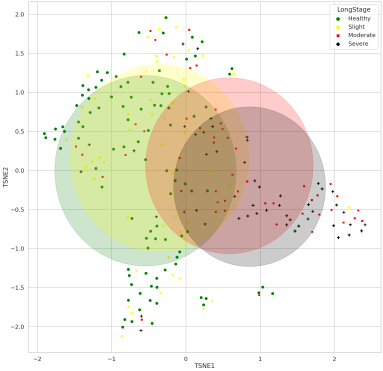


Figure: Weighted Unifrac Distance with DADA2

Beta-diversity V

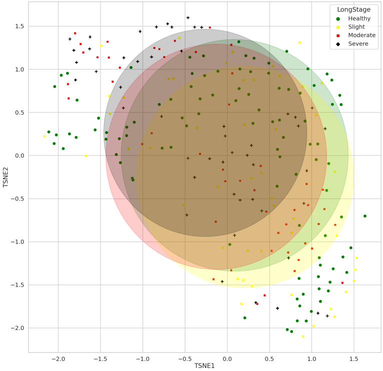


Figure: Bray-Curtis Distance with Deblur

Beta-diversity VI

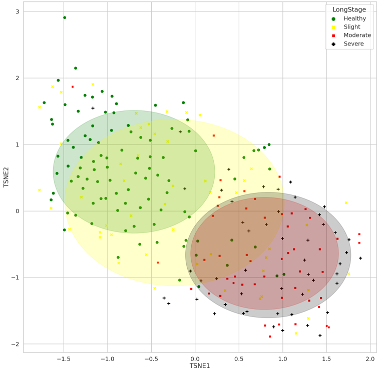


Figure: Jaccard Distance with Deblur

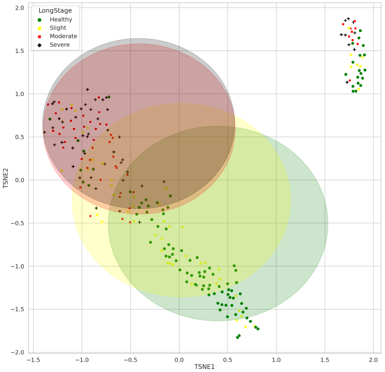


Figure: Unweighted Unifrac Distance with Deblur

Beta-diversity VIII

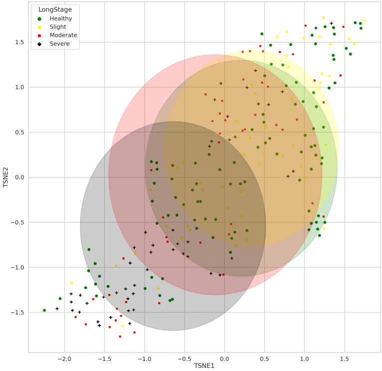
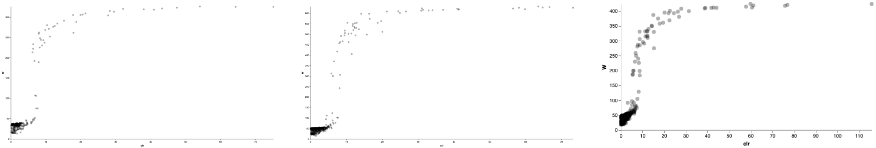


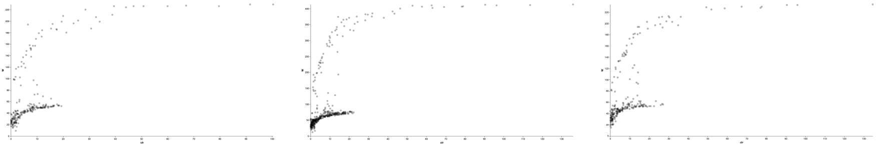
Figure: Weighted Unifrac Distance with Deblur

ANCOM



(a) GG (15 Taxa) (b) SILVA (23 Taxa) (c) HOMD (20 Taxa)

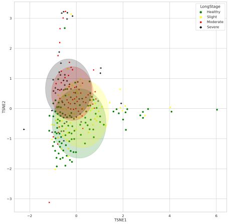
Figure: ANCOM Volcano Plot with DADA2



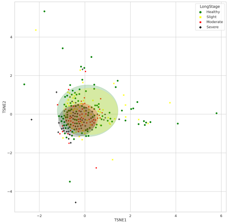
(a) GG (27 Taxa) (b) SILVA (20 Taxa) (c) HOMD (28 Taxa)

Figure: ANCOM Volcano Plot with Deblur

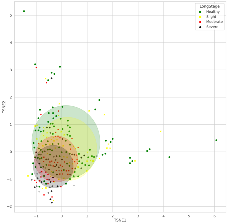
t-SNE with Whole Microbiome I



(a) GG (328 Taxa)



(b) SILVA (633 Taxa)



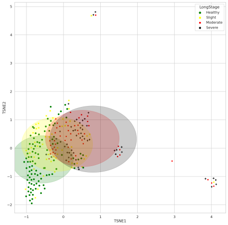
(c) HOMD (425 Taxa)

Figure: t-SNE Plot with Whole Microbiome from DADA2

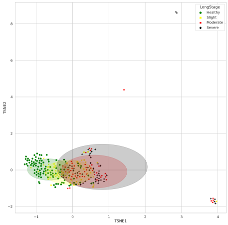
(a) GG (232 Taxa) (b) SILVA (414 Taxa) (c) HOMD (235 Taxa)

Figure: t-SNE Plot with Whole Microbiome from Deblur

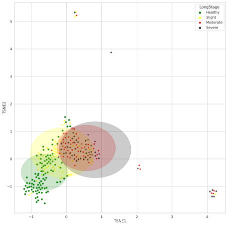
t-SNE with ANCOM Selected I



(a) GG (15 Taxa)



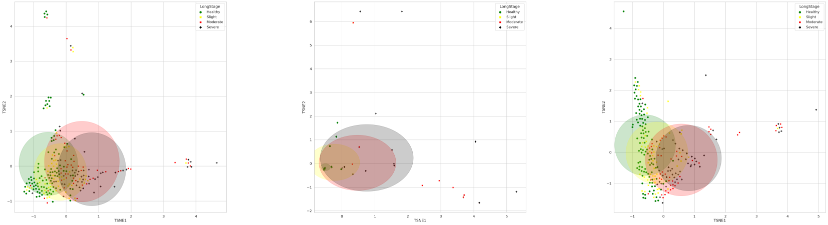
(b) SILVA (23 Taxa)



(c) HOMD (20 Taxa)

Figure: t-SNE Plot with ANCOM Selected from DADA2

t-SNE with ANCOM Selected II



(a) GG (27 Taxa) (b) SILVA (20 Taxa) (c) HOMD (28 Taxa)

Figure: t-SNE Plot with ANCOM Selected from Deblur

Random Forest Classifier – Every Classes I

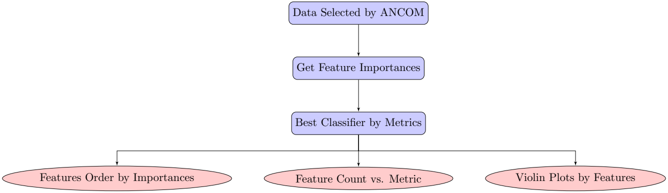


Figure: Random Forest Classifier Workflow

Deblur + HOMD gives the best result with balanced accuracy (0.778 with 13 taxa).

Random Forest Classifier – Every Classes II

Table: Features Order by Importances

Order	Taxonomy (Genus [Species])
0	*Porphyromonas gingivalis*
1	*Actinomyces*
2	*Pasteurellaceae*
3	*Veillonella denticariosi*
4	*Oribacterium sinus*
5	*Peptostreptococcus anaerobius*
6	*Prevotella nanceiensis*
7	*Treponema*
8	*Parvimonas sp. HMT 393*
9	*Filifactor alocis*
...	...

Random Forest Classifier – Every Classes III

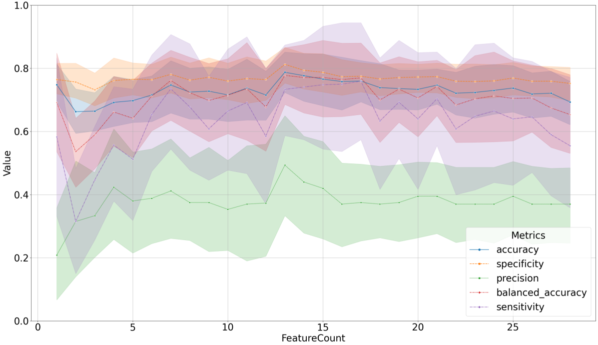
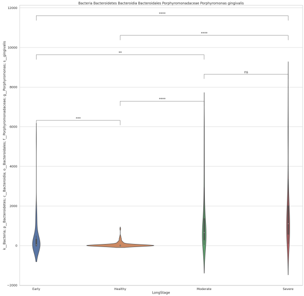
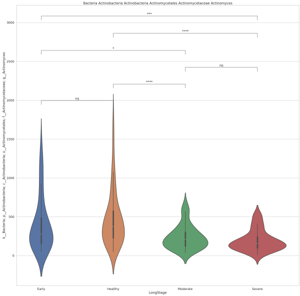


Figure: Metrics by Feature Count

Random Forest Classifier – Every Classes IV



(a) *Porphyromonas gingivalis*



(b) *Actinomyces*

Figure: Violin Plot by Features

Random Forest Classifier – Merge (Healthy+Early) I

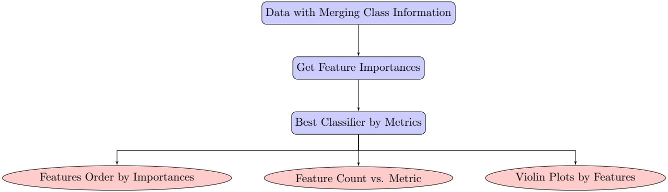


Figure: Random Forest Classifier with Merging Workflow

DADA2 + HOMD gives the best result with balanced accuracy (0.938 with 11 taxa).

Random Forest Classifier – Merge (Healthy+Early) II

Table: Features Order by Importances

Order	Taxonomy (Genus [Species])
0	*Porphyromonas gingivalis*
1	*Actinomyces*
2	*Prevotella intermedia*
3	*Filifactor alocis*
4	*Treponema sp. HMT 260*
5	*Campylobacter showae*
6	*Porphyromonas sp. HMT 285*
7	*Peptostreptococcaceae XIG-6 nodatum*
8	*Prevotella sp. HMT 304*
9	*Actinomyces graevenitzii*
10	*Tannerella forsythia*

Random Forest Classifier – Merge (Healthy+Early) III

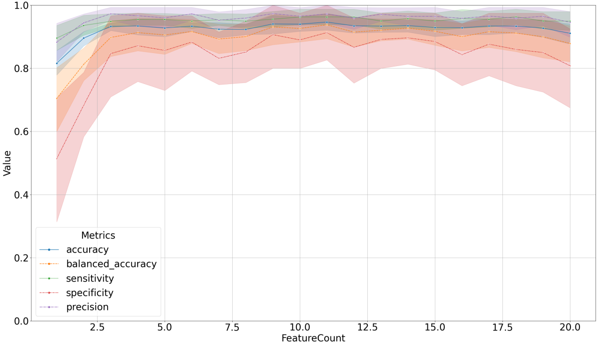
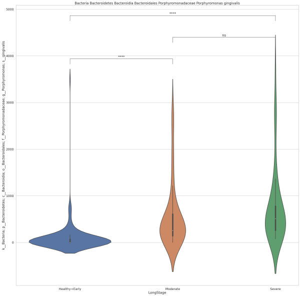
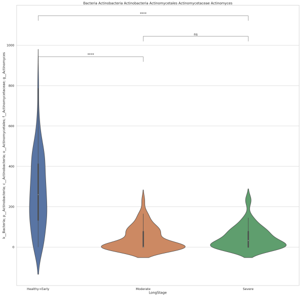


Figure: Metrics by Feature Count

Random Forest Classifier – Merge (Healthy+Early) IV



(a) *Porphyromonas gingivalis*



(b) *Actinomyces*

Figure: Violin Plot by Features

Random Forest Classifier – Merge (Moderate + Severe) I

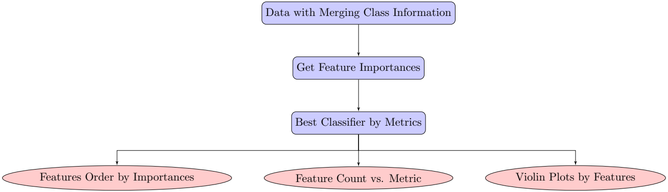


Figure: Random Forest Classifier with Merging Workflow

Deblur + HOMD gives the best result with balanced accuracy (0.777 with 28 taxa).

Random Forest Classifier – Merge (Moderate + Severe) II

Table: Features Order by Importances

Order	Taxonomy (Genus [Species])
0	*Porphyromonas gingivalis*
1	*Actinomyces*
2	*Pasteurellaceae*
3	*Veillonella denticariosi*
4	*Oribacterium sinus*
5	*Peptostreptococcus anaerobius*
6	*Treponema*
7	*Prevotella nanceiensis*
8	*Porphyromonas sp. HMT 285*
9	*Parvimonas sp. HMT 393*
...	...

Random Forest Classifier – Merge (Moderate + Severe) III

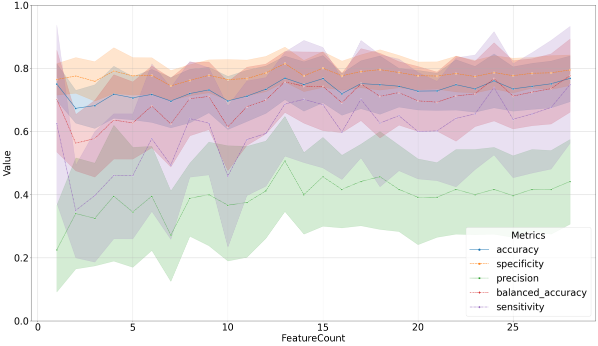
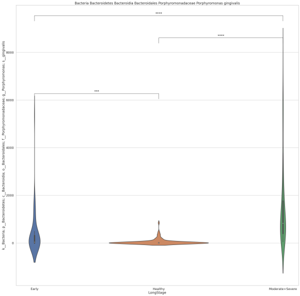
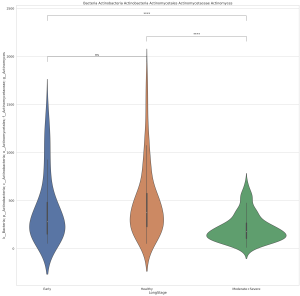


Figure: Metrics by Feature Count

Random Forest Classifier – Merge (Moderate + Severe) IV



(a) *Porphyromonas gingivalis*



(b) *Actinomyces*

Figure: Violin Plot by Features

Random Forest Classifier – Merge (H + E) & (M + S) I

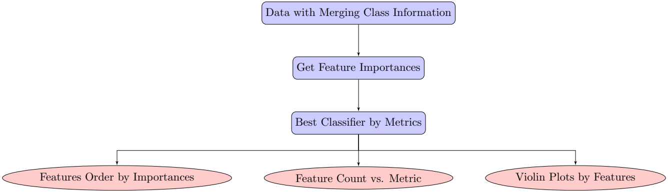


Figure: Random Forest Classifier with Merging Workflow

DADA2 + SILVA gives the best result with balanced accuracy (0.934 with 15 taxa).

Random Forest Classifier – Merge (H + E) & (M + S) II

Table: Features Order by Importances

Order	Taxonomy (Genus [Species])
0	*Actinomyces*
1	*Treponema denticola*
2	*Porphyromonas gingivalis*
3	*Prevotella intermedia*
4	*Filifactor alocis*
5	*Schaalia odontolytica*
6	*Tannerella forsythia*
7	*Eubacterium nodatum*
8	*Corynebacterium durum*
9	*Oribacterium*
...	...

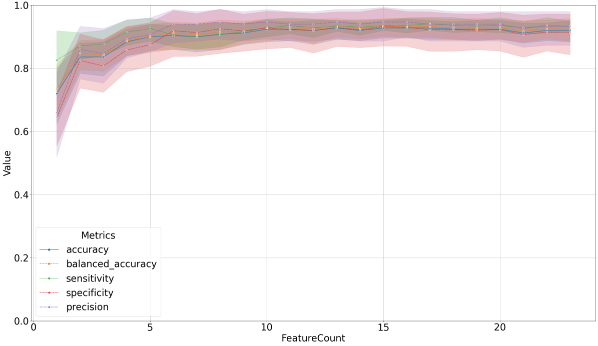
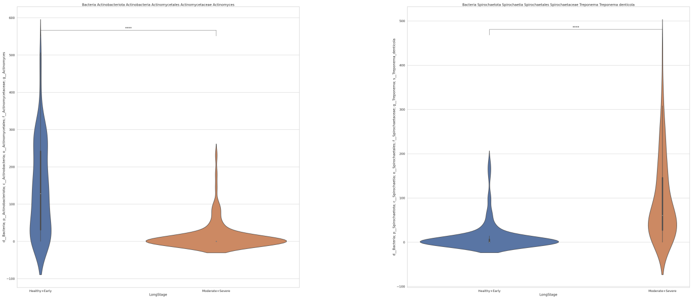


Figure: Metrics by Feature Count

Random Forest Classifier – Merge (H + E) & (M + S) IV



(a) *Actinomyces*　　(b) *Treponema denticola*

Figure: Violin Plot by Features

Random Forest Classifier – Only (H + E) I

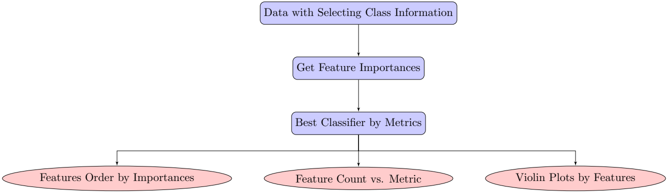


Figure: Random Forest Classifier with Merging Workflow

DADA2 + HOMD gives the best result with balanced accuracy (0.771 with 4 taxa).

Random Forest Classifier – Only (H + E) II

Table: Features Order by Importances

Order	Taxonomy (Genus [Species])
0	*Actinomyces*
1	*Porphyromonas gingivalis*
2	*Actinomyces graevenitzii*
3	*Filifactor alocis*

Random Forest Classifier – Only (H + E) III

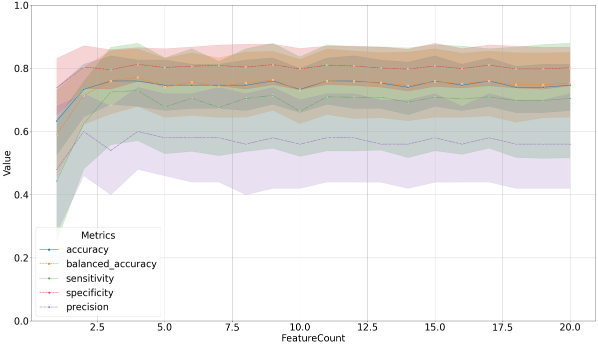
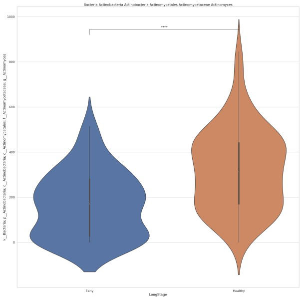
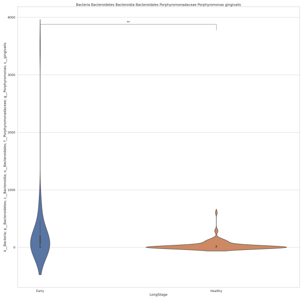


Figure: Metrics by Feature Count



(a) *Actinomyces*



(b) *Porphyromonas gingivalis*

Figure: Violin Plot by Features

Discussion

Porphyromonas Genus

Actinomyces Genus

Treponema Genus

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.

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.

References III

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.

References IV

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.

References V

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.

References VI

Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., ... others (2017). A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews*, *92*(2), 698–715.

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.