

Doctoral Thesis

Microbiota in Human Diseases

Jaewoong Lee

Department of Biomedical Engineering

Ulsan National Institute of Science and Technology

2024

Microbiota in Human Diseases

Jaewoong Lee

Department of Biomedical Engineering

Ulsan National Institute of Science and Technology



CHURCH OF THE FLYING SPAGHETTI MONSTER

February 09, 2021

Letter of Good Standing

Dear Sir or Madam:

I am pleased to verify that _____

JAEWOONG LEE

is an ordained minister of the Church of the Flying Spaghetti Monster and recognized
within our organization as a member in good standing.

We hereby consent to this minister performing ceremonies and request that they are
granted all privileges and respect appropriate to a spiritual leader.

Any questions can be directed to the undersigned.

Representative,
Church of the Flying Spaghetti Monster
Bobby Henderson



CHURCH OF THE FLYING SPAGHETTI MONSTER

February 09, 2021

Letter of Good Standing

Dear Sir or Madam:

I am pleased to verify that _____

JAEWOONG LEE

is an ordained minister of the Church of the Flying Spaghetti Monster and recognized
within our organization as a member in good standing.

We hereby consent to this minister performing ceremonies and request that they are
granted all privileges and respect appropriate to a spiritual leader.

Any questions can be directed to the undersigned.

Representative,
Church of the Flying Spaghetti Monster
Bobby Henderson

Abstract

Your abstract should be here.

Contents

1	Abstract	2
2	Predicting preterm birth using machine learning techniques in oral microbiome	3
2.1	Introduction	3
2.2	Materials and methods	5
2.3	Results	7
2.4	Discussion	15
3	Prediction model for periodontitis severity based on the salivary microbiomes	17
3.1	Introduction	17
3.2	Materials and methods	18
3.3	Results	19
3.4	Discussion	29
4	Lung microbiome	30
4.1	Introduction	30
4.2	Materials and methods	31
4.3	Results	32
5	General conclusion and future perspective	34
5.1	General conclusions	34

5.2	Plan for future	35
5.3	Future perspective	36
	References	37
	Acknowledgments	40

List of Figures

1	DAT volcano plot	9
2	Oral microbiome compositions over DAT	10
3	Machine learning evaluations over DAT	11
4	Diversity indices	12
5	PROM-related DAT volcano plot	13
6	Heatmap plot of PTB classification with validation data	14
7	Diversity indices	21
8	Differentially abundant taxa	22
9	Correlation heatmap	23
10	Random forest classification metrics	24
11	Rarefaction curves for alpha diversity indices	25
12	Absolute abundance of salivary bacterial taxa in the different periodontal statuses at the species level	26
13	Correlation plots for differentially abundant taxa	27
14	Random forest classification metrics from external datasets	28

List of Tables

1	Baseline clinical characteristics of study subjects	8
2	Clinical characteristics of the study subjects	20

List of Abbreviations

ACC Accuracy

ASV Amplicon sequence variant

AUC Area-under-curve

BA Balanced accuracy

DAT Differentially abundant taxa

F1 F1 score

Faith PD Faith's phylogenetic diversity

FTB Full-term birth

GA Gestational age

MWU test Mann-Whitney U-test

PRE Precision

PROM Prelabor rupture of membrane

PTB Preterm birth

ROC curve Receiver-operating characteristics curve

rRNA Ribosomal RNA

SEN Sensitivity

SPE Specificity

t-SNE t-distributed stochastic neighbor embedding

1 Abstract

This is my abstract ...

This doctoral dissertation is an addition based on the following papers that the author has already published:

- Hong, Y. M., Lee, J., Cho, D. H., Jeon, J. H., Kang, J., Kim, M. G., ... & Kim, J. K. (2023). Predicting preterm birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1), 21105.

2 Predicting preterm birth using machine learning techniques in oral microbiome

This section includes the published contents:

Hong, Y. M., Lee, J., Cho, D. H., Jeon, J. H., Kang, J., Kim, M. G., ... & Kim, J. K. (2023). Predicting preterm birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1), 21105.

2.1 Introduction

Preterm birth (PTB), defined as delivery of newborns before 37 weeks of gestation, is a leading cause of morbidity and mortality in neonates (Blencowe et al., 2012). Established risk factors for PTB include genitourinary tract infections, short cervical length, and multiple pregnancies (Goldenberg, Culhane, Iams, & Romero, 2008). However, there is still disagreement regarding the magnitude of these factors' effects on birth outcomes. Early identification of pregnant women at high risk of PTB can facilitate the implementation of strategies to prolong gestation and improve birth outcomes (Iams & Berghella, 2010).

Despite increased understanding of risk factors contributing to PTB, there remains a considerable lack of sensitivity in predictive models that can serve as a framework for intervention strategies (Sotiriadis, Papatheodorou, Kavvadias, & Makrydimas, 2010). Numerous attempts have been made to predict PTB using machine learning techniques combined with data from health records, inflammatory markers, and vaginal microbiome (Berghella, 2012). Fetal fibronectin is widely used clinically due to its affordability and simplicity. However, it has a low prediction rate with a sensitivity of only 56% (Honest et al., 2009). Cervical length measurement also has limitations due to the hassle and inaccuracy of the procedure and the need for a skilled specialist (Leitich & Kaider, 2003).

Approximately 70% of PTBs result from spontaneous onset of preterm labor and preterm pre-labor rupture of membranes (PROM) due to intrauterine infection and inflammation (Romero, Dey, & Fisher, 2014). However, the mechanism of PTB cannot be fully explained by inflammatory and infectious pathways as anti-inflammatory and antibiotic treatments could not reduce PTB incidence rates (Romero, Hassan, et al., 2014). With advancements in molecular genetic technology, studies on maternal microbiomes using 16S ribosomal RNA (rRNA) sequencing have emerged to explore unknown pathways of PTB (Fettweis et al., 2019).

Microorganisms associated with PTB have been postulated to originate from one of two places: the reproductive or genitourinary tract ascending through the cervix or a hematogenous route (Han & Wang, 2013). Recent research has identified vaginal microbial signatures in women who later experience PTB and attempted to predict PTB using cervicovaginal fluid (Kindinger et al., 2017). Although existing reports have verified a potential relationship of vaginal microbiome with PTB, they can only explain an ascending route.

Decades of epidemiological research studies have suggested that periodontitis is an independent risk factor for various adverse birth outcomes, including PTB (Offenbacher et al., 1996). Based on these precedents, it is expected that the oral microbiome can explain another hematogenous route. However,

prenatal oral microbiome is not well understood.

Thus, this study aimed to compare oral microbiome compositions between a PTB group and a full-term birth group, to identify oral microbiome associated with PTB, and to develop a machine learning prediction model of PTB based on oral microbiome compositions.

2.2 Materials and methods

2.2.1 Study design and participants

This study was conducted on singleton pregnant women admitted for delivery at Jeonbuk National University Hospital between 2019 and 2021. This study received approval from the Ethical Research Committee (IRB file No. 2019-01-024). All participants provided written informed consent. Eligible participants included women admitted for induction delivery, elective cesarean section, and those who were hospitalized due to symptoms of preterm labor or preterm pre-labor rupture of membranes.

2.2.2 Data collection and grouping

Data on current and historical pregnancy outcomes were collected from questionnaires and electronic medical records. This information encompassed demographic factors (gestational age, birth weight, sex) and maternal risk factors (maternal age at delivery, cesarean section, preterm pre-labor rupture of membranes, previous preterm delivery history, gestational or overt diabetes mellitus, pregnancy-induced or chronic hypertension, and pre-pregnancy overweight or gestational weight gain). All subjects were divided into a PTB group or a full-term birth (FTB) group, with PTB defined as delivery before 37 weeks of gestation.

2.2.3 Oral microbiome sample collection

Oral microbiome samples were collected using mouthwash within 24 h before delivery. Standard sterile techniques were employed. Medical staff supervised all sample collection procedures. Participants were instructed to avoid brushing their teeth, eating, or drinking 30 min before sampling. Saliva samples were obtained by rinsing the mouth with 12 mL of a gargle solution (E-zen Gargle, JN Pharm, Pyeongtaek, Korea) for 30 s. Sample were labeled with the subject's anonymous ID and stored at 4 °C until further processing. The resuspended sample was transferred to a microcentrifuge tube. Genomic DNA was extracted using an ExgeneTM Clinic SV kit (GeneAll Biotechnology, Seoul, Korea) according to the manufacturer's instructions and stored at -20 °C.

2.2.4 16s rRNA gene sequencing and taxonomy assignment

Specimens were sent to Department of Biomedical Engineering, Ulsan National Institute of Science and Technology for taxonomy assignment. Then 16S rRNA sequencing was performed using an Illumina MiSeq Reagent Kit v3 (Illumina, San Diego, CA, USA) commissioned by Macrogen (Macrogen, Seoul, Korea). Library protocols for amplifying V3 and V4 regions were used. Pooled library was sequenced using a v3 600 cycle chemistry after diluting the sample to a final concentration of 6 pM with 20% PhiX control to generate 300 bp paired end reads.

2.2.5 Bioinformatics and statistical analysis

Independent t-test and chi-square test were used to compare differences between the PTB group and the FTB group. SPSS (version 20.0) was used for all data analyses (Spss et al., 2011). Statistical significance was considered at $p < 0.05$.

16S rRNA sequences from study subjects were imported with QIIME2 (version 2022.2) for further processing (Bolyen et al., 2019). Sequences were filtered with DADA2 (Callahan et al., 2016). Amplicon sequence variants (ASV) were assigned taxonomy with the Human Oral Microbiome Database (version 15.22) (Chen et al., 2010). To measure richness of microbiomes, two diversity indices were calculated. Alpha diversity, a measure of species within a particular community, was calculated using the Faith's phylogenetic diversity (Faith PD) index within the QIIME2 platform. Communities numerically dominated by a few species will exhibit a low Faith PD index, whereas communities in which abundance is distributed equally among species will exhibit a high evenness. Mann–Whitney–Wilcoxon test was used to find statistically significant differences in Faith PD index. Beta diversity measures dissimilarity between pairs of communities. It was calculated using the Hamming diversity index. PERMANOVA multivariate test was used to calculate statistically significant differences in the Hamming diversity index.

To identify differentially abundant taxa (DAT) with distinct abundance differences between the PTB group and the FTB group, DESeq2 was implemented using `DESeqDataSetFromMatrix` method as described in the package tutorial (Love, Huber, & Anders, 2014). Taxa with $|\log_2 \text{FoldChange}| > 1$ and adjusted $p < 0.05$ are considered as significant different.

2.2.6 Machine learning prediction model development

Following qualitative and quantitative analyses of associations between the PTB group and each bacterium, a random forest classifier was used to find the criteria to predict PTB with oral microbiome data. Random forest classification is an ensemble machine learning algorithm that summarizes many decision trees to improve classification evaluations and robustness (Breiman, 2001). Random forest classifier was implemented to predict PTB based on oral microbiome compositions. To assure consistence and reliable classification results (Wong & Yeh, 2019), we performed stratified k -fold cross-validation ($k = 5$). Moreover, to decide the best features that could maximize classification evaluations, we performed random forest classification evaluations only with some DAT selected by their importances. Evaluations for classification included accuracy (ACC), balanced accuracy (BA), precision (PRE), sensitivity (SEN), and specificity (SPE).

2.2.7 Ethics approval and consent to participate

The research protocol was approved by the Institutional Review Board of Jeonbuk National University Hospital (No. 2019-01-024) and was performed according to the Declaration of Helsinki (Goodyear, Krleza-Jeric, & Lemmens, 2007).

2.3 Results

2.3.1 Study participant demographics

In this study, a total of 69 volunteer mothers were initially recruited. However, one participant with incomplete data and nine individuals with twin pregnancies were excluded from the study cohort. As a result, 59 women (30 in the PTB group and 29 in the FTB group) were included in the final analysis. Demographic and clinical characteristics of subjects in the PTB group and the FTB group are summarized in Table 1. Because PROM is a major cause of preterm birth, it was significantly higher in PTB group. There was no significant difference in other maternal clinical characteristics between the PTB group and the FTB group. There were no cases with a history of smoking or concurrent periodontal disease in both groups.

2.3.2 Comparison of oral microbiomes

The oral microbiome comprised 13,953,804 sequences from 59 oral microbiome samples, with $102,305.95 \pm 19,095.60$ and $64,823.41 \pm 15,841.65$ reads per sample before and after filtering low-quality reads and trimming extra-long tails, respectively. After filtering low-quality reads and trimming extra-long tails, remaining representative reads were clustered in ASVs with their exact sequence match. There were no significant differences in measures of alpha diversity (Faith PD) and beta diversity (Hamming distance) indices for samples between the PTB group and the FTB group (Figure 4).

Of 465 genera and species analyzed, 32 DAT between the PTB group and the FTB group were selected by DESeq2 (Love et al., 2014), including 26 FTB-enriched DAT and six PTB-enriched DAT. In order to mitigate the confounding effect of PROM, we excluded 7 PROM-related DATs from these 32 DAT (Figure 5). There were a total of 25 DATs between the PTB group and the FTB group, with 22 DATs enriched in the FTB group and three in the PTB group, as depicted in Figure 1.

Figure 2 displays DAT volcano plot in the oral microbiome sorted by differences between FTB-enriched DAT and PTB-enriched DAT proportions, indicating a decrease in gestational age for FTB-enriched DAT proportions. Pearson correlation analysis revealed a strong negative correlation ($r = -0.542$ and $p = 7.8e - 6$) between gestational age and difference between PTB-enriched DAT and FTB-enriched DAT proportions.

2.3.3 Random forest classification

Random forest classifiers were established to classify PTB based on DAT. The best BA(0.765 ± 0.071) was achieved using the nine most important taxa (Figure 3a). The Random forest model calculated the importance of each DAT (Figure 3b). Overall ACC, PRE, SEN, and SPE were 0.714 ± 0.061 , 0.700 ± 0.194 , 0.728 ± 0.058 , and 0.743 ± 0.138 , respectively. To validate the performance of our machine learning prediction model, we conducted a validation test on nine twin pregnancies that were excluded from the paper (Figure 6). On PTB subjects of these twin samples, the machine learning classifications have 87.5% accuracy, comparable to the machine learning classification on the singleton study subjects.

Table 1: **Baseline clinical characteristics of study subjects.** Continuous variables: Mean \pm standard deviation. Categorical variables: count (proportion). Continuous variable for independent *t*-test. Categorical variable for Pearson's χ^2 -square test.

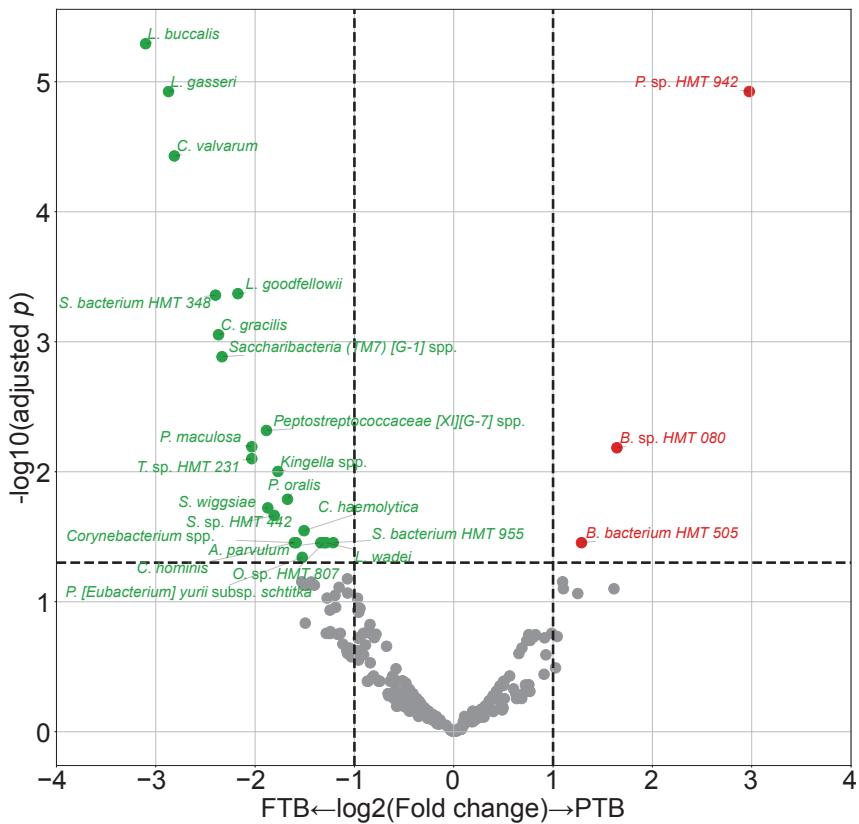


Figure 1: **DAT volcano plot.** DAT volcano plot shows DAT, with PTB-enriched DAT shown as red dots and FTB-enriched DAT shown as green dots. Taxa with $|\log_2 \text{FoldChange}| > 1$ and adjusted $p < 0.05$ are considered as significant different.

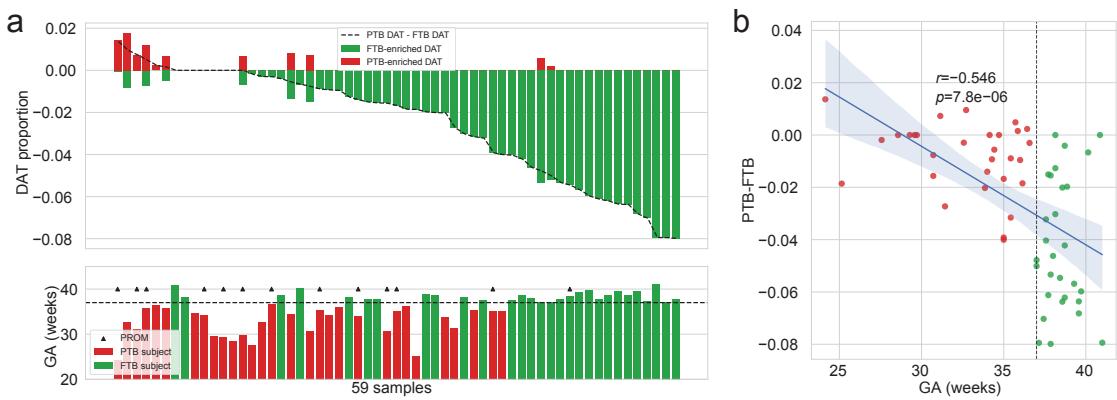


Figure 2: Oral microbiome compositions over DAT. (a) Proportions of DAT of study subjects. Samples are sorted by difference between PTB-enriched DAT proportion and FTB-enriched DAT proportion. GA of samples are shown, matched with order of upper panel. PTB: red bar, FTB: green bar. PROM: arrow head. (b) Correlation plot with GA and difference between PTB-enriched DAT proportion and FTB-enriched DAT proportion. Pearson correlation shows strong negative coefficient ($r = -0.542$ and $p = 7.8e - 06$) between GA and difference between PTB-enriched DAT proportion and FTB-enriched DAT proportion.

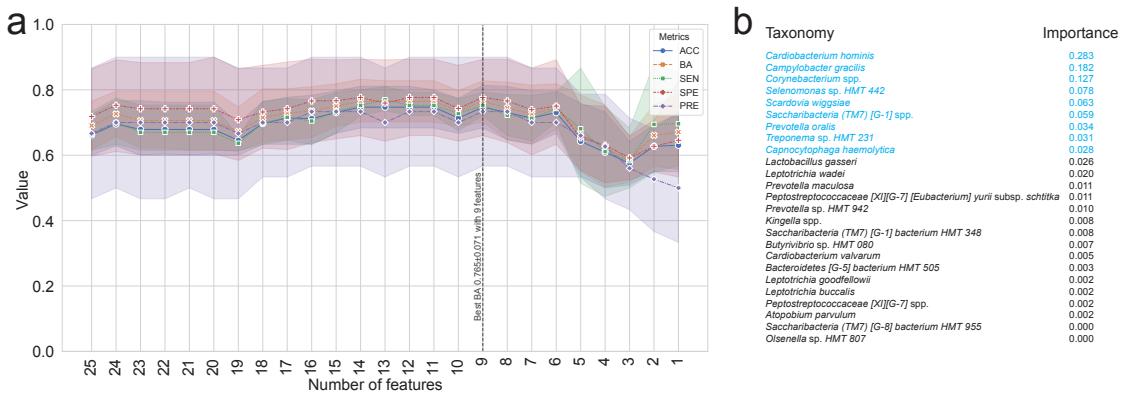


Figure 3: **Machine learning evaluations over DAT.** (a) Machine learning evaluation upon number of features (DAT). Random Forest classifier has the best balanced accuracy (mean \pm standard deviation.; 0.765 ± 0.071) with the nine most important DAT. (b) Importance of DAT is shown. Note that $0 \leq$ importance ≤ 1 , and \sum importance = 1. The 20 most important DAT that give the best balanced accuracy is marked with blue.

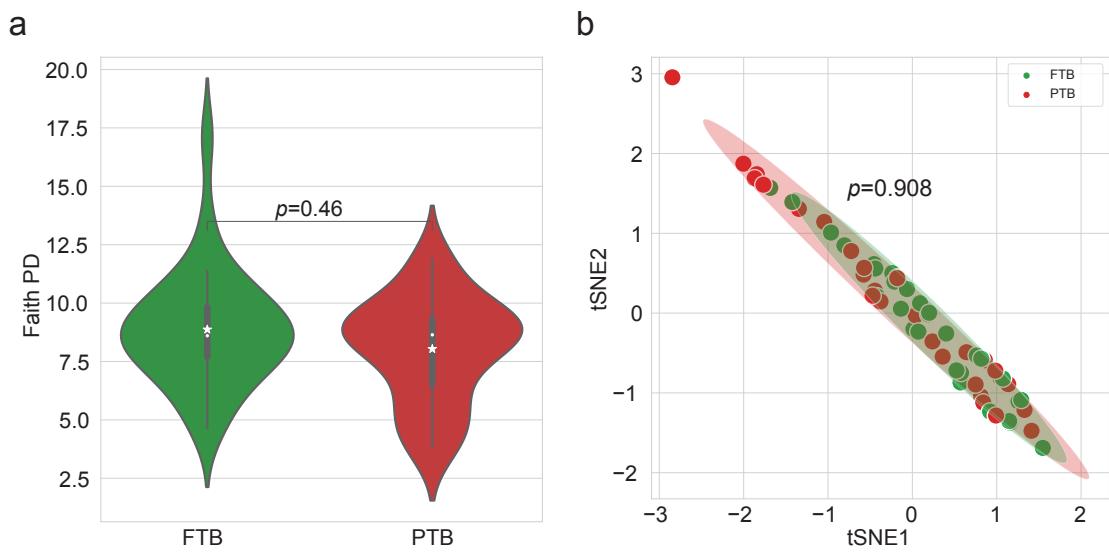


Figure 4: **Diversity indices.** (a) Alpha diversity index (Faith PD). Mann-Whitney-Wilcoxon test did not find statistically significant difference between the PTB group and the FTB group. Mean values are marked with star-point. (b) t-SNE plot with beta diversity index (Hamming distance). PERMANOVA multivariate test did not find statistically significant difference between the PTB group and the FTB group.

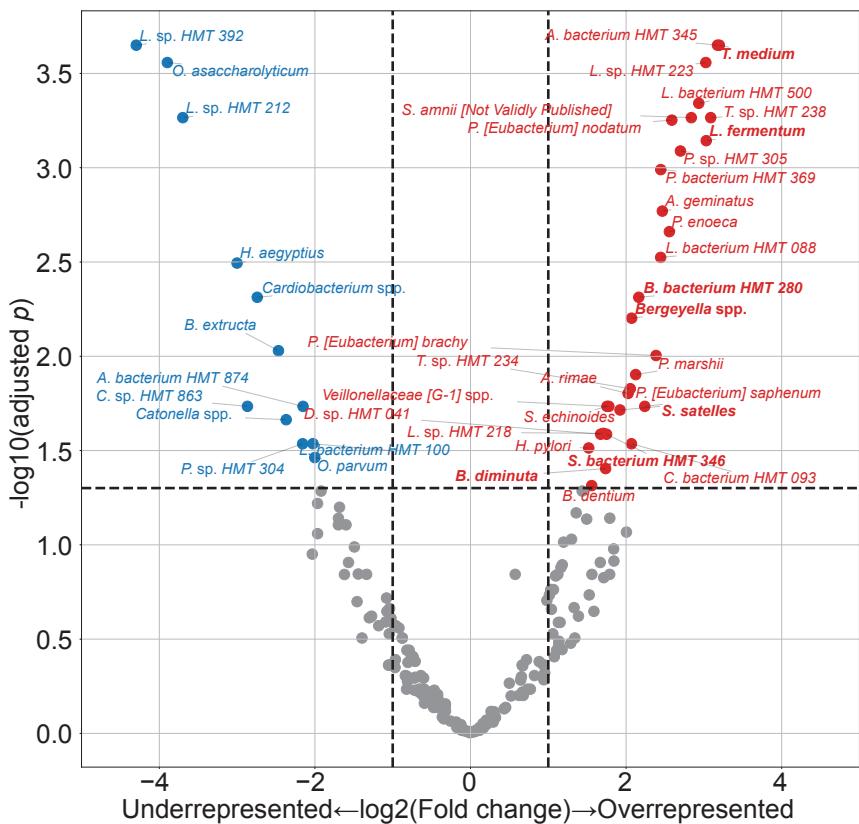


Figure 5: **PROM-related DAT volcano plot.** This is a subgroup analysis between 12 participants in PTB group with PROM and 18 participants in PTB group without PROM, 42 PROM-related DAT were selected between these two groups. Out of these 42 PROM-related DAT, only 7 DAT overlapped with PTB-related DAT, as it indicated by the bold marking. Volcano plot shows PROM-related DAT, with 12 PROM-underrepresented DAT shown as blue dots and 30 PROM-overrepresented DAT shown as red dots. Taxa with $|\log_2 \text{FoldChange}| > 1$ and adjusted $p < 0.05$ are considered as significant different.

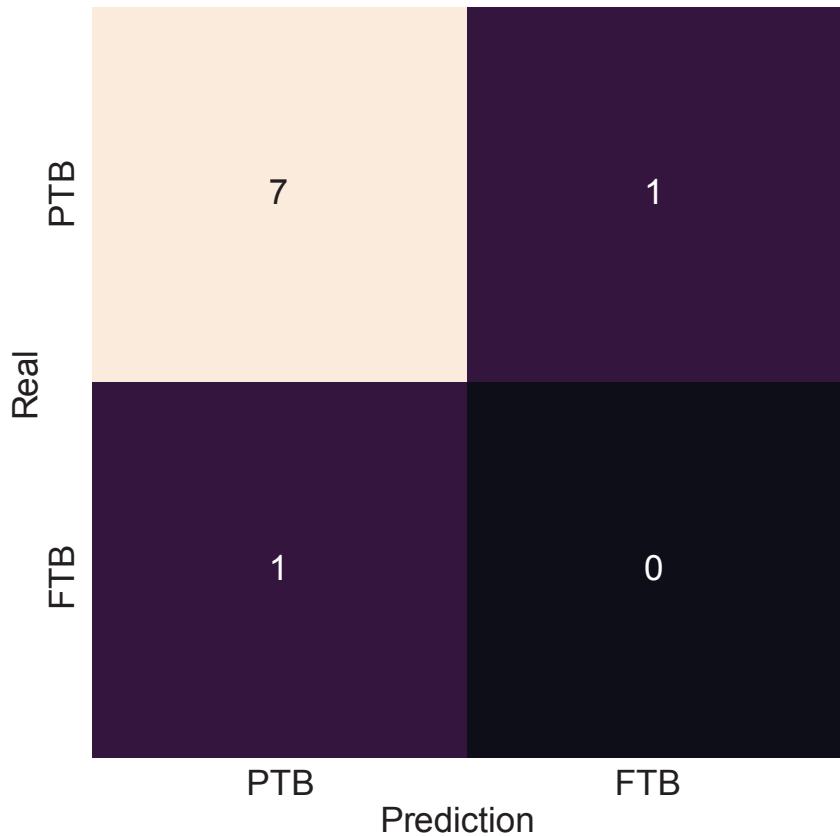


Figure 6: **Heatmap plot of PTB classification with validation data.** A validation test on 9 twin pregnancies that had excluded from the paper was conducted. They consist of 8 PTB subjects and 1 FTB subject. As twin pregnancies have a 7-10 times higher PTB rate than singleton pregnancies, resulting in a majority of the test data falling into the PTB group. The machine learning classifications have 87.5% accuracy, comparable to the machine learning classification on the singleton study subjects (Mean \pm standard deviation. 0.714 ± 0.061)

2.4 Discussion

In this study, we developed a method for predicting PTB based on random forest classifier using oral microbiome compositions. Recently, several sporadic reports have suggested a bidirectional relationship between oral microbiome and pregnancy (Han & Wang, 2013). However, prenatal oral microbiome is not well understood yet. Some research has shown that oral microbial dysbiosis combined with gingival inflammation can lead to adverse pregnancy outcomes, including low birth weight, PTB, pre-eclampsia, and miscarriages (Ide & Papapanou, 2013). Nevertheless, these results have been inconsistent due to methodologies employed in studies that only target known pathogens.

Fusobacterium nucleatum is the most prevalent oral microbiome studied (Han, 2015; Brennan & Garrett, 2019; Bolstad, Jensen, & Bakken, 1996). *Fusobacterium nucleatum* is a Gramnegative, anaerobic, filamentous oral microbiome. It is considered one of the most abundant species in the oral microbiome. It can also be isolated from vaginal microbiome (Vander Haar, So, Gyamfi-Bannerman, & Han, 2018; Witkin, 2019). Intra-amniotic *Fusobacterium nucleatum* infection leading to PTB has been reported in human and animal studies (Doyle et al., 2014). Other studies have shown that other oral pathogens including *Porphyromonas gingivalis* and intrauterine *Bergeyella* spp. can be isolated from the placenta of women who deliver prematurely (León et al., 2007; Katz, Chegini, Shiverick, & Lamont, 2009). In the present study, although *Bergeyella* spp. was overrepresented in the PTB group with PROM, it was excluded in the finally selected 25 DAT. Furthermore, *Campylobacter gracilis* was one of the FTB-enriched DAT that can aid colonization by periodontal pathogens including *Porphyromonas gingivalis* in subgingival microbiome (Yang et al., 2022). *Lactobacillus gasseri* was also one of the FTB-enriched DAT. It is known that *Lactobacillus gasseri* in vaginal microbiome can decrease early PTB risk (Basavaprabhu, Sonu, & Prabha, 2020; Payne et al., 2021).

We found that decisive species differentiating between two groups were mainly abundant in the FTB group, with DAT consisting of 22 FTB-enriched DAT and three PTB-enriched DAT. We hypothesize that deficiency of species having a protective impact might have triggered the pathophysiology of PTB. Two different mechanisms have been proposed to explain the relationship between unhealthy microbiota composition and adverse pregnancy outcomes. The first mechanism proposed that periodontal bacteria originating in the gingival biofilm could translocate from the unhealthy oral cavity and cross the placenta, reach the intra-amniotic fluid and fetal circulation and directly affect the fetoplacental unit, resulting in bacteremia (Hajishengallis, 2015). The second mechanisms proposed that systemic dissemination of endotoxins and/or inflammatory mediators derived from periodontal plaque and secreted by the subgingival inflammatory site could be carried to the fetoplacental unit (Stout et al., 2013; Aagaard et al., 2014). Although certain microbiota has the same species, their subgroups can have both positive and negative influences on pregnancy outcomes. Following this line of thought, we believe that composition or dysbiosis of the oral microbiome is more important than the presence of specific microbiota.

It is worth noting that microbial changes occurring during pregnancy might be nature consequences of a healthy pregnancy. Three reasons can explain the susceptibility to oral diseases such as periodontitis during pregnancy. These diseases are common in pregnant women due to hormonally driven hyper-

reactivity of the gingiva to bacteria in the subgingival biofilm. Other factors that increase the risk of poor oral health during pregnancy include changes in dietary habits (frequent snacking or increased consumption of carbohydrate-rich foods), stomach acids from nausea and vomiting that contribute to the breakdown of tooth enamel, and a decreased likelihood of seeking dental care during pregnancy. We plan to implement pathway analyses to investigate direct link between the microbiome and PTB.

Even though there was limited power resulting from a small number of participants and restricted validation sample size, our study verified that oral microbiota might provide potential biomarkers for predicting pregnancy complications using machine learning methods including random forest classification. Additionally, the fact that the entire microbiome was not analyzed was a limitation of this study because our analysis only used relative values measured by 16S rRNA sequencing, not 16S metagenome sequencing. We did not investigate other factors could impact the oral microbiome, such as participant's diet and socioeconomic status.

Despite these limitations, this prospective study demonstrated the potential of a PTB prediction model using oral microbiome in mouthwash. Further multi-center and larger-scale studies are needed to confirm our results before applying techniques developed in this study in the clinical field.

3 Prediction model for periodontitis severity based on the salivary microbiomes

This section includes the published contents:

3.1 Introduction

3.2 Materials and methods

3.2.1 Study subjects and clinical procedure

3.2.2 Bioinformatics analysis

3.3 Results

- 3.3.1 Summary of study subjects and sequencing data**
- 3.3.2 Diversity indices reveal differences in diversity among the periodontal statuses**
- 3.3.3 Differentially abundant taxa among healthy and different periodontitis stages and their correlation**
- 3.3.4 Classification of periodontitis stages by random forest models**

Table 2: **Clinical characteristics of the study subjects.** Significant differences were assessed using the Kruskal-Wallis test. NA: Not applicable.

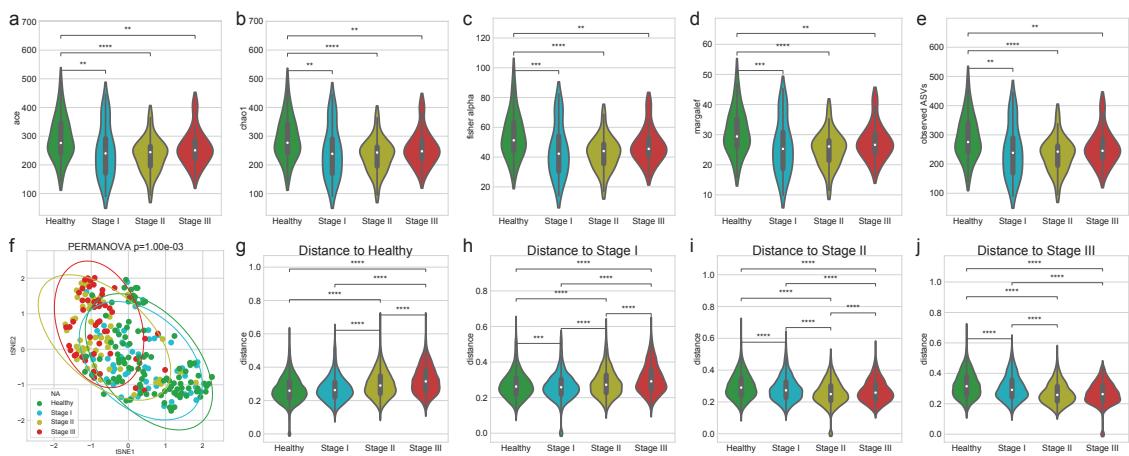


Figure 7: Diversity indices. Comparisons of salivary microbiomes among healthy controls and patients with various periodontitis stages. Alpha-diversity indices (**a-e**) indicate that healthy controls have increased heterogeneity than periodontitis stages as measured by: **(a)** ace **(b)** chao1 **(c)** Fisher alpha **(d)** Margalef, and **(e)** observed ASVs. **(f)** The beta-diversity index (weighted UniFrac) was visualized using a tSNE-transformed plot. The confidence ellipses are shown to display the distribution of each periodontitis stage. The distance to each stage demonstrated that each periodontitis stage was distinguished from the other periodontitis stages: **(g)** distance to Healthy **(h)** distance to Stage I **(i)** distance to Stage II, and **(j)** distance to Stage III. Statistical significance determined by the Mann-Whitney U-test (MWU) test: $p \leq 0.01$ (**) and $p \leq 0.0001$ (****).

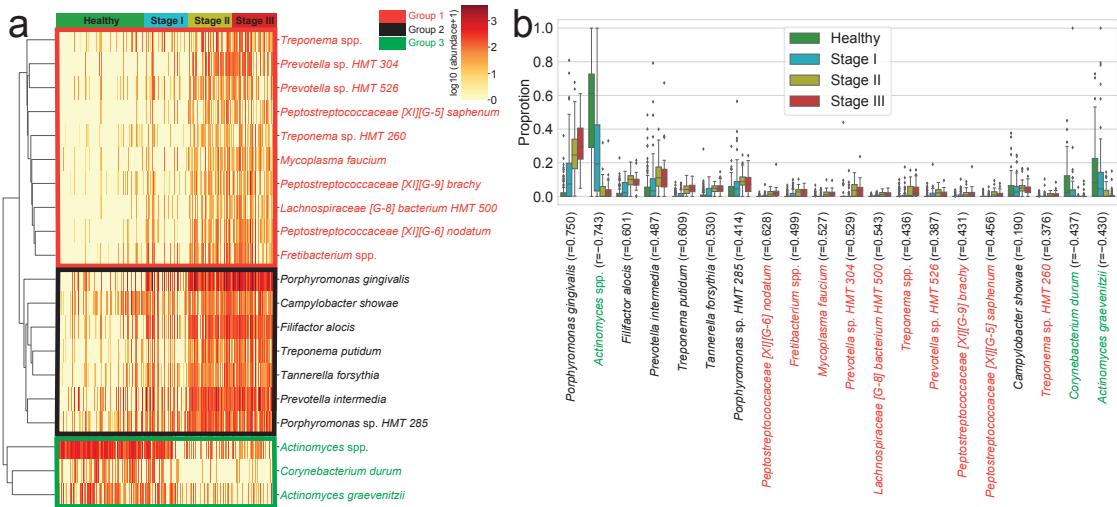


Figure 8: **Differentially abundant taxa**. Differentially abundant taxa (DAT) that were identified by ANCOM. **(a)** Heatmap of clustered DAT with similar distribution among subjects. Group 1, Group 2, and Group 3 are marked in red, black, and green, respectively. **(b)** Box plots showing the proportions of DAT. Taxa were sorted by their importance according to ANCOM.

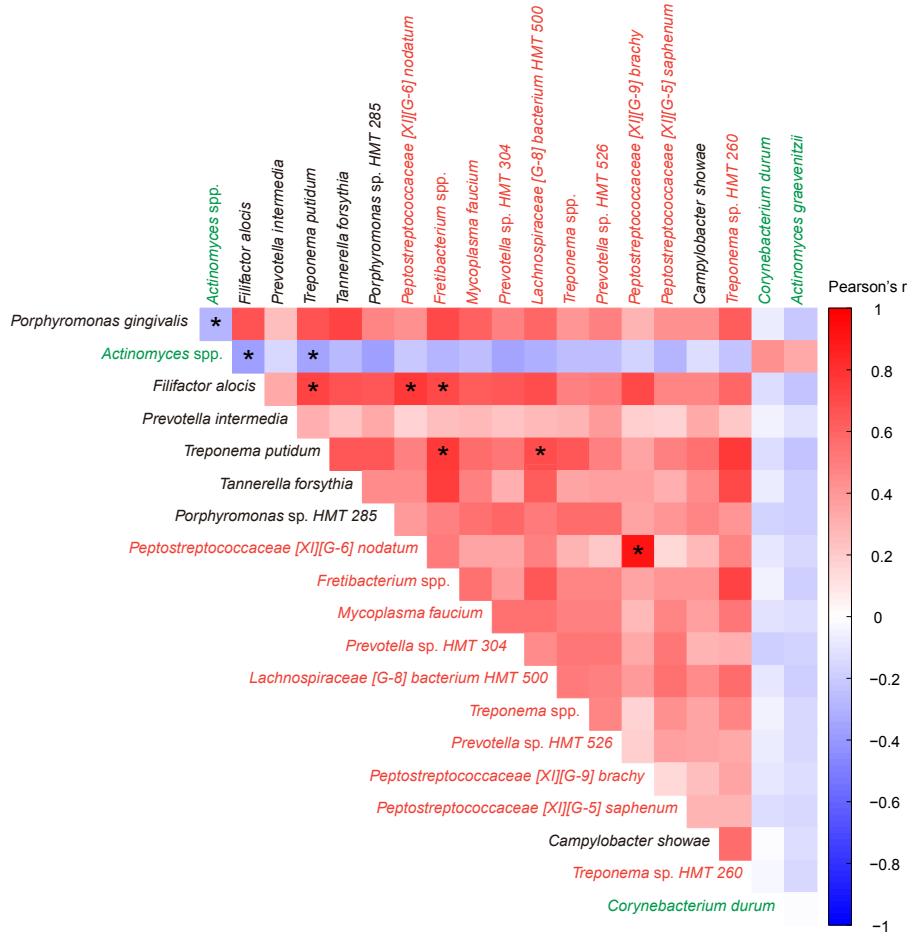


Figure 9: **Correlation heatmap.** Pearson's correlations between differentially abundant taxa in healthy status and periodontitis stages. Statistical significance was determined by strong correlation, i.e., $| \text{coefficient} | \geq 0.5$ (*).

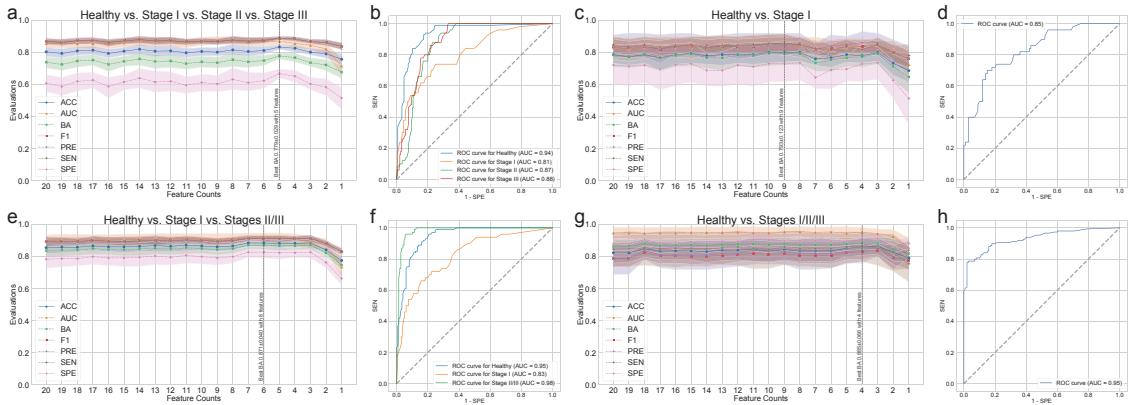


Figure 10: Random forest classification metrics. The classification metrics in the random forest classifications were as follows: accuracy (ACC), area-under-curve (AUC), balanced accuracy (BA), F1 score (F1), precision (PRE), sensitivity (SEN), and specificity (SPE). Every classification metric ranges from [0, 1], with higher values indicating better performance. **(a)** Classification performance for healthy vs. stage I vs. stage II vs. stage III. **(b)** Receiver-operating characteristics (ROC) curve for the highest BA of (a). **(c)** Classification performance for healthy vs. stage I. **(d)** ROC curve on the highest BA of (c). **(e)** Classification performance for healthy vs. stage I vs. stages II/III. **(f)** ROC curve for the highest BA of (e). **(g)** Classification performance for healthy vs. stages I/II/III. **(h)** ROC curve for the highest BA of (h).

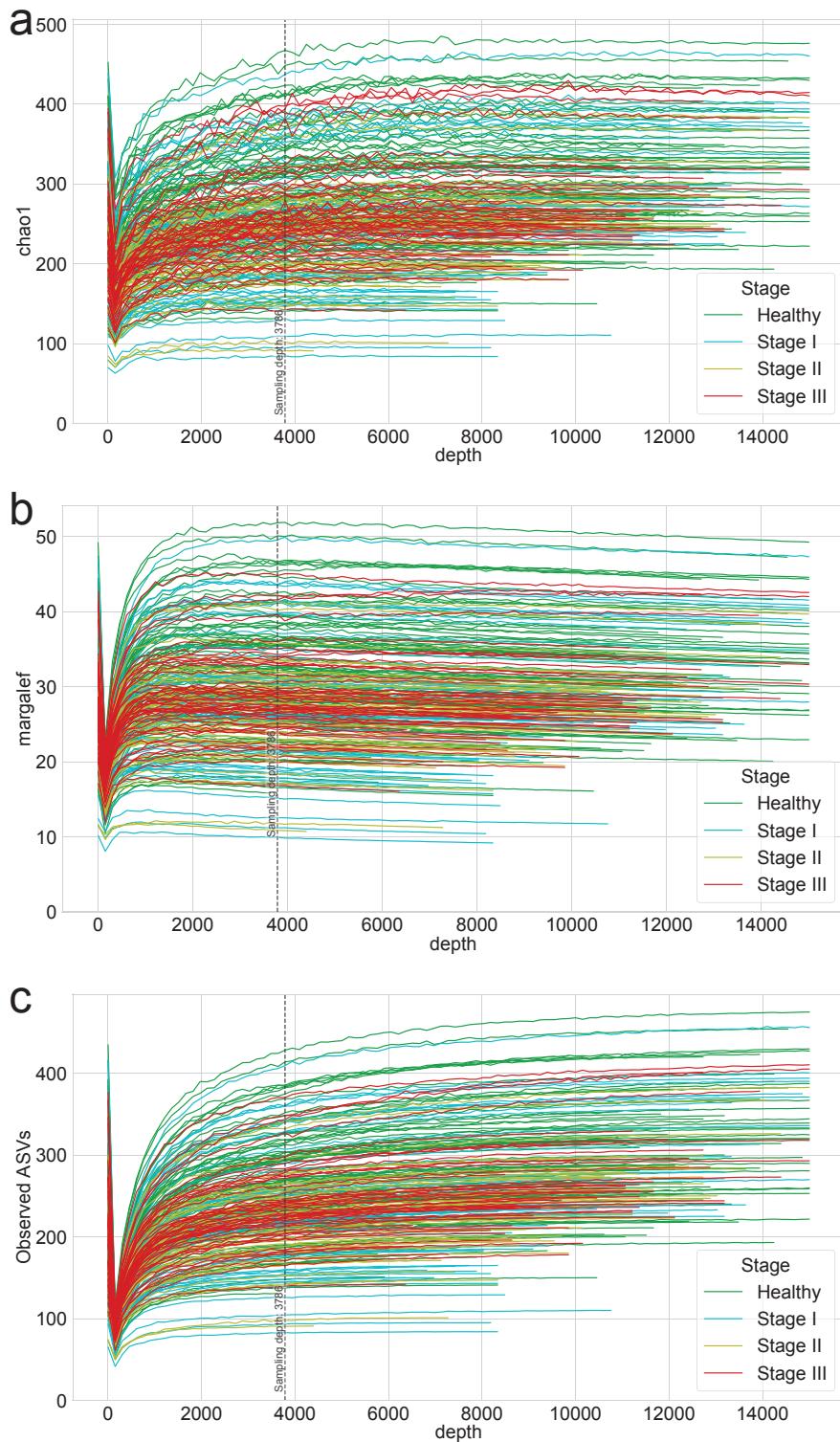


Figure 11: Rarefaction curves for alpha diversity indices. Rarefaction of (a) chao1 (b) margalef, and (c) observed ASVs were generated to measure species richness and determine the sampling depth of each sample.

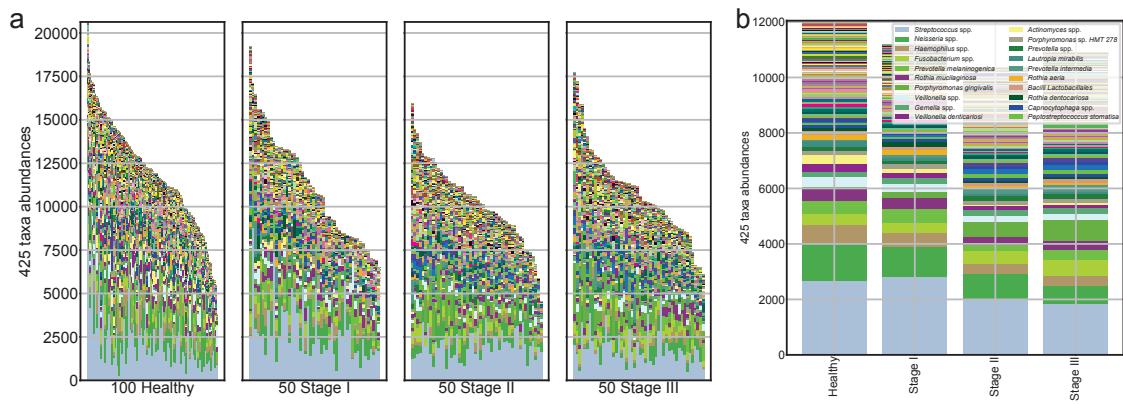


Figure 12: Absolute abundance of salivary bacterial taxa in the different periodontal statuses at the species level. Stacked bar plot of the absolute abundance of bacterial species for all samples **(a)** and the mean absolute abundance of bacterial species in the healthy, stage I, stage II, and stage III groups **(b)**.

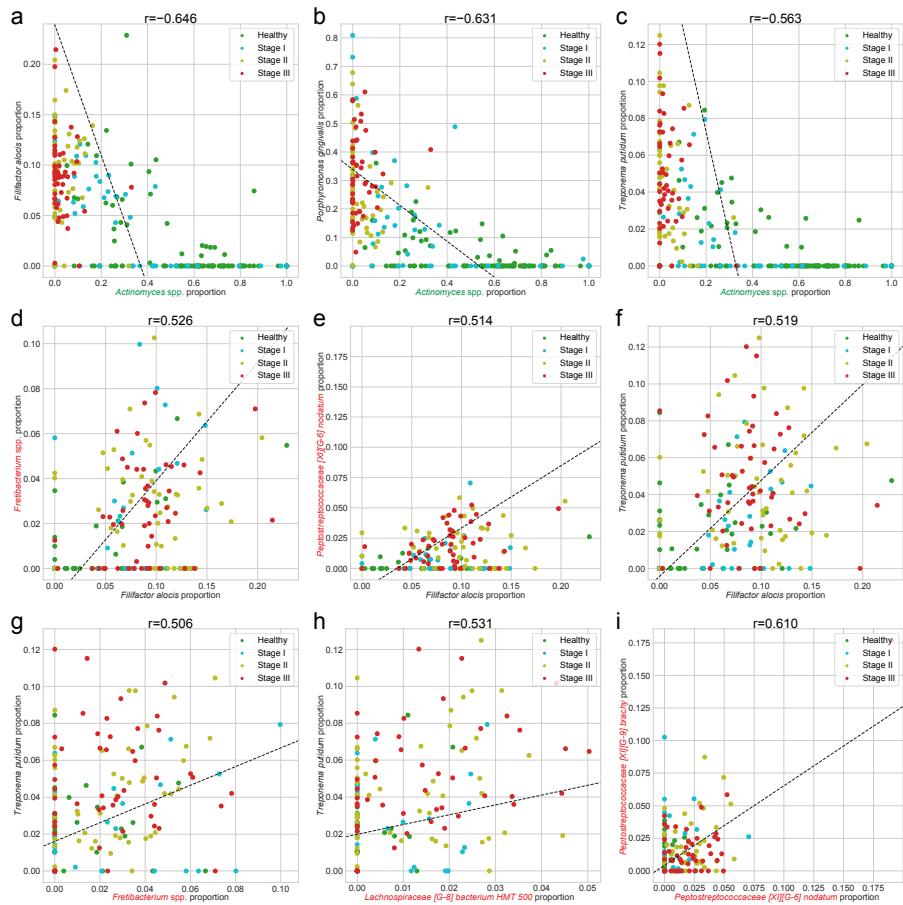


Figure 13: Correlation plots for differentially abundant taxa. We selected the combinations of DAT with absolute Spearman correlation coefficients greater than 0.5. The color represents periodontal healthy periodontal statuses (green: healthy, cyan: stage I, yellow: stage II, and red: stage III).

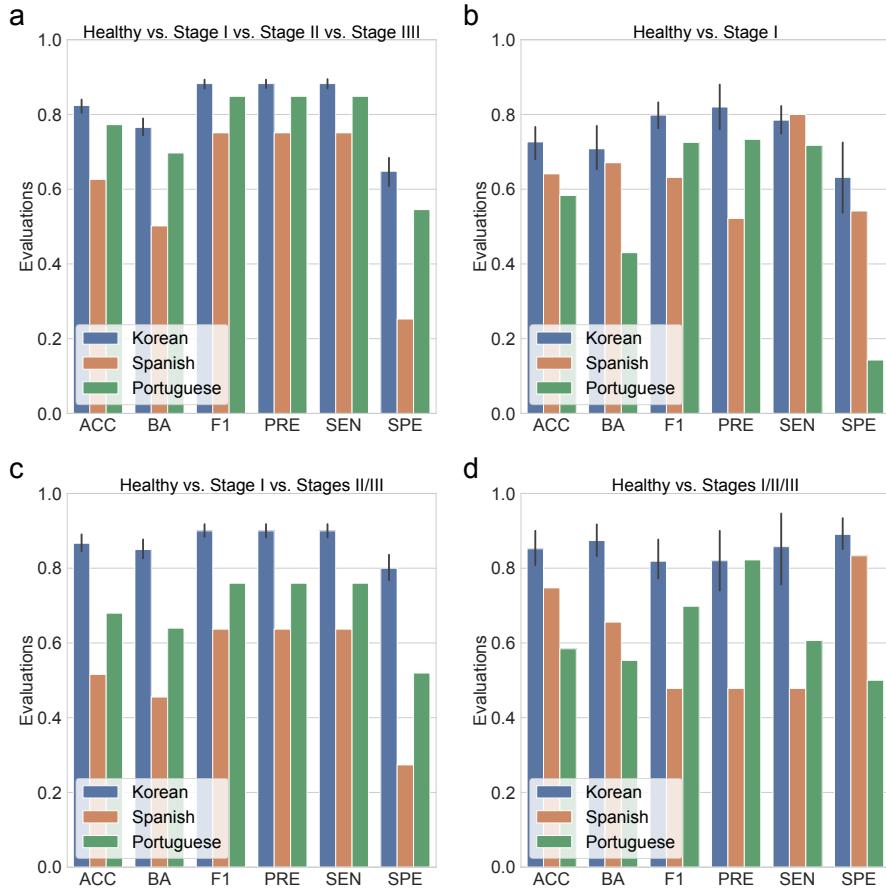


Figure 14: Random forest classification metrics from external datasets. The classification metrics in the random forest classification included accuracy (ACC), balanced accuracy (BA), F1 score (F1), precision (PRE), sensitivity (SEN), and specificity (SPE). Every classification metric ranges from [0, 1], with higher values indicating better performance. **(a)** Classification performance for healthy vs. stage I vs. stage II vs. stage III. **(b)** Classification performance for healthy vs. stage I. **(c)** Classification performance for healthy vs. stage I vs. stages II/III. **(d)** Classification performance for healthy vs. stages I/II/III.

3.4 Discussion

4 Lung microbiome

4.1 Introduction

4.2 Materials and methods

4.3 Results

4.3.1 Discussion

5 General conclusion and future perspective

5.1 General conclusions

In conclusion, the research described in this doctoral dissertation was conducted to identify significant

...

In the Section 2, I show that

5.2 Plan for future

5.3 Future perspective

References

- Aagaard, K., Ma, J., Antony, K. M., Ganu, R., Petrosino, J., & Versalovic, J. (2014). The placenta harbors a unique microbiome. *Science translational medicine*, 6(237), 237ra65–237ra65.
- Basavaprabhu, H., Sonu, K., & Prabha, R. (2020). Mechanistic insights into the action of probiotics against bacterial vaginosis and its mediated preterm birth: An overview. *Microbial pathogenesis*, 141, 104029.
- Berghella, V. (2012). Universal cervical length screening for prediction and prevention of preterm birth. *Obstetrical & gynecological survey*, 67(10), 653–657.
- Blencowe, H., Cousens, S., Oestergaard, M. Z., Chou, D., Moller, A.-B., Narwal, R., ... others (2012). National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The lancet*, 379(9832), 2162–2172.
- Bolstad, A., Jensen, H. B., & Bakken, V. (1996). Taxonomy, biology, and periodontal aspects of fusobacterium nucleatum. *Clinical microbiology reviews*, 9(1), 55–71.
- 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, 5–32.
- Brennan, C. A., & Garrett, W. S. (2019). Fusobacterium nucleatum—symbiont, opportunist and onacobacterium. *Nature Reviews Microbiology*, 17(3), 156–166.
- 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.
- Doyle, R., Alber, D., Jones, H., Harris, K., Fitzgerald, F., Peebles, D., & Klein, N. (2014). Term and preterm labour are associated with distinct microbial community structures in placental membranes which are independent of mode of delivery. *Placenta*, 35(12), 1099–1101.
- Fettweis, J. M., Serrano, M. G., Brooks, J. P., Edwards, D. J., Girerd, P. H., Parikh, H. I., ... others (2019). The vaginal microbiome and preterm birth. *Nature medicine*, 25(6), 1012–1021.

- Goldenberg, R. L., Culhane, J. F., Iams, J. D., & Romero, R. (2008). Epidemiology and causes of preterm birth. *The lancet*, 371(9606), 75–84.
- Goodey, M. D., Krleza-Jeric, K., & Lemmens, T. (2007). *The declaration of helsinki* (Vol. 335) (No. 7621). British Medical Journal Publishing Group.
- Hajishengallis, G. (2015). Periodontitis: from microbial immune subversion to systemic inflammation. *Nature reviews immunology*, 15(1), 30–44.
- Han, Y. W. (2015). Fusobacterium nucleatum: a commensal-turned pathogen. *Current opinion in microbiology*, 23, 141–147.
- Han, Y. W., & Wang, X. (2013). Mobile microbiome: oral bacteria in extra-oral infections and inflammation. *Journal of dental research*, 92(6), 485–491.
- Honest, H., Forbes, C., Durée, K., Norman, G., Duffy, S., Tsourapas, A., ... others (2009). Screening to prevent spontaneous preterm birth: systematic reviews of accuracy and effectiveness literature with economic modelling. *Health Technol Assess*, 13(43), 1–627.
- Hong, Y. M., Lee, J., Cho, D. H., Jeon, J. H., Kang, J., Kim, M.-G., ... Kim, J. K. (2023). Predicting preterm birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1), 21105.
- Iams, J. D., & Berghella, V. (2010). Care for women with prior preterm birth. *American journal of obstetrics and gynecology*, 203(2), 89–100.
- Ide, M., & Papapanou, P. N. (2013). Epidemiology of association between maternal periodontal disease and adverse pregnancy outcomes—systematic review. *Journal of clinical periodontology*, 40, S181–S194.
- Katz, J., Chegini, N., Shiverick, K., & Lamont, R. (2009). Localization of p. gingivalis in preterm delivery placenta. *Journal of dental research*, 88(6), 575–578.
- Kindinger, L. M., Bennett, P. R., Lee, Y. S., Marchesi, J. R., Smith, A., Caciato, S., ... MacIntyre, D. A. (2017). The interaction between vaginal microbiota, cervical length, and vaginal progesterone treatment for preterm birth risk. *Microbiome*, 5, 1–14.
- Leitich, H., & Kaider, A. (2003). Fetal fibronectin—how useful is it in the prediction of preterm birth? *BJOG: An International Journal of Obstetrics & Gynaecology*, 110, 66–70.
- León, R., Silva, N., Ovalle, A., Chaparro, A., Ahumada, A., Gajardo, M., ... Gamonal, J. (2007). Detection of porphyromonas gingivalis in the amniotic fluid in pregnant women with a diagnosis of threatened premature labor. *Journal of periodontology*, 78(7), 1249–1255.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15, 1–21.
- Offenbacher, S., Katz, V., Fertik, G., Collins, J., Boyd, D., Maynor, G., ... Beck, J. (1996). Periodontal infection as a possible risk factor for preterm low birth weight. *Journal of periodontology*, 67, 1103–1113.
- Payne, M. S., Newnham, J. P., Doherty, D. A., Furfaro, L. L., Pendal, N. L., Loh, D. E., & Keelan, J. A. (2021). A specific bacterial dna signature in the vagina of australian women in midpregnancy predicts high risk of spontaneous preterm birth (the predict1000 study). *American journal of*

obstetrics and gynecology, 224(2), 206–e1.

Romero, R., Dey, S. K., & Fisher, S. J. (2014). Preterm labor: one syndrome, many causes. *Science*, 345(6198), 760–765.

Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., ... others (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome*, 2, 1–19.

Sotiriadis, A., Papatheodorou, S., Kavvadias, A., & Makrydimas, G. (2010). Transvaginal cervical length measurement for prediction of preterm birth in women with threatened preterm labor: a meta-analysis. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 35(1), 54–64.

Spss, I., et al. (2011). IBM SPSS statistics for windows, version 20.0. New York: IBM Corp, 440, 394.

Stout, M. J., Conlon, B., Landeau, M., Lee, I., Bower, C., Zhao, Q., ... Mysorekar, I. U. (2013). Identification of intracellular bacteria in the basal plate of the human placenta in term and preterm gestations. *American journal of obstetrics and gynecology*, 208(3), 226–e1.

Vander Haar, E. L., So, J., Gyamfi-Bannerman, C., & Han, Y. W. (2018). Fusobacterium nucleatum and adverse pregnancy outcomes: epidemiological and mechanistic evidence. *Anaerobe*, 50, 55–59.

Witkin, S. (2019). Vaginal microbiome studies in pregnancy must also analyse host factors. *BJOG: An International Journal of Obstetrics & Gynaecology*, 126(3), 359–359.

Wong, T.-T., & Yeh, P.-Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8), 1586–1594.

Yang, I., Claussen, H., Arthur, R. A., Hertzberg, V. S., Geurs, N., Corwin, E. J., & Dunlop, A. L. (2022). Subgingival microbiome in pregnancy and a potential relationship to early term birth. *Frontiers in cellular and infection microbiology*, 12, 873683.

Acknowledgments

I would like to disclose my earnest appreciation for my advisor, Professor Semin Lee, who provided solicitous supervision and cherished opportunities throughout the course of my research. His advice and consultation encouraged me to become as a researcher and to receive all humility and gentleness. I am also grateful to all of my committee members, Professor AAA, Professor BBB, Professor CCC, and Professor DDD, for their critical and meaningful mentions and suggestions.

I extend my deepest gratitude to my Lord, *the Flying Spaghetti Monster*, His Noodly Appendage has guided me through the twist and turns of this academic journey. His presence, ever comforting and mysterious, has been a source of strength and humor during both highs and lows. In moments of doubt, I found solace in the belief that you were there, gently reminding me to keep faith in the process. His Holy Noodle has nourished my mind, and for that, I am truly overwhelmed. May His Holy Noodle continue to guide me in all my future endeavors. R’Amen.

(Professors)

I would like to extend my heartfelt gratitude to my colleagues of the Computational Biology Lab @ UNIST, whose collaboration, friendship, brotherhood, and support have been an invaluable part of my journey. Your willingness to share insights, engage in thoughtful discussions, and offer encouragement during the challenging moments of research has significantly shaped my academic experience. The camaraderie in the Computational Biology Lab made even the most demanding days more enjoyable, and I am deeply grateful for the collaborative environment we created together. I appreciate you for standing by my side throughout this Ph.D. journey.

I would like to express my heartfelt gratitude to my family, whose unwavering support has been the foundation of everything I have achieved. Your love, encouragement, and belief in me have sustained me through every challenge, and I could not have come this far without you. From your words of wisdom to your patience and understanding, each of you has played a vital role in helping me navigate this journey. The strength and comfort I have drawn from our family bond have been my greatest source of resilience. Your presence, both near and far, has filled my life with warmth and motivation. I am deeply grateful for your unconditional love and for always being there when I needed you the most. Thank you for being my constant source of strength and inspiration.

I am incredibly pleased to my friends, especially my GSHS alumni (오]망특), for their unwavering support and encouragement throughout this journey. The bonds we formed back in our school days have only grown stronger over the years, and I am fortunate to have had such loyal and understanding friends by my side. Your constant words of motivation, and even moments of levity during stressful times have helped keep me grounded. Whether it was a late-night conversations, a shared laugh, or a simple message of reassurance, you all have played a vital role in keeping me focused and motivated. I am relieved for the ways you celebrated each small achievement with me and how you patiently listened to my worries. The memories of our shared past provided me with comfort and a sense of stability when the road ahead felt uncertain. I could not have reached this point without the love and friendship that you all have generously given. Each of your, in your unique way, has contributed to this dissertation, even

if indirectly, and for that, I am forever beholden. I look forward to continuing our friendship as we all grow in our individual paths, knowing that the support we share is something truly special.

I would like to express my sincere gratitude to the amazing members of my animal protection groups, DRDR (두루두루) and UNIMALS (유니멀스), whose dedication and compassion have been a constant source of motivation. Your unwavering commitment to improving the lives of animals has inspired me throughout this journey. I am also thankful for the beautiful cats we have cared for, whose presence brought both joy and purpose to our allegiance. Their playful spirits and gentle companionship served as daily reminders of why we continue to fight for animal rights. The bond we share, both with each other and with the animals we protect, has enriched my life in countless ways. I appreciate you all again for your support, dedication, and for being part of this meaningful cause.

I would like to express my deepest gratitude to everyone I have had the honor of meeting throughout this journey. Your kindness, encouragement, and support have carried me through both the challenging and rewarding moments of my life. Whether through a kind word, thoughtful advice, or simply being there when I needed it most, your presence has made all the difference. I am incredibly fortunate to have received such generosity and warmth from those around me, and I do not take it for granted. Every act of kindness, no matter how big or small, has been a source of strength and motivation for me. To all my friends, colleagues, mentors, and beloved ones, thank you for your unwavering support. I am truly grateful for each of you, and your kindness has left an indelible mark on my journey.

My Lord, *the Flying Spaghetti Monster*,
give us grace to accept with serenity the things that cannot be changed,
courage to change the things that should be changed,
and the wisdom to distinguish the one from the other.

Glory be to *the Meatball*, to *the Sauce*, and to *the Holy Noodle*.

As it was in the beginning, is now, and ever shall be.

R'Amen.

