Doctoral Thesis

Microbiota in Human Diseases

Jaewoong Lee

Department of Biomedical Engineering

Ulsan National Institute of Science and Technology

2025

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.

Bobby Henderson

Representative,
Church of the Flying Spaghetti Monster
Bobby Henderson

1935 nw Calkins Ave | Roseburg, OR 97471 | 503-482-8008
mail@spaghettimonster.org | www.spaghettimonster.org

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.

Bobby Henderson

Representative,
Church of the Flying Spaghetti Monster
Bobby Henderson

1935 nw Calkins Ave | Roseburg, OR 97471 | 503-482-8008
mail@spaghettimonster.org | www.spaghettimonster.org

Abstract

(Microbiome)

(PTB) Section 2 introduces...

(Periodontitis) Section 3 describes...

(Colon) Setion 4...

(Conclusion)

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

- Hong, Y. M., **Lee, Jaewoong**, 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.

Contents

List of Figures

List of Tables

List of Abbreviations

ACC Accuracy

ASV Amplicon sequence variant

AUC Area-under-curve

BA Balanced accuracy

C-section Cesarean section

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

SD Standard deviation

SEN Sensitivity

SPE Specificity

t-SNE t-distributed stochastic neighbor embedding

1 Introduction

The microbiome refers to the complex community of microorganisms, including bacteria, viruses, fungi, and other microbes, that inhabit various environment within living organisms (Ursell, Metcalf, Parfrey, & Knight, 2012; Gilbert et al., 2018). In humans, the microbiome plays a crucial role in maintaining health (Lloyd-Price, Abu-Ali, & Huttenhower, 2016), influencing processes such as digestion (Lim, Park, Tong, & Yu, 2020), immune response (Thaiss, Zmora, Levy, & Elinav, 2016; Kogut, Lee, & Santin, 2020; C. H. Kim, 2018), and even mental health (Mayer, Tillisch, Gupta, et al., 2015; X. Zhu et al., 2017; X. Chen, D'Souza, & Hong, 2013). These microbial communities are not static nor constant, but rather dynamic ecosystem that interacts with their host and respond to environmental changes. Recent studies have revealed that imbalances in the microbiome, known as dysbiosis, can contribute to a wide range of diseases, including obesity (John & Mullin, 2016; Tilg, Kaser, et al., 2011; Castaner et al., 2018), diabetes (Barlow, Yu, & Mathur, 2015; Hartstra, Bouter, Bäckhed, & Nieuwdorp, 2015; Sharma & Tripathi, 2019), infections (Whiteside, Razvi, Dave, Reid, & Burton, 2015; Alverdy, Hyoju, Weigerinck, & Gilbert, 2017), inflammatory conditions (Francescone, Hou, & Grivennikov, 2014; Peirce & Alviña, 2019; Honda & Littman, 2012), and cancers (Helmink, Khan, Hermann, Gopalakrishnan, & Wargo, 2019; Cullin, Antunes, Straussman, Stein-Thoeringer, & Elinav, 2021; Sepich-Poore et al., 2021; Schwabe & Jobin, 2013). Thus, understanding the composition of the human microbiomes is essential for developing new therapeutic approaches that target these microbial populations to promote health and prevent diseases.

(Brain-gut axis)

16S ribosomal RNA (rRNA) gene sequencing is one of the most extensively applied methods for characterizing microbial communities by targeting the conserved 16S rRNA gene, which contains both highly conserved and variable regions in bacteria (Tringe & Hugenholtz, 2008; Janda & Abbott, 2007). The conserved regions enable universal primer binding, while the variable regions provide the specificity needed to differentiate microbial taxa. Among these regions, the V3-V4 region is frequently selected for sequencing due to its balance between phylogenetic resolution and sequencing efficiency (Johnson et al., 2019). Therefore, the V3-V4 region offers sufficient variability to classify a wide range of bacteria taxa while maintaining compatibility with widely used sequencing platforms.

Diversity indices are essential techniques for evaluating the complexity and variety of microbial communities,in ecological and microbiological research (Tucker et al., 2017; Hill, 1973). Alpha-diversity index attributes to the heterogeneity within a specific community, obtaining the number of different taxa and the distribution of taxa among the individuals, i.e., richness and evenness. On the other hand, beta-diversity index measures the variations in microbiome compositions between the individuals, highlighting differences among the microbiome compositions of the study participants. Altogether, by providing a thorough understanding of microbiome compositions, diversity indices, e.g. alpha-diversity and beta-diversity, allow us to investigate factors that affecting community variability and structure.

(DAT selection)

Classification is one of the supervised machine learning techniques used to categorized data into predefined classes based on features within the data. In other words, the method learns the relationship between input features and their corresponding output classes through the process of training a classification model using labeled data. Classification models are essential for advising choices in a wide range of applications, including medical diagnostics. Thus, researchers could uncover sophisticated connections in input features and corresponding classes and produce reliable prediction by utilizing machine learning classification.

Random forest classification is one of the ensemble machine learning methods that constructs several decision trees during training and aggregates their results to provide classification predictions (Breiman, 2001). A portion of the features and classes–known as bootstrapping (Jiang & Simon, 2007; Champagne, McNairn, Daneshfar, & Shang, 2014; J.-H. Kim, 2009) and feature bagging (Bryll, Gutierrez-Osuna, & Quek, 2003; Alelyani, 2021; Yaman & Subasi, 2019)–are utilized to construct each tree in the forest. The majority vote from each tree determines the final classification, which lowers the possibility of overfitting in comparison to a single decision tree. Furthermore, random forest classifier offers several advantages, including its robustness to outliers and its ability to calculate the feature importance.

Evaluating the performance of a machine learning classification model is essential to ensure its reliability and effectiveness in real-world solutions and applications. A confusion matrix is a tabular representation of predictions of classification, showing the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (Table 1). From this matrix, evaluations can be derived: accuracy (ACC; Equation 1), balanced accuracy (BA; Equation 2), F1 score (F1; Equation 3), sensitivity (SEN; Equation 4), specificity (SPE; Equation 5), and precision (PRE; Equation 6). These metrics are in $[0, 1]$ range and high metrics are good metrics. The confusion matrix also helps in identifying specific types of errors, such as a tendency to produce false positive or false negatives, offering valuable insights for improving the classification model. By combining the confusion matrix with other evaluation metrics, researchers can comprehensively assess the classification metrics and refine it for real-world solutions and applications.

(AUC)

(Limitation & Novelty)

Table 1: Confusion matrix

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

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}} \tag{1}$$

$$\mathrm{BA} = \frac{1}{2} \times (\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} + \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}) \tag{2}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} \tag{5}$$

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

2 Predicting preterm birth using random forest classifier in salivary microbiome

This section includes the published contents:
Hong, Y. M., **Lee, Jaewoong**, 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), characterized by the delivery of neonates prior to 37 weeks of gestation, is one of the major cause to neonatal mortality and morbidity (Blencowe et al., 2012). Multiple pregnancies including twins, short cervical length, and infection on genitourinary tract are known risk factor for PTB (Goldenberg, Culhane, Iams, & Romero, 2008). Nevertheless, the extent to which these aspects affect birth outcomes is still up for debate. Henceforth, strategies to boost gestation and enhance delivery outcomes can be more conveniently implemented when pregnant women at high risk of PTB are identified early (Iams & Berghella, 2010).

Prediction models that can be utilized as a foundation for intervention methods still have an unacceptable amount of classification evaluations, including accuracy, sensitivity, and specificity, despite a great awareness of the risk factors that trigger PTB (Sotiriadis, Papatheodorou, Kavvadias, & Makrydimas, 2010). Several attempts have been made to predict PTB through integrating data such as human microbiome composition, inflammatory markers, and prior clinical data with predictive machine learning methods (Berghella, 2012). Because it is affordable and straightforward to use, fetal fibronectin is commonly used in medical applications. However, with a sensitivity of only 56% that merely similar to random prediction, it has a low classification evaluation (Honest et al., 2009). Due to the difficulty and imprecision of the method in general, as well as the requirement for a qualified specialist cervical length measuring is also restricted (Leitich & Kaider, 2003).

Preterm prelabor rupture of membranes (PROM) brought on by gestational inflammation and infection contribute to about 70% of PTB cases (Romero, Dey, & Fisher, 2014). Nevertheless, as antibiotics and anti-inflammatory therapeutic strategies were ineffective to decrease PTB occurrence rates, the pathology of PTB has not been entirely elucidated by inflammatory and infectious pathways (Romero, Hassan, et al., 2014). Recent researches on maternal microbiomes were beginning to examine unidentified connections of PTB as a consequence of developmental processes in molecular biological technology (Fettweis et al., 2019).

However, as anti-inflammatory and antibiotic therapies were insufficient to lower PTB occurrence rates, infectious and inflammatory processes are insufficient to exhaustively clarify the pathogenesis and pathophysiology of PTB. It has been hypothesized that the microbiota linked to PTB originate from either a hematogenous pathway or the female genitourinary tract increasing through the vagina and/or cervix. (Han & Wang, 2013). Vaginal microbiome compositions have been found in women who eventually acquire PTB, and recent studies have tried to predict PTB risk using cervico-vaginal fluid (Kindinger et al., 2017). Even though previous investigation have confirmed the potential relationships between the vaginal microbiome compositions and PTB, these studies are only able to clarify an upward trajectory.

Multiple unfavorable birth outcomes, including PROM and PTB, have been linked to periodontitis as an independence risk factor, according to numerous epidemiological researches (Offenbacher et al., 1996). It is expected that the oral microbiome will be able to explain additional hematogenous pathways in light of these precedents; however, the oral microbiome composition of fetuses is limited understood.

Hence, in order to identify the salivary microbiome linked to PTB and to establish a machine learning prediction model of PTB determined by oral microbiome compositions, this study examined the salivary microbiome compositions of PTB study participants with a full-term birth (FTB) study participants.

2.2 Materials and methods

2.2.1 Study design and study participants

Between 2019 and 2021, singleton pregnant women who received treatment to Jeonbuk National University Hospital for childbirth were the participants of this study. This study was conducted according to the Declaration of Helsinki (Goodyear, Krleza-Jeric, & Lemmens, 2007). The Institutional Review Board authorized this study (IRB file No. 2019-01-024). Participants who were admitted for elective cesarean sections (C-sections) or induction births, as well as those who had written informed consent obtained with premature labor or PROM, were eligible.

2.2.2 Clinical data collection and grouping

Questionnaires and electronic medical records were implemented to gather information on both previous and current pregnancy outcomes. The following clinical data were analyzed:

- maternal age at delivery
- diabetes mellitus
- hypertension
- overweight and obesity
- C-section
- history PROM or PTB
- gestational week on delivery
- birth weight
- sex

2.2.3 Salivary microbiome sample collection

Salivary microbiome samples were collected 24 hours before to delivery using mouthwash. The standard methods of sterilizing were performed. Medical experts oversaw each stage of the sample collecting procedure. Participants received instruction not to eat, drink, or brush their teeth for 30 minutes before sampling salivary microbiome. Saliva samples were gathered by washing the mouth for 30 seconds with 12 mL of a mouthwash solution (E-zen Gargle, JN Pharm, Pyeongtaek, Gyeonggi, Korea). The samples were tagged with the anonymous ID for each participant and kept at 4 °Cuntil they underwent further processing. Genomic DNA was extracted using an ExgeneTM Clinic SV kit (GeneAll Biotechnology, Seoul, Korea) following with the manufacturer instructions and store at -20 °C.

2.2.4 16s rRNA gene sequencing

Salivary microbiome samples were transported to the Department of Biomedical Engineering of the Ulsan National Institute of Science and Technology . 16S rRNA sequencing was then carried out using a commissioned Illumina MiSeq Reagent Kit v3 (Illumina, San Diego, CA, USA). Library methods were utilized to amplify the V3-V4 areas. 300 base-pair paired-end reads were produced by sequencing the pooled library using a v3 $\times$600 cycle chemistry after the samples had been diluted to a final concentration of 6 pM with a 20% PhiX control.

2.2.5 Bioinformatics analysis

The independent t-test was utilized to evaluate the differences of continuous values between from the PTB participants than the FTB participants; χ-square test was applied to decide statistical differences of categorical values. Clinical measurement comparisons were conducted using SPSS (version 20.0) (Spss et al., 2011). At $p < 0.05$, statistical significance was taken into consideration.

QIIME2 (version 2022.2) was implemented to import 16S rRNA gene sequences from salivary microbiome samples of study participants for additional bioinformatics processing (Bolyen et al., 2019). DADA2 was used to verify the qualities of raw sequences (Callahan et al., 2016). The remain sequences were clustered into amplicon sequence variants (ASVs). Diversity indices, namely Faith PD for alpha diversity index (Faith, 1992) and Hamming distance for beta diversity index (Hamming, 1950), were calculated. MWU test (Mann & Whitney, 1947), and PERMANOVA multivariate test were evaluated for measuring statistical significance (Anderson, 2014; Kelly et al., 2015).

Taxonomic assignment were implemented with HOMD (version 15.22) (T. Chen et al., 2010). Afterward, DESeq2 was implemented to identify differentially abundant taxa (DAT) that could distinguish between salivary microbiome from PTB and FTB participants (Love, Huber, & Anders, 2014). Taxa with $|\log_2 \text{FoldChange}| > 1$ and $p < 0.05$ were considered as statistically significant.

The taxa for predicting PTB using salivary microbiome data were determined using a random forest classifier (Breiman, 2001). Through stratified k-fold cross-validation ($k = 5$) that preserves the existence rate of PTB and FTB participants, consistency and trustworthy classification were ensured (Wong & Yeh, 2019).

2.2.6 Data and code availability

All sequences from the 59 study participants have been added to the Sequence Read Archives (project ID PRJNA985119): `https://dataview.ncbi.nlm.nih.gov/object/PRJNA985119?reviewer=6fdj2e9c8gp9vtf52n330e2h8j`. Docker image that employed throughout this study is available in the DockerHub: `https://hub.docker.com/r/fumire/helixco_premature`. Every code used in this study can be found on GitHub: `https://github.com/CompbioLabUnist/Helixco_Premature`.

2.3 Results

2.3.1 Overview of clinical information

In the beginning, 69 volunteer mothers were recruited for this study. However, due to insufficient clinical information or twin pregnancies, 10 participants were excluded from the study participants. Demographic and clinical information of the study participants are displayed in Table 2. Because PROM is one of the leading factors of PTB, it was prevalent in the PTB group than the FTB group. Other maternal clinical factors did not significantly differ between the FTB and PTB groups. There were no cases in both groups that had a history of simultaneous periodontal disease or cigarette smoking.

2.3.2 Comparison of salivary microbiomes composition

The salivary microbiome composition was composed of 13953804 sequences from 59 study participants, with 102305.95±19095.60 and 64823.41±15841.65 (mean±SD) reads/sample before and following the quality-check stage, accordingly. There was not a significant distinction between the PTB and FTB groups with regard to on alpha diversity nor beta diversity metrics (Figure 4).

DESeq2 was used to select 32 DAT that distinguish between the PTB and FTB groups out of the 465 species that were examined (Love et al., 2014): 26 FTB-enriched DAT and six PTB-enriched DAT. Seven PROM-related DAT were removed from these 32 PTB-related DAT to lessen the confounding effect of PROM (Figure 5). Therefore, there were a total of 25 PTB-related DAT: 22 FTB-enriched DAT and three PTB-enriched DAT (Figure 1).

A significant negative correlation was found using Pearson correlation analysis between GW and differences between PTB-enriched DAT and FTB-enriched DAT (r = -0.542 and p = 7.8e-6; Figure 5).

2.3.3 Random forest classification to predict PTB risk

To classify PTB according to DAT, random forest classifiers were constructed. The nine most significant DAT were used to obtain the best BA (0.765±0.071; Figure 3a). Moreover, random forest classification model determined each DAT's importance (Figure 3b). We conducted a validation procedure on nine twin pregnancies that were excluded in the initial study design in order to confirm the reliability and dependability of our random forest-based PTB prediction model (Figure 6). Comparable to the PTB prediction model on the 59 initial singleton study participants, the validation classification on PTB risk of these twin participants have an accuracy of 87.5%.

Table 2: **Standard clinical information of study participants**.

Continuous variable for independent t-test. Categorical variable for Pearson's χ-square test. Continuous variable: mean±SD. Categorical vaiable: count (proprotion)

	PTB (n=30)	FTB (n=29)	p-value
Maternal age (years)	31.8±5.2	33.7±4.5	0.687
C-section	20 (66.7%)	24 (82.7%)	0.233
Previous PTB history	4 (13.3%)	1 (3.4%)	0.353
PROM	12 (40.0%)	1 (3.4%)	0.001
Pre-pregnant overweight	8 (26.7%)	7 (24.1%)	1.000
Gestational weight gain (kg)	9.0±5.9	11.5±4.6	0.262
Diabetes	2 (6.7%)	2 (6.9%)	1.000
Hypertension	11 (36.7%)	4 (13.8%)	0.072
Gestational age (weeks)	32.5±3.4	38.3±1.1	≤0.001
Birth weight (g)	1973.4±686.6	3283.4±402.7	≤0.001
Male	14 (46.7%)	13 (44.8%)	1.000

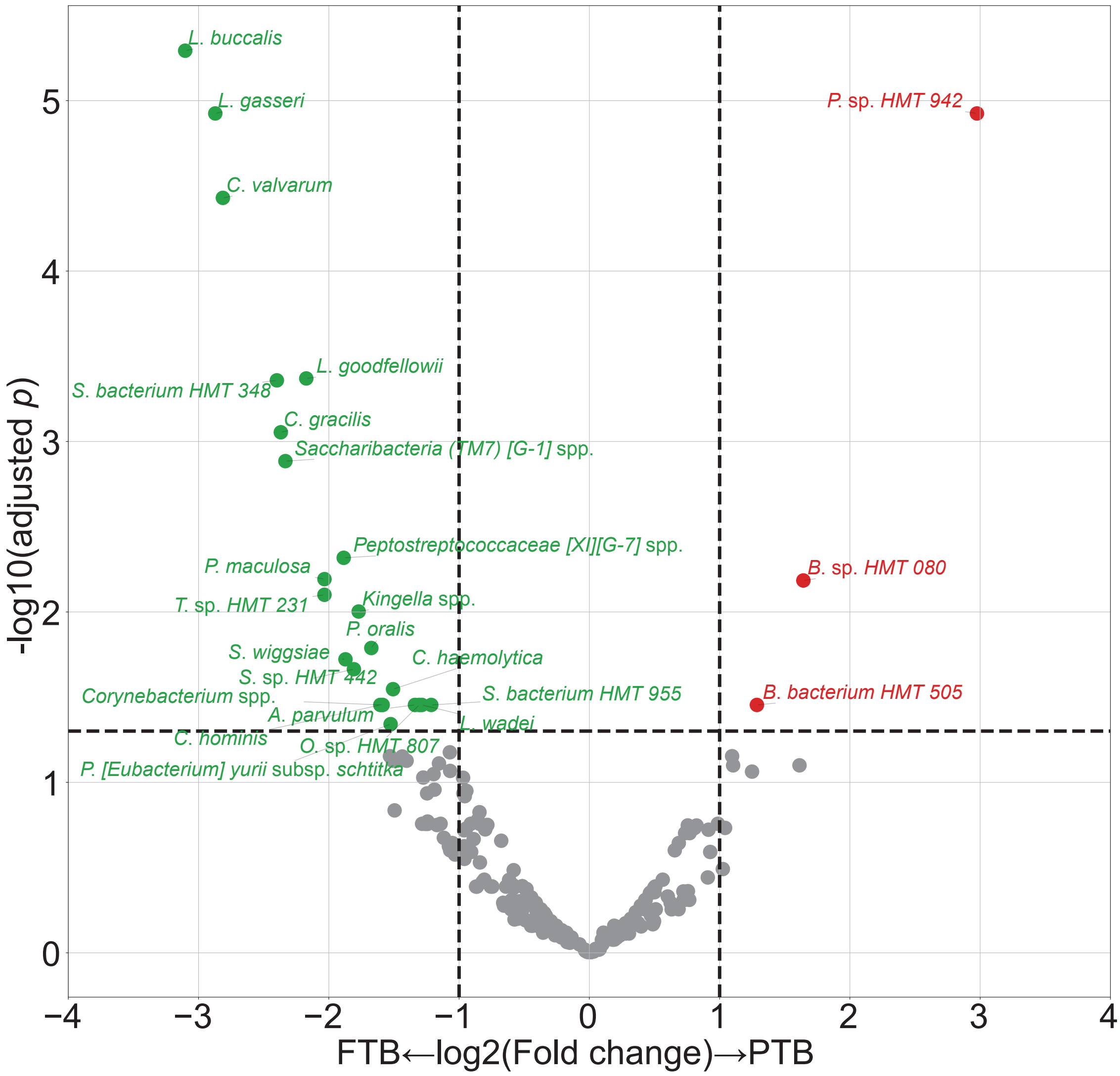

Figure 1: **DAT volcano plot**.

Red dots represent PTB-enriched DAT, while green dots represent FTB-enriched DAT.

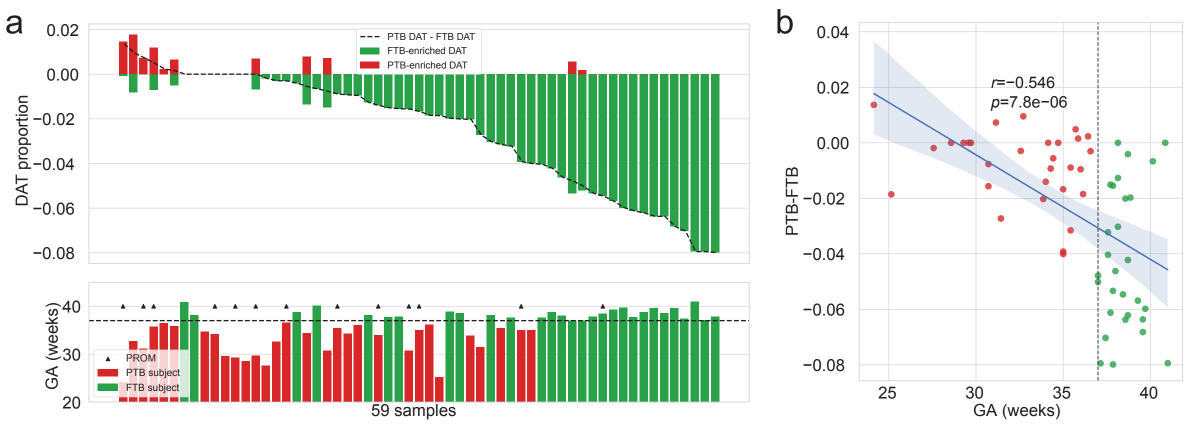

Figure 2: **Salivary microbiome compositions over DAT**.

(a) Frequencies of DAT of study subjects. The study participants are arranged in respect of (PTB-enriched DAT – FTB-enriched DAT). The study participants' GA is displayed in accordance with the upper panel's order (PTB: red bar, FTB: green bar. PROM: arrow head.) **(b)** Correlation plot with GA and (PTB-enriched DAT – FTB-enriched DAT). Strong negative correlation is found with Pearson correlation.

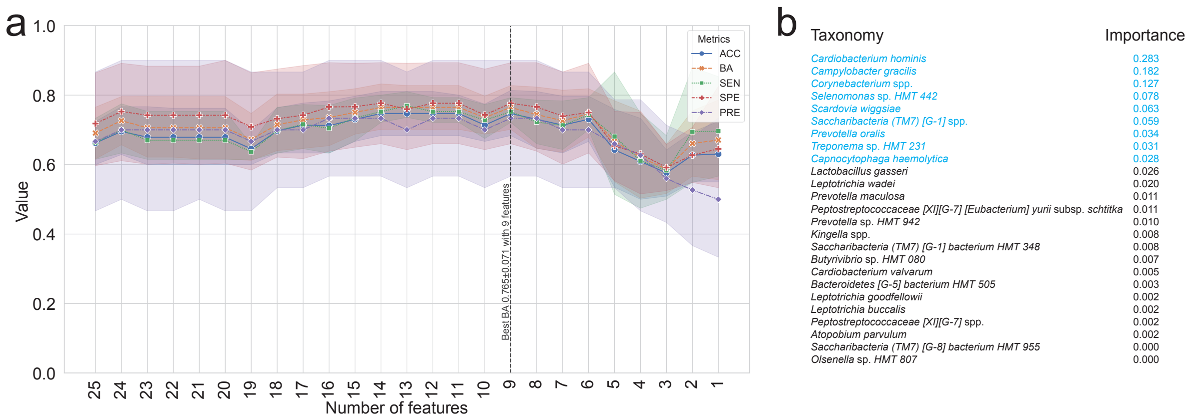

Figure 3: **Random forest-based PTB prediction model**.

(a) Machine learning evaluations upon number of features (DAT). Random Forest classifier has the best BA (0.765±0.071; Mean±SD) with the nine most important DAT. **(b)** Importance of DAT.

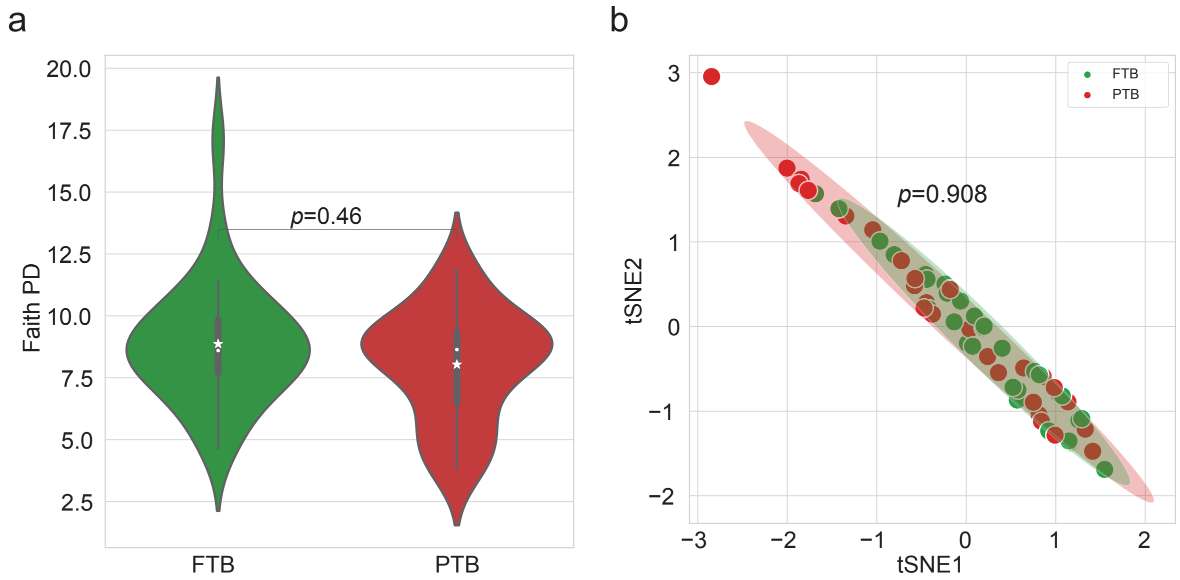

Figure 4: **Diversity indices**.

(a) Alpha diversity index (Faith PD). There is no statistically significant difference between the PTB and FTB group (MWU test $p = 0.46$). **(b)** t-SNE plot with beta diversity index (Hamming distance). There is no statistically significant difference between the PTB and FTB group (PERMANOVA test $p = 0.908$)

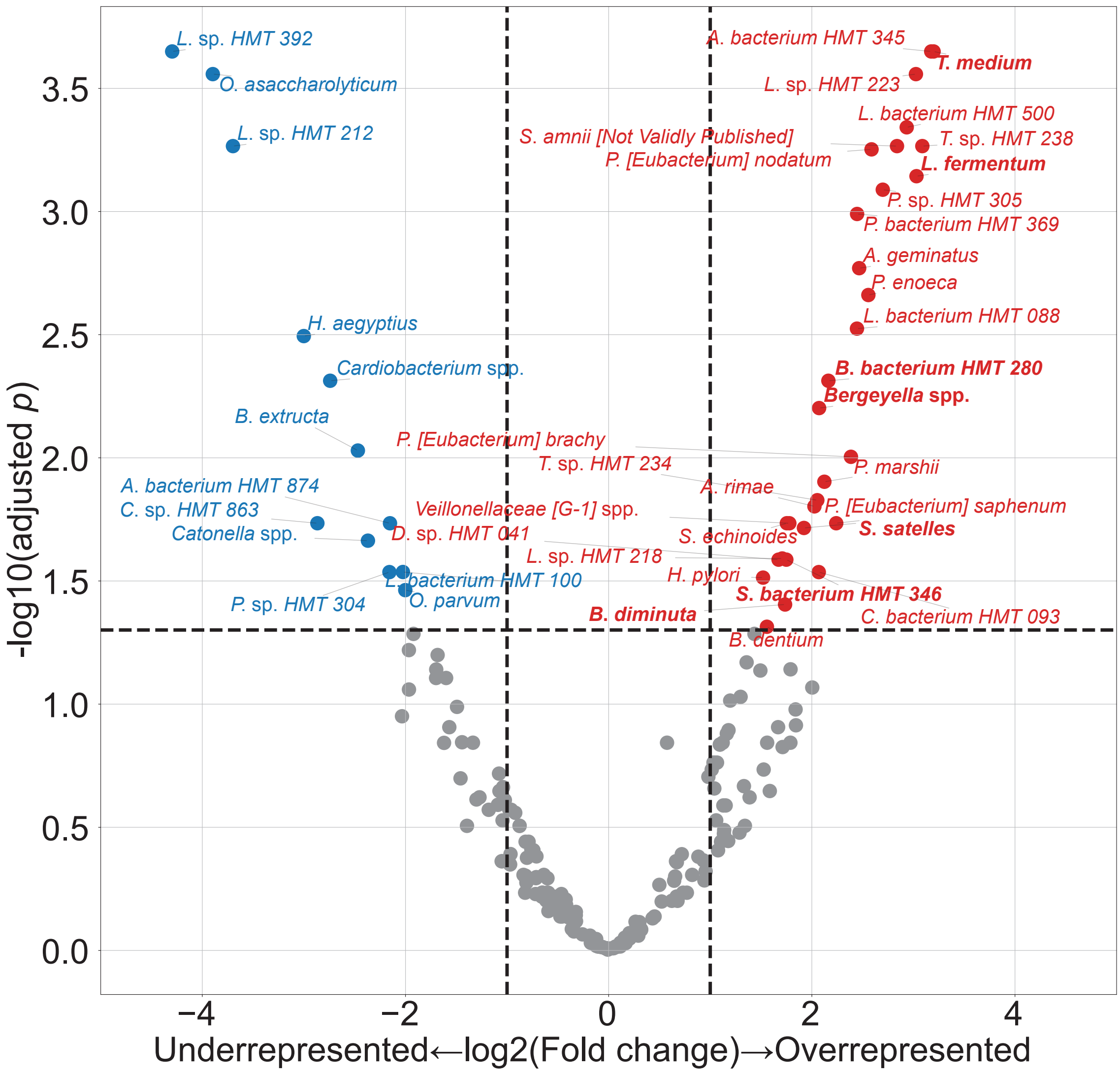

Figure 5: **PROM-related DAT**.

Only seven of these 42 PROM-related DAT overlapped with PTB-related DAT (bold text). Blue dots represented PROM-underrepresented DAT, while red dots represented PROM-overrepresented DAT.

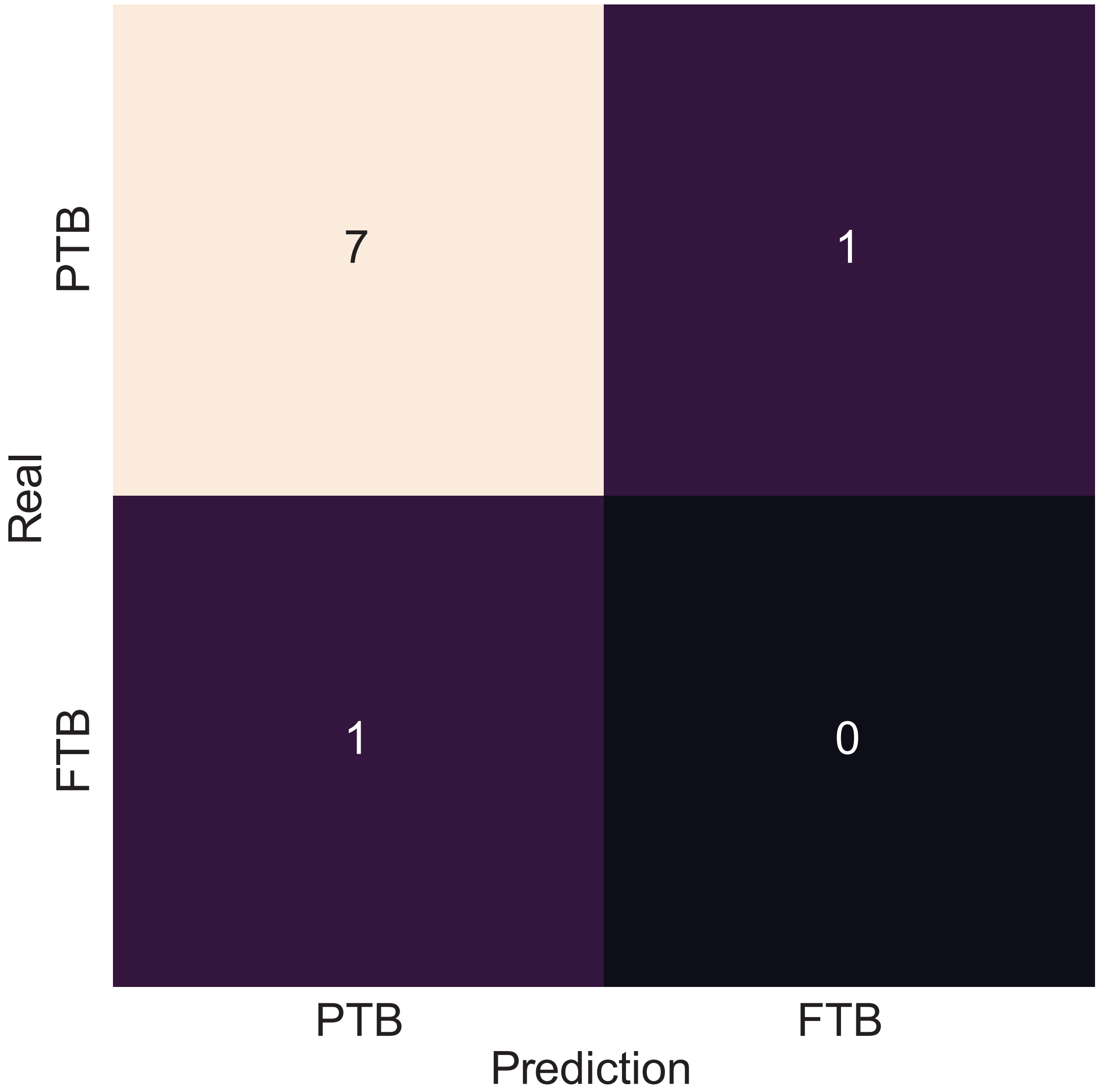

Figure 6: **Validation of random forest-based PTB prediction model**.
Nine twin pregnancies (eight PTB subjects and a FTB subject) that were excluded in the initial study subjects were subjected to a validation procedure. The random forest-based PTB prediction model shows 87.5% accuracy, comparable to the PTB classification evaluations on the singleton study subjects (0.714±0.061. Mean±SD)

2.4 Discussion

In this study, we employed salivary microbiome compositions to develop the random forest-based PTB prediction models to estimate PTB risks. Previous reports have indicated bidirectional associations between pregnancy outcomes and salivary microbiome compositions (Han & Wang, 2013). Nevertheless, the salivary microbiome composition is not yet elucidated. Salivary microbial dysbiosis, including gingival inflammation and periodontitis, have been connected to unfavorable pregnancy outcomes, such as PTB (Ide & Papapanou, 2013). However, the techniques utilized in recent research that primarily focus on recognized infections have led to inconsistent outcomes.

One of the most common salivary taxa that has been examined is *Fusobacterium nucleatum* (Han, 2015; Brennan & Garrett, 2019; Bolstad, Jensen, & Bakken, 1996), that is a Gram-negative, anaerobic, and filamentous bacteria. *Fusobacterium nucleatum* can be separated from not only the salivary microbiome but also the vaginal microbiome (Vander Haar, So, Gyamfi-Bannerman, & Han, 2018; Witkin, 2019). In both animal and human investigation, *Fusobacterium nucleatum* infection has been linke to risk of PTB (Doyle et al., 2014). According to recent researches, the placenta women who give birth prematurely may include additional salivary microbiome dysbiosis, such as *Bergeyella* spp. and *Porphyromonas gingivalis* (León et al., 2007; Katz, Chegini, Shiverick, & Lamont, 2009). Although *Bergeyella* spp. were one of the PROM-overrepresented DAT (Figure 5), it was excluded in the final 25 PTB-related DAT. Furthermore, *Porpyhromonas gingivalis* and *Campylobacter gracilis* were pathogens of periodontitis in sub-gingival microbiome (Yang et al., 2022). *Lactobacillus gasseri* was also one of the FTB-enriched DAT (Figure 1), and it is well established that early PTB risk can be reduced by *Lactobacillus gasseri* in the vaginal microbiome (Basavaprabhu, Sonu, & Prabha, 2020; Payne et al., 2021).

With DAT comprising 22 FTB-enriched DAT and three PTB-enriched DAT (Figure 1), we discovered that the FTB study participants had the majority of the essential DAT that distinguished between the PTB and FTB groups. Thus, we hypothesize that the pathogenesis and pathophysiology of PTB may have been triggered by an absence of species with protective characteristics. The association between unfavorable pregnancy outcomes and a dysfunctional microbiome has been explained through two distinct processes. According to the first hypothesis, periodontal pathogens originating in the gingival biofilm might spread from the infected salivary microbiome over the placenta microbiome, invade the intra-amniotic fluid and fetal circulation, and then have a direct impact on the fetoplacental unit, leading to bacteremia (Hajishengallis, 2015). Based on the second hypothesis, inflammatory mediators and endotoxins that generated by the sub-gingival inflammation and derived from dental plaque of periodontitis may spread throughout the body and reach the fetoplacental unit (Stout et al., 2013; Aagaard et al., 2014). Despite belonging to the same species, some subgroups of the salivary microbiome may influence pregnancy outcomes in both favorable and adverse manners. Following this line of argumentation, the salivary microbiome composition or their dysbiosis are more significant than the existence of particular bacteria.

Notably, microbial alteration that take place throughout pregnancy may be expected results of a healthy pregnancy. Those pregnancy-related vulnerabilities to dental problem like periodontitis can be explained by three factors. Because of hormone-driven gingival hyper-reactivity to the salivary microbiome in the oral biofilm including sub-gingival biofilm, these conditions are prevalent in pregnant women. For insight at the relationship between the salivary microbiome compositions and PTB, further studies with pathway analysis are warranted.

Our study confirmed that salivary microbiome composition could provide potential biomarkers for predicting pregnancy complications including PTB risks using random forest-based classification models, despite a limited number of study participants and a tiny validation sample size. Another limitation of our study was 16S rRNA sequencing. In other words, unlike the shotgun sequencing, 16S rRNA gene sequencing only focused on bacteria, not viruses nor fungi. We did not delve into other variables like nutrition status and socioeconomic statuses of study participants that might affect the salivary microbiome composition.

Notwithstanding these limitations, this prospective examination showed the promise of the random forest-based PTB prediction models based on mouthwash-derived salivary microbiome composition. Before applying the methods developed in this study in a clinical context, more multi-center and extensive research is warranted to validate our findings.

3 Random forest prediction model for periodontitis statuses based on the salivary microbiomes

This section includes the published contents:

3.1 Introduction

Saliva microbial dysbiosis brought on by the accumulation of plaque results in periodontitis, a chronic inflammatory disease of the tissue that surrounds the tooth (Kinane, Stathopoulou, & Papapanou, 2017). Loss of periodontal attachment is a consequence of periodontitis, which may lead to irreversible bone loss and, eventually, permanent tooth loss if left untreated. A new classification criterion of periodontal diseases was created in 2018, about 20 years after the 1999 statements of the previous one (Papapanou et al., 2018). Even with this evolution, radiographic and clinical markers of periodontitis progression remain the primary methods for diagnosing periodontitis (Papapanou et al., 2018). Such tools, nevertheless, frequently demonstrate the prior damage from periodontitis rather than its present condition. Certain individuals have a higher risk of periodontitis, a higher chance of developing severe generalized periodontitis, and a worse response to common salivary bacteria control techniques utilized to prevent and treat periodontitis. As a result, the 2017 framework for diagnosing periodontitis additionally allows for the potential development of biomarkers to enhance diagnosis and treatment of periodontitis (Tonetti, Greenwell, & Kornman, 2018). Instead of only depending on the progression of periodontitis, a new etiological indication based on the current state must be introduced in order to enable appropriate intervention through early detection of periodontitis. Thus, the current clinical diagnostic techniques that rely on periodontal probing can be uncomfortable for patients with periodontitis (Canakci & Canakci, 2007).

Due to the development of salivaomics, in this manner, the examination of saliva has emerged as a significant alternative to the conventional ways of identifying periodontitis (Altıngöz et al., 2021; Melguizo-Rodríguez, Costela-Ruiz, Manzano-Moreno, Ruiz, & Illescas-Montes, 2020). Given that saliva sampling is non-invasive, painless, and accessible to non-specialists, it may be a valuable instrument for diagnosing periodontitis (Zhang et al., 2016). Furthermore, much research has suggested that periodontitis could be a trigger in the development and exacerbation of metabolic syndrome (Morita et al., 2010; Nesbitt et al., 2010). Consequently, alteration in these levels of salivary microbiome markers may serve as high effective diagnostic, prognostic, and therapeutic indicators for periodontitis and other systemic diseases (Miller, Ding, Dawson III, & Ebersole, 2021; Čižmárová et al., 2022). The pathogenesis of periodontitis typically comprises qualitative as well as quantitative alterations in the salivary microbial community, despite that it is a complex disease impacted by a number of contributing factors including age, smoking status, stress, and nourishment (Abusleme, Hoare, Hong, & Diaz, 2021; Lafaurie et al., 2022). Depending on the severity of periodontitis, the salivary microbial community's diversity and characteristics vary (Abusleme et al., 2021), indicating that a new etiological diagnostic standards might be microbial community profiling based on clinical diagnostic criteria. As a consequence, salivary microbiome compositions have been characterized in numerous research in connection with periodontitis. High-throughput sequencing, including 16S rRNA gene sequencing, has recently used in multiple studies to identify variations in the bacterial composition of sub-gingival plaque collections from periodontal healthy individuals and patients with periodontitis (Altabtbaei et al., 2021; Iniesta et al., 2023; Nemoto et al., 2021). This realization has rendered clear that alterations in the salivary microbial community–especially, shifts to dysbiosis–are significant contributors to the pathogenesis and development of periodontitis (Lamont, Koo, & Hajishengallis, 2018). Yet most of these research either focused only on the microbiome alterations in sub-gingival plaque collection, comprised a limited number of periodontitis study participants, or did not account for the impact of multiple severities of periodontitis.

For the objective of diagnosing periodontitis, previous research has developed machine learning-based prediction models based on oral microbiome compositions, such as the sub-gingival microbial dysbiosis index (T. Chen, Marsh, & Al-Hebshi, 2022; Chew, Tan, Chen, Al-Hebshi, & Goh, 2024), which have demonstrated good diagnostic evaluation and could be applied to individual saliva collection. Despite offering valuable details, these indicators are frequently restricted by their limited emphasis on classifying the multiple severities of periodontitis. Furthermore, many of these machine learning models currently in practice are trained solely upon the existence of periodontitis rather than on the multiple severities of periodontitis.

Recently, we employed multiplex quantitative-PCR and machine learning-based classification model to predict the severity of periodontitis based on the amount of nine pathogens of periodontitis from saliva collections (E.-H. Kim et al., 2020). On the other hand, the fact that we focused merely at nine pathogens for periodontitis and neglected the variety bacterial species associated to the various severities of periodontitis constrained the breadth of our investigation. By developing a machine learning model that could classify multiple severities of periodontitis based on the salivary microbiome composition, this study aims to fill these knowledge gaps and produce more accurate and therapeutically useful guidance to evaluate progression of periodontitis. Hence, in order to examine the salivary microbiome composition of both healthy controls and patients with periodontitis in multiple stages, we applied 16S rRNA gene sequencing. Furthermore, employing the 2018 classification criteria, we sought to find biomarkers (species) for the precise prediction of periodontitis severities (Papapanou et al., 2018; Chapple et al., 2018).

3.2 Materials and methods

3.2.1 Study participants enrollment

Between 2018-08 and 2019-03, 250 study participants–100 healthy controls, 50 patients with stage I periodontitis, 50 patients with stage II periodontitis, and 50 patients with stage III periodontitis–visited visited the Department of Periodontics at Pusan National University Dental Hospital. The Institutional Review Board of the Pusan National University Dental Hospital accepted this study protocol and design (IRB No. PNUDH-2016-019). Every study participants provided their written informed authorization after being fully informed about this study's objectives and methodologies. Exclusion criteria for the study participants are followings:

1. People who, throughout the previous six months, underwent periodontal therapy, including root planing and scaling.
2. People who struggle with systemic conditions that may affect periodontitis developments, such as diabetes.
3. People who, throughout the previous three months, were prescribed anti-inflammatory medications or antibiotics.
4. Women who were pregnant or breastfeeding.
5. People who have persistent mucosal lesions, e.g. pemphigus or pemphigoid, or acute infection, e.g. herpertic gingivostomatitis.
6. Patient with grade C periodontitis or localized periodontitis ($<$ 30% of teeth involved).

3.2.2 Periodontal clinical parameter diagnosis

A skilled periodontist conducted each clinical procedure. Six sites per tooth were used to quantify gingival recession and probing depth: mesiobuccal, midbuccal, distobuccal, mesiolingual, midlingual, and distolingual (Huang et al., 2007). A periodontal probe (Hu-Friedy, IL, USA) was placed parallel to the major axis of the tooth at each tooth location in order to gather measurements. The cementoenamel junction of the tooth was analyzed to determine the clinical attachment level, and the deepest point of probing was taken to determine the periodontal pocket depth from the marginal gingival level of the tooth. Plaque index was measured by probing four surfaces per tooth: mesial, distal, buccal, and palatal or lingual. Plaque index was scored by the following criteria:

0. No plaque present.
1. A thin layer of plaque that adheres to the surrounding tissue of the tooth and free gingival margin. Only through the use of a periodontal probe on the tooth surface can the plaque be existed.
2. Significant development of soft deposits that are visible within the gingival pocket, which is a region between the tooth and gingival margin.

3. Considerable amount of soft matter on the tooth, the gingival margin, and the gingival pocket.

The arithmetic average of the plaque indices collected from every tooth was determined to calculate plaque index of each study participant. By probing four surfaces per tooth, mesial, distal, buccal, and palatal or lingual, to assess gingival bleeding, the gingival index was scored by the following criteria:

0. Normal gingiva: without inflammation nor discoloration.
1. Mild inflammation: minimal edema and slight color changes, but no bleeding on probing.
2. Moderate inflammation: edema, glazing, redness, and bleeding on probing.
3. Severe inflammation: significant edema, ulceration, redness, and spontaneous bleeding.

The arithmetic average of the gingival indices collected from every tooth was determined to calculate gingival index of each study participant. The relevant data was not displayed, despite that furcation involvement and bleeding on probing were thoroughly utilized into account during the diagnosis process.

Periodontitis was diagnosed in respect to the 2018 classification criteria (Papapanou et al., 2018; Chapple et al., 2018). An experienced periodontist diagnosed the periodontitis severity by considering complexity, depending on clinical examinations including radiographic images and periodontal probing. Periodontitis is categorized into healthy, stage I, stage II, and stage III with the following criteria:

- Healthy:
  1. Bleeding sites $< 10\%$
  2. Probing depth: ≤ 3 mm
- Stage I:
  1. No tooth loss because of periodontitis.
  2. Inter-dental clinical attachment level at the site of the greatest loss: 1-2 mm
  3. Radiographic bone loss: $< 15\%$
- Stage II:
  1. No tooth loss because of periodontitis.
  2. Inter-dental clinical attachment level at the site of the greatest loss: 3-4 mm
  3. Radiographic bone loss: 15-33%
- Stage III:
  1. Teeth loss because of periodontitis: $\leq$ teeth
  2. Inter-dental clinical attachment level at the site of the greatest loss: ≥ 5 mm
  3. Radiographic bone loss: $> 33\%$

3.2.3 Saliva sampling and DNA extraction procedure

All study participants received instructions to avoid eating, drinking, brushing, and using mouthwash for at least an hour prior to the saliva sample collection process. These collections were conducted between 09:00 and 11:00. Mouth rinse was collected by rinsing the mouth for 30 seconds with 12 mL of a solution (E-zen Gargle, JN Pharm, Korea). All saliva samples were tagged with anonymous ID and stored at -4 °C.

Bacteria DNA was extracted from saliva samples using an Exgene™Clinic SV DNA extraction kit (GeneAll, Seoul, Korea), and quality and quantity of bacterial DNA was measured using a NanoDrop spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA). Hyper-variable regions (V3-V4) of the 16S rRNA gene were amplified using the following primer:

- Forward: `5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3'`
- Reverse: `5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'`

The standard protocols of the Illumina 16S Metagenomic Sequencing Library Preparation were followed in the preparation of the libraries.The PCR conditions were as follows:

1. Heat activation for 30 seconds at 95 °C.
2. 25 cycles for 30 seconds at 95 °C.
3. 30 seconds at 55 °C.
4. 30 seconds at 72 °C.

NexteraXT Indexed Primer was applied to amplification 10 μL of the purified initial PCR products for the final library creation. The second PCR used the same conditions as the first PCR conditions but with 10 cycles. 16S rRNA gene sequencing was performed via 2$\times$300 bp paired-end sequencing at Macrogen Inc. (Macrogen, Seoul, Korea) using Illumina MiSeq platform (Illumina, San Diego, CA, USA).

3.2.4 Bioinformatics analysis

We computed alpha-diversity and beta-diversity indices to quantify the divergence of phylogenetic information. Following alpha-diversity indices were calculated using the scikit-bio Python package (version 0.5.5) (Rideout et al., 2018), and these alpha-diversity indices were compared using the MWU test:

- Abundance-based Coverage Estimator (ACE) (Chao & Lee, 1992)
- Chao1 (Chao, 1984)
- Fisher (Fisher, Corbet, & Williams, 1943)
- Margalef (Magurran, 2021)
- Observed ASVs (DeSantis et al., 2006)
- Berger-Parker *d* (Berger & Parker, 1970)
- Gini index (Gini, 1912)

- Shannon (Weaver, 1963)
- Simpson (Simpson, 1949)

Aitchison index for a beta-diversity index was calculated using QIIME2 (version 2020.8) (Aitchison, Barceló-Vidal, Martín-Fernández, & Pawlowsky-Glahn, 2000; Bolyen et al., 2019). We employed the t-SNE algorithm to illustrate multi-dimensional data from the beta-diversity index computation (Van der Maaten & Hinton, 2008). The beta-diversity index was compared using the PERMANOVA test (Anderson, 2014; Kelly et al., 2015) and MWU test.

DAT between multiple periodontitis stages were identified by ANCOM (Lin & Peddada, 2020). The log-transformed absolute abundances of DAT were analyzed by hierarchical clustering in order to identify sub-groups with similar abundance patterns on periodontitis severities. Additionally, we examined the relative proportions among the 20 DAT in order to reduce the effect of salivary bacteria that differ insignificantly across the multiple severities of periodontitis.

Differentially abundant taxa (DAT) among multiple periodontitis severities were selected from the salivary microbiome compositions by ANCOM (Lin & Peddada, 2020). In contrast to conventional techniques that examine raw abundance counts, ANCOM applies log-ratio between taxa to account for the salivary microbiome composition data. The log-transformed abundances of DAT were subjected to hierarchical clustering to discover subgroups of DAT with similar patterns on periodontitis severities. Furthermore, we examined the relative proportion among the DAT in order to reduce the effects of other salivary bacteria that differ non-non-significantly across the multiple periodontitis severities.

As previously stated (E.-H. Kim et al., 2020), we used stratified k-fold cross-validation ($k = 10$) by severity of periodontitis to achieve consistent and trustworthy classification results (Wong & Yeh, 2019). Additionally, we utilized various features with confusion matrices and their derivations to evaluate the classification outcomes in order to identify which features optimize classification evaluations and decrease sequencing efforts. Using the DAT discovered by ANCOM, we iteratively removed the least significant taxa from the input features (taxa) of the random forest classification models using the backward elimination method.

We investigated external datasets from Spanish individuals (Iniesta et al., 2023) and Portuguese individuals (Relvas et al., 2021) to confirm that our random forest classification was consistent. To ascertain repeatability and dependability, the external datasets were processed using the same pipeline and parameters as those used for our study participants.

3.2.5 Data and code availability

All sequences from the 250 study participants have been added to the Sequence Read Archives (project ID PRJNA976179): `https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA976179`. Docker image that employed throughout this study is available in the DockerHub: `https://hub.docker.com/repository/docker/fumire/periodontitis_16s`. Every code used in this study can be found on GitHub: `https://github.com/CompbioLabUnist/Periodontitis_16S`.

3.3 Results

3.3.1 Summary of clinical information and sequencing data

Among clinical information of the study participants, clinical attachment level, probing depth, plaque index, and gingival index, were significantly increased with periodontitis severity (Kruskal-Wallis test $p < 0.001$), while sex were observed no significant difference (Table 2). Notably, clinical attachment level and probing depth have significant differences among the periodontitis severities (MWU test $p < 0.01$; Figure 15). Additionally, 71461.00±11792.30 and 45909.78±11404.65 reads per sample were obtained before and after filtering low-quality reads and trimming extra-long tails, respectively (Figure 16).

3.3.2 Diversity indices reveal differences among the periodontitis severities

Rarefaction curves showed that the sequencing depth was sufficient (Figure 12). Alpha-diversity indices indicated significant differences between the healthy and the periodontitis stages (MWU test $p < 0.01$;Figure 7a-e); however, there were no significant differences between the periodontitis stages. This emphasizes how essential it is to classify the salivary microbiome compositions and distinguish between the stages of periodontitis using machine learning approaches.

The confidence ellipses of the tSNE-transformed beta-diversity index (Aitchison index) indicated distinct distributions among the periodontitis severities (PERMANOVA $p \leq 0.001$; Figure 7f). Aitchison index demonstrated significant differences every pairwise of the periodontitis severities (PERMANOVA test $p \leq 0.001$; Table 7). Significant differences in the distances between periodontitis severities further demonstrated the uniqueness of each severity of periodontitis (MWU test $p \leq 0.05$; Figure 7g-j).

3.3.3 DAT among multiple periodontitis severities and their correlation

Of the 425 total taxa that identified in the salivary microbiome composition (Figure 13), 20 DAT were identified (Table 5). Three separate subgroups were formed from the participants-level abundances of the DAT using a hierarchical clustering methodology (Figure 8a):

- Group 1
  1. *Treponema* spp.
  2. *Prevotella* sp. *HMT 304*
  3. *Prevotella* sp. *HMT 526*
  4. *Peptostreptococcaceae [XI][G-5] saphenum*
  5. *Treponema* sp. *HMT 260*
  6. *Mycoplasma faucium*
  7. *Peptostreptococcaceae [XI][G-9] brachy*
  8. *Lachnospiraceae [G-8] bacterium HMT 500*
  9. *Peptostreptococcaceae [XI][G-6] nodatum*
  10. *Fretibacterium* spp.

- Group 2
    1. *Porphyromonas gingivalis*
    2. *Campylobacter showae*
    3. *Filifactor alocis*
    4. *Treponema putidum*
    5. *Tannerella forsythia*
    6. *Prevotella intermedia*
    7. *Porphyromonas* sp. *HMT 285*
- Group 3
    1. *Actinomyces* spp.
    2. *Corynebacterium durum*
    3. *Actinomyces graevenitzii*

Ten DAT that were significant enriched in stage II and stage III, but deficient in healthy formed Group 1. Furthermore, in comparison to the healthy, the seven DAT of Group 2 were significantly enriched in each of the stages of periodontitis. On the other hand, three DAT in Group 3 were deficient in stage II and stage III, but significantly enriched in healthy. The relative proportions of the DAT further supported these findings (Figure 8b), suggesting that the DAT is primarily linked to periodontitis rather than other salivary bacteria.

Correlation analysis from the DAT showed that DAT from Group 3 was negatively correlated with Group 1 and Group 2 (Figure 9), and strong correlations were observed the nine pairs of DAT (Figure 14).

3.3.4 Classification of periodontitis severities by random forest models

Based on the proportion of DAT, random forest classifier were trained to classify the periodontitis severities (Table 6). First of all, we conducted multi-label classification for the multiple periodontitis severities, namely healthy, stage I, stage II, and stage III. In this setting, we classified multiple periodontitis severities with the highest BA of 0.779±0.029 (Table 4). AUC ranged between 0.81 and 0.94 (Figure 10b).

Second, since timely detection in dentistry is demanding (Tonetti et al., 2018), we implemented a random forest classification for both healthy and stage I. Remarkably, the random forest classifier had the highest BA at 0.793±0.123 (Table 4). In this setting, this model showed high AUC value for the classifying of stage I from healthy (AUC=0.85; Figure 10d).

Third, based on the findings that the salivary microbiome composition in stage II is more comparable to those in stage III than to other severities (Figure 7f and Figure 7j), we combined stage II and stage III to perform a multi-label classification.

Table 3: **Clinical characteristics of the study subjects**.

Significant differences were assessed using the Kruskal-Wallis test. NA: Not applicable.

Index	**Healthy**	**Stage I**	**Stage II**	**Stage III**	**p-value**
Age (year)	33.83±13.04	43.30±14.28	50.26±11.94	51.08±11.13	6.18E-17
Gender (Male)	44 (44.0%)	22 (44.0%)	25 (50.0%)	25 (50.0%)	NA
Smoking (Never)	83 (83.0%)	36 (72.0%)	34 (68.0%)	29 (58.0%)	NA
Smoking (Ex)	12 (12.0%)	7 (14.0%)	9 (18.0%)	10 (20.0%)	NA
Smoking (Current)	2 (2.0%)	7 (14.0%)	7 (14.0%)	10 (20.0%)	NA
Number of teeth	28.03±2.23	27.36±1.80	26.72±2.89	25.74±4.34	8.07E-05
Attachment level (mm)	2.45±0.29	2.75±0.38	3.64±0.83	4.54±1.14	1.82E-35
Probing depth (mm)	2.42±0.29	2.61±0.40	3.27±0.76	3.95±0.88	6.43E-28
Plaque index	17.66±16.21	35.46±23.75	54.40±23.79	58.30±25.25	3.23E-22
Gingival index	0.09±0.16	0.44±0.46	0.85±0.52	1.06±0.52	2.59E-32

Table 4: **Feature combinations and their evaluations**

Classification performance with the most important taxon, the two most important taxa, and taxa with the best-balanced accuracy. *P.gingivalis* and *Act.* are *Porphyromonas gingivalis* and *Actinomyces* spp., respectively.

Classification	Features	ACC	AUC	BA	F1	PRE	SEN	SPE
Healthy vs. Stage I vs. Stage II vs. Stage III	*P.gingivalis*	0.758±0.051	0.716±0.177	0.677±0.068	0.839±0.034	0.839±0.034	0.839±0.034	0.516±0.102
	P.gingivalis+*Act.*	0.792±0.043	0.822±0.105	0.723±0.057	0.861±0.029	0.861±0.029	0.861±0.029	0.584±0.086
	Top 5 taxa	0.834±0.022	0.870±0.079	0.779±0.029	0.889±0.015	0.889±0.015	0.889±0.015	0.668±0.033
Healthy vs. Stage I	*Act.*	0.687±0.116	0.725±0.145	0.647±0.159	0.762±0.092	0.760±0.128	0.781±0.116	0.513±0.224
	Act.+*P.gingivalis*	0.733±0.119	0.831±0.081	0.713±0.122	0.797±0.097	0.800±0.126	0.798±0.082	0.627±0.191
	Top 9 taxa	0.800±0.103	0.852±0.103	0.793±0.123	0.849±0.080	0.850±0.112	0.857±0.090	0.730±0.193
Healthy vs. Stage I vs. Stages II/III	*P.gingivalis*	0.776±0.042	0.736±0.196	0.748±0.047	0.832±0.031	0.832±0.031	0.832±0.031	0.664±0.062
	P.gingivalis+*Act.*	0.843±0.035	0.876±0.109	0.823±0.039	0.882±0.026	0.882±0.026	0.882±0.026	0.764±0.052
	Top 6 taxa	0.885±0.036	0.914±0.027	0.871±0.038	0.914±0.027	0.914±0.025	0.914±0.025	0.828±0.051
Healthy vs. Stages I/II/III	*P.gingivalis*	0.792±0.114	0.856±0.105	0.819±0.088	0.776±0.089	0.840±0.092	0.756±0.175	0.883±0.054
	P.gingivalis+*Act.*	0.828±0.121	0.926±0.074	0.847±0.116	0.797±0.123	0.800±0.126	0.830±0.191	0.864±0.074
	Top 4 taxa	0.860±0.078	0.953±0.049	0.885±0.066	0.832±0.079	0.840±0.128	0.864±0.157	0.905±0.070

Table 5: **List of DAT among healthy status and periodontitis stages**

No.	Taxonomy	ANCOM W score
1	*Porphyromonas gingivalis*	424
2	*Actinomyces* spp.	424
3	*Filifactor alocis*	421
4	*Prevotella intermedia*	419
5	*Treponema putidum*	418
6	*Tannerella forsythia*	415
7	*Porphyromonas* sp. *HMT 285*	412
8	*Peptostreptococcaceae [XI][G-6] nodatum*	412
9	*Fretibacterium* spp.	411
10	*Mycoplasma faucium*	411
11	*Prevotella* sp. *HMT 304*	411
12	*Lachnospiraceae [G-8] bacterium HMT 500*	409
13	*Treponema spp.*	408
14	*Prevotella* sp. *HMT 526*	401
15	*Peptostreptococcaceae [XI][G-9] brachy*	400
16	*Peptostreptococcaceae [XI][G-5] saphenum*	398
17	*Campylobacter showae*	395
18	*Treponema* sp. *HMT 260*	393
19	*Corynebacterium durum*	393
20	*Actinomyces graevenitzii*	387

Table 6: **Feature the importance of taxa in the classification of different periodontal statuses**

Taxa are ranked in descending order of importance; from most important to least important.

Condition	**Healthy vs. Stage I vs. Stage II vs. Stage III**		**Healthy vs. Stage I**		**Healthy vs. Stage I vs. Stage II/III**		**Healthy vs. Stage I/II/III**	
Rank	Taxa	Importance	Taxa	Importance	Taxa	Importance	Taxa	Importance
1	*Porphyromonas gingivalis*	0.297	*Actinomyces* spp.	0.360	*Porphyromonas gingivalis*	0.426	*Porphyromonas gingivalis*	0.461
2	*Actinomyces* spp.	0.195	*Porphyromonas gingivalis*	0.125	*Actinomyces* spp.	0.244	*Actinomyces* spp.	0.257
3	*Prevotella intermedia*	0.054	*Actinomyces graevenitzii*	0.095	*Actinomyces graevenitzii*	0.049	*Actinomyces graevenitzii*	0.059
4	*Actinomyces graevenitzii*	0.052	*Porphyromonas* sp. *HMT 285*	0.062	Corynebacterium durum	0.046	Corynebacterium durum	0.035
5	*Filifactor alocis*	0.050	*Lachnospiraceae [G-8] bacterium HMT 500*	0.052	*Filifactor alocis*	0.036	*Filifactor alocis*	0.032
6	*Campylobacter showae*	0.042	*Campylobacter showae*	0.050	*Prevotella intermedia*	0.033	*Campylobacter showae*	0.023
7	*Porphyromonas* sp. *HMT 285*	0.040	*Filifactor alocis*	0.039	*Tannerella forsythia*	0.025	*Porphyromonas* sp. *HMT 285*	0.022
8	Corynebacterium durum	0.032	Corynebacterium durum	0.038	*Campylobacter showae*	0.023	*Prevotella intermedia*	0.022
9	*Treponema* spp.	0.032	*Treponema* spp.	0.037	*Porphyromonas* sp. *HMT 285*	0.021	*Treponema* spp.	0.022
10	*Tannerella forsythia*	0.026	*Tannerella forsythia*	0.029	*Treponema* spp.	0.018	*Peptostreptococcaceae [XI][G-9] brachy*	0.015
11	*Treponema putidum*	0.025	*Prevotella intermedia*	0.026	*Peptostreptococcaceae [XI][G-9] brachy*	0.014	*Lachnospiraceae [G-8] bacterium HMT 500*	0.010
12	*Fretibacterium* spp.	0.023	*Fretibacterium* spp.	0.018	*Lachnospiraceae [G-8] bacterium HMT 500*	0.011	*Tannerella forsythia*	0.009
13	*Peptostreptococcaceae [XI][G-9] brachy*	0.021	*Peptostreptococcaceae [XI][G-9] brachy*	0.018	*Peptostreptococcaceae [XI][G-6] nodatum*	0.010	*Fretibacterium* spp.	0.009
14	*Treponema* sp. *HMT 260*	0.019	*Treponema putidum*	0.014	*Treponema putidum*	0.009	*Treponema putidum*	0.006
15	*Prevotella* sp. *HMT 526*	0.018	*Prevotella* sp. *HMT 526*	0.011	*Prevotella* sp. *HMT 526*	0.008	*Peptostreptococcaceae [XI][G-6] nodatum*	0.004
16	*Peptostreptococcaceae [XI][G-6] nodatum*	0.018	*Treponema* sp. *HMT 260*	0.008	*Fretibacterium* spp.	0.008	*Treponema* sp. *HMT 260*	0.004
17	*Prevotella* sp. *HMT 304*	0.017	*Peptostreptococcaceae [XI][G-6] nodatum*	0.008	*Treponema* sp. *HMT 260*	0.005	*Mycoplasma faucium*	0.004
18	*Mycoplasma faucium*	0.014	*Mycoplasma faucium*	0.004	*Prevotella* sp. *HMT 304*	0.005	*Prevotella* sp. *HMT 526*	0.003
19	*Peptostreptococcaceae [XI][G-5] saphenum*	0.014	*Prevotella* sp. *HMT 304*	0.003	*Mycoplasma faucium*	0.005	*Peptostreptococcaceae [XI][G-5] saphenum*	0.002
20	*Lachnospiraceae [G-8] bacterium HMT 500*	0.013	*Peptostreptococcaceae [XI][G-5] saphenum*	0.003	*Peptostreptococcaceae [XI][G-5] saphenum*	0.004	*Prevotella* sp. *HMT 304*	0.001

Table 7: **Beta-diversity pairwise comparisons on the periodontitis statuses**
Statistically significant (p-value) was determined by the PERMANOVA test.

Group 1	Group 2	p-value
Healthy	Stage I	0.001
Healthy	Stage II	0.001
Healthy	Stage III	0.001
Stage I	Stage II	0.001
Stage I	Stage III	0.001
Stage II	Stage III	0.737

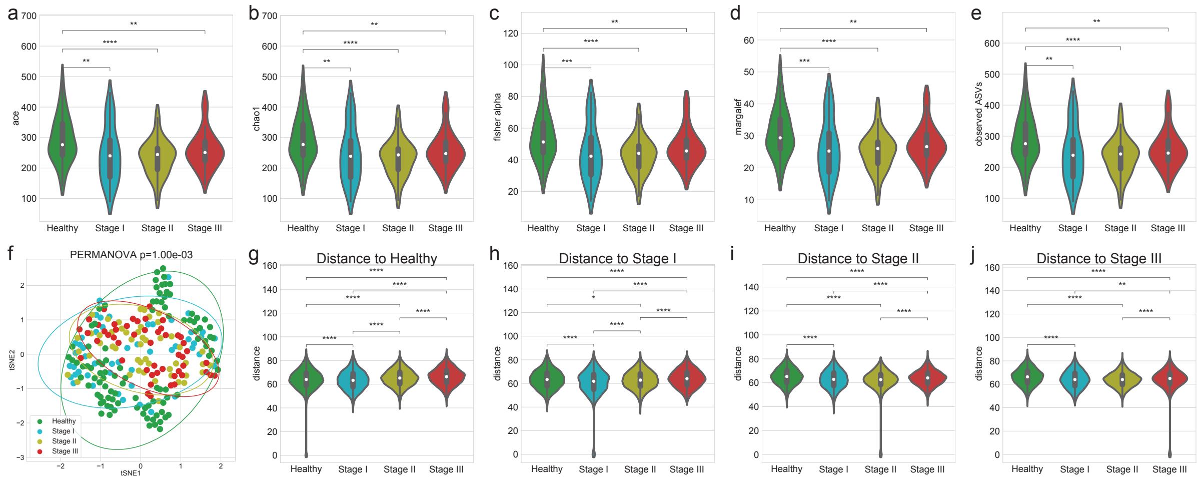

Figure 7: **Diversity indices**.

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 MWU test and the PERMANOVA test: $p \leq 0.01$ (**) and $p \leq 0.0001$ (****).

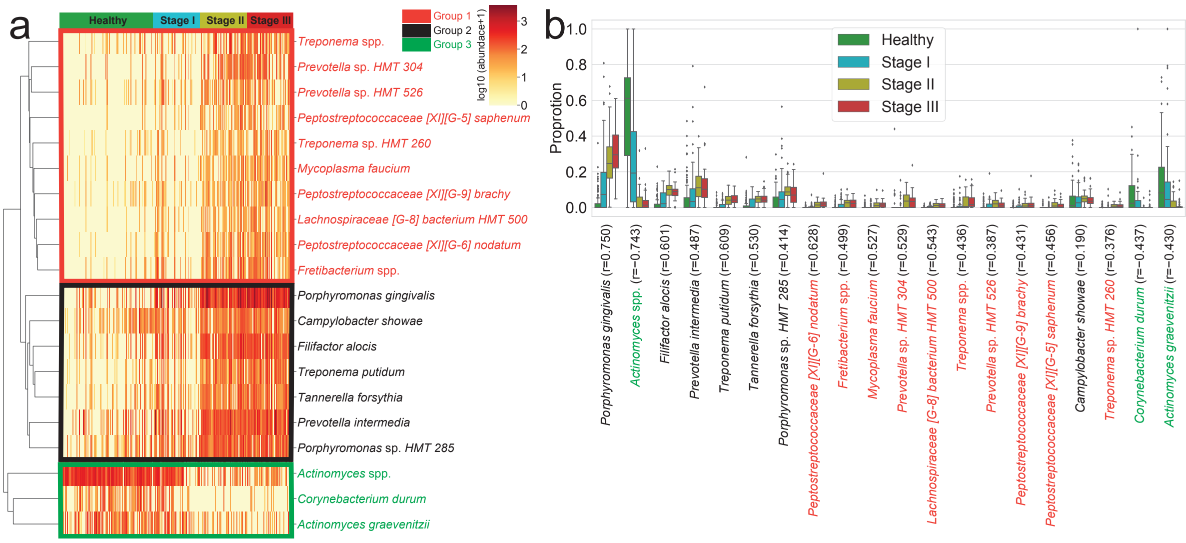

Figure 8: **Differentially abundant taxa (DAT).**

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.

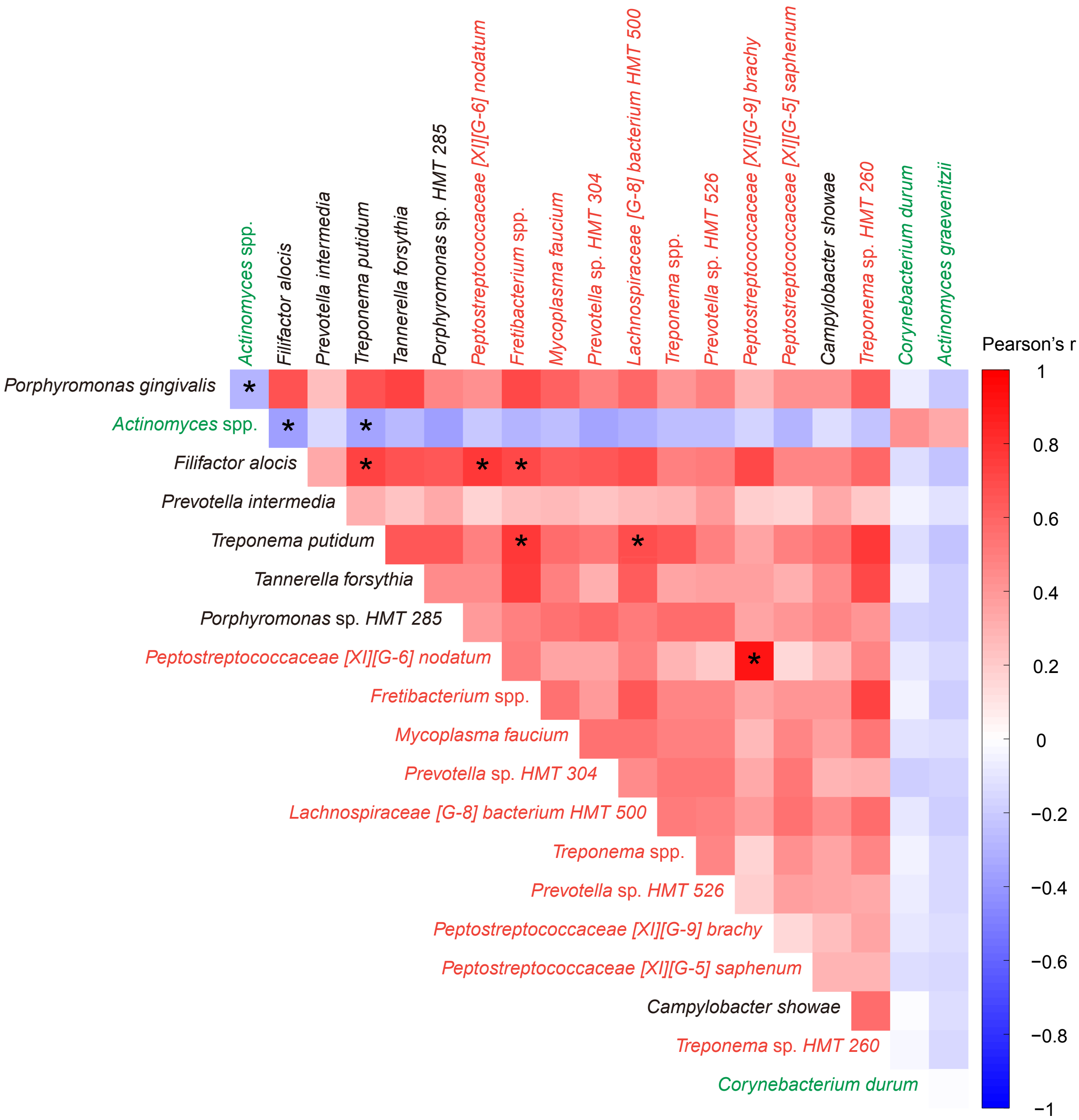

Figure 9: **Correlation heatmap**.

Pearson's correlations between DAT in healthy status and periodontitis stages. Statistical significance was determined by strong correlation, i.e., |coefficient| ≥ 0.5 (*).

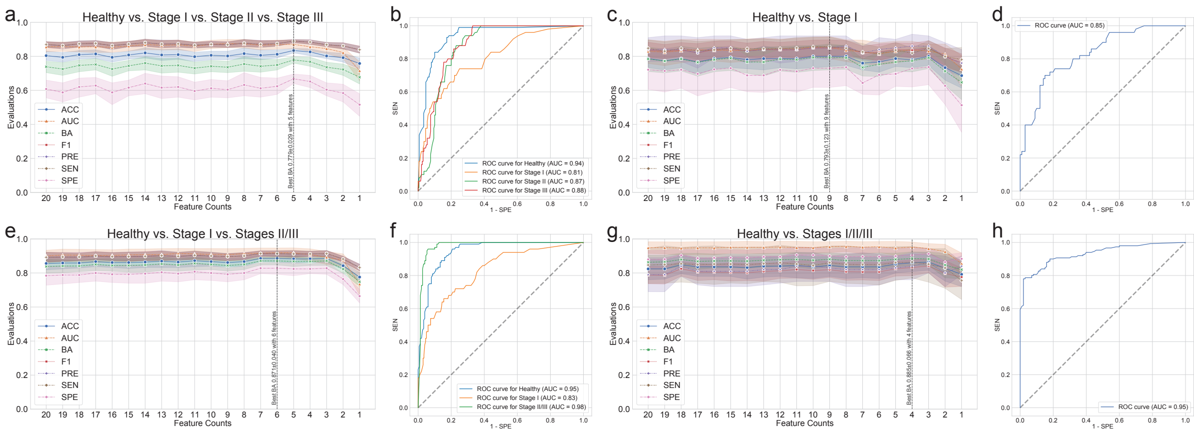

Figure 10: **Random forest classification metrics**.

The classification metrics in the random forest classifications were as follows: ACC, AUC, BA, F1, PRE, SEN, and SPE. **(a)** Classification performance for healthy vs. stage I vs. stage II vs. stage III. **(b)** 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).

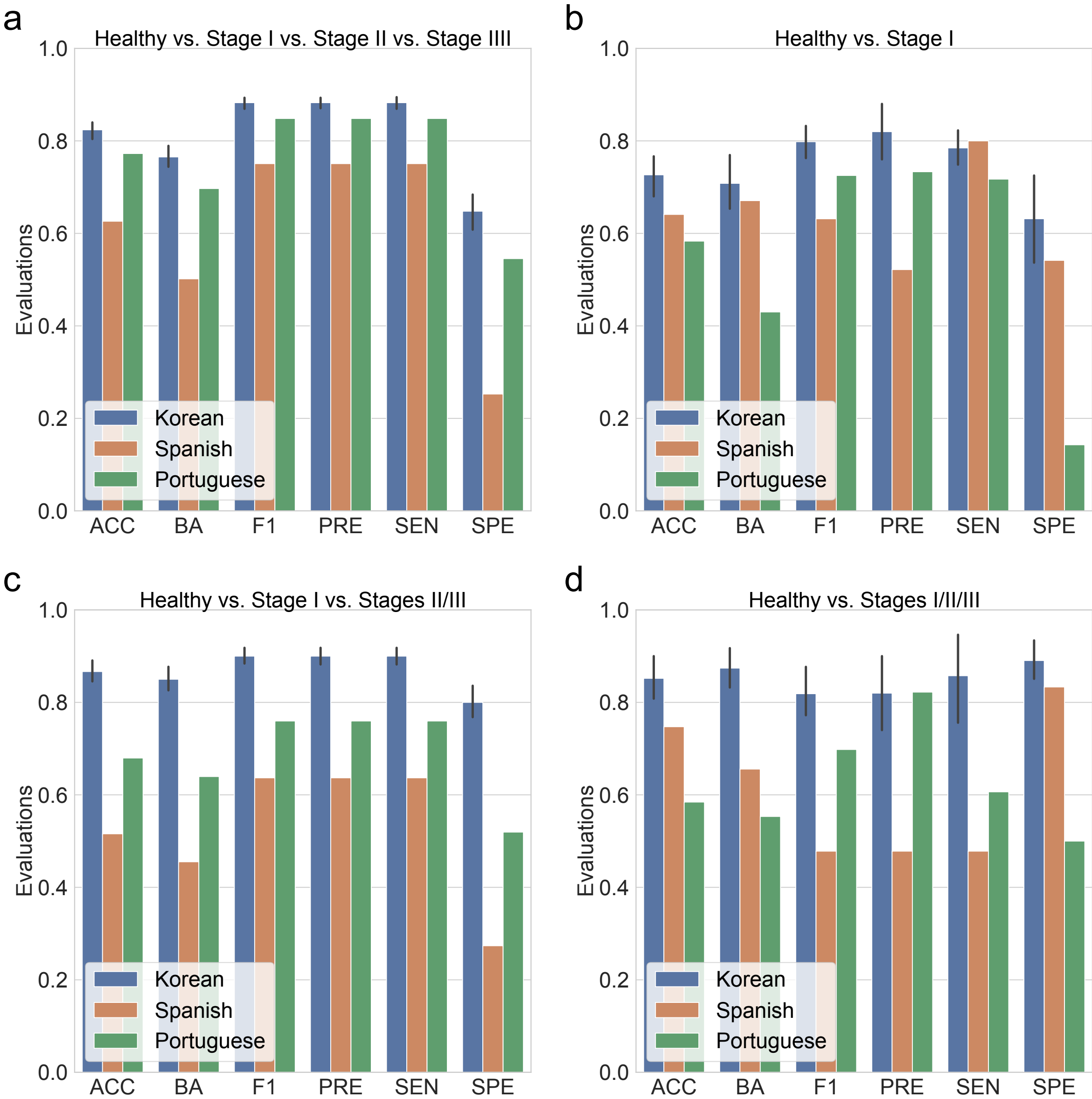

Figure 11: **Random forest classification metrics from external datasets**.

The classification metrics in the random forest classifications were as follows: ACC, AUC, BA, F1, PRE, SEN, and SPE. **(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.

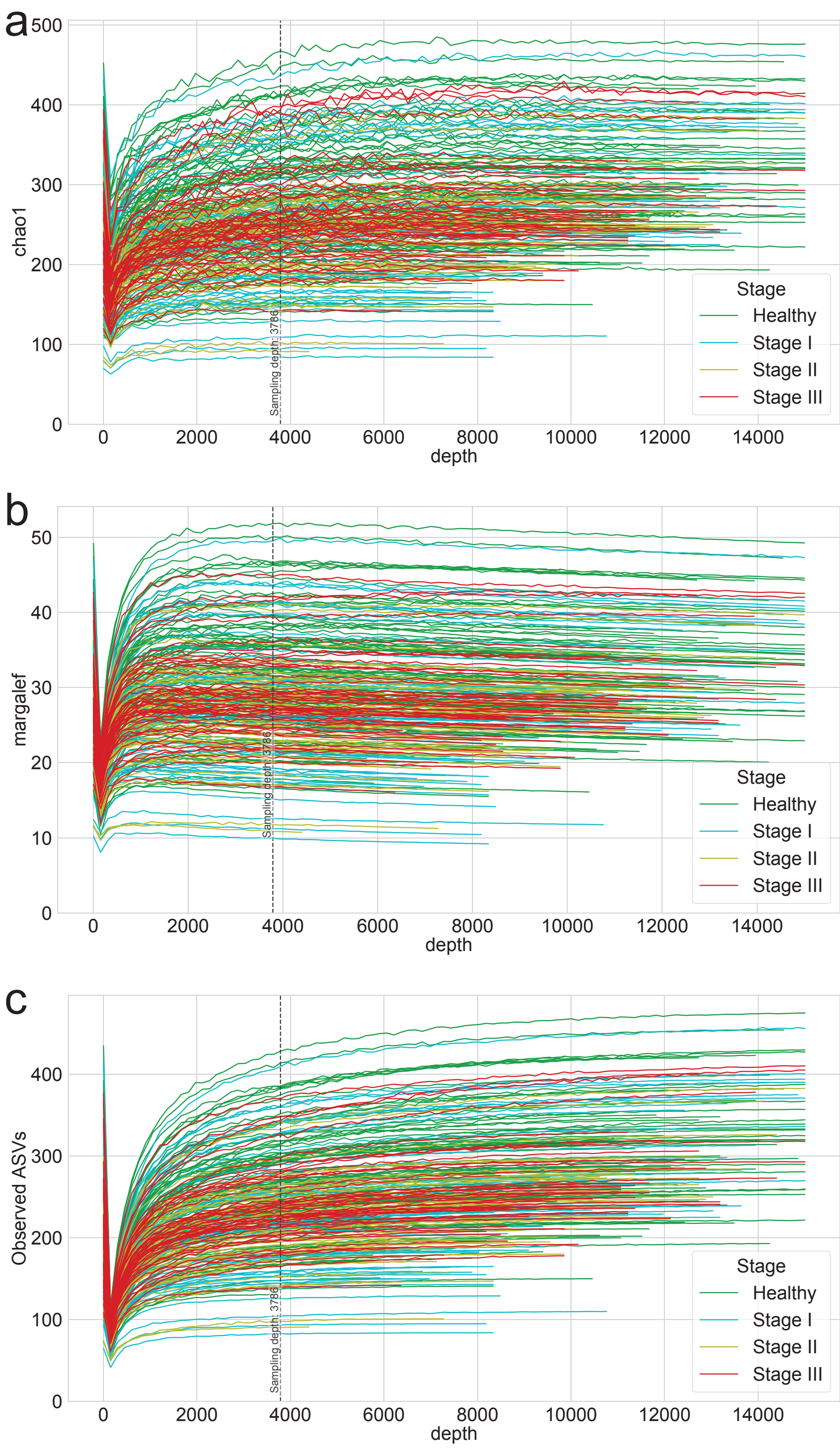

Figure 12: **Rarefaction curves for alpha-diversity indices**.

Rarefaction of **(a)** chao1 **(b)** margalef, and **(c)** observed ASVs were generated to measure species richness and determine the sampling depth of each sample.

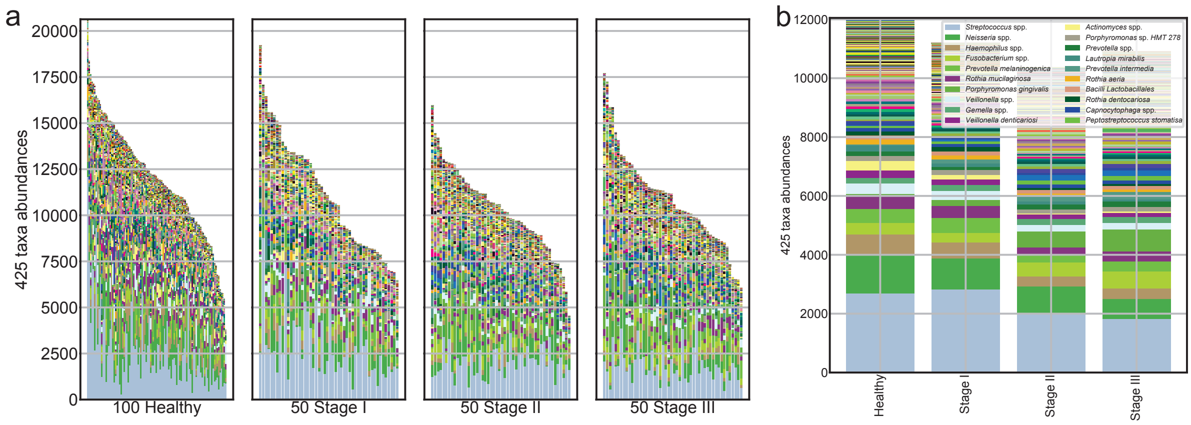

Figure 13: **Salivary microbiome compositions in the different periodontal statuses**.
Stacked bar plot of the absolute abundance of bacterial species for all samples **(a)** and the mean absolute abundance of bacterial species in the healthy, stage I, stage II, and stage III groups **(b)**.

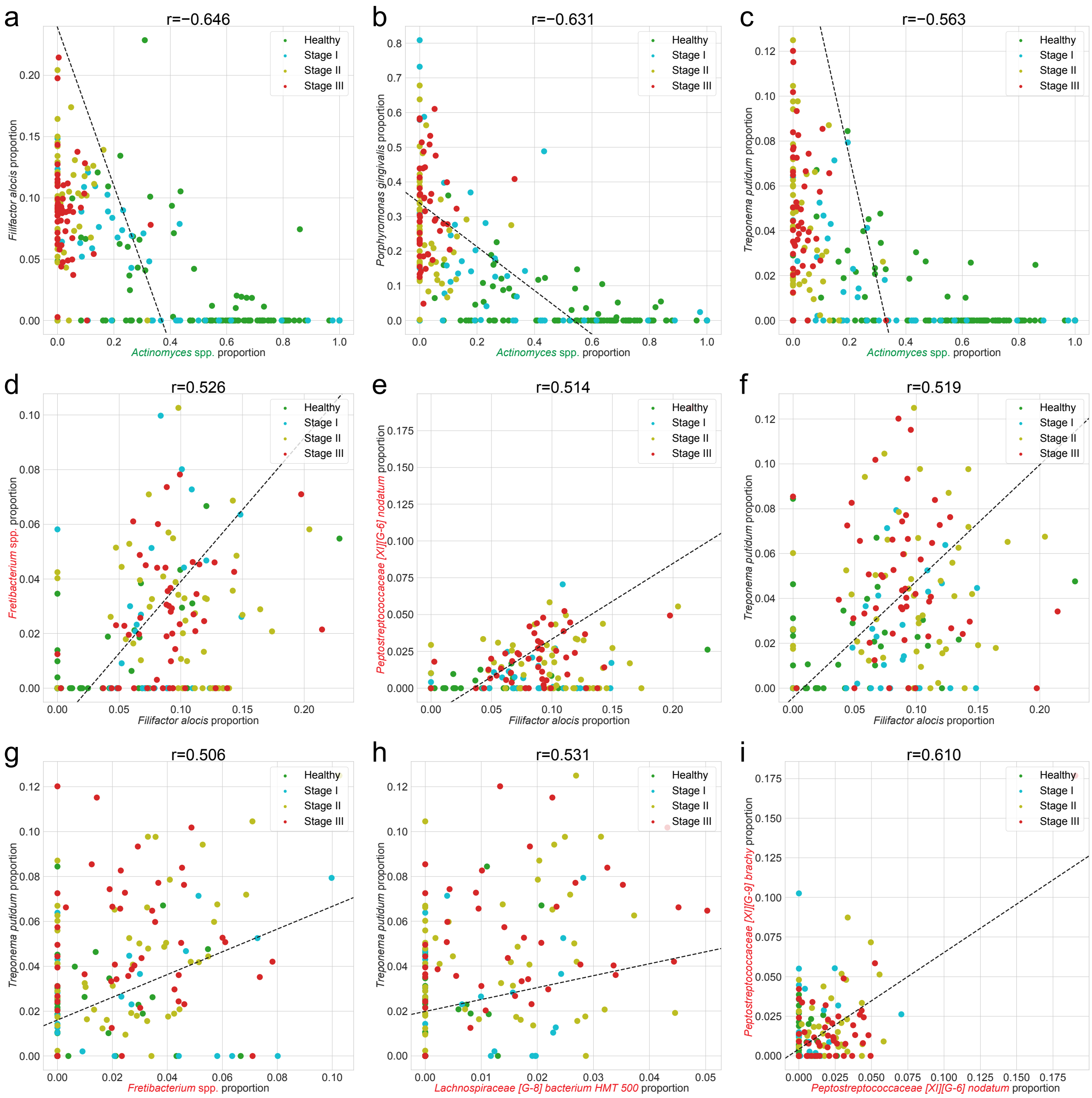

Figure 14: **Correlation plots for differentially abundant taxa**.

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).

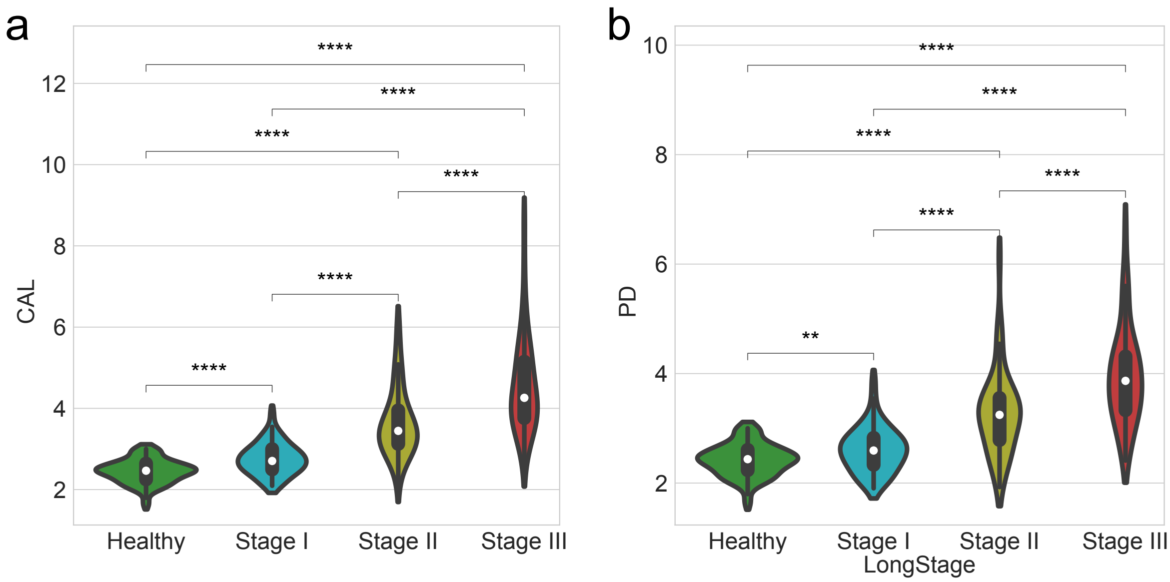

Figure 15: **Clinical measurements by the periodontitis statuses**.
Comparisons of clinical measurement among healthy controls and patients with various periodontitis stages. **(a)** Clinical attachment level **(b)** Probing depth. Statistical significance determined by the MWU test: $p \leq 0.01$ (**) and $p \leq 0.0001$ (****).

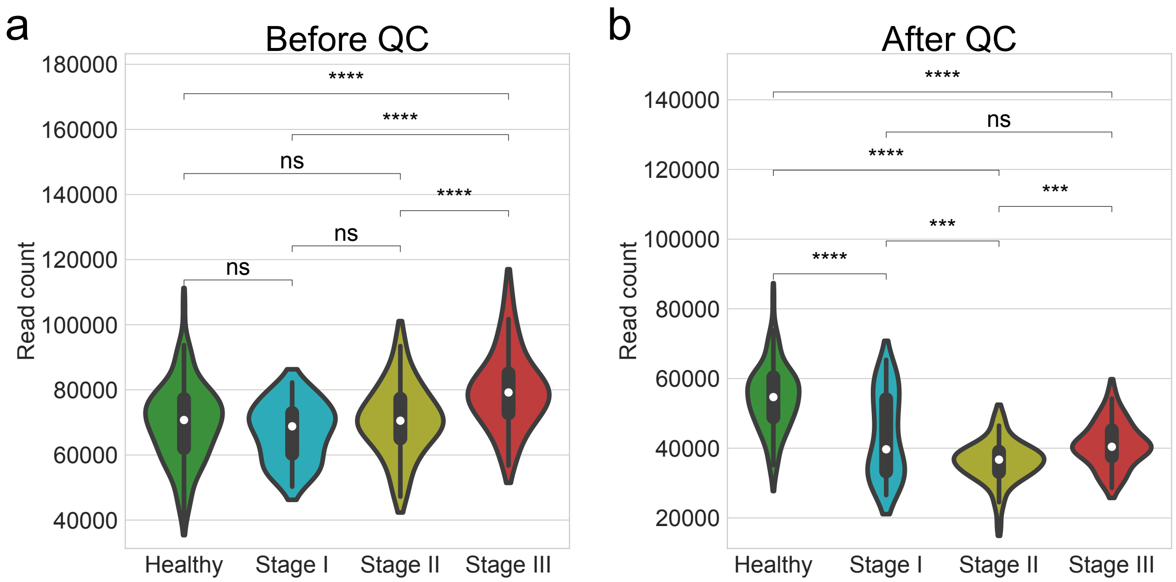

Figure 16: **Number of read counts by the periodontitis statuses**.
Comparisons of the number of read counts among healthy controls and patients with various periodontitis stages. **(a)** Before quality check **(b)** After quality check. Statistical significance determined by the MWU test: $p > 0.05$ (ns), $p \leq 0.001$ (***), and $p \leq 0.0001$ (****).

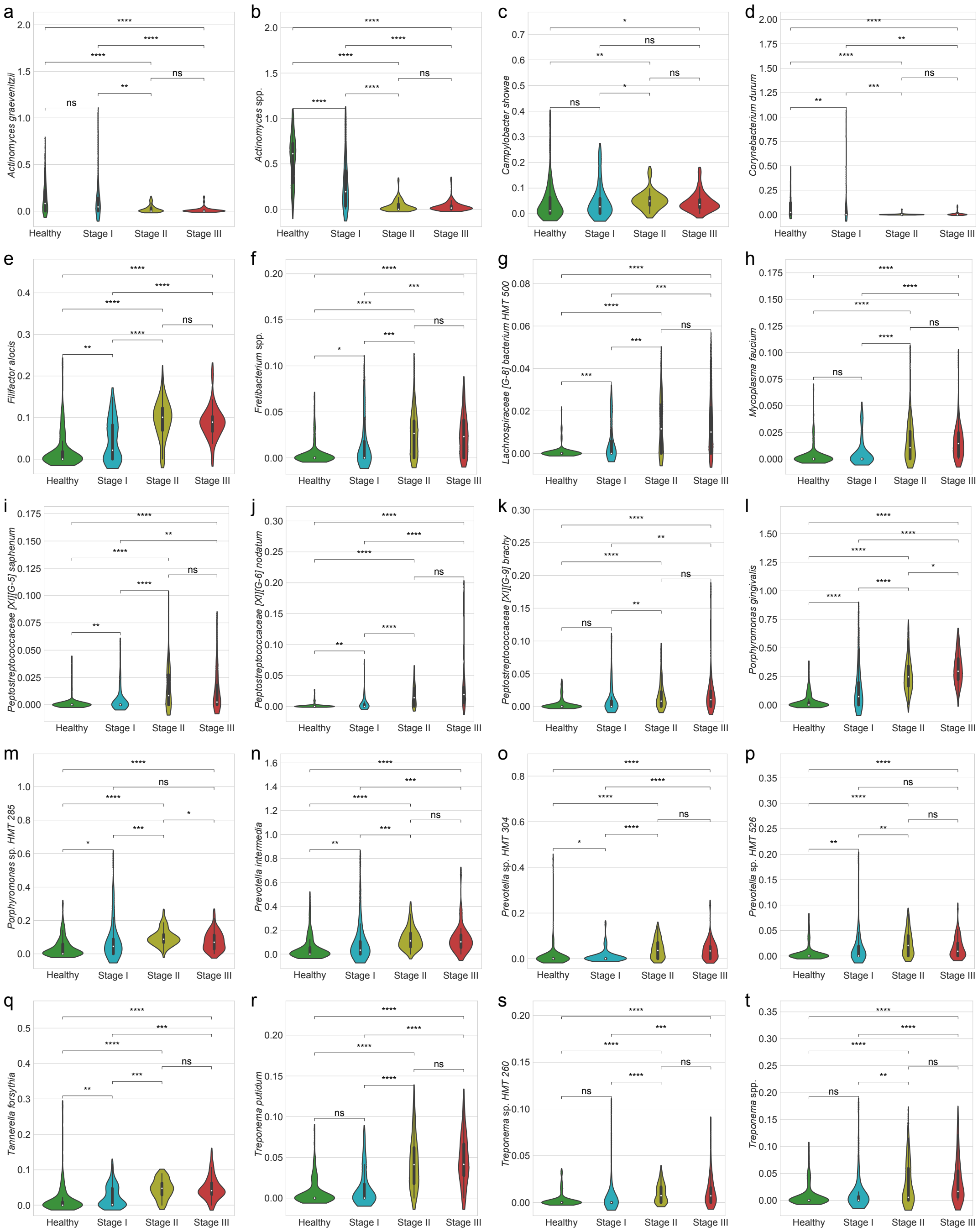

Figure 17: **Proportion of DAT.**

(a) *Actinomyces graevenitzii* **(b)** *Actinomyces* spp. **(c)** *Campylobacter showae* **(d)** *Corynebacterium durum* **(e)** *Filifactor alocis* **(f)** *Fretibacterium* spp. **(g)** *Lachnospiraceae [G-8] bacterium HMT 500* **(h)** *Mycoplasma faucium* **(i)** *Peptostreptococcaceae [XI][G-5] saphenum* **(j)** *Peptostreptococcaceae [XI][G-6] nodatum* **(k)** *Peptostreptococcaceae [XI][G-9] brachy* **(l)** *Porphyromonas gingivalis* **(m)** *Porphyromonas* sp. *HMT 285* **(n)** *Prevotella intermedia* **(o)** *Prevotella* sp. *HMT 304* **(p)** *Prevotella* sp. *HMT 526* **(q)** *Tannerella forsythia* **(r)** *Treponema putidum* **(s)** *Treponema* sp. *HMT 260* **(t)** *Treponema* spp. Statistical significance determined by the MWU test: $p > 0.05$ (ns), $p \leq 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***), and $p \leq 0.0001$ (****).

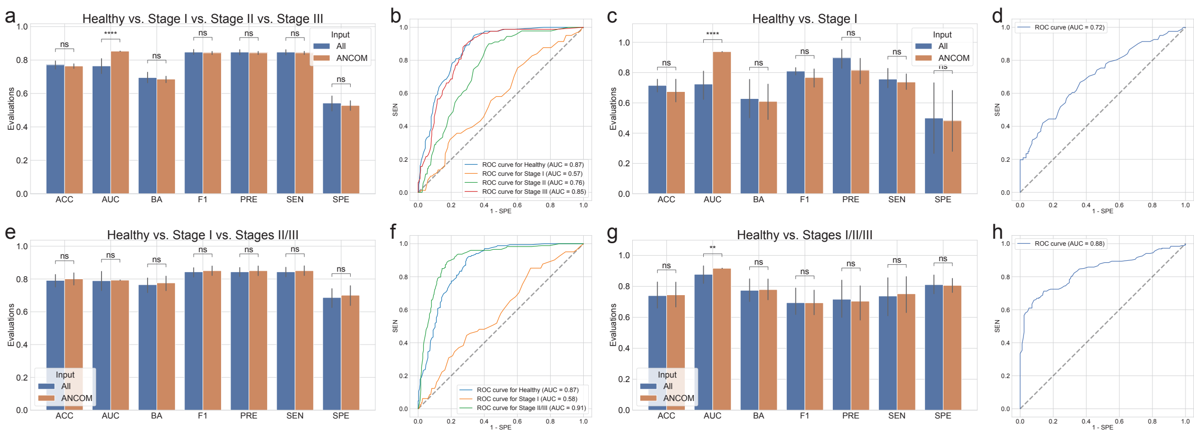

Figure 18: **Random forest classification metrics with the full microbiome compositions and ANCOM-selected DAT compositions**.

The classification metrics in the random forest classifications were as follows: ACC, AUC, BA, F1, PRE, SEN, and SPE. **(a)** Classification performance for healthy vs. stage I vs. stage II vs. stage III. **(b)** 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 (g). Statistical significance determined by the MWU test: $p > 0.05$ (ns), $p \leq 0.01$ (**), and $p \leq 0.0001$ (****).

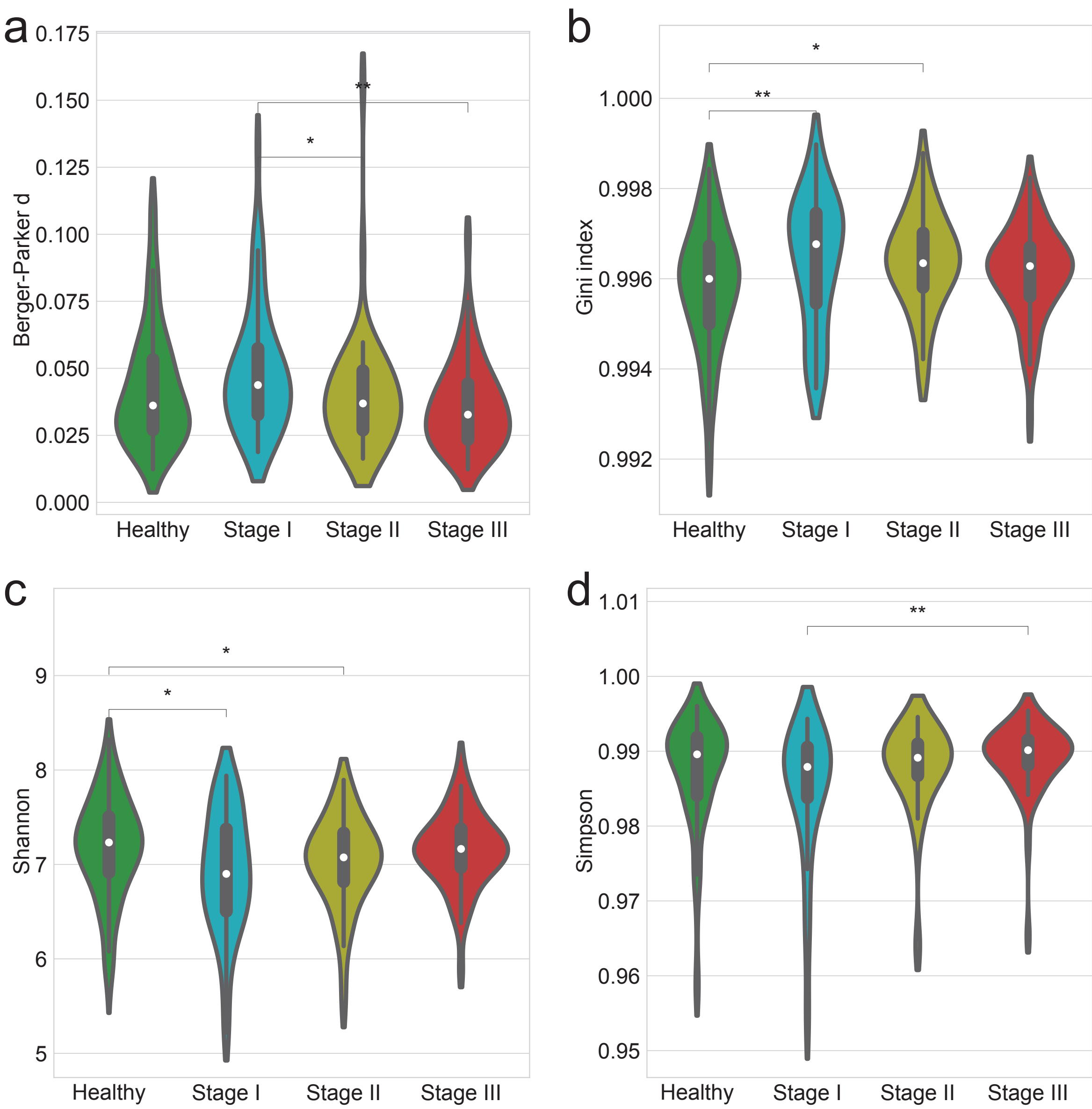

Figure 19: **Alpha-diversity indices account for evenness**.
Alpha-diversity indices **(a-d)** indicate that the heterogeneity between the periodontitis stages as measured by: **(a)** Berger-Parker *d* **(b)** Gini **(c)** Shannon **(d)** Simpson. Statistical significance determined by the MWU test: $p \leq 0.05$ (*) and $p \leq 0.01$ (**)

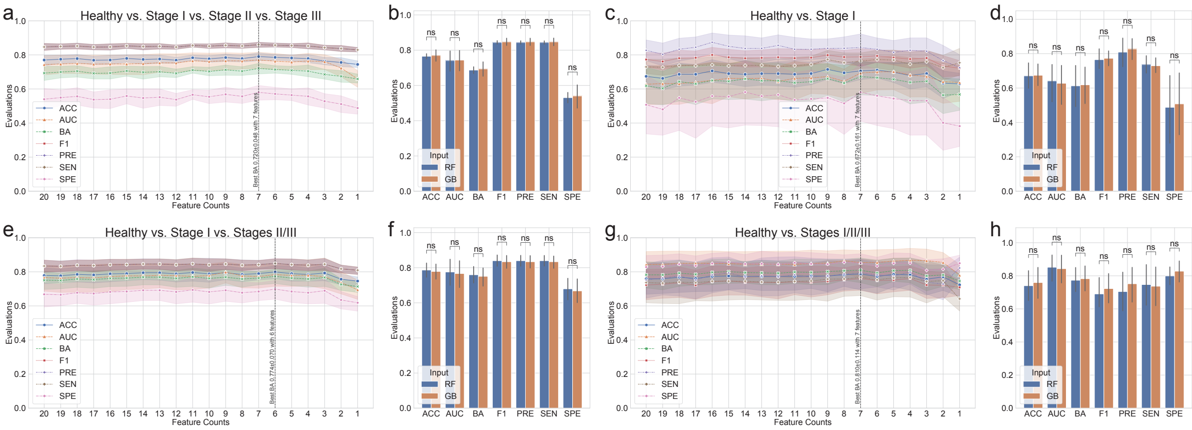

Figure 20: **Gradient Boosting classification metrics**.

The classification metrics in the random forest classifications were as follows: ACC, AUC, BA, F1, PRE, SEN, and SPE. The feature counts mean that the classification model trained on the most important n features as the Table 5. **(a)** Comparison of Random forest (RF) and Gradient boosting (GB) for healthy vs. stage I vs. stage II vs. stage III. **(b)** Comparison of RF and GB for the highest BA of (a). **(c)** Classification performance for healthy vs. stage I. **(d)** Comparison of RF and GB for healthy vs. stage I vs. stages II/III. **(e)** Comparison of RF and GB for the highest BA of (d). **(f)** Comparison of RF and GB for Healthy vs. Stage I vs. Stages II/III. **(g)** Classification performance for healthy vs. stages I/II/III. **(h)** Comparison of RF and GB for Healthy vs. Stages I/II/III.

3.4 Discussion

In order to investigate at potential alterations in the salivary microbiome compositions based on periodontal statuses, including healthy, stage I, stage II, and stage III, we employed 16S rRNA gene sequencing to perform a cross-sectional periodontitis analysis. In this study, the 2018 periodontitis classification served as the basis for the classification of periodontitis severities (Papapanou et al., 2018). There were notable variations in the salivary microbiome composition among the multiple severities of periodontitis (Figure 13). Furthermore, our random forest classification model based on the proportions of DAT in the salivary microbiome compositions across study participants to predict multiple periodontitis statuses with high AUC of 0.870 ± 0.079 (Table 4).

Previous research identified the red complex as the primary pathogens of periodontitis (Listgarten, 1986): *Porphyromonas gingivalis*, *Tannerella forsythia*, and *Treponema denticola*. Other studies, however, have shown that periodontal pathogens communicate with other bacteria in the salivary microbiome networks to generate dental plaque prior to the pathogenesis and development of periodontitis (Lamont & Jenkinson, 2000; Rosan & Lamont, 2000; Yoshimura, Murakami, Nishikawa, Hasegawa, & Kawaminami, 2009).

Using subgingival plaque collections, recent researches have suggested a connection between the periodontitis severity and the salivary microbiome compositions (Altabtbaei et al., 2021; Iniesta et al., 2023; Nemoto et al., 2021). Therefore, we have examined the salivary microbiome compositions of patients with multiple severities of periodontitis and periodontally healthy controls, extending on earlier studies.

According to our findings, the salivary microbiome compositions have 425 taxa (Figure 13). We computed the alpha-diversity indices to determine the variability within each salivary microbiome composition, including ace (Chao & Lee, 1992), chao1 (Chao, 1984), fisher alpha (Fisher et al., 1943), margalef (Magurran, 2021), observed ASVs (DeSantis et al., 2006), Berger-Parker *d* (Berger & Parker, 1970), Gini index (Gini, 1912), Shannon (Weaver, 1963), and Simpson (Simpson, 1949) (Figure 7 and Figure 19). Alpha-diversity indices suggested that the microbial richness of periodontally healthy controls was higher than that of patients with periodontitis (Figure 7a-e and Figure 19). These results are in line with findings with that patients with advanced periodontitis, namely stage II and stage III, have less diversified communities than periodontally healthy controls (Jorth et al., 2014). Recognizing that the periodontitis severity increases the amount of *Porphyromonas gingivalis*, the salivary microbiome compositions from periodontally healthy controls conserved microbial networks dominated by *Streptococcus* spp. (Figure 13). *Porphyromonas gingivalis* is one of the known periodontal pathogen that could cause dysbiosys in the salivary microbiomes, suggesting in the pathophysiology of periodontitis. Despite this finding, earlier research found that subgingival microbiome of patients with periodontitis had a greater alpha-diversity index (observed ASVs) than that of healthy controls (Iniesta et al., 2023), might due to the different sampling sites between saliva and subgingival plaque. On the other hand, another research has addressed significant discrepancies in alpha-diversity indices from subgingival plaque, saliva, and tongue biofilms from healthy controls and periodontitis patients, resulting the highest alpha-diversity index in saliva collections (Belstrøm et al., 2021). Moreover, early-stage periodontitis, namely stage I, did not determine statisticall ysignificant differences in alpha-diversity indices compared to advanced periodontitis, including stage II and stage III (Figure 7a-e). Accordingly, saliva collection of stage I periodontitis may exhibit heterogeneity, indicating a midpoint condition between a healthy state and advanced periodontitis (stage II and stage III). Likewise, gingivitis is often associated with low abundances of the majority of periodontal pathogens, including *Porphyromonas gingivalis*, *Tannerella forsythia*, and *Treponema denticola* (Abusleme et al., 2021). Compared to healthy controls, patients with stage I periodontitis have higher detection rates of *Porphyromonas gingivalis* and *Tannerella forsythia* (Tanner et al., 2006, 2007).

Therefore, we calculated beta-diversity indices to analyze the differences between the study participants. The distances for the multiple stages of periodontitis, including stage I, stage II, and stage III, as well as healthy controls (Figure 4g-j and Table 7), suggesting notable differences among the multiple periodontitis severities. In other words, the composition of the salivary microbiome compositions varies depending on the periodontitis stages, so that supporting the findings from a previous study (Iniesta et al., 2023). Taken together that it is nearly impossible to fully restore the attachment level after it has been lost due to the progression and development of periodontitits, the ability to rapidly screen for periodontitis in its early phases using saliva collections would be highly beneficial for effective disease management and treatment.

Of the total of 425 taxa in the salivary microbiome composition that have been identified (Figure 13), ANCOM was applied to select 20 taxa as the DAT that indicated notable abundance variation among the periodontitis severities (Figure 8 and Table 5). Three sub-groups were formed from the DAT using hierarchical clustering (Figure 8a). Surprisingly, two of the red complex pathogens (Rôças, Siqueira Jr, Santos, Coelho, & de Janeiro, 2001), *Porphyromonas gingivalis* and *Tannerella forsythia*, were classified in Group 2 and were more prevalent in stage II and stage II periodontitis compared to healthy controls. *Campylobacter showae* was additionally placed in Group 2 of the orange complex pathogens (Gambin et al., 2021). Furthermoe, some of the DAT in Group 2 have reported their crucial roles in pathogenesis and development of periodontitis: *Filifactor alocis* (Aruni et al., 2015), *Treponema putidum* (Wyss et al., 2004), *Tannerella forsythia* (Stafford, Roy, Honma, & Sharma, 2012; W. Zhu & Lee, 2016), and *Prevotella intermedia* (Karched, Bhardwaj, Qudeimat, Al-Khabbaz, & Ellepola, 2022). Taken together, this indicates that DAT in Group 2 is essential to periodontitis. The portion of some Group 1 DAT, including *Peptostreptococcaceae[XI][G-5] saphenum*, *Peptostreptococcaceae[XI][G-6] nodatum*, and *Peptostreptococcaceae[XI][G-9] brachy*, in healthy controls and patients with periodontitis significantly differed, according to earlier research (Lafaurie et al., 2022). These outcomes support our research, implying that Group 1 DAT are also essential to the etiology and progression of periodontitis. However, in contrast to patients with periodontitis, Group 3 DAT, namely *Corynebacterium durum* and *Actinomyces graevenitzii*, were enriched in healthy controls, which is consistent with earlier research (Redanz et al., 2021; Nibali et al., 2020).

In our correlation analysis (Figure 9), we have discovered strongly negative correlations (coefficient ≤ -0.5) between DAT of Group 3 and these of Group 1 and Group 2; we have also identified nine DAT pairs with strong correlations (coefficient $\leq -0.5 \vee$ coefficient ≥ 0.5) (Figure 14). Interestingly, there were strongly negative correlations (coefficient ≤ -0.5) between Group 2 DAT and *Actinomyces* spp., taxa which belong to Group 3: *Filifactor alocis* (Figure 14a), *Porphyromonas gingivalis* (Figure 14b), and *Treponema putidum* (Figure 14c). Taken together that pathogens, including *Filifactor alocis* (Aja, Mangar, Fletcher, & Mishra, 2021; Hiranmayi, Sirisha, Rao, & Sudhakar, 2017), *Porphyromonas gingivalis* (Rôças et al., 2001), and *Treponema putidum* (Wyss et al., 2004), become dominant taxa in patients with stage III periodontitis. On the other hand, commensal salivary bacteria, such as *Actinomyces* spp., gradually declined. Additionally, several DAT from Group 1 and Group 2 exhibited strong positive correlations (coefficient ≥ 0.5) (Figure 14d-i). It has been established that all of these DAT from Group 1 and Group 2 are periodontal pathogens: *Filifactor alocis* (Aja et al., 2021; Hiranmayi et al., 2017), *Fretibacterium* spp. (Teles, Wang, Hajishengallis, Hasturk, & Marchesan, 2021), *Lachnospiraceae[G-8] bacterium HMT 500* (Lafaurie et al., 2022), *Peptostreptococcaceae[XI][G-6] nodatum* (Lafaurie et al., 2022; Haffajee, Teles, & Socransky, 2006), *Peptostreptococcaceae[XI][G-9] brachy* (Lafaurie et al., 2022), and Treponema putidum (Wyss et al., 2004). Thus, these fundamental roles of identified periodontal pathogens in the pathophysiology and progression of periodontitis are further supported by these strong positive correlations (coefficient ≥ 0.5), suggesting that advanced periodontitis, *i.e.*, stage III, might arise from the additional DAT from Group 1 and Group 2.

Moreover, to predict peridontitis statuses from salivary microbiome composition, we have constructed machine-learning classification models based on random forest for four classification settings:

1. healthy vs. stage I vs. stage II vs. stage III
2. healthy vs. stage I
3. healthy vs. stage I vs. stages II/III
4. healthy vs. stages I/II/III

Porphyromonas gingivalis and *Actinomyces* spp. were the two most important taxa (feature) in all classification settings. This finding aligns with a recent study that identifies *Actinomyces* spp. as the most prevalent bacteria in both the healthy gingivitis controls, while *Porphyromonas gingivalis* is recognized as the most predominant taxon within the periodontitis subjects, based on analyses of subgingival plaque samples (Nemoto et al., 2021). We have previously developed machine learning models for the classification of periodontitis, with the objective of predicting the severities of chronic periodontitis by analyzing the copy numbers of nine known salivary bacteria species. We classified healthy controls and patients with periodontitis utilizing bacterial combinations in conjunction with a random forest model (E.-H. Kim et al., 2020):

- AUC: 94%
- BA: 84%
- SEN: 95%
- SPE: 72%

Another study established a machine-learning model for the classification of periodontitis, employing 266 species derived from the buccal microbiome (Na et al., 2020):

- AUC: 92%

- BA: 84%
- SEN: 94%
- SPE: 74%

By separating patients with periodontitis from healthy controls using only four DAT, *e.g. Actinomyces graevenitzii*, *Actinomyces* spp., *Corynebacterium durum*, and *Porphyromonas gingivalis*, our machine learning model performed better than previously published models (Figure 10, Table 4, and Table 6):

- AUC: 95.3%±4.9%
- BA: 88.5%±6.6%
- SEN: 86.4%±15.7%
- SPE: 90.5%±7.0%

This result showed that by detecting Group 3 bacteria that were substantially abundant in health controls than patients with periodontitis, our study increased BA by at least 5% and SPE by at least 17%.

Furthermore, we have validated our machine-learning prediction model using openly accessible 16S gene rRNA sequencing data from Portuguese (Iniesta et al., 2023) and Spanish participants (Relvas et al., 2021) in order to ensure the consistency of our random forest classification model (Figure 11). Our classification models employed in this study were primarily developed and assessed on Korean study participants, which may limit their generalizability to other ethnic groups with different salivary microbiome compositions (Premaraj et al., 2020; Renson et al., 2019). Therefore, the evaluations of this periodontitis classification models can be affected by ethnic-specific variances and differences, highlighting the necessity for additional validation and adjustment across a spectrum of ethnic backgrounds.

Regarding the clinical characteristics and potential confounders influencing the analysis of salivary microbiome compositions connected with periodontitis severity, this study had a number of limitations that were pointed out. We did not offer clinical information, such as the percentage of teeth, the percentage of bleeding on probing, nor dental furcation involvement, even though we did gather information on attachment level, probing depth, plaque index, and gingival index; this might have it challenging to present thorough and in-depth data about periodontal health. Moreover, the broad age range may make it tougher to evaluate the relationship between age and periodontitis statuses, providing the necessity for future studies to consider into account more comprehensive clinical characteristics associated with periodontitis. Additionally, potential confounders–*e.g.* body mass index and e-cigarette use–which might have affected dental health and salivary microbiome composition were disregarding consideration in addition to smoking status and systemic diseases. Thus, future research incorporating these components would offer a more thorough knowledge of how lifestyle factors interact and affect the salivary microbiome composition and periodontal health. Throughout, resolving these limitations will advance our understanding in pathogenesis and development of periodontitis, offering significant novel insights on the causal connection between systemic diseases and the salivary microbiome compositions.

4 Colon microbiome

4.1 Introduction

4.2 Materials and methods

4.3 Results

4.4 Discussion

5 Conclusion

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

In the section 2, I show that

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.

Abusleme, L., Hoare, A., Hong, B.-Y., & Diaz, P. I. (2021). Microbial signatures of health, gingivitis, and periodontitis. *Periodontology 2000*, *86*(1), 57–78.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., & Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Mathematical geology*, *32*, 271–275.

Aja, E., Mangar, M., Fletcher, H., & Mishra, A. (2021). Filifactor alocis: recent insights and advances. *Journal of dental research*, *100*(8), 790–797.

Alelyani, S. (2021). Stable bagging feature selection on medical data. *Journal of Big Data*, *8*(1), 11.

Altabtbaei, K., Maney, P., Ganesan, S. M., Dabdoub, S. M., Nagaraja, H. N., & Kumar, P. S. (2021). Anna karenina and the subgingival microbiome associated with periodontitis. *Microbiome*, *9*, 1–15.

Altıngöz, S. M., Kurgan, Ş., Önder, C., Serdar, M. A., Ünlütürk, U., Uyanık, M., … Günhan, M. (2021). Salivary and serum oxidative stress biomarkers and advanced glycation end products in periodontitis patients with or without diabetes: A cross-sectional study. *Journal of periodontology*, *92*(9), 1274–1285.

Alverdy, J., Hyoju, S., Weigerinck, M., & Gilbert, J. (2017). The gut microbiome and the mechanism of surgical infection. *Journal of British Surgery*, *104*(2), e14–e23.

Anderson, M. J. (2014). Permutational multivariate analysis of variance (permanova). *Wiley statsref: statistics reference online*, 1–15.

Aruni, A. W., Mishra, A., Dou, Y., Chioma, O., Hamilton, B. N., & Fletcher, H. M. (2015). Filifactor alocis–a new emerging periodontal pathogen. *Microbes and infection*, *17*(7), 517–530.

Barlow, G. M., Yu, A., & Mathur, R. (2015). Role of the gut microbiome in obesity and diabetes mellitus. *Nutrition in clinical practice*, *30*(6), 787–797.

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.

Belstrøm, D., Constancias, F., Drautz-Moses, D. I., Schuster, S. C., Veleba, M., Mahé, F., & Givskov, M. (2021). Periodontitis associates with species-specific gene expression of the oral microbiota. *npj Biofilms and Microbiomes*, *7*(1), 76.

Berger, W. H., & Parker, F. L. (1970). Diversity of planktonic foraminifera in deep-sea sediments.

Science, *168*(3937), 1345–1347.

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 oncobacterium. *Nature Reviews Microbiology*, *17*(3), 156–166.

Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern recognition*, *36*(6), 1291–1302.

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.

Canakci, V., & Canakci, C. F. (2007). Pain levels in patients during periodontal probing and mechanical non-surgical therapy. *Clinical oral investigations*, *11*, 377–383.

Castaner, O., Goday, A., Park, Y.-M., Lee, S.-H., Magkos, F., Shiow, S.-A. T. E., & Schröder, H. (2018). The gut microbiome profile in obesity: a systematic review. *International journal of endocrinology*, *2018*(1), 4095789.

Champagne, C., McNairn, H., Daneshfar, B., & Shang, J. (2014). A bootstrap method for assessing classification accuracy and confidence for agricultural land use mapping in canada. *International Journal of Applied Earth Observation and Geoinformation*, *29*, 44–52.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, 265–270.

Chao, A., & Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the American statistical Association*, *87*(417), 210–217.

Chapple, I. L., Mealey, B. L., Van Dyke, T. E., Bartold, P. M., Dommisch, H., Eickholz, P., … others (2018). Periodontal health and gingival diseases and conditions on an intact and a reduced periodontium: Consensus report of workgroup 1 of the 2017 world workshop on the classification of periodontal and peri-implant diseases and conditions. *Journal of periodontology*, *89*, S74–S84.

Chen, T., Marsh, P., & Al-Hebshi, N. (2022). Smdi: an index for measuring subgingival microbial dysbiosis. *Journal of dental research*, *101*(3), 331–338.

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*.

Chen, X., D'Souza, R., & Hong, S.-T. (2013). The role of gut microbiota in the gut-brain axis: current challenges and perspectives. *Protein & cell*, *4*, 403–414.

Chew, R. J. J., Tan, K. S., Chen, T., Al-Hebshi, N. N., & Goh, C. E. (2024). Quantifying periodontitis-associated oral dysbiosis in tongue and saliva microbiomes—an integrated data analysis. *Journal of Periodontology*.

Čižmárová, B., Tomečková, V., Hubková, B., Hurajtová, A., Ohlasová, J., & Birková, A. (2022). Salivary redox homeostasis in human health and disease. *International Journal of Molecular Sciences*, *23*(17), 10076.

Cullin, N., Antunes, C. A., Straussman, R., Stein-Thoeringer, C. K., & Elinav, E. (2021). Microbiome and cancer. *Cancer Cell*, *39*(10), 1317–1341.

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.

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.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological conservation*, *61*(1), 1–10.

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.

Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 42–58.

Francescone, R., Hou, V., & Grivennikov, S. I. (2014). Microbiome, inflammation, and cancer. *The Cancer Journal*, *20*(3), 181–189.

Gambin, D. J., Vitali, F. C., De Carli, J. P., Mazzon, R. R., Gomes, B. P., Duque, T. M., & Trentin, M. S. (2021). Prevalence of red and orange microbial complexes in endodontic-periodontal lesions: a systematic review and meta-analysis. *Clinical Oral Investigations*, 1–14.

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current understanding of the human microbiome. *Nature medicine*, *24*(4), 392–400.

Gini, C. (1912). Variabilità e mutabilità (variability and mutability). *Tipografia di Paolo Cuppini, Bologna, Italy*, *156*.

Goldenberg, R. L., Culhane, J. F., Iams, J. D., & Romero, R. (2008). Epidemiology and causes of preterm birth. *The lancet*, *371*(9606), 75–84.

Goodyear, M. D., Krleza-Jeric, K., & Lemmens, T. (2007). *The declaration of helsinki* (Vol. 335) (No. 7621). British Medical Journal Publishing Group.

Haffajee, A., Teles, R., & Socransky, S. (2006). Association of eubacterium nodatum and treponema denticola with human periodontitis lesions. *Oral microbiology and immunology*, *21*(5), 269–282.

Hajishengallis, G. (2015). Periodontitis: from microbial immune subversion to systemic inflammation. *Nature reviews immunology*, *15*(1), 30–44.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*, *29*(2), 147–160.

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.

Hartstra, A. V., Bouter, K. E., Bäckhed, F., & Nieuwdorp, M. (2015). Insights into the role of the microbiome in obesity and type 2 diabetes. *Diabetes care*, *38*(1), 159–165.

Helmink, B. A., Khan, M. W., Hermann, A., Gopalakrishnan, V., & Wargo, J. A. (2019). The microbiome, cancer, and cancer therapy. *Nature medicine*, *25*(3), 377–388.

Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, *54*(2), 427–432.

Hiranmayi, K. V., Sirisha, K., Rao, M. R., & Sudhakar, P. (2017). Novel pathogens in periodontal microbiology. *Journal of Pharmacy and Bioallied Sciences*, *9*(3), 155–163.

Honda, K., & Littman, D. R. (2012). The microbiome in infectious disease and inflammation. *Annual review of immunology*, *30*(1), 759–795.

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., … J. K. (2023). Predicting preterm birth using machine learning techniques in oral microbiome. *Scientific Reports*, *13*(1), 21105.

Huang, R.-Y., Lin, C.-D., Lee, M.-S., Yeh, C.-L., Shen, E.-C., Chiang, C.-Y., … Fu, E. (2007). Mandibular disto-lingual root: a consideration in periodontal therapy. *Journal of periodontology*, *78*(8), 1485–1490.

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.

Iniesta, M., Chamorro, C., Ambrosio, N., Marín, M. J., Sanz, M., & Herrera, D. (2023). Subgingival microbiome in periodontal health, gingivitis and different stages of periodontitis. *Journal of Clinical Periodontology*, *50*(7), 905–920.

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.

Jiang, W., & Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. *Statistics in medicine*, *26*(29), 5320–5334.

John, G. K., & Mullin, G. E. (2016). The gut microbiome and obesity. *Current oncology reports*, *18*, 1–7.

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., ... others (2019). Evaluation of 16s rrna gene sequencing for species and strain-level microbiome analysis. *Nature communications*, *10*(1), 5029.

Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N., & Whiteley, M. (2014). Metatranscriptomics of the human oral microbiome during health and disease. *MBio*, *5*(2), 10–1128.

Karched, M., Bhardwaj, R. G., Qudeimat, M., Al-Khabbaz, A., & Ellepola, A. (2022). Proteomic analysis of the periodontal pathogen prevotella intermedia secretomes in biofilm and planktonic lifestyles. *Scientific Reports*, *12*(1), 5636.

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.

Kelly, B. J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J. D., Collman, R. G., ... Li, H. (2015). Power and sample-size estimation for microbiome studies using pairwise distances and permanova. *Bioinformatics*, *31*(15), 2461–2468.

Kim, C. H. (2018). Immune regulation by microbiome metabolites. *Immunology*, *154*(2), 220–229.

Kim, E.-H., Kim, S., Kim, H.-J., Jeong, H.-o., Lee, J., Jang, J., ... others (2020). Prediction of chronic periodontitis severity using machine learning models based on salivary bacterial copy number. *Frontiers in Cellular and Infection Microbiology*, *10*, 571515.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, *53*(11), 3735–3745.

Kinane, D. F., Stathopoulou, P. G., & Papapanou, P. N. (2017). Periodontal diseases. *Nature reviews Disease primers*, *3*(1), 1–14.

Kindinger, L. M., Bennett, P. R., Lee, Y. S., Marchesi, J. R., Smith, A., Cacciatore, S., ... MacIntyre, D. A. (2017). The interaction between vaginal microbiota, cervical length, and vaginal progesterone treatment for preterm birth risk. *Microbiome*, *5*, 1–14.

Kogut, M. H., Lee, A., & Santin, E. (2020). Microbiome and pathogen interaction with the immune system. *Poultry science*, *99*(4), 1906–1913.

Lafaurie, G. I., Neuta, Y., Ríos, R., Pacheco-Montealegre, M., Pianeta, R., Castillo, D. M., ... others (2022). Differences in the subgingival microbiome according to stage of periodontitis: A comparison of two geographic regions. *PloS one*, *17*(8), e0273523.

Lamont, R. J., & Jenkinson, H. F. (2000). Subgingival colonization by porphyromonas gingivalis. *Oral Microbiology and Immunology: Mini-review*, *15*(6), 341–349.

Lamont, R. J., Koo, H., & Hajishengallis, G. (2018). The oral microbiota: dynamic communities and host interactions. *Nature reviews microbiology*, *16*(12), 745–759.

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.

Lim, J. W., Park, T., Tong, Y. W., & Yu, Z. (2020). The microbiome driving anaerobic digestion and microbial analysis. In *Advances in bioenergy* (Vol. 5, pp. 1–61). Elsevier.

Lin, H., & Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nature communications*, *11*(1), 3514.

Listgarten, M. A. (1986). Pathogenesis of periodontitis. *Journal of clinical periodontology*, *13*(5), 418–425.

Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome medicine*, *8*, 1–11.

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.

Magurran, A. E. (2021). Measuring biological diversity. *Current Biology*, *31*(19), R1174–R1177.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

Mayer, E. A., Tillisch, K., Gupta, A., et al. (2015). Gut/brain axis and the microbiota. *The Journal of clinical investigation*, *125*(3), 926–938.

Melguizo-Rodríguez, L., Costela-Ruiz, V. J., Manzano-Moreno, F. J., Ruiz, C., & Illescas-Montes, R. (2020). Salivary biomarkers and their application in the diagnosis and monitoring of the most common oral pathologies. *International journal of molecular sciences*, *21*(14), 5173.

Miller, C. S., Ding, X., Dawson III, D. R., & Ebersole, J. L. (2021). Salivary biomarkers for discriminating periodontitis in the presence of diabetes. *Journal of clinical periodontology*, *48*(2), 216–225.

Morita, T., Yamazaki, Y., Mita, A., Takada, K., Seto, M., Nishinoue, N., … Maeno, M. (2010). A cohort study on the association between periodontal disease and the development of metabolic syndrome. *Journal of periodontology*, *81*(4), 512–519.

Na, H. S., Kim, S. Y., Han, H., Kim, H.-J., Lee, J.-Y., Lee, J.-H., & Chung, J. (2020). Identification of potential oral microbial biomarkers for the diagnosis of periodontitis. *Journal of clinical medicine*, *9*(5), 1549.

Nemoto, T., Shiba, T., Komatsu, K., Watanabe, T., Shimogishi, M., Shibasaki, M., … others (2021). Discrimination of bacterial community structures among healthy, gingivitis, and periodontitis statuses through integrated metatranscriptomic and network analyses. *Msystems*, *6*(6), e00886–21.

Nesbitt, M. J., Reynolds, M. A., Shiau, H., Choe, K., Simonsick, E. M., & Ferrucci, L. (2010). Association of periodontitis and metabolic syndrome in the baltimore longitudinal study of aging. *Aging clinical and experimental research*, *22*, 238–242.

Nibali, L., Sousa, V., Davrandi, M., Spratt, D., Alyahya, Q., Dopico, J., & Donos, N. (2020). Differences in the periodontal microbiome of successfully treated and persistent aggressive periodontitis. *Journal of Clinical Periodontology*, *47*(8), 980–990.

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.

Papapanou, P. N., Sanz, M., Buduneli, N., Dietrich, T., Feres, M., Fine, D. H., ... others (2018). Periodontitis: Consensus report of workgroup 2 of the 2017 world workshop on the classification of periodontal and peri-implant diseases and conditions. *Journal of periodontology*, *89*, S173–S182.

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.

Peirce, J. M., & Alviña, K. (2019). The role of inflammation and the gut microbiome in depression and anxiety. *Journal of neuroscience research*, *97*(10), 1223–1241.

Premaraj, T. S., Vella, R., Chung, J., Lin, Q., Hunter, P., Underwood, K., ... Zhou, Y. (2020). Ethnic variation of oral microbiota in children. *Scientific reports*, *10*(1), 14788.

Redanz, U., Redanz, S., Treerat, P., Prakasam, S., Lin, L.-J., Merritt, J., & Kreth, J. (2021). Differential response of oral mucosal and gingival cells to corynebacterium durum, streptococcus sanguinis, and porphyromonas gingivalis multispecies biofilms. *Frontiers in cellular and infection microbiology*, *11*, 686479.

Relvas, M., Regueira-Iglesias, A., Balsa-Castro, C., Salazar, F., Pacheco, J., Cabral, C., ... Tomás, I. (2021). Relationship between dental and periodontal health status and the salivary microbiome: bacterial diversity, co-occurrence networks and predictive models. *Scientific reports*, *11*(1), 929.

Renson, A., Jones, H. E., Beghini, F., Segata, N., Zolnik, C. P., Usyk, M., ... others (2019). Sociodemographic variation in the oral microbiome. *Annals of epidemiology*, *35*, 73–80.

Rideout, J. R., Caporaso, G., Bolyen, E., McDonald, D., Baeza, Y. V., Alastuey, J. C., ... Sharma, K. (2018, December). *biocore/scikit-bio: scikit-bio 0.5.5: More compositional methods added.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.2254379` doi: 10.5281/zenodo.2254379

Rôças, I. N., Siqueira Jr, J. F., Santos, K. R., Coelho, A. M., & de Janeiro, R. (2001). "red complex"(bacteroides forsythus, porphyromonas gingivalis, and treponema denticola) in endodontic infections: a molecular approach. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, *91*(4), 468–471.

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.

Rosan, B., & Lamont, R. J. (2000). Dental plaque formation. *Microbes and infection*, *2*(13), 1599–1607.

Schwabe, R. F., & Jobin, C. (2013). The microbiome and cancer. *Nature Reviews Cancer*, *13*(11), 800–812.

Sepich-Poore, G. D., Zitvogel, L., Straussman, R., Hasty, J., Wargo, J. A., & Knight, R. (2021). The microbiome and human cancer. *Science*, *371*(6536), eabc4552.

Sharma, S., & Tripathi, P. (2019). Gut microbiome and type 2 diabetes: where we are and where to go?

The Journal of nutritional biochemistry, *63*, 101–108.

Simpson, E. (1949). Measurement of diversity. *Nature*, *163*.

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.

Stafford, G., Roy, S., Honma, K., & Sharma, A. (2012). Sialic acid, periodontal pathogens and tannerella forsythia: stick around and enjoy the feast! *Molecular Oral Microbiology*, *27*(1), 11–22.

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.

Tanner, A. C., Kent Jr, R., Kanasi, E., Lu, S. C., Paster, B. J., Sonis, S. T., ... Van Dyke, T. E. (2007). Clinical characteristics and microbiota of progressing slight chronic periodontitis in adults. *Journal of clinical periodontology*, *34*(11), 917–930.

Tanner, A. C., Paster, B. J., Lu, S. C., Kanasi, E., Kent Jr, R., Van Dyke, T., & Sonis, S. T. (2006). Subgingival and tongue microbiota during early periodontitis. *Journal of dental research*, *85*(4), 318–323.

Teles, F., Wang, Y., Hajishengallis, G., Hasturk, H., & Marchesan, J. T. (2021). Impact of systemic factors in shaping the periodontal microbiome. *Periodontology 2000*, *85*(1), 126–160.

Thaiss, C. A., Zmora, N., Levy, M., & Elinav, E. (2016). The microbiome and innate immunity. *Nature*, *535*(7610), 65–74.

Tilg, H., Kaser, A., et al. (2011). Gut microbiome, obesity, and metabolic dysfunction. *The Journal of clinical investigation*, *121*(6), 2126–2132.

Tonetti, M. S., Greenwell, H., & Kornman, K. S. (2018). Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. *Journal of periodontology*, *89*, S159–S172.

Tringe, S. G., & Hugenholtz, P. (2008). A renaissance for the pioneering 16s rrna gene. *Current opinion in microbiology*, *11*(5), 442–446.

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.

Ursell, L. K., Metcalf, J. L., Parfrey, L. W., & Knight, R. (2012). Defining the human microbiome. *Nutrition reviews*, *70*(suppl_1), S38–S44.

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.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Weaver, W. (1963). *The mathematical theory of communication*. University of Illinois Press.

Whiteside, S. A., Razvi, H., Dave, S., Reid, G., & Burton, J. P. (2015). The microbiome of the urinary tract—a role beyond infection. *Nature Reviews Urology*, *12*(2), 81–90.

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.

Wyss, C., Moter, A., Choi, B.-K., Dewhirst, F., Xue, Y., Schüpbach, P., ... Guggenheim, B. (2004). Treponema putidum sp. nov., a medium-sized proteolytic spirochaete isolated from lesions of human periodontitis and acute necrotizing ulcerative gingivitis. *International journal of systematic and evolutionary microbiology*, *54*(4), 1117–1122.

Yaman, E., & Subasi, A. (2019). Comparison of bagging and boosting ensemble machine learning methods for automated emg signal classification. *BioMed research international*, *2019*(1), 9152506.

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.

Yoshimura, F., Murakami, Y., Nishikawa, K., Hasegawa, Y., & Kawaminami, S. (2009). Surface components of porphyromonas gingivalis. *Journal of periodontal research*, *44*(1), 1–12.

Zhang, C.-Z., Cheng, X.-Q., Li, J.-Y., Zhang, P., Yi, P., Xu, X., & Zhou, X.-D. (2016). Saliva in the diagnosis of diseases. *International journal of oral science*, *8*(3), 133–137.

Zhu, W., & Lee, S.-W. (2016). Surface interactions between two of the main periodontal pathogens: Porphyromonas gingivalis and tannerella forsythia. *Journal of periodontal & implant science*, *46*(1), 2–9.

Zhu, X., Han, Y., Du, J., Liu, R., Jin, K., & Yi, W. (2017). Microbiota-gut-brain axis and the central nervous system. *Oncotarget*, *8*(32), 53829.

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 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 (이망톡), 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.

(Girlfriend)

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.

May your progress be evident to all