

Metagenome Analysis of Preterm Birth

Jaewoong Lee Semin Lee

Department of Biomedical Engineering
Ulsan National Institute of Science and Technology

jwlee230@unist.ac.kr

2022-11-16

Overview

1 Introduction

2 Materials

3 Methods

4 Results

5 Discussion

6 References

1. Introduction

Microbiome

- Microbiota: the microorganisms which live inside & on humans (Turnbaugh et al., 2007)
- Microbiome: 10^{13} to 10^{14} microorganisms whose 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)

Preterm Birth (PTB)

PTB:

- ① PTB < 37 GW (Gestational week)
- ② Normal ≥ 37 GW

Detailed PTB:

- ① Early PTB < 34 GW
- ② 34 GW \leq Late PTB < 37 GW
- ③ Normal ≥ 37 GW

(J. Tucker & McGuire, 2004; Voronkov, Solonovych, Liashenko, & Revenko, 2018)

2. Materials

16S rRNA Sequencing

16S rRNA sequencing is the *reference method* for bacterial taxonomy & identification (Mignard & Flandrois, 2006)

Three main reasons (Janda & Abbott, 2007):

- 16S rRNA exists in almost all bacteria
- Functions of the 16S rRNA has not changed over evolution.
- 16S rRNA is large enough for bioinformatics

Data Composition

- JBNU/Helixco data
 - First data
 - Second data
 - Stool data

Table: Sample Information

Data	Participants	Samples	Remarks
First	24	107	-
Second	35	288	-
Third	10	106	-
Stool	63	126	Stool

3. Methods

Qiime 2 Workflow

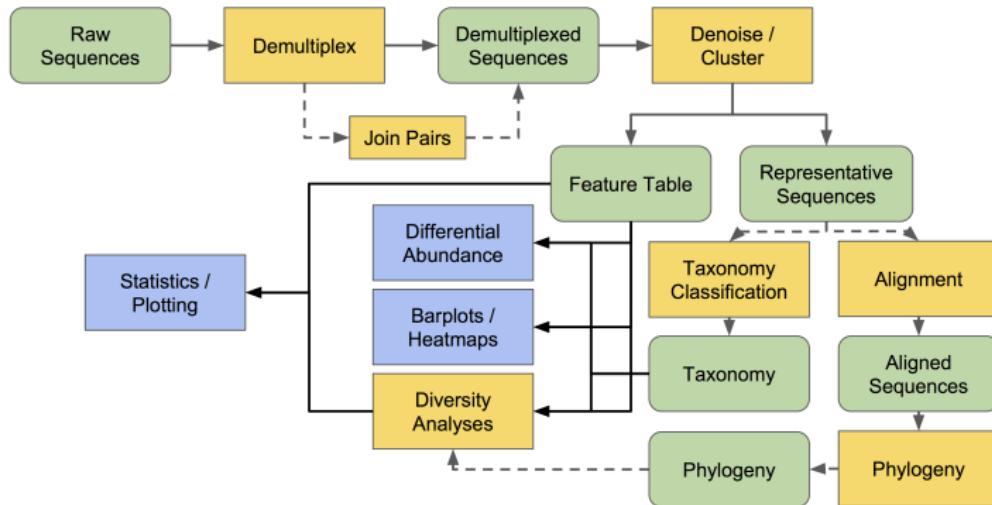


Figure: QIIME 2 workflow (Bolyen et al., 2019; Mandal, Van Treuren, White, Eggesbø, et al., 2015; McDonald et al., 2012)

4. Results

4. Results

4.1. Data Processing with Qiime

Filtering with Quality Score

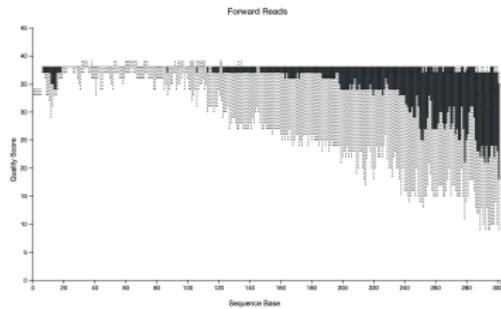
Drawback between:

- Longer sequence read
- Higher quality value

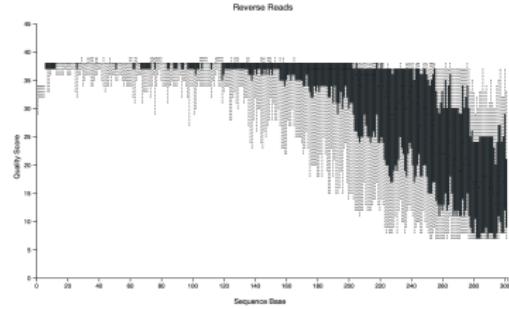
∴ Select the maximum length n where:

$$\begin{aligned} \forall n_i \in \{n_k | \text{MedianQualityScore} \geq 30\} \\ \exists! n \in \{n_i\} : n \geq n_i \end{aligned} \tag{1}$$

Quality Score from JBU/Helixco Data



(a) Forward



(b) Reverse

Figure: Quality Score Plot

- $\ell_{Forward} = 300$
- $\ell_{Reverse} = 245$

Denoising Techniques

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

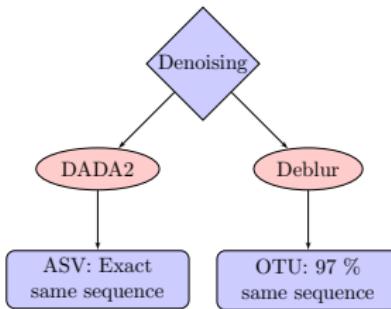


Figure: Denoising Algorithms

Taxonomy Classification

- Greengenes (GG) (DeSantis et al., 2006)
- SILVA (Pruesse et al., 2007; Quast et al., 2012)
- Human Oral Microbiome Database (HOMD) (Chen et al., 2010)

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

4. Results

4.2. Taxonomy Overview

4. Results

4.2. Taxonomy Overview

4.2.1. Abundance

Abundance Distribution

Abundance

- Minimum: 0
- Mean: 1.8
- Median: 0.0
- Maximum: 8848.0

Abundance without Zero

- Minimum: 1.0
- Mean: 189.6
- Median: 28.0
- Maximum: 8848.0

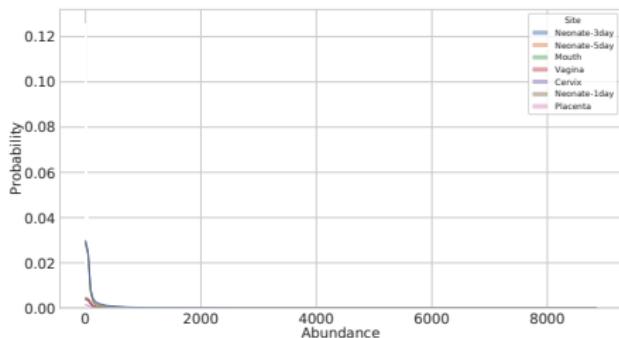


Figure: Abundance distribution

Clustermap with Abundance

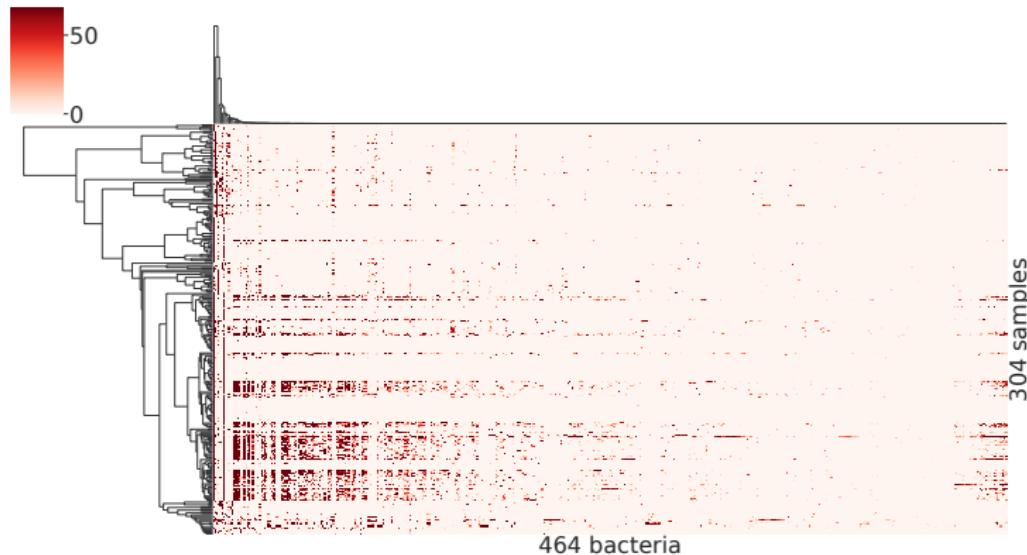
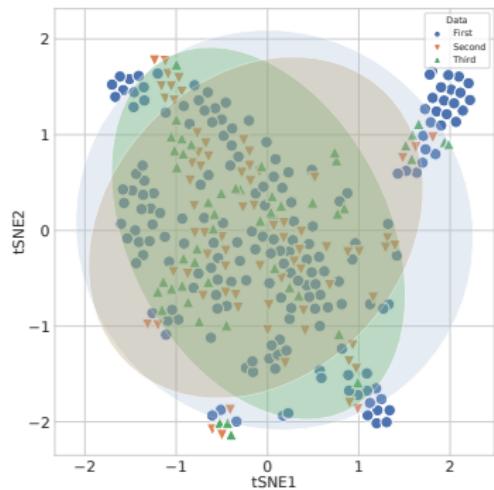
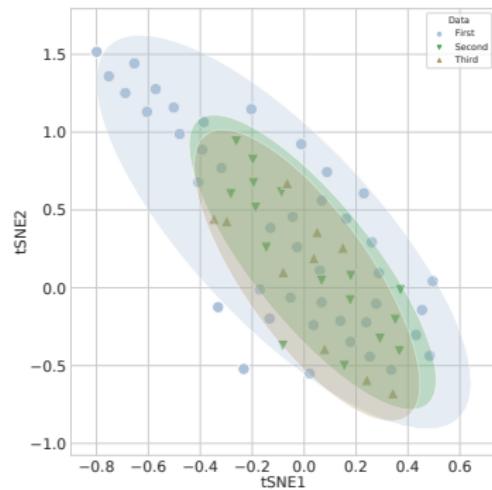


Figure: Clutermap with Abundance

t-SNE with Abundance I



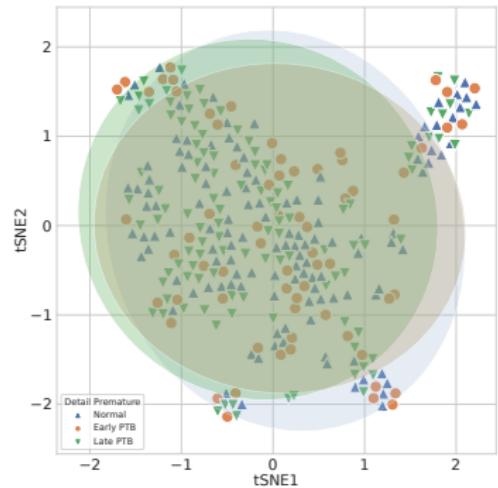
(a) All



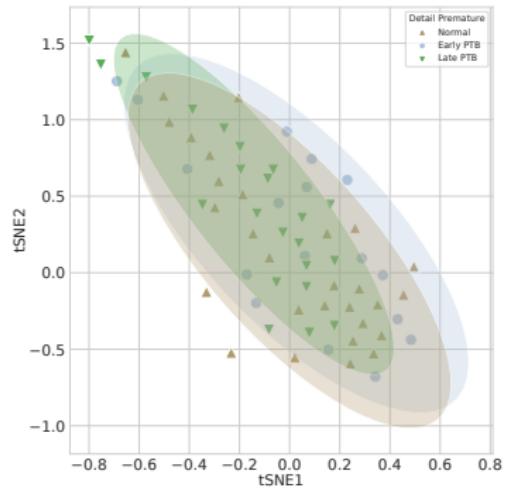
(b) Mother Mouth

Figure: t-SNE plot of Abundance with Data Batch

t-SNE with Abundance II



(a) All



(b) Mother Mouth

Figure: t-SNE plot of Abundance with PTB

4. Results

4.2. Taxonomy Overview

4.2.2. Proportion

Proportion Distribution

Proportion

- Minimum: 0.0
- Mean: 0.00008
- Median: 0.0
- Maximum: 0.793

Proportion without Zero

- Minimum: 0.00002
- Mean: 0.00008
- Median: 0.00153
- Maximum: 0.793

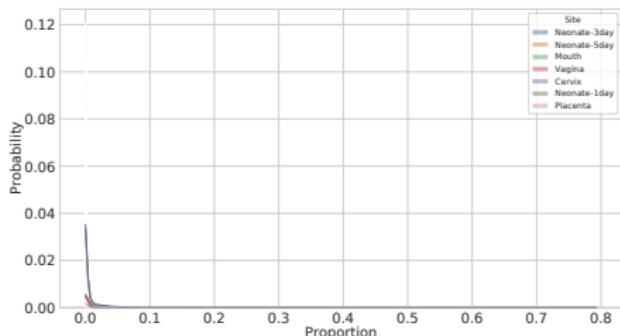


Figure: Proportion distribution

Clustermap with Proportion

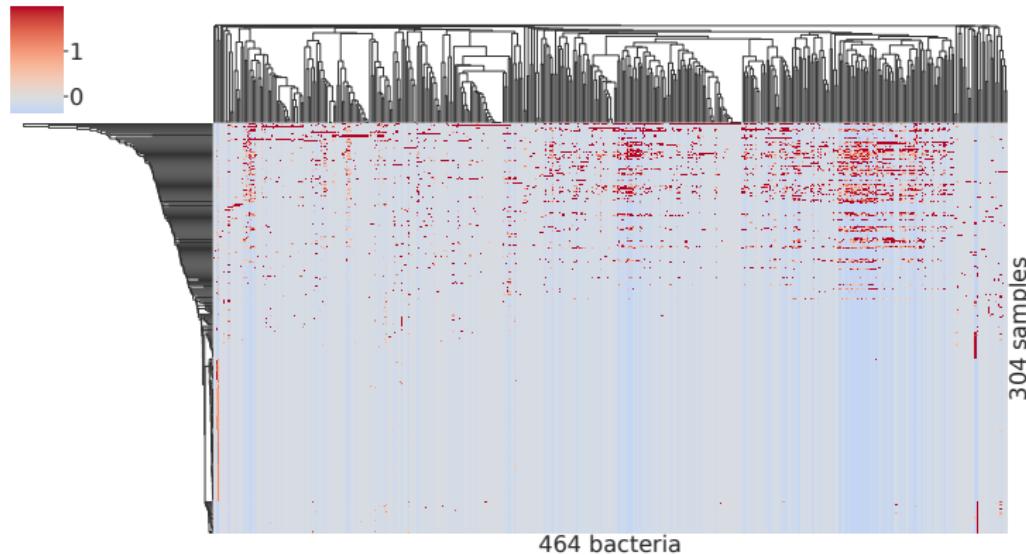
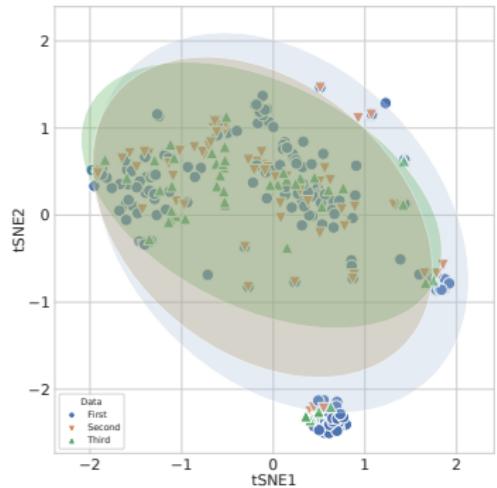
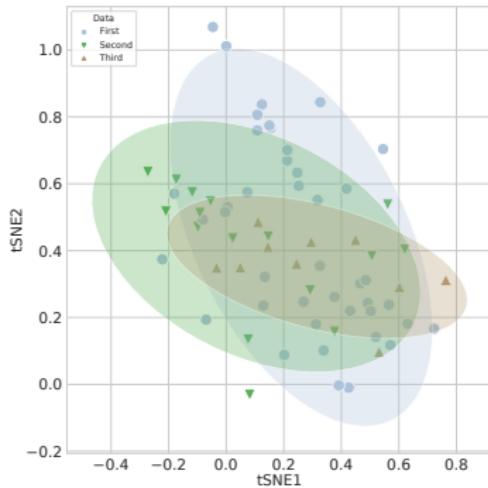


Figure: Clustermap with Proportion

t-SNE with Proportion I



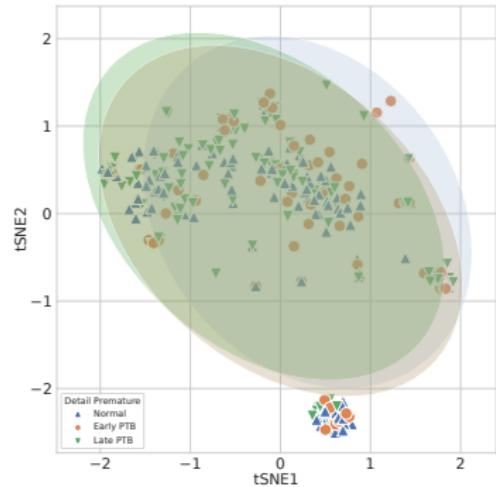
(a) All



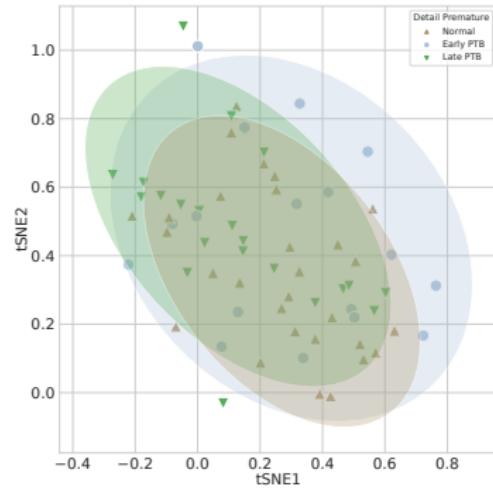
(b) Mother Mouth

Figure: t-SNE plot of Proportion with Data Batch

t-SNE with Proportion II



(a) All



(b) Mother Mouth

Figure: t-SNE plot of Proportion with PTB

4. Results

4.3. Diversity Index

Diversity Index

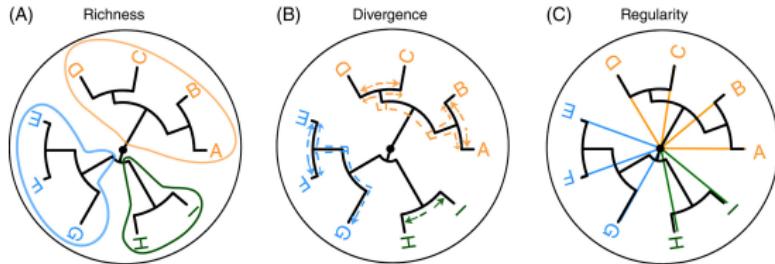


Figure: Three dimensions of phylogenetic information (C. M. Tucker et al., 2017)

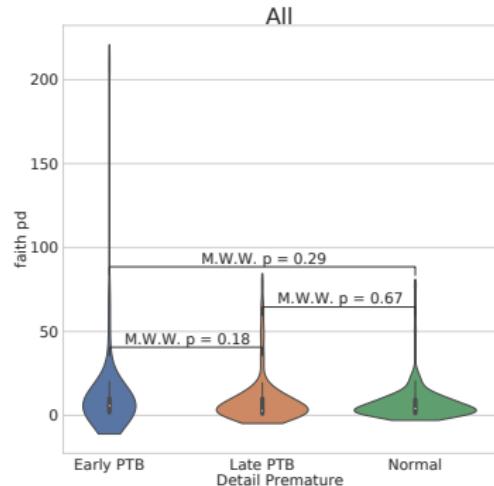
- A quantitative measure that shows richness, divergence, and regularity (C. M. Tucker et al., 2017)
- Alpha diversity: the richness of taxa **at a single community**
- Beta diversity: taxonomy differentiation **between communities**

4. Results

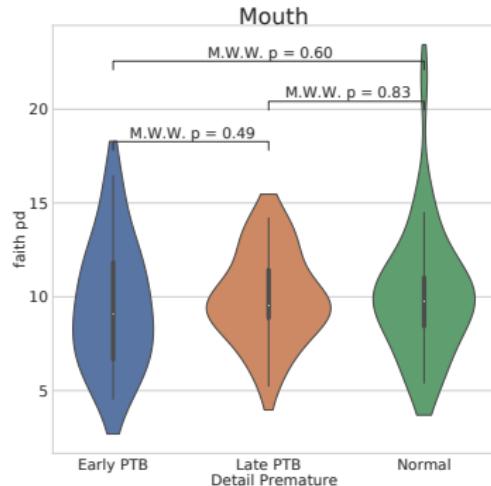
4.3. Diversity Index

4.3.1. Alpha-diversity

Violin Plot with Alpha-diversity I



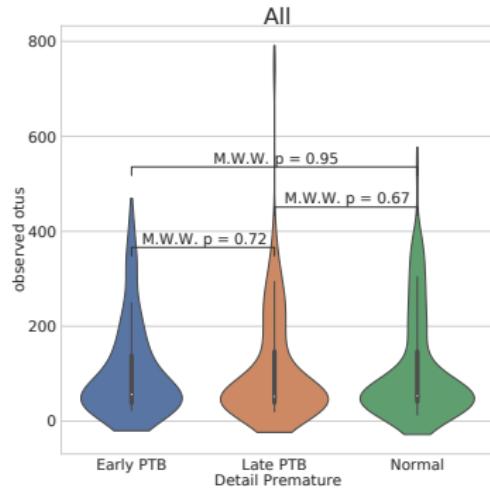
(a) All



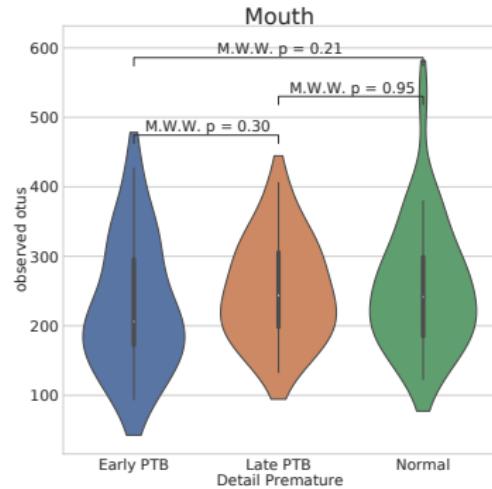
(b) Mother Mouth

Figure: Detail premature & Faith's PD

Violin Plot with Alpha-diversity II



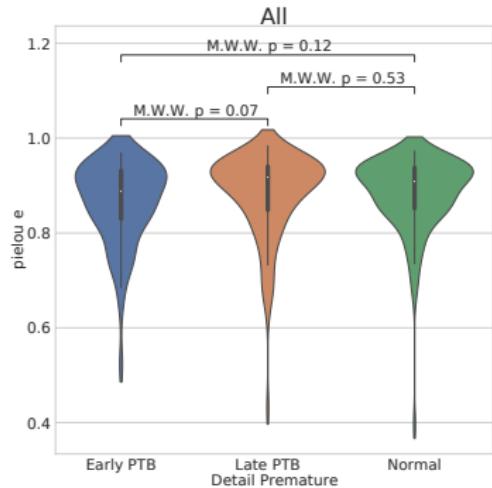
(a) All



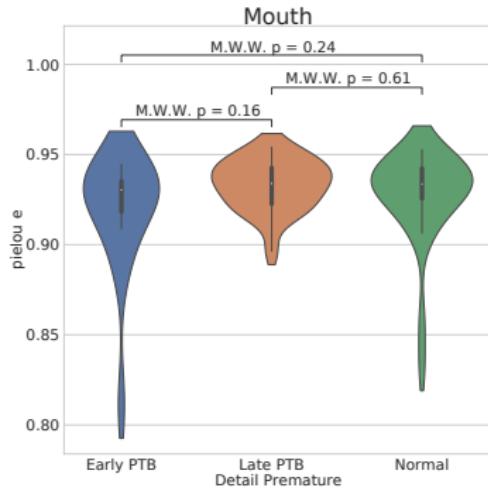
(b) Mother Mouth

Figure: Detail premature & Observed OTUs

Violin Plot with Alpha-diversity III



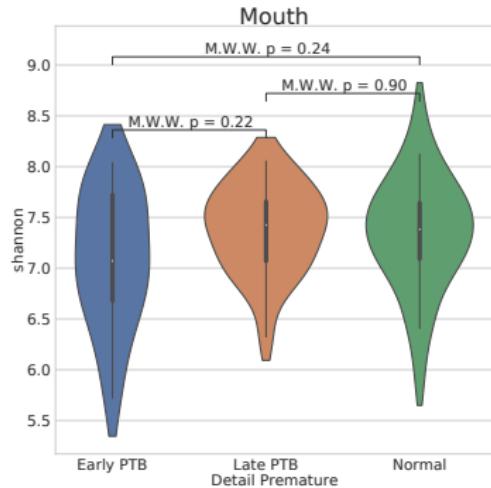
(a) All



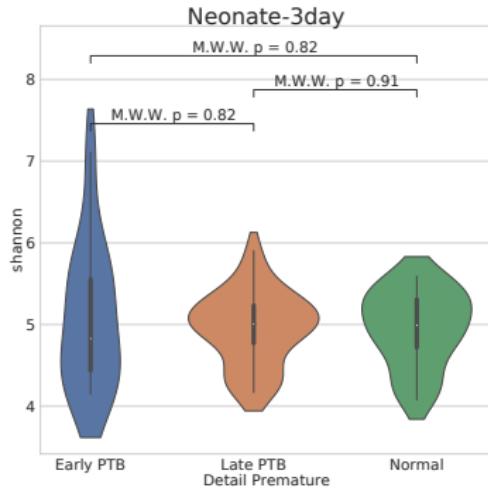
(b) Mother Mouth

Figure: Detail premature & Pielou Evenness

Violin Plot with Alpha-diversity IV



(a) All



(b) Mother Mouth

Figure: Detail premature & Shannon Entropy

4. Results

4.3. Diversity Index

4.3.2. Beta-diversity

Cluster map with Beta-diversity I

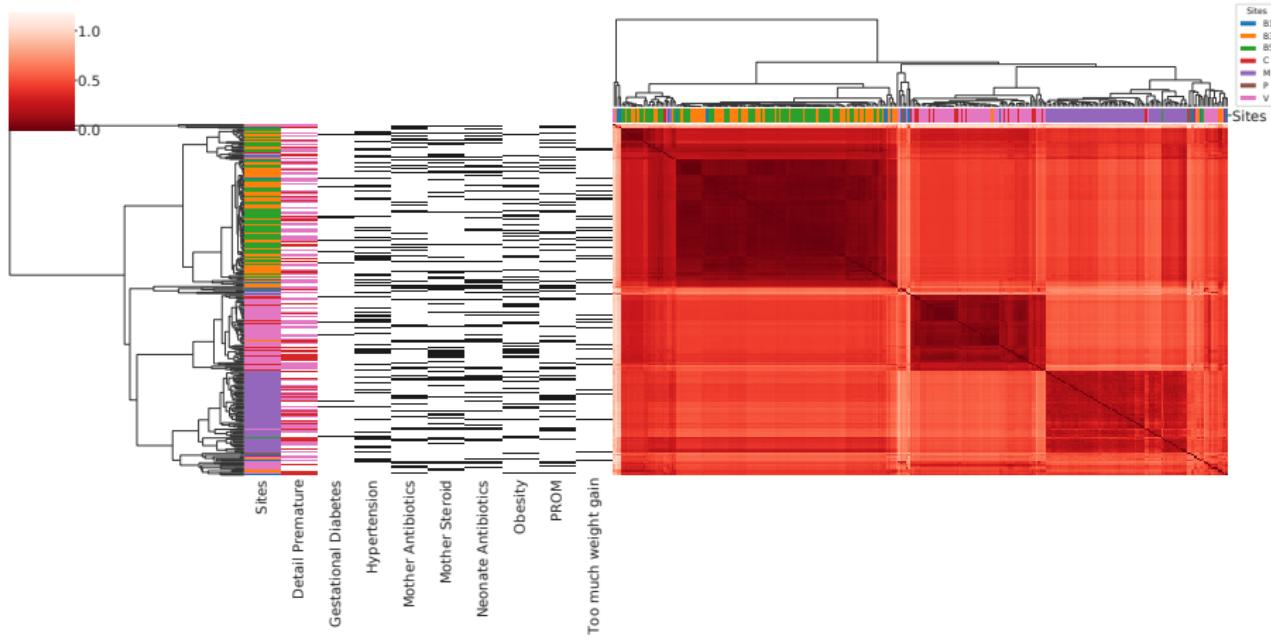
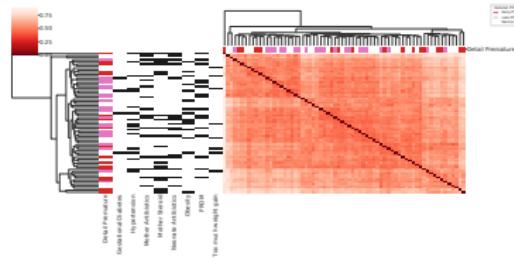
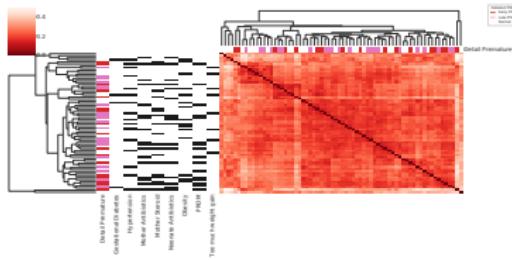


Figure: Cluster map with Weighted UniFrac distance index for DADA2

Cluster map with Beta-diversity II



(a) Unweighted UniFrac



(b) Weighted UniFrac

Figure: Clustermap of Abundance in Maternal Mouth

4. Results

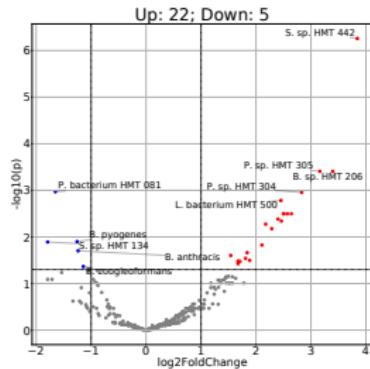
4.4. Taxonomy Analyses

4. Results

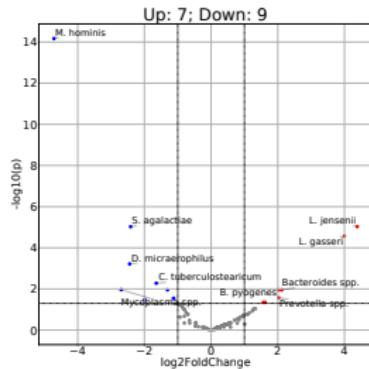
4.4. Taxonomy Analyses

4.4.1. Differentially Abundant Taxa

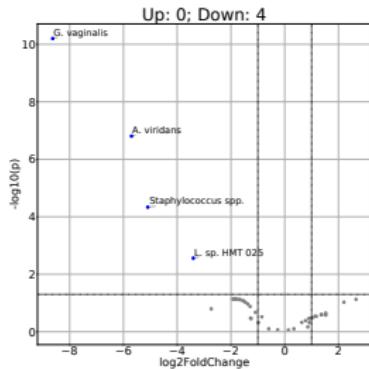
Volcano plot



(a) Mouth



(b) Vagina



(c) Cervix

Figure: Differentially abundant taxa

4. Results

4.4. Taxonomy Analyses

4.4.2. Correlation with Clinical Data

Correlation between Taxonomy & Clinical data I

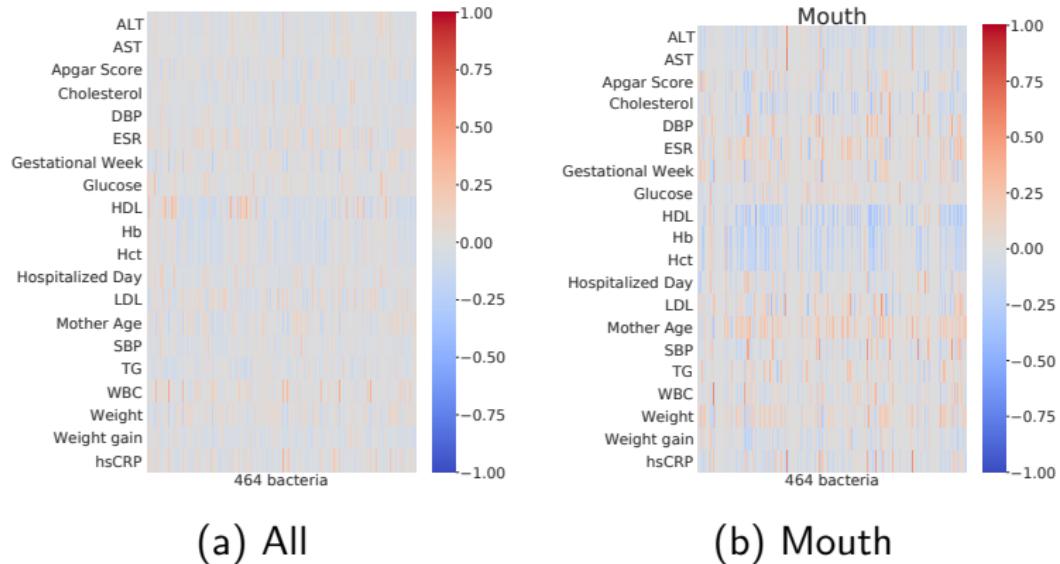


Figure: Pearson correlation on Taxonomy Abundance

Correlation between Taxonomy & Clinical data II

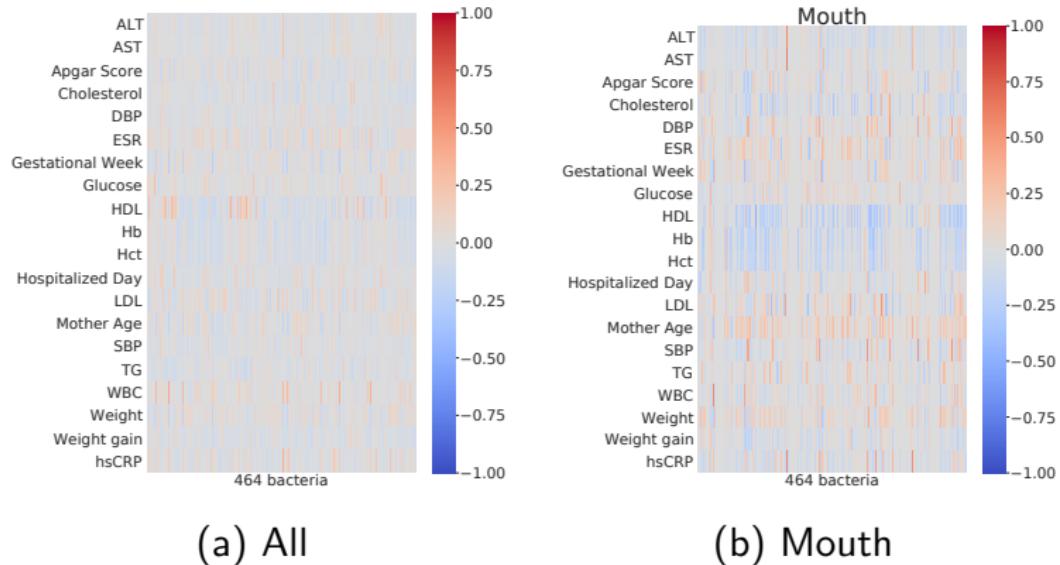


Figure: Pearson correlation on Taxonomy Proportion

4. Results

4.5. Machine Learning

ML algorithm comparison

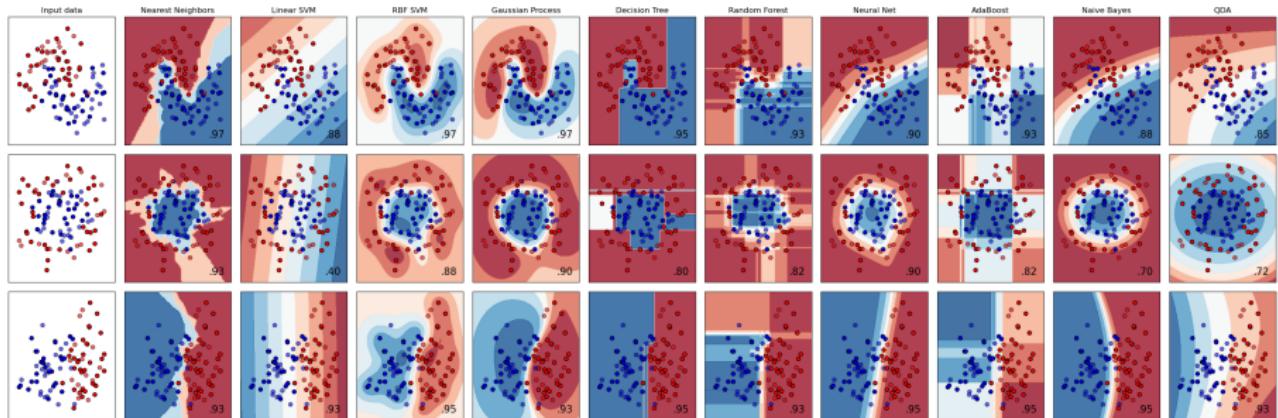


Figure: Classification Comparison (Pedregosa et al., 2011)

Oversampling

Naive Random Oversampling

- ① Randomly pick a sample from under-represented class.
- ② Duplicate it.
- ③ Repeat until all classes are even.

4. Results

4.5. Machine Learning

4.5.1. Random Forest Classifier on Abundance

Random Forest with (Early vs. Late vs. Full) I

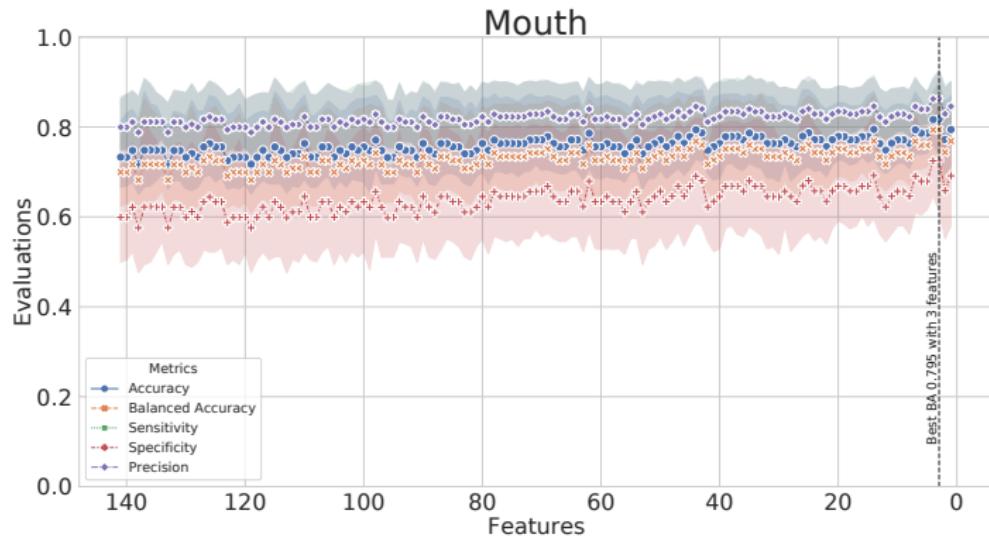


Figure: RF evaluations with feature counts

Random Forest with (Early vs. Late vs. Full) II

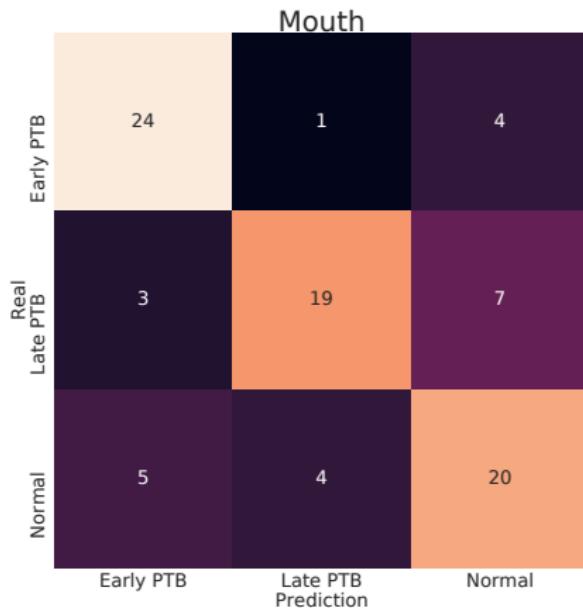


Figure: RF confusion matrix

Random Forest with (Early vs. Late vs. Full) III

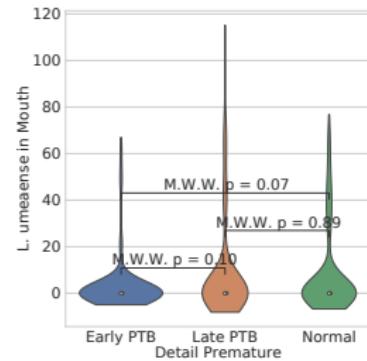
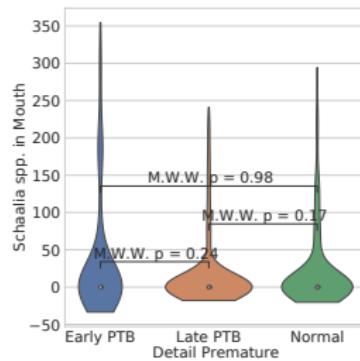
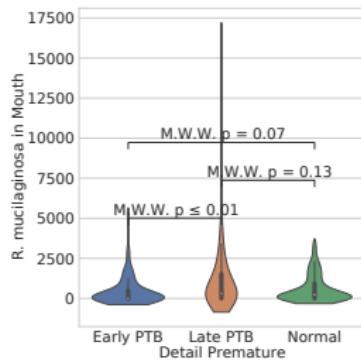


Figure: RF most important taxa

Random Forest with (Early vs. Late + Full) I

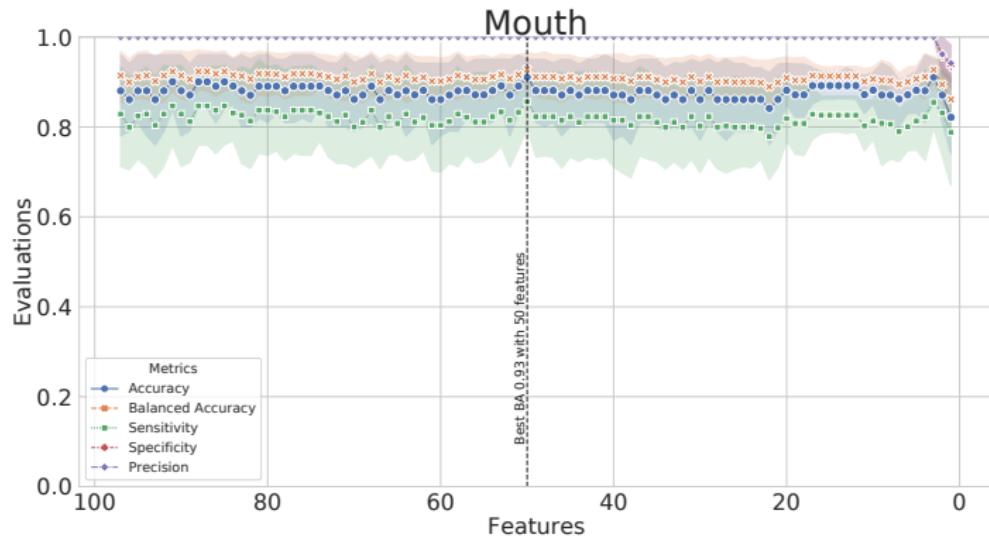


Figure: RF evaluations with feature counts

Random Forest with (Early vs. Late + Full) II

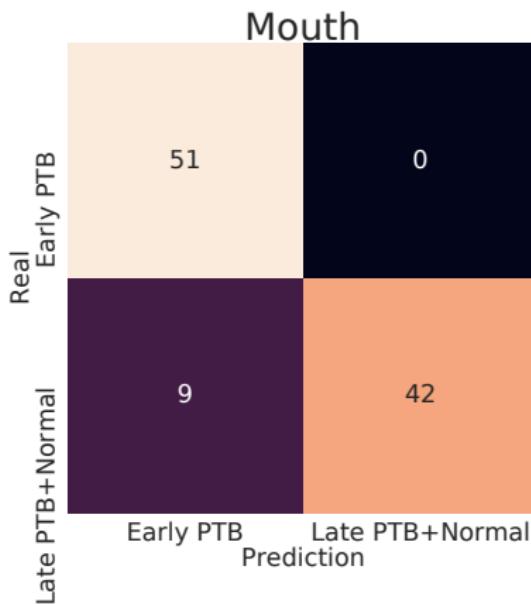


Figure: RF confusion matrix

Random Forest with (Early vs. Late + Full) III

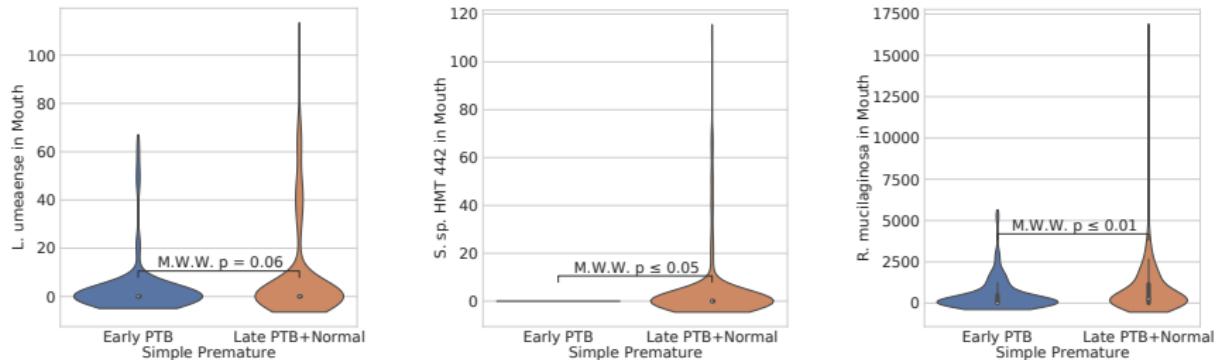


Figure: RF most important taxa

4. Results

4.5. Machine Learning

4.5.2. Random Forest Classifier on Proportion

Random Forest with (Early vs. Late vs. Full) I

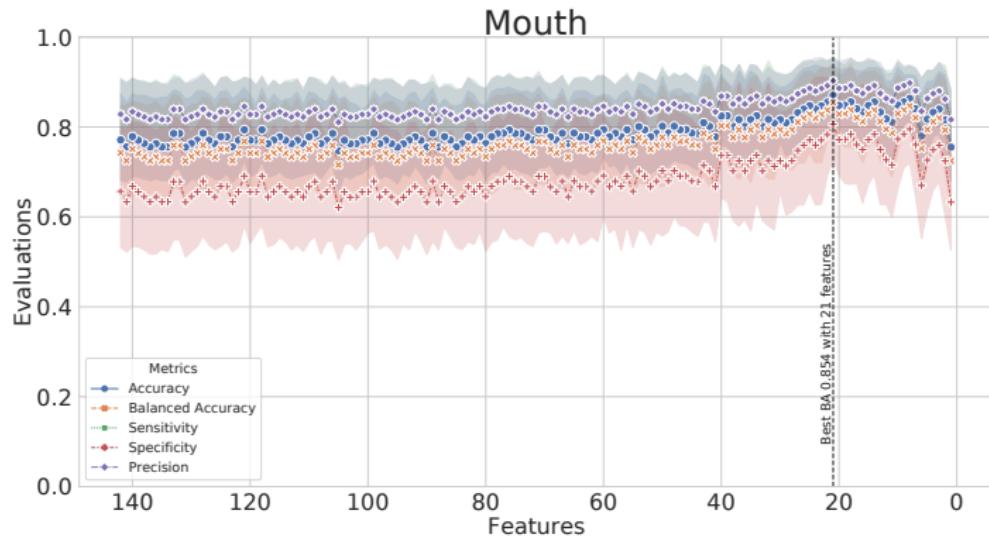


Figure: RF evaluations with feature counts

Random Forest with (Early vs. Late vs. Full) II

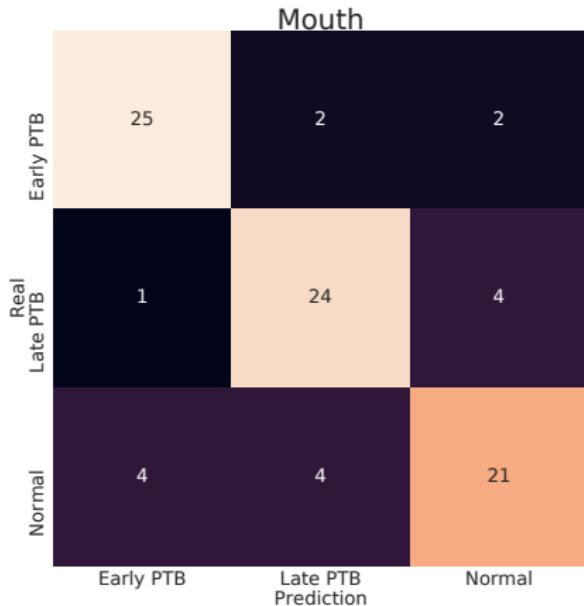


Figure: RF confusion matrix

Random Forest with (Early vs. Late vs. Full) III

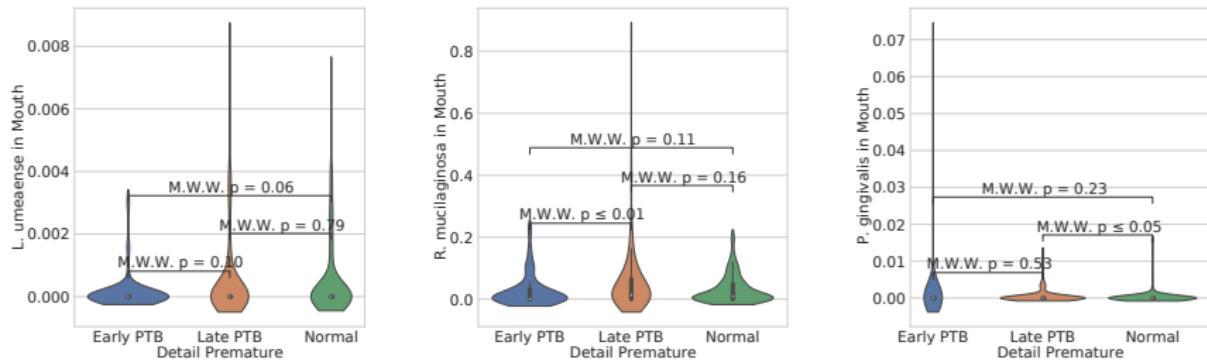


Figure: RF most important taxa

Random Forest with (Early vs. Late + Full) I

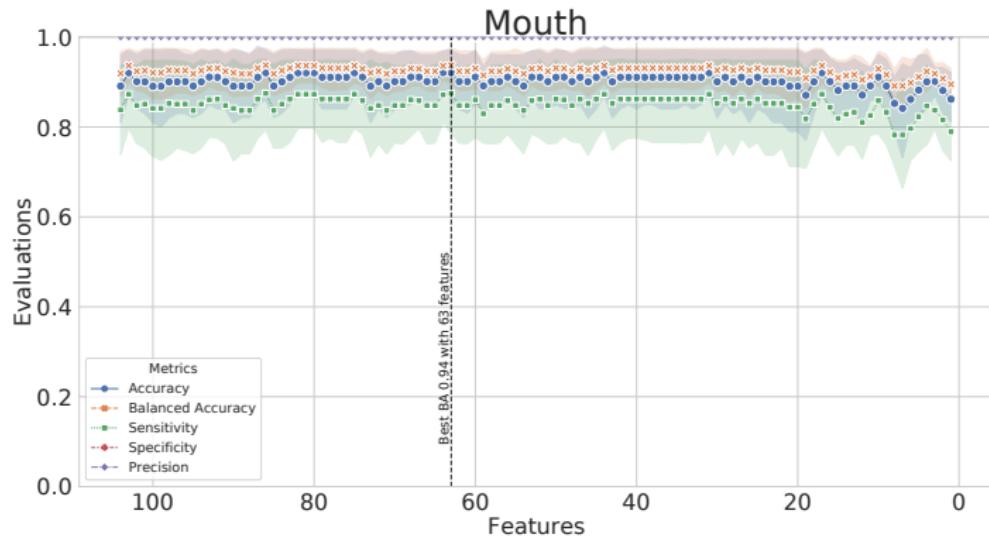


Figure: RF evaluations with feature counts

Random Forest with (Early vs. Late + Full) II

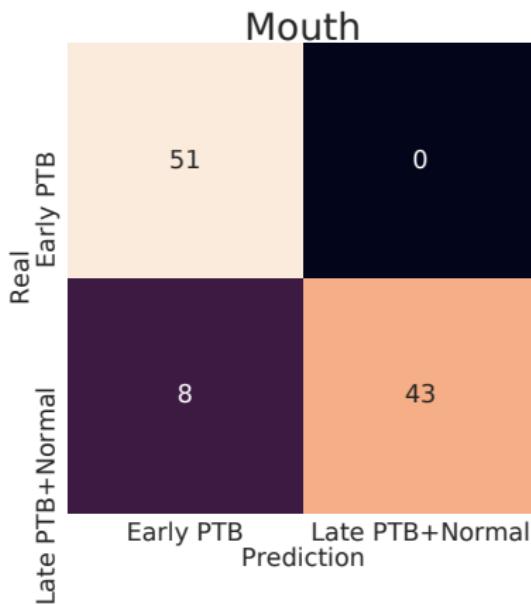


Figure: RF confusion matrix

Random Forest with (Early vs. Late + Full) III

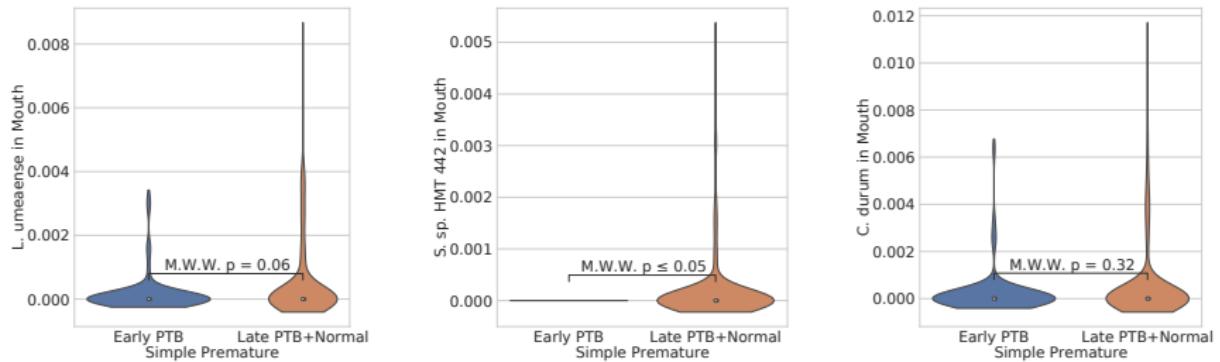


Figure: RF most important taxa

5. Discussion

6. References

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).
- Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature Biotechnology*, 37(8), 852-857. Retrieved from <https://doi.org/10.1038/s41587-019-0209-9> doi: 10.1038/s41587-019-0209-9
- 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.

References II

- 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.
- 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.
- Janda, J. M., & Abbott, S. L. (2007). 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9), 2761–2764.
- 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.
- 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. doi: 10.3402/mehd.v26.27663

References IV

- McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., ... Caporaso, J. G. (2012). The biological observation matrix (biom) format or: how i learned to stop worrying and love the ome-ome. *GigaScience*, 1(1), 7. doi: 10.1186/2047-217X-1-7
- Mignard, S., & Flandrois, J.-P. (2006). 16s rrna sequencing in routine bacterial identification: a 30-month experiment. *Journal of microbiological methods*, 67(3), 574–581.
- 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.

References V

- 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.
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... Glöckner, F. O. (2012). The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(D1), D590–D596.
- Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome biology*, 12(6), 1–18.
- 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.

References VI

- Tucker, J., & McGuire, W. (2004). Epidemiology of preterm birth. *Bmj*, 329(7467), 675–678.
- 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.
- Voronkov, L., Solonovych, A., Liashenko, A., & Revenko, I. (2018). Prognostic value of cognitive tests and their combination in patients with chronic heart failure and reduced left ventricular ejection fraction. *Eureka: health sciences*(6), 36–45.