

¹

Doctoral Thesis

²

Microbiota in Human Diseases

³

Jaewoong Lee

⁴

Department of Biomedical Engineering

⁵

Ulsan National Institute of Science and Technology

⁶

2025

⁷

Microbiota in Human Diseases

⁸

Jaewoong Lee

⁹

Department of Biomedical Engineering

¹⁰

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.

A handwritten signature in black ink that reads "Bobby Henderson".

Representative,
Church of the Flying Spaghetti Monster
Bobby Henderson



CHURCH OF THE FLYING SPAGHETTI MONSTER

February 09, 2021

Letter of Good Standing

Dear Sir or Madam:

I am pleased to verify that _____

JAEWOONG LEE

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

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

Any questions can be directed to the undersigned.

A handwritten signature in black ink that reads "Bobby Henderson".

Representative,
Church of the Flying Spaghetti Monster
Bobby Henderson

13

Abstract

14 (Microbiome)

15 (PTB) Section 2 introduces...

16 (Periodontitis) Section 3 describes...

17 (Lung)

18 (Conclusion)

19

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

- 22 • Hong, Y. M., **Lee, Jaewoong**, Cho, D. H., Jeon, J. H., Kang, J., Kim, M. G., ... & Kim, J. K. (2023).
23 Predicting preterm birth using machine learning techniques in oral microbiome. *Scientific Reports*,
24 13(1), 21105.

Contents

26	1	Introduction	2
27	2	Predicting preterm birth using random forest classifier in salivary microbiome	6
28	2.1	Introduction	6
29	2.2	Materials and methods	8
30	2.2.1	Study design and study participants	8
31	2.2.2	Clinical data collection and grouping	8
32	2.2.3	Salivary microbiome sample collection	8
33	2.2.4	16s rRNA gene sequencing	9
34	2.2.5	Bioinformatics analysis	9
35	2.2.6	Data and code availability	9
36	2.3	Results	10
37	2.3.1	Overview of clinical information	10
38	2.3.2	Comparison of salivary microbiomes composition	10
39	2.3.3	Random forest classification to predict PTB risk	10
40	2.4	Discussion	18
41	3	Random forest prediction model for periodontitis statuses based on the salivary microbiomes	20
42	3.1	Introduction	20
43	3.2	Materials and methods	22
44	3.2.1	Study participants enrollment	22
45	3.2.2	Periodontal clinical parameter diagnosis	22
46	3.2.3	Saliva sampling and DNA extraction procedure	24
47	3.2.4	Bioinformatics analysis	24
48	3.2.5	Data and code availability	25
49	3.3	Results	27

50	3.3.1	Summary of clinical information and sequencing data	27
51	3.3.2	Diversity indices reveal differences among the periodontitis severities .	27
52	3.3.3	DAT among multiple periodontitis severities and their correlation . . .	27
53	3.3.4	Classification of periodontitis severities by random forest models . . .	28
54	3.4	Discussion	48
55	4	Lung microbiome	51
56	4.1	Introduction	51
57	4.2	Materials and methods	52
58	4.3	Results	53
59	4.4	Discussion	54
60	5	Conclusion	55
61	References	56
62	Acknowledgments	65

63

List of Figures

64	1	DAT volcano plot	12
65	2	Salivary microbiome compositions over DAT	13
66	3	Random forest-based PTB prediction model	14
67	4	Diversity indices	15
68	5	PROM-related DAT	16
69	6	Validation of random forest-based PTB prediction model	17
70	7	Diversity indices	34
71	8	Differentially abundant taxa (DAT)	35
72	9	Correlation heatmap	36
73	10	Random forest classification metrics	37
74	11	Random forest classification metrics from external datasets	38
75	12	Rarefaction curves for alpha-diversity indices	39
76	13	Salivary microbiome compositions in the different periodontal statuses	40
77	14	Correlation plots for differentially abundant taxa	41
78	15	Clinical measurements by the periodontitis statuses	42
79	16	Number of read counts by the periodontitis statuses	43
80	17	Proportion of DAT	44

81	18	Random forest classification metrics with the full microbiome compositions and ANCOM-selected DAT compositions	45
82			
83	19	Alpha-diversity indices account for evenness	46
84	20	Gradient Boosting classification metrics	47

List of Tables

86	1	Confusion matrix	4
87	2	Standard clinical information of study participants	11
88	3	Clinical characteristics of the study subjects	29
89	4	Feature combinations and their evaluations	30
90	5	List of DAT among the periodontally healthy and periodontitis stages	31
91	6	Feature the importance of taxa in the classification of different periodontal statuses.	32
92	7	Beta-diversity pairwise comparisons on the periodontitis statuses	33

93

List of Abbreviations

- 94 **ACC** Accuracy
95 **ASV** Amplicon sequence variant
96 **AUC** Area-under-curve
97 **BA** Balanced accuracy
98 **C-section** Cesarean section
99 **DAT** Differentially abundant taxa
100 **F1** F1 score
101 **Faith PD** Faith's phylogenetic diversity
102 **FTB** Full-term birth
103 **GA** Gestational age
104 **MWU test** Mann-Whitney U-test
105 **PRE** Precision
106 **PROM** Prelabor rupture of membrane
107 **PTB** Preterm birth
108 **ROC curve** Receiver-operating characteristics curve
109 **rRNA** Ribosomal RNA
110 **SD** Standard deviation
111 **SEN** Sensitivity
112 **SPE** Specificity
113 **t-SNE** t-distributed stochastic neighbor embedding

¹¹⁴ 1 Introduction

¹¹⁵ The microbiome refers to the complex community of microorganisms, including bacteria, viruses, fungi,
¹¹⁶ and other microbes, that inhabit various environment within living organisms (Ursell, Metcalf, Parfrey,
¹¹⁷ & Knight, 2012; Gilbert et al., 2018). In humans, the microbiome plays a crucial role in maintaining
¹¹⁸ health (Lloyd-Price, Abu-Ali, & Huttenhower, 2016), influencing processes such as digestion (Lim, Park,
¹¹⁹ Tong, & Yu, 2020), immune response (Thaiss, Zmora, Levy, & Elinav, 2016; Kogut, Lee, & Santin, 2020;
¹²⁰ C. H. Kim, 2018), and even mental health (Mayer, Tillisch, Gupta, et al., 2015; X. Zhu et al., 2017;
¹²¹ X. Chen, D'Souza, & Hong, 2013). These microbial communities are not static nor constant, but rather
¹²² dynamic ecosystem that interacts with their host and respond to environmental changes. Recent studies
¹²³ have revealed that imbalances in the microbiome, known as dysbiosis, can contribute to a wide range of
¹²⁴ diseases, including obesity (John & Mullin, 2016; Tilg, Kaser, et al., 2011; Castaner et al., 2018), diabetes
¹²⁵ (Barlow, Yu, & Mathur, 2015; Hartstra, Bouter, Bäckhed, & Nieuwdorp, 2015; Sharma & Tripathi, 2019),
¹²⁶ infections (Whiteside, Razvi, Dave, Reid, & Burton, 2015; Alverdy, Hyoju, Weigerinck, & Gilbert, 2017),
¹²⁷ inflammatory conditions (Francescone, Hou, & Grivennikov, 2014; Peirce & Alviña, 2019; Honda &
¹²⁸ Littman, 2012), and cancers (Helmink, Khan, Hermann, Gopalakrishnan, & Wargo, 2019; Cullin, Antunes,
¹²⁹ Straussman, Stein-Thoeringer, & Elinav, 2021; Sepich-Poore et al., 2021; Schwabe & Jobin, 2013). Thus,
¹³⁰ understanding the composition of the human microbiomes is essential for developing new therapeutic
¹³¹ approaches that target these microbial populations to promote health and prevent diseases.

¹³² (Brain-gut axis)

¹³³ 16S ribosomal RNA (rRNA) gene sequencing is one of the most extensively applied methods for
¹³⁴ characterizing microbial communities by targeting the conserved 16S rRNA gene, which contains both
¹³⁵ highly conserved and variable regions in bacteria (Tringe & Hugenholtz, 2008; Janda & Abbott, 2007).
¹³⁶ The conserved regions enable universal primer binding, while the variable regions provide the specificity
¹³⁷ needed to differentiate microbial taxa. Among these regions, the V3-V4 region is frequently selected for
¹³⁸ sequencing due to its balance between phylogenetic resolution and sequencing efficiency (Johnson et al.,
¹³⁹ 2019). Therefore, the V3-V4 region offers sufficient variability to classify a wide range of bacteria taxa
¹⁴⁰ while maintaining compatibility with widely used sequencing platforms.

¹⁴¹ Diversity indices are essential techniques for evaluating the complexity and variety of microbial
¹⁴² communities, in ecological and microbiological research (Tucker et al., 2017; Hill, 1973). Alpha-diversity
¹⁴³ index attributes to the heterogeneity within a specific community, obtaining the number of different taxa
¹⁴⁴ and the distribution of taxa among the individuals, i.e., richness and evenness. On the other hand, beta-
¹⁴⁵ diversity index measures the variations in microbiome compositions between the individuals, highlighting
¹⁴⁶ differences among the microbiome compositions of the study participants. Altogether, by providing a
¹⁴⁷ thorough understanding of microbiome compositions, diversity indices, e.g. alpha-diversity and beta-
¹⁴⁸ diversity, allow us to investigate factors that affecting community variability and structure.

¹⁴⁹ (DAT selection)

¹⁵⁰ Classification is one of the supervised machine learning techniques used to categorized data into
¹⁵¹ predefined classes based on features within the data. In other words, the method learns the relationship

152 between input features and their corresponding output classes through the process of training a classifica-
153 tion model using labeled data. Classification models are essential for advising choices in a wide range of
154 applications, including medical diagnostics. Thus, researchers could uncover sophisticated connections in
155 input features and corresponding classes and produce reliable prediction by utilizing machine learning
156 classification.

157 Random forest classification is one of the ensemble machine learning methods that constructs several
158 decision trees during training and aggregates their results to provide classification predictions (Breiman,
159 2001). A portion of the features and classes—known as bootstrapping (Jiang & Simon, 2007; Champagne,
160 McNairn, Daneshfar, & Shang, 2014; J.-H. Kim, 2009) and feature bagging (Bryll, Gutierrez-Osuna, &
161 Quek, 2003; Alelyani, 2021; Yaman & Subasi, 2019)—are utilized to construct each tree in the forest. The
162 majority vote from each tree determines the final classification, which lowers the possibility of overfitting
163 in comparison to a single decision tree. Furthermore, random forest classifier offers several advantages,
164 including its robustness to outliers and its ability to calculate the feature importance.

165 Evaluating the performance of a machine learning classification model is essential to ensure its
166 reliability and effectiveness in real-world solutions and applications. A confusion matrix is a tabular
167 representation of predictions of classification, showing the counts of true positives (TP), true negatives
168 (TN), false positives (FP), and false negatives (FN) (Table 1). From this matrix, evaluations can be derived:
169 accuracy (ACC; Equation 1), balanced accuracy (BA; Equation 2), F1 score (F1; Equation 3), sensitivity
170 (SEN; Equation 4), specificity (SPE; Equation 5), and precision (PRE; Equation 6). These metrics are in
171 [0, 1] range and high metrics are good metrics. The confusion matrix also helps in identifying specific
172 types of errors, such as a tendency to produce false positive or false negatives, offering valuable insights
173 for improving the classification model. By combining the confusion matrix with other evaluation metrics,
174 researchers can comprehensively assess the classification metrics and refine it for real-world solutions
175 and applications.

176 (AUC)

177 (Limitation & Novelty)

Table 1: Confusion matrix

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

178

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \quad (1)$$

179

$$\text{BA} = \frac{1}{2} \times \left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} \right) \quad (2)$$

180

$$\text{F1} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \quad (3)$$

181

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

182

$$\text{SPE} = \frac{\text{TN}}{\text{TN} + \text{FN}} \quad (5)$$

$$\text{PRE} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6)$$

183 **2 Predicting preterm birth using random forest classifier in salivary mi-**
184 **crobiome**

185 **This section includes the published contents:**

186 Hong, Y. M., **Lee, Jaewoong**, Cho, D. H., Jeon, J. H., Kang, J., Kim, M. G., ... & Kim, J. K. (2023).
187 Predicting preterm birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1),
188 21105.

189 **2.1 Introduction**

190 Preterm birth (PTB), characterized by the delivery of neonates prior to 37 weeks of gestation, is one
191 of the major cause to neonatal mortality and morbidity (Blencowe et al., 2012). Multiple pregnancies
192 including twins, short cervical length, and infection on genitourinary tract are known risk factor for
193 PTB (Goldenberg, Culhane, Iams, & Romero, 2008). Nevertheless, the extent to which these aspects
194 affect birth outcomes is still up for debate. Henceforth, strategies to boost gestation and enhance delivery
195 outcomes can be more conveniently implemented when pregnant women at high risk of PTB are identified
196 early (Iams & Berghella, 2010).

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

207 Preterm prelabor rupture of membranes (PROM) brought on by gestational inflammation and infection
208 contribute to about 70% of PTB cases (Romero, Dey, & Fisher, 2014). Nevertheless, as antibiotics and
209 anti-inflammatory therapeutic strategies were ineffective to decrease PTB occurrence rates, the pathology
210 of PTB has not been entirely elucidated by inflammatory and infectious pathways (Romero, Hassan, et al.,
211 2014). Recent researches on maternal microbiomes were beginning to examine unidentified connections
212 of PTB as a consequence of developmental processes in molecular biological technology (Fettweis et al.,
213 2019).

214 However, as anti-inflammatory and antibiotic therapies were insufficient to lower PTB occurrence
215 rates, infectious and inflammatory processes are insufficient to exhaustively clarify the pathogenesis and
216 pathophysiology of PTB. It has been hypothesized that the microbiota linked to PTB originate from either
217 a hematogenous pathway or the female genitourinary tract increasing through the vagina and/or cervix.
218 (Han & Wang, 2013). Vaginal microbiome compositions have been found in women who eventually

219 acquire PTB, and recent studies have tried to predict PTB risk using cervico-vaginal fluid (Kindinger et
220 al., 2017). Even though previous investigation have confirmed the potential relationships between the
221 vaginal microbiome compositions and PTB, these studies are only able to clarify an upward trajectory.

222 Multiple unfavorable birth outcomes, including PROM and PTB, have been linked to periodontitis
223 as an independence risk factor, according to numerous epidemiological researches (Offenbacher et al.,
224 1996). It is expected that the oral microbiome will be able to explain additional hematogenous pathways
225 in light of these precedents; however, the oral microbiome composition of fetuses is limited understood.

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

229 **2.2 Materials and methods**

230 **2.2.1 Study design and study participants**

231 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
232 Declaration of Helsinki (Goodyear, Krleza-Jeric, & Lemmens, 2007). The Institutional Review Board
233 authorized this study (IRB file No. 2019-01-024). Participants who were admitted for elective cesarean
234 sections (C-sections) or induction births, as well as those who had written informed consent obtained
235 with premature labor or PROM, were eligible.
236

237 **2.2.2 Clinical data collection and grouping**

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

- 240 • maternal age at delivery
- 241 • diabetes mellitus
- 242 • hypertension
- 243 • overweight and obesity
- 244 • C-section
- 245 • history PROM or PTB
- 246 • gestational week on delivery
- 247 • birth weight
- 248 • sex

249 **2.2.3 Salivary microbiome sample collection**

250 Salivary microbiome samples were collected 24 hours before to delivery using mouthwash. The standard
251 methods of sterilizing were performed. Medical experts oversaw each stage of the sample collecting
252 procedure. Participants received instruction not to eat, drink, or brush their teeth for 30 minutes before
253 sampling salivary microbiome. Saliva samples were gathered by washing the mouth for 30 seconds with
254 12 mL of a mouthwash solution (E-zen Gargle, JN Pharm, Pyeongtaek, Gyeonggi, Korea). The samples
255 were tagged with the anonymous ID for each participant and kept at 4 °C until they underwent further
256 processing. Genomic DNA was extracted using an ExgeneTM Clinic SV kit (GeneAll Biotechnology,
257 Seoul, Korea) following with the manufacturer instructions and store at -20 °C.

258 **2.2.4 16s rRNA gene sequencing**

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

265 **2.2.5 Bioinformatics analysis**

266 The independent *t*-test was utilized to evaluate the differences of continuous values between from the
267 PTB participants than the FTB participants; χ^2 -square test was applied to decide statistical differences of
268 categorical values. Clinical measurement comparisons were conducted using SPSS (version 20.0) (Spss
269 et al., 2011). At $p < 0.05$, statistical significance was taken into consideration.

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

277 Taxonomic assignment were implemented with HOMD (version 15.22) (T. Chen et al., 2010).
278 Afterward, DESeq2 was implemented to identify differentially abundant taxa (DAT) that could distinguish
279 between salivary microbiome from PTB and FTB participants (Love, Huber, & Anders, 2014). Taxa with
280 $|\log_2 \text{FoldChange}| > 1$ and $p < 0.05$ were considered as statistically significant.

281 The taxa for predicting PTB using salivary microbiome data were determined using a random forest
282 classifier (Breiman, 2001). Through stratified *k*-fold cross-validation (*k* = 5) that preserves the existence
283 rate of PTB and FTB participants, consistency and trustworthy classification were ensured (Wong & Yeh,
284 2019).

285 **2.2.6 Data and code availability**

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

291 **2.3 Results**

292 **2.3.1 Overview of clinical information**

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

300 **2.3.2 Comparison of salivary microbiomes composition**

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

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

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

311 **2.3.3 Random forest classification to predict PTB risk**

312 To classify PTB according to DAT, random forest classifiers were constructed. The nine most significant
313 DAT were used to obtain the best BA (0.765 ± 0.071 ; Figure 3a). Moreover, random forest classification
314 model determined each DAT's importance (Figure 3b). We conducted a validation procedure on nine
315 twin pregnancies that were excluded in the initial study design in order to confirm the reliability and
316 dependability of our random forest-based PTB prediction model (Figure 6). Comparable to the PTB
317 prediction model on the 59 initial singleton study participants, the validation classification on PTB risk of
318 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^2 -square test. Continuous variable: mean \pm SD. Categorical variable: count (proportion)

	PTB (n=30)	FTB (n=29)	p-value
Maternal age (years)	31.8 \pm 5.2	33.7 \pm 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 \pm 5.9	11.5 \pm 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 \pm 3.4	38.3 \pm 1.1	\leq 0.001
Birth weight (g)	1973.4 \pm 686.6	3283.4 \pm 402.7	\leq 0.001
Male	14 (46.7%)	13 (44.8%)	1.000

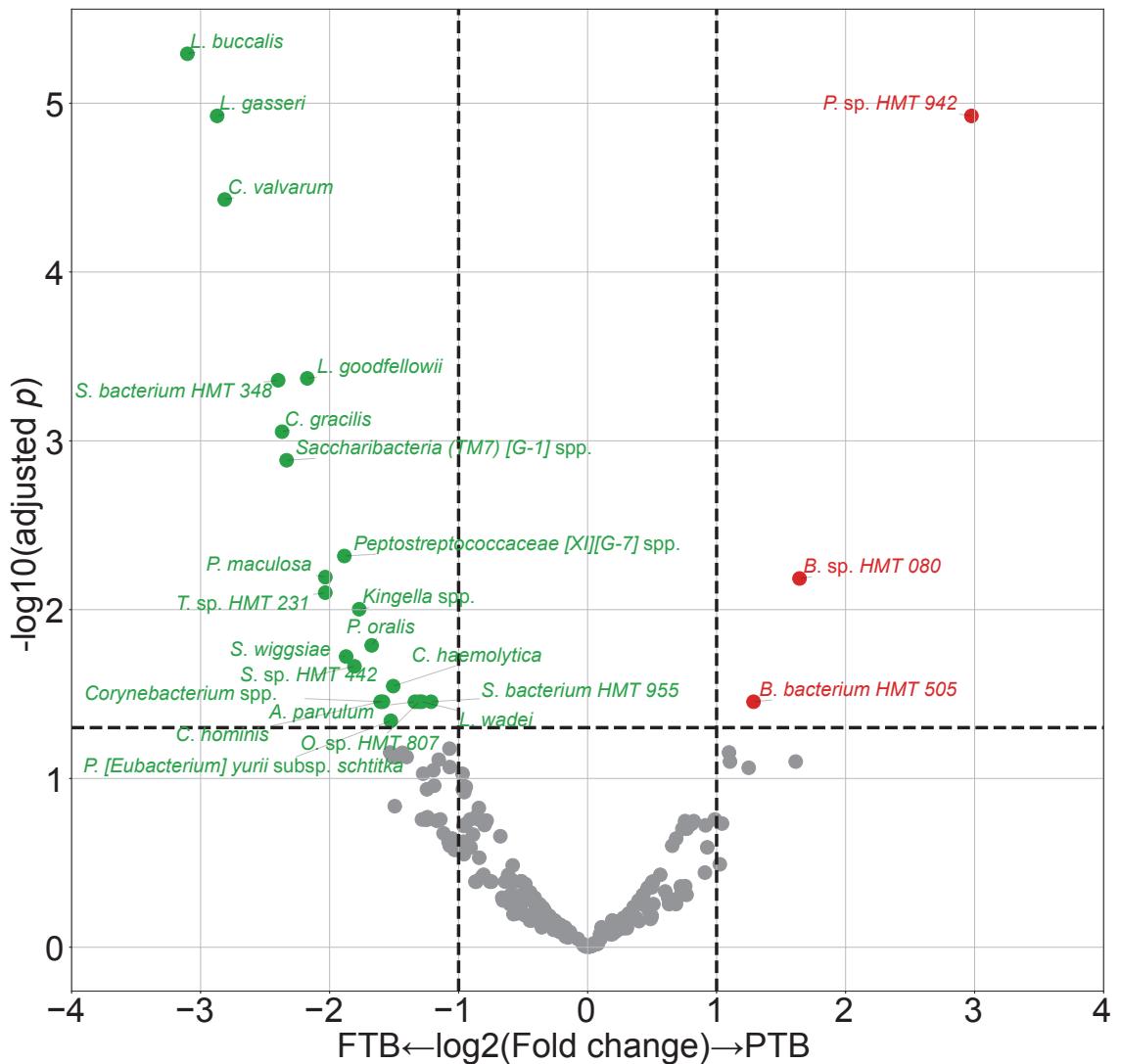


Figure 1: DAT volcano plot.

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

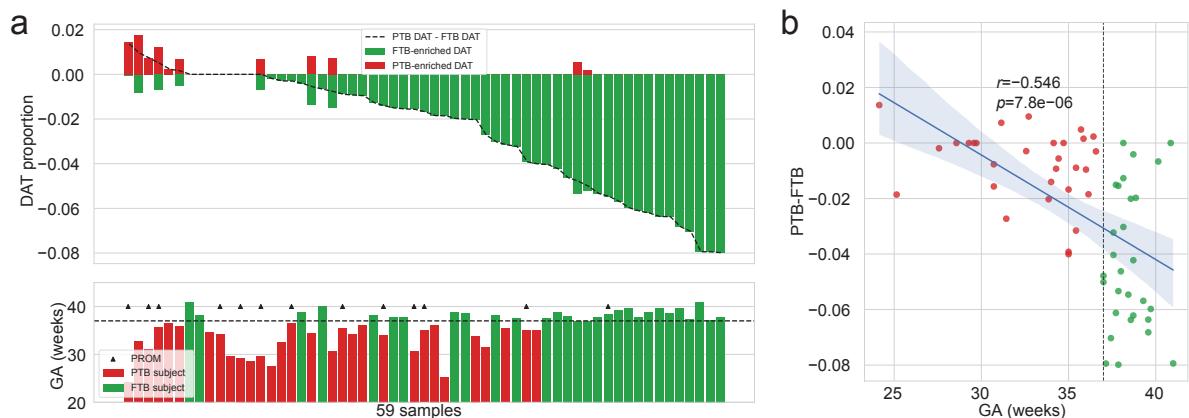


Figure 2: **Salivary microbiome compositions over DAT.**

(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.

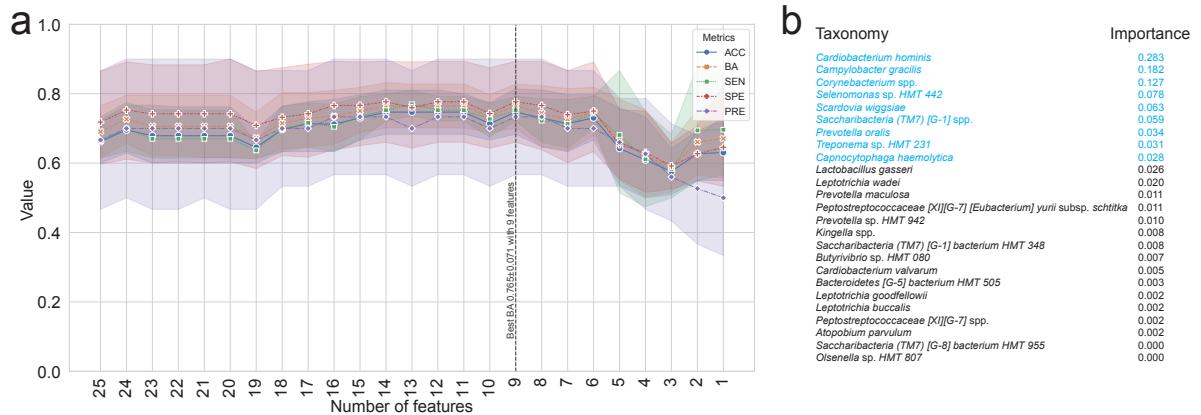


Figure 3: **Random forest-based PTB prediction model.**

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

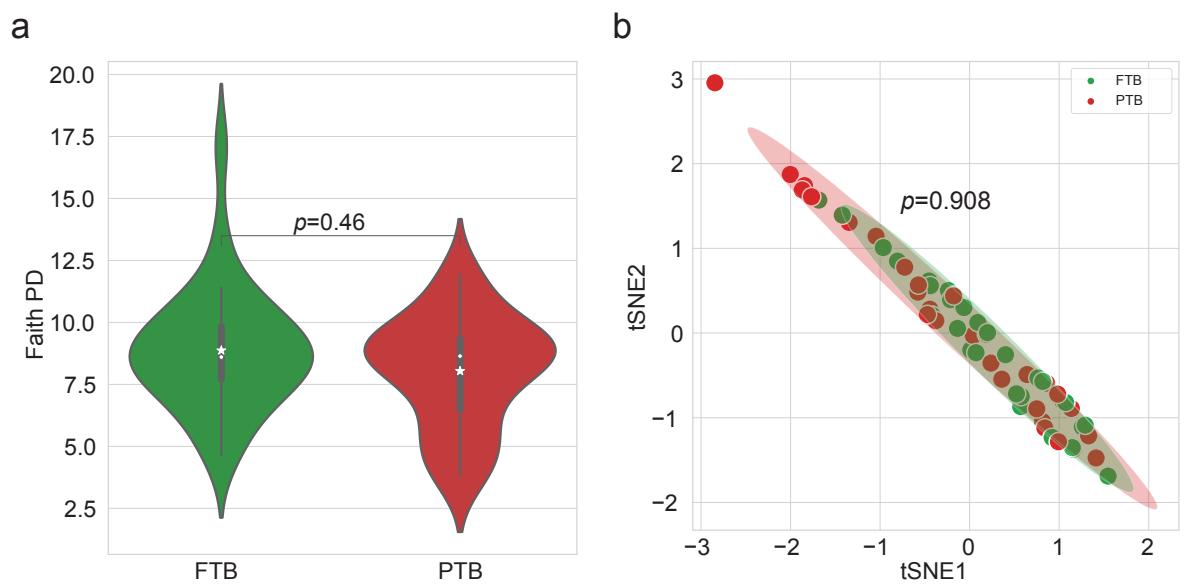


Figure 4: **Diversity indices.**

(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 $p = 0.908$)

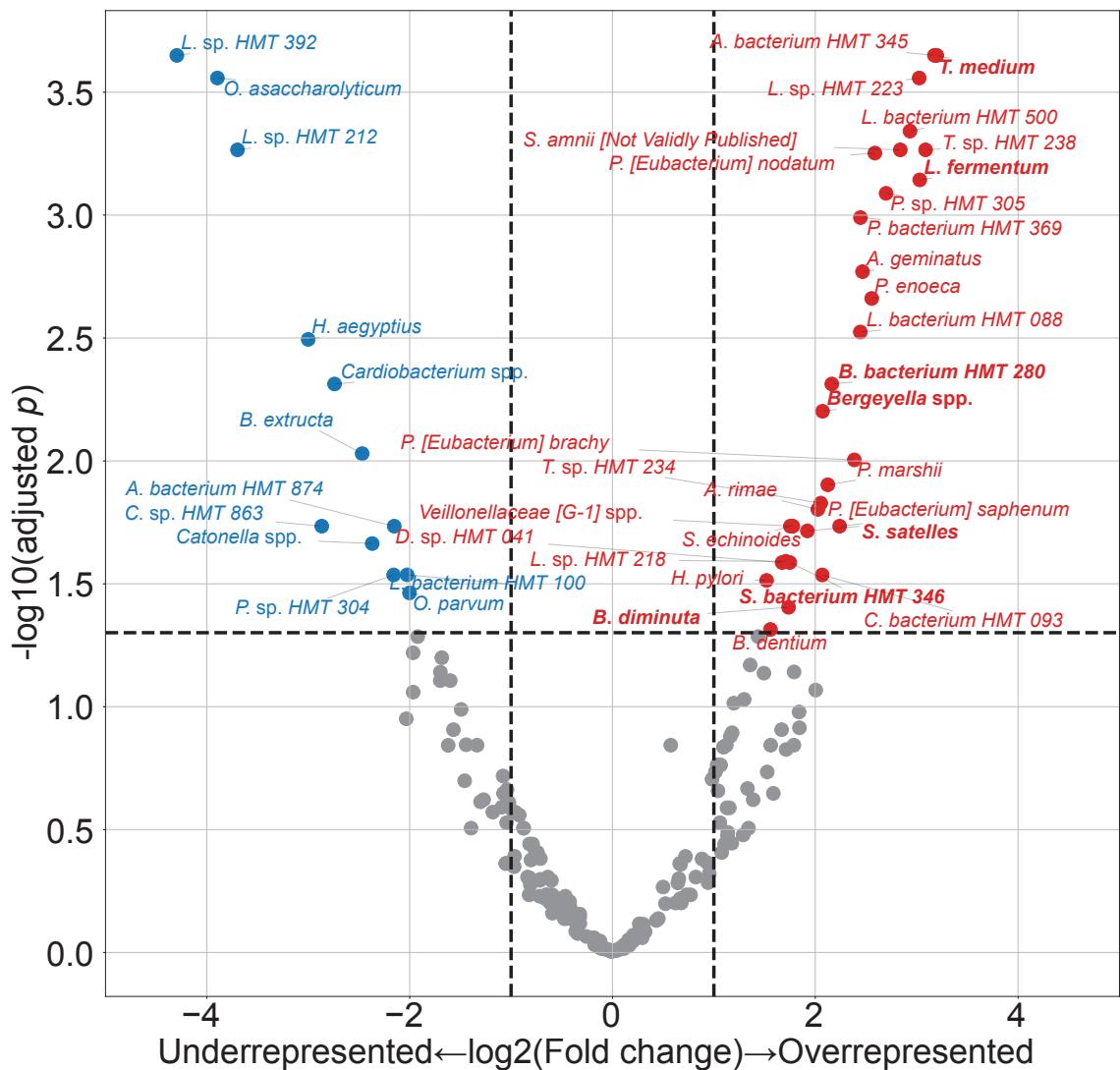


Figure 5: **PROM-related DAT**.

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.

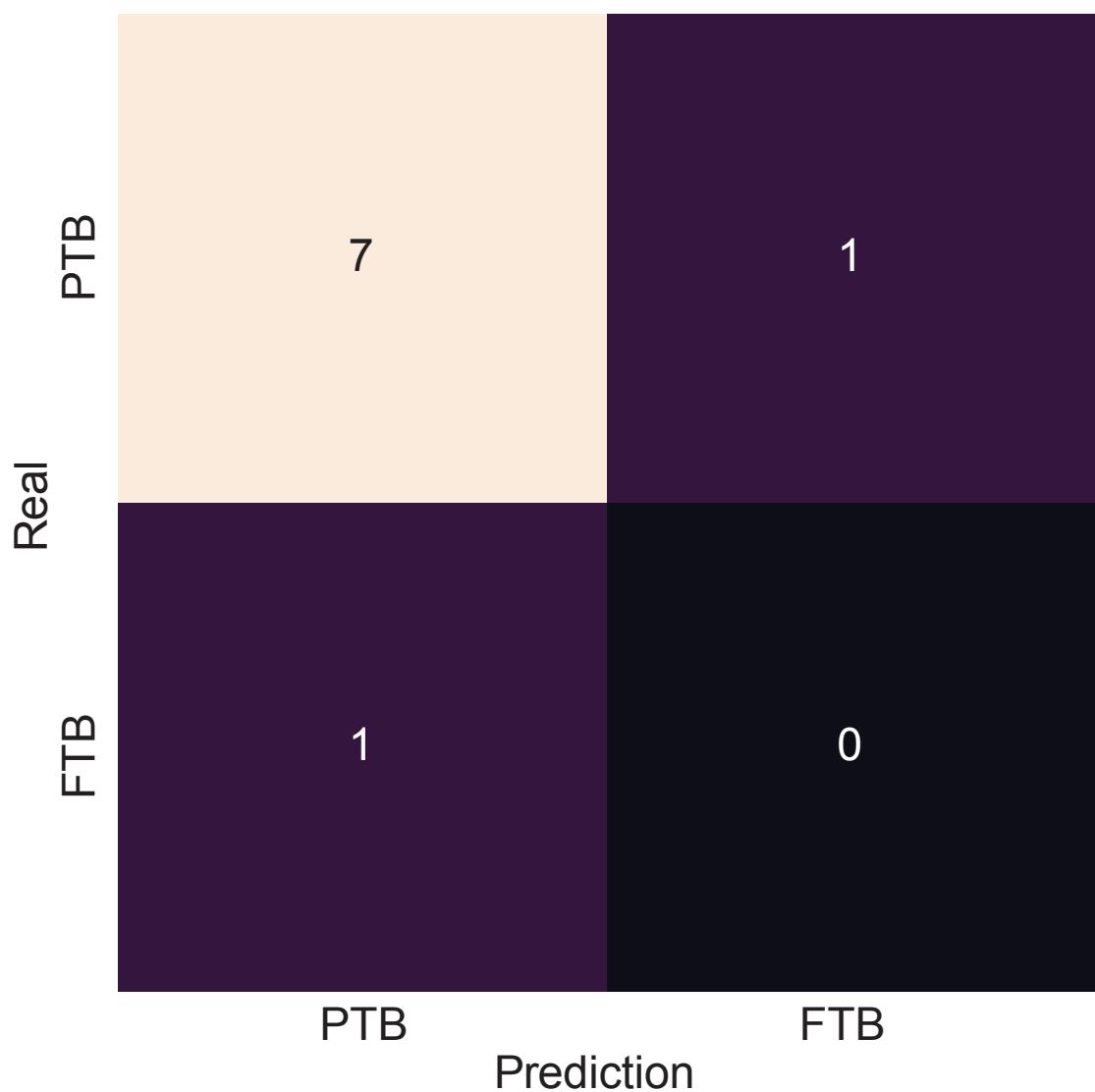


Figure 6: Validation of random forest-based PTB prediction model.

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 \pm SD)

319 **2.4 Discussion**

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

327 One of the most common salivary taxa that has been examined is *Fusobacterium nucleatum* (Han,
328 2015; Brennan & Garrett, 2019; Bolstad, Jensen, & Bakken, 1996), that is a Gram-negative, anaerobic, and
329 filamentous bacteria. *Fusobacterium nucleatum* can be separated from not only the salivary microbiome
330 but also the vaginal microbiome (Vander Haar, So, Gyamfi-Bannerman, & Han, 2018; Witkin, 2019). In
331 both animal and human investigation, *Fusobacterium nucleatum* infection has been linked to risk of PTB
332 (Doyle et al., 2014). According to recent researches, the placenta women who give birth prematurely may
333 include additional salivary microbiome dysbiosis, such as *Bergeyella* spp. and *Porphyromonas gingivalis*
334 (León et al., 2007; Katz, Chegini, Shiverick, & Lamont, 2009). Although *Bergeyella* spp. were one of the
335 PROM-overrepresented DAT (Figure 5), it was excluded in the final 25 PTB-related DAT. Furthermore,
336 *Porphyromonas gingivalis* and *Campylobacter gracilis* were pathogens of periodontitis in sub-gingival
337 microbiome (Yang et al., 2022). *Lactobacillus gasseri* was also one of the FTB-enriched DAT (Figure
338 1), and it is well established that early PTB risk can be reduced by *Lactobacillus gasseri* in the vaginal
339 microbiome (Basavaprabhu, Sonu, & Prabha, 2020; Payne et al., 2021).

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

354 Notably, microbial alteration that take place throughout pregnancy may be expected results of a healthy
355 pregnancy. Those pregnancy-related vulnerabilities to dental problem like periodontitis can be explained
356 by three factors. Because of hormone-driven gingival hyper-reactivity to the salivary microbiome in the

357 oral biofilm including sub-gingival biofilm, these conditions are prevalent in pregnant women. For insight
358 at the relationship between the salivary microbiome compositions and PTB, further studies with pathway
359 analysis are warranted.

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

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

371 **3 Random forest prediction model for periodontitis statuses based on the**
372 **salivary microbiomes**

373 **This section includes the published contents:**

374

375 **3.1 Introduction**

376 Saliva microbial dysbiosis brought on by the accumulation of plaque results in periodontitis, a chronic
377 inflammatory disease of the tissue that surrounds the tooth (Kinane, Stathopoulou, & Papapanou, 2017).
378 Loss of periodontal attachment is a consequence of periodontitis, which may lead to irreversible bone loss
379 and, eventually, permanent tooth loss if left untreated. A new classification criterion of periodontal diseases
380 was created in 2018, about 20 years after the 1999 statements of the previous one (Papapanou et al.,
381 2018). Even with this evolution, radiographic and clinical markers of periodontitis progression remain the
382 primary methods for diagnosing periodontitis (Papapanou et al., 2018). Such tools, nevertheless, frequently
383 demonstrate the prior damage from periodontitis rather than its present condition. Certain individuals have
384 a higher risk of periodontitis, a higher chance of developing severe generalized periodontitis, and a worse
385 response to common salivary bacteria control techniques utilized to prevent and treat periodontitis. As a
386 result, the 2017 framework for diagnosing periodontitis additionally allows for the potential development
387 of biomarkers to enhance diagnosis and treatment of periodontitis (Tonetti, Greenwell, & Kornman, 2018).
388 Instead of only depending on the progression of periodontitis, a new etiological indication based on the
389 current state must be introduced in order to enable appropriate intervention through early detection of
390 periodontitis. Thus, the current clinical diagnostic techniques that rely on periodontal probing can be
391 uncomfortable for patients with periodontitis (Canakci & Canakci, 2007).

392 Due to the development of salivaomics, in this manner, the examination of saliva has emerged as
393 a significant alternative to the conventional ways of identifying periodontitis (Altingöz et al., 2021;
394 Melguizo-Rodríguez, Costela-Ruiz, Manzano-Moreno, Ruiz, & Illescas-Montes, 2020). Given that saliva
395 sampling is non-invasive, painless, and accessible to non-specialists, it may be a valuable instrument for
396 diagnosing periodontitis (Zhang et al., 2016). Furthermore, much research has suggested that periodontitis
397 could be a trigger in the development and exacerbation of metabolic syndrome (Morita et al., 2010; Nesbitt
398 et al., 2010). Consequently, alteration in these levels of salivary microbiome markers may serve as high
399 effective diagnostic, prognostic, and therapeutic indicators for periodontitis and other systemic diseases
400 (Miller, Ding, Dawson III, & Ebersole, 2021; Čižmárová et al., 2022). The pathogenesis of periodontitis
401 typically comprises qualitative as well as quantitative alterations in the salivary microbial community,
402 despite that it is a complex disease impacted by a number of contributing factors including age, smoking
403 status, stress, and nourishment (Abusleme, Hoare, Hong, & Diaz, 2021; Lafaurie et al., 2022). Depending
404 on the severity of periodontitis, the salivary microbial community's diversity and characteristics vary
405 (Abusleme et al., 2021), indicating that a new etiological diagnostic standards might be microbial
406 community profiling based on clinical diagnostic criteria. As a consequence, salivary microbiome

407 compositions have been characterized in numerous research in connection with periodontitis. High-
408 throughput sequencing, including 16S rRNA gene sequencing, has recently used in multiple studies to
409 identify variations in the bacterial composition of sub-gingival plaque collections from periodontal healthy
410 individuals and patients with periodontitis (Altabtbaei et al., 2021; Iniesta et al., 2023; Nemoto et al., 2021).
411 This realization has rendered clear that alterations in the salivary microbial community—especially, shifts to
412 dysbiosis—are significant contributors to the pathogenesis and development of periodontitis (Lamont, Koo,
413 & Hajishengallis, 2018). Yet most of these research either focused only on the microbiome alterations in
414 sub-gingival plaque collection, comprised a limited number of periodontitis study participants, or did not
415 account for the impact of multiple severities of periodontitis.

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

424 Recently, we employed multiplex quantitative-PCR and machine learning-based classification model
425 to predict the severity of periodontitis based on the amount of nine pathogens of periodontitis from
426 saliva collections (E.-H. Kim et al., 2020). On the other hand, the fact that we focused merely at nine
427 pathogens for periodontitis and neglected the variety bacterial species associated to the various severities
428 of periodontitis constrained the breadth of our investigation. By developing a machine learning model
429 that could classify multiple severities of periodontitis based on the salivary microbiome composition,
430 this study aims to fill these knowledge gaps and produce more accurate and therapeutically useful
431 guidance to evaluate progression of periodontitis. Hence, in order to examine the salivary microbiome
432 composition of both healthy controls and patients with periodontitis in multiple stages, we applied
433 16S rRNA gene sequencing. Furthermore, employing the 2018 classification criteria, we sought to find
434 biomarkers (species) for the precise prediction of periodontitis severities (Papapanou et al., 2018; Chapple
435 et al., 2018).

436 **3.2 Materials and methods**

437 **3.2.1 Study participants enrollment**

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

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

455 **3.2.2 Periodontal clinical parameter diagnosis**

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

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

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

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

473 0. Normal gingiva: without inflammation nor discoloration.

474 1. Mild inflammation: minimal edema and slight color changes, but no bleeding on probing.

475 2. Moderate inflammation: edema, glazing, redness, and bleeding on probing.

476 3. Severe inflammation: significant edema, ulceration, redness, and spontaneous bleeding.

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

480 Periodontitis was diagnosed in respect to the 2018 classification criteria (Papapanou et al., 2018;
481 Chapple et al., 2018). An experienced periodontist diagnosed the periodontitis severity by considering
482 complexity, depending on clinical examinations including radiographic images and periodontal probing.

483 Periodontitis is categorized into healthy, stage I, stage II, and stage III with the following criteria:

484 • Healthy:

485 1. Bleeding sites < 10%

486 2. Probing depth: \leq 3 mm

487 • Stage I:

488 1. No tooth loss because of periodontitis.

489 2. Inter-dental clinical attachment level at the site of the greatest loss: 1-2 mm

490 3. Radiographic bone loss: < 15%

491 • Stage II:

492 1. No tooth loss because of periodontitis.

493 2. Inter-dental clinical attachment level at the site of the greatest loss: 3-4 mm

494 3. Radiographic bone loss: 15-33%

495 • Stage III:

496 1. Teeth loss because of periodontitis: \leq teeth

497 2. Inter-dental clinical attachment level at the site of the greatest loss: \geq 5 mm

498 3. Radiographic bone loss: > 33%

499 **3.2.3 Saliva sampling and DNA extraction procedure**

500 All study participants received instructions to avoid eating, drinking, brushing, and using mouthwash for
501 at least an hour prior to the saliva sample collection process. These collections were conducted between
502 09:00 and 11:00. Mouth rinse was collected by rinsing the mouth for 30 seconds with 12 mL of a solution
503 (E-zen Gargle, JN Pharm, Korea). All saliva samples were tagged with anonymous ID and stored at -4 °C.

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

- 508 • Forward: 5' -TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3'
509 • Reverse: 5' -GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'

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

- 512 1. Heat activation for 30 seconds at 95 °C.
513 2. 25 cycles for 30 seconds at 95 °C.
514 3. 30 seconds at 55 °C.
515 4. 30 seconds at 72 °C.

516 NexteraXT Indexed Primer was applied to amplification 10 µL of the purified initial PCR products for
517 the final library creation. The second PCR used the same conditions as the first PCR conditions but with
518 10 cycles. 16S rRNA gene sequencing was performed via 2×300 bp paired-end sequencing at Macrogen
519 Inc. (Macrogen, Seoul, Korea) using Illumina MiSeq platform (Illumina, San Diego, CA, USA).

520 **3.2.4 Bioinformatics analysis**

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

- 525 • Abundance-based Coverage Estimator (ACE) (Chao & Lee, 1992)
526 • Chao1 (Chao, 1984)
527 • Fisher (Fisher, Corbet, & Williams, 1943)
528 • Margalef (Magurran, 2021)
529 • 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-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/>

566 repository/docker/fumire/periodontitis_16s. Every code used in this study can be found on
567 GitHub: https://github.com/CompbioLabUnist/Periodontitis_16S.

568 **3.3 Results**

569 **3.3.1 Summary of clinical information and sequencing data**

570 Among clinical information of the study participants, clinical attachment level, probing depth, plaque
571 index, and gingival index, were significantly increased with periodontitis severity (Kruskal-Wallis test
572 $p < 0.001$), while sex were observed no significant difference (Table 2). Notably, clinical attachment level
573 and probing depth have significant differences among the periodontitis severities (MWU test $p < 0.01$;
574 Figure 15). Additionally, 71461.00 ± 11792.30 and 45909.78 ± 11404.65 reads per sample were obtained
575 before and after filtering low-quality reads and trimming extra-long tails, respectively (Figure 16).

576 **3.3.2 Diversity indices reveal differences among the periodontitis severities**

577 Rarefaction curves showed that the sequencing depth was sufficient (Figure 12). Alpha-diversity in-
578 dices indicated significant differences between the healthy and the periodontitis stages (MWU test
579 $p < 0.01$; Figure 7a-e); however, there were no significant differences between the periodontitis stages.
580 This emphasizes how essential it is to classify the salivary microbiome compositions and distinguish
581 between the stages of periodontitis using machine learning approaches.

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

587 **3.3.3 DAT among multiple periodontitis severities and their correlation**

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

- 591 • Group 1
- 592 1. *Treponema* spp.
- 593 2. *Prevotella* sp. HMT 304
- 594 3. *Prevotella* sp. HMT 526
- 595 4. *Peptostreptococcaceae [XI][G-5]* saphenum
- 596 5. *Treponema* sp. HMT 260
- 597 6. *Mycoplasma faecium*
- 598 7. *Peptostreptococcaceae [XI][G-9]* brachy
- 599 8. *Lachnospiraceae [G-8]* bacterium HMT 500
- 600 9. *Peptostreptococcaceae [XI][G-6]* nodatum

601 10. *Fretibacterium* spp.

602 • Group 2

- 603 1. *Porphyromonas gingivalis*
604 2. *Campylobacter showae*
605 3. *Filifactor alocis*
606 4. *Treponema putidum*
607 5. *Tannerella forsythia*
608 6. *Prevotella intermedia*
609 7. *Porphyromonas* sp. HMT 285

610 • Group 3

- 611 1. *Actinomyces* spp.
612 2. *Corynebacterium durum*
613 3. *Actinomyces graevenitzii*

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

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

622 3.3.4 Classification of periodontitis severities by random forest models

623 Based on the proportion of DAT, random forest classifier were trained to classify the periodontitis
624 severities (Table 6). First of all, we conducted multi-label classification for the multiple periodontitis
625 severities, namely healthy, stage I, stage II, and stage III. In this setting, we classified multiple periodontitis
626 severities with the highest BA of 0.779 ± 0.029 (Table 4). AUC ranged between 0.81 and 0.94 (Figure
627 10b).

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

632 Third, based on the findings that the salivary microbiome composition in stage II is more comparable
633 to those in stage III than to other severities (Figure 7f and Figure 7j), we combined stage II and stage III
634 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	<i>P.gingivalis</i>	0.758±0.051	0.716±0.177	0.677±0.068	0.839±0.034	0.839±0.034	0.516±0.102	
	<i>P.gingivalis+Act.</i>	0.792±0.043	0.822±0.105	0.723±0.057	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	0.668±0.033	
Healthy vs. Stage I	<i>Act.</i>	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
	<i>Act.+P.gingivalis</i>	0.733±0.119	0.831±0.081	0.713±0.122	0.797±0.097	0.797±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	<i>P.gingivalis</i>	0.776±0.042	0.736±0.196	0.748±0.047	0.832±0.031	0.832±0.031	0.664±0.062	
	<i>P.gingivalis+Act.</i>	0.843±0.035	0.876±0.109	0.823±0.039	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	0.828±0.051	
Healthy vs. Stages I/II/III	<i>P.gingivalis</i>	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
	<i>P.gingivalis+Act.</i>	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	<i>Porphyromonas gingivalis</i>	424
2	<i>Actinomyces</i> spp.	424
3	<i>Filifactor alocis</i>	421
4	<i>Prevotella intermedia</i>	419
5	<i>Treponema putidum</i>	418
6	<i>Tannerella forsythia</i>	415
7	<i>Porphyromonas</i> sp. HMT 285	412
8	<i>Peptostreptococcaceae [XI][G-6] nodatum</i>	412
9	<i>Fretibacterium</i> spp.	411
10	<i>Mycoplasma faecium</i>	411
11	<i>Prevotella</i> sp. HMT 304	411
12	<i>Lachnospiraceae [G-8] bacterium</i> HMT 500	409
13	<i>Treponema</i> spp.	408
14	<i>Prevotella</i> sp. HMT 526	401
15	<i>Peptostreptococcaceae [XI][G-9] brachy</i>	400
16	<i>Peptostreptococcaceae [XI][G-5] saphenum</i>	398
17	<i>Campylobacter showae</i>	395
18	<i>Treponema</i> sp. HMT 260	393
19	<i>Corynebacterium durum</i>	393
20	<i>Actinomyces graevenitzii</i>	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 III/IV		
	Rank	Taxa	Importance	Taxa	Importance	Taxa	Importance	Taxa	Importance	Taxa	Importance	
1	<i>Porphyromonas gingivalis</i>	0.297	<i>Actinomyces spp.</i>	0.195	<i>Porphyromonas gingivalis</i>	0.360	<i>Porphyromonas gingivalis</i>	0.426	<i>Porphyromonas gingivalis</i>	0.461		
2	<i>Actinomyces spp.</i>	0.195	<i>Actinomyces graevenitzii</i>	0.054	<i>Actinomyces spp.</i>	0.125	<i>Actinomyces spp.</i>	0.244	<i>Actinomyces spp.</i>	0.257		
3	<i>Prevotella intermedia</i>	0.054	<i>Actinomyces graevenitzii</i>	0.052	<i>Porphyromonas sp. HMT 285</i>	0.055	<i>Actinomyces graevenitzii</i>	0.049	<i>Actinomyces spp.</i>	0.059		
4	<i>Actinomyces graevenitzii</i>	0.052	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.050	<i>Porphyromonas sp. HMT 285</i>	0.062	<i>Corynebacterium durum</i>	0.046	<i>Corynebacterium durum</i>	0.035		
5	<i>Filifactor alocis</i>	0.050	<i>Campylobacter showae</i>	0.042	<i>Campylobacter showae</i>	0.052	<i>Filifactor alocis</i>	0.036	<i>Filifactor alocis</i>	0.032		
6	<i>Campylobacter showae</i>	0.042	<i>Porphyromonas sp. HMT 285</i>	0.040	<i>Corynebacterium durum</i>	0.052	<i>Prevotella intermedia</i>	0.033	<i>Campylobacter showae</i>	0.023		
7	<i>Porphyromonas sp. HMT 285</i>	0.040	<i>Treponema spp.</i>	0.032	<i>Treponema spp.</i>	0.038	<i>Tannerella forsythia</i>	0.025	<i>Porphyromonas sp. HMT 285</i>	0.022		
8	<i>Corynebacterium durum</i>	0.032	<i>Tannerella forsythia</i>	0.026	<i>Tannerella forsythia</i>	0.037	<i>Prevotella intermedia</i>	0.023	<i>Prevotella intermedia</i>	0.022		
9	<i>Treponema spp.</i>	0.032	<i>Prevotella intermedia</i>	0.025	<i>Prevotella intermedia</i>	0.029	<i>Treponema spp.</i>	0.021	<i>Treponema spp.</i>	0.022		
10	<i>Tannerella forsythia</i>	0.026	<i>Prevotella intermedia</i>	0.025	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.026	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.018	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.015		
11	<i>Treponema putidum</i>	0.025	<i>Freibacterium spp.</i>	0.023	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.018	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.014	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.010		
12	<i>Freibacterium spp.</i>	0.023	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.021	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.018	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.011	<i>Tannerella forsythia</i>	0.009		
13	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.021	<i>Treponema putidum</i>	0.019	<i>Treponema putidum</i>	0.014	<i>Treponema putidum</i>	0.010	<i>Freibacterium spp.</i>	0.009		
14	<i>Treponema sp. HMT 260</i>	0.019	<i>Prevotella sp. HMT 526</i>	0.018	<i>Prevotella sp. HMT 526</i>	0.011	<i>Prevotella sp. HMT 526</i>	0.009	<i>Prevotella sp. HMT 526</i>	0.006		
15	<i>Prevotella sp. HMT 526</i>	0.018	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.018	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.008	<i>Freibacterium spp.</i>	0.008	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.004		
16	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.018	<i>Prevotella sp. HMT 304</i>	0.017	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.008	<i>Treponema sp. HMT 260</i>	0.008	<i>Treponema sp. HMT 260</i>	0.004		
17	<i>Prevotella sp. HMT 304</i>	0.017	<i>Mycoplasma faecium</i>	0.014	<i>Mycoplasma faecium</i>	0.004	<i>Prevotella sp. HMT 304</i>	0.005	<i>Mycoplasma faecium</i>	0.003		
18	<i>Mycoplasma faecium</i>	0.014	<i>Prevotella sp. HMT 304</i>	0.014	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.003	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.005	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.002		
19	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.014	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.013	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.003	<i>Prevotella sp. HMT 304</i>	0.004	<i>Prevotella sp. HMT 304</i>	0.001		
20	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.013										

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

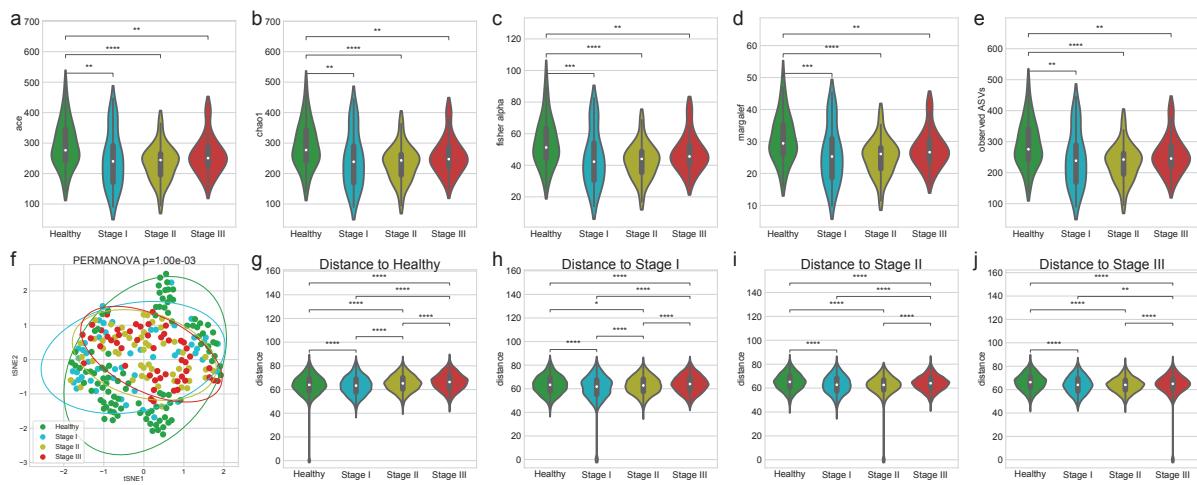
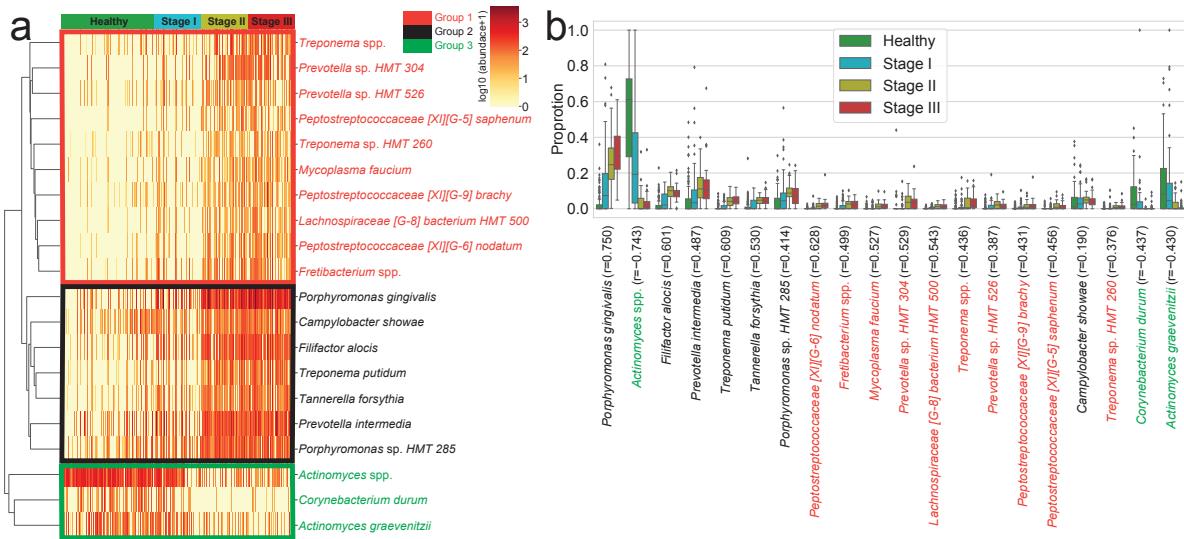


Figure 7: Diversity indices.

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 $p \leq 0.0001$ (****).



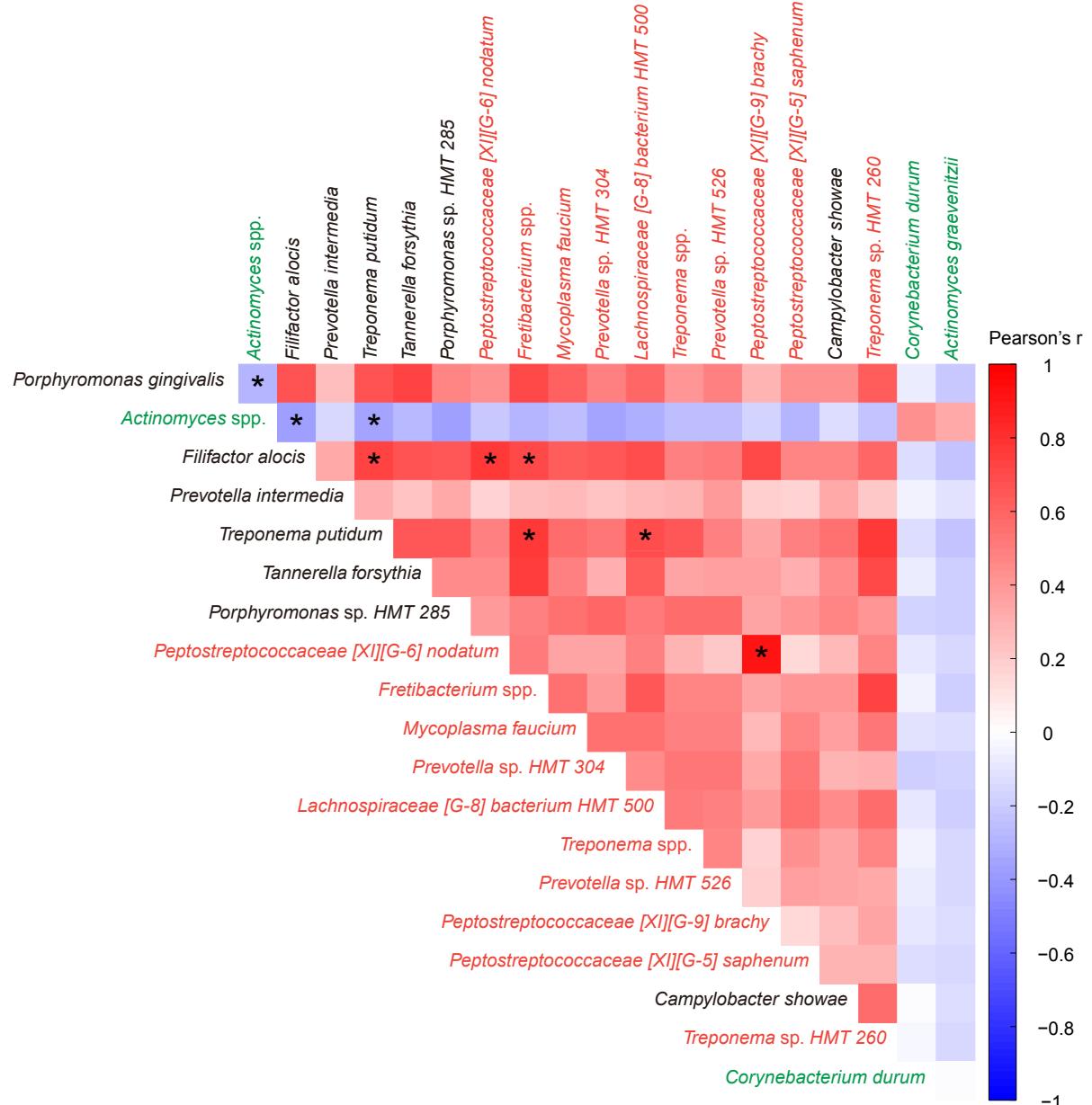


Figure 9: Correlation heatmap.

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

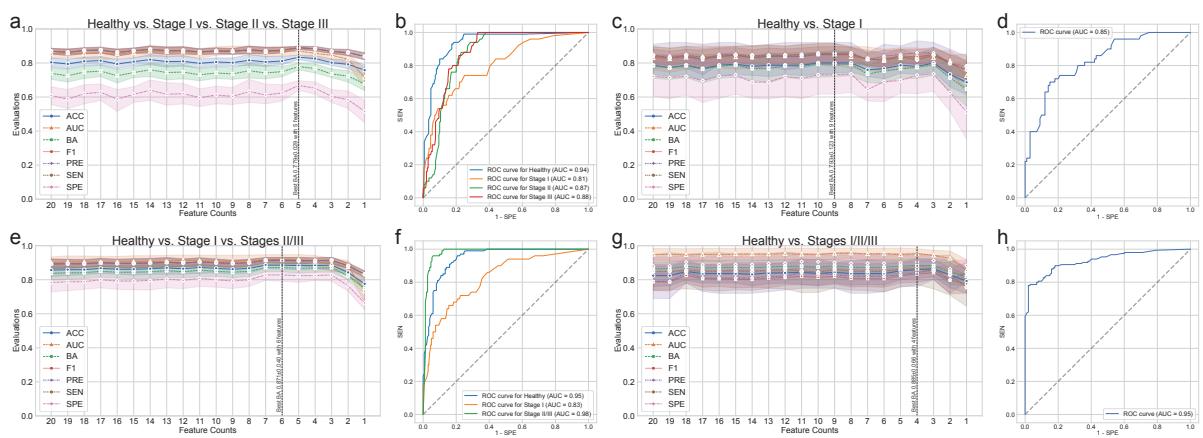


Figure 10: Random forest classification metrics.

The classification metrics in the random forest classifications were as follows: 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).

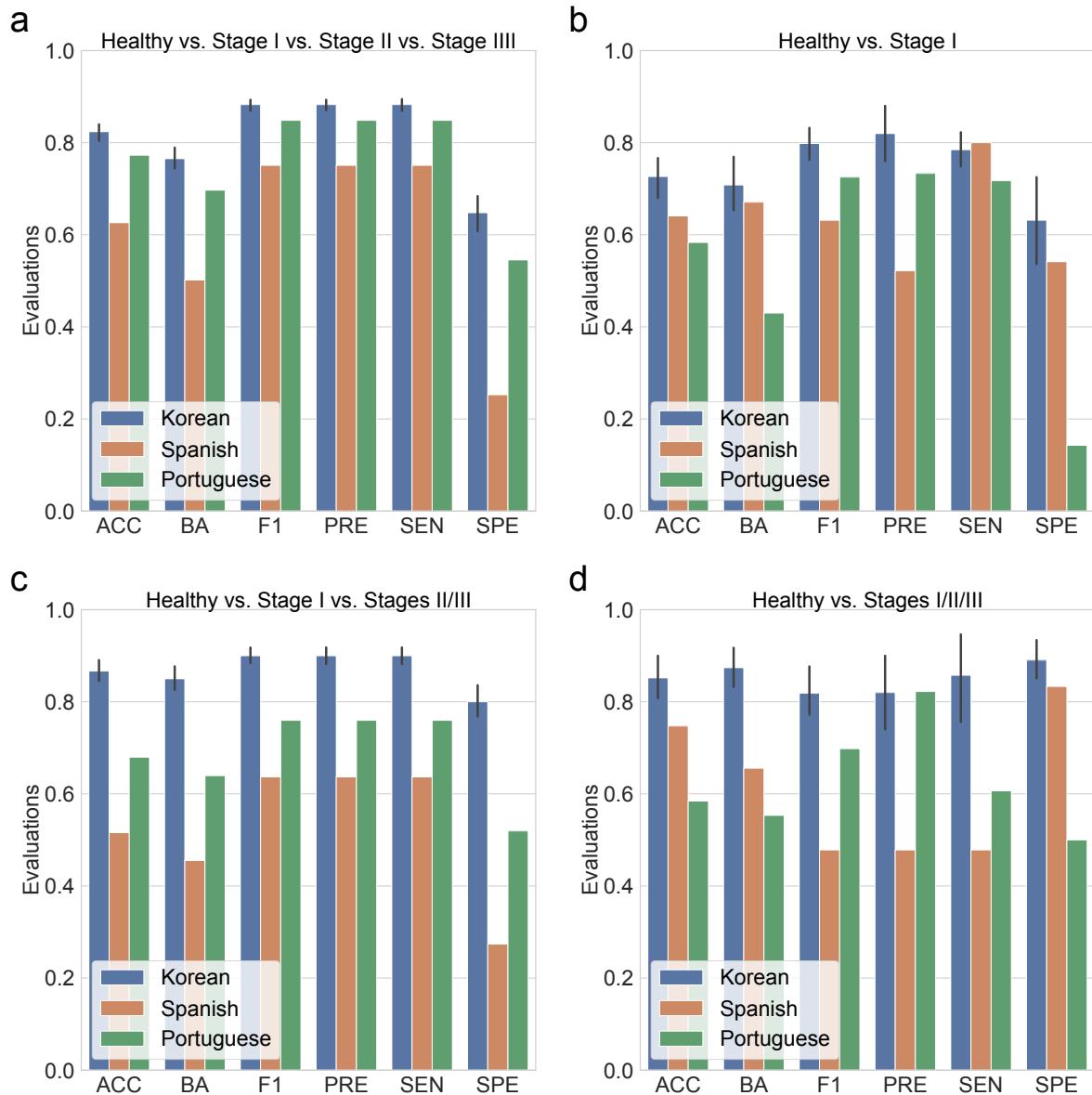


Figure 11: **Random forest classification metrics from external datasets.**

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.

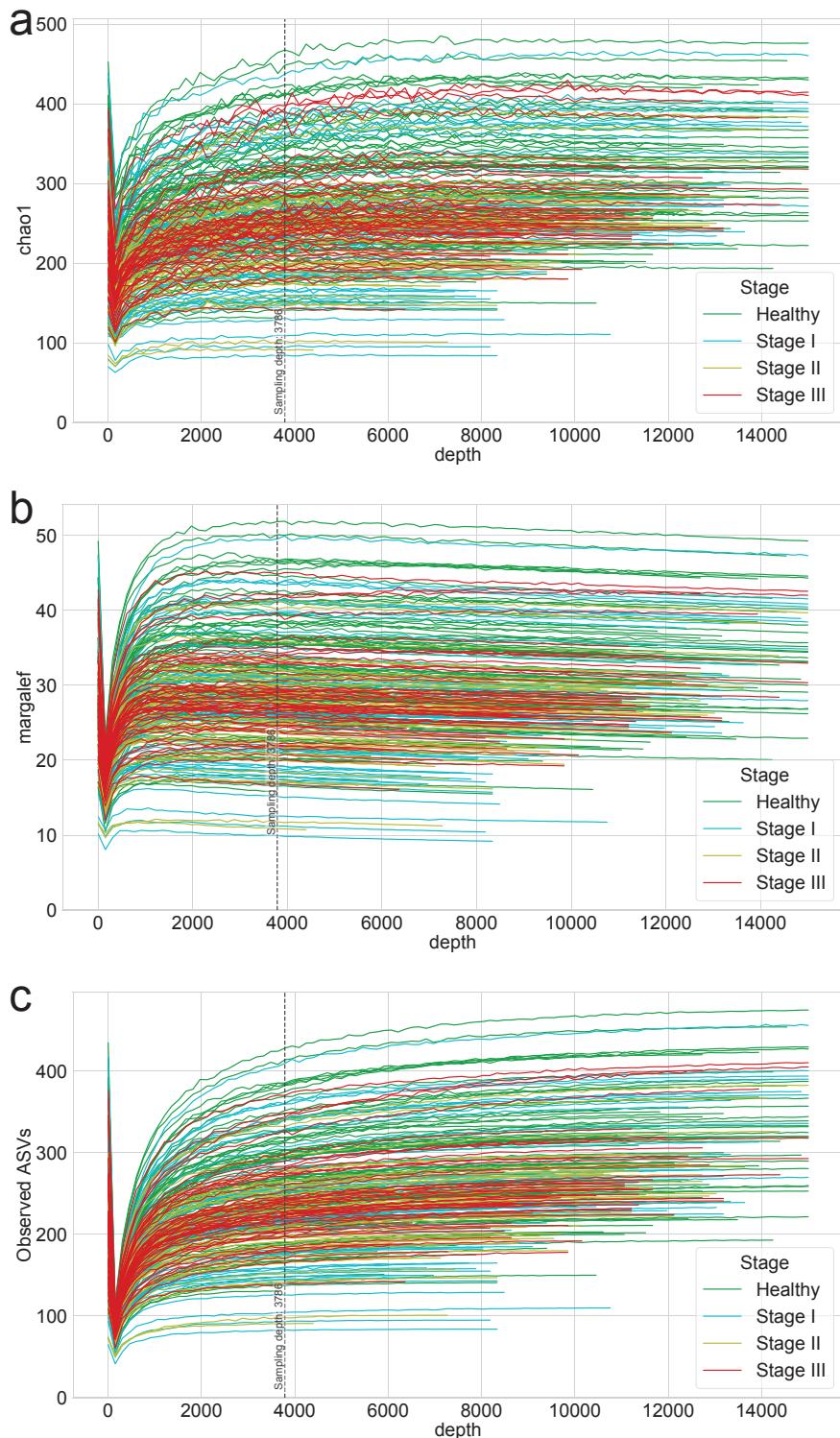


Figure 12: **Rarefaction curves for alpha-diversity indices.**

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

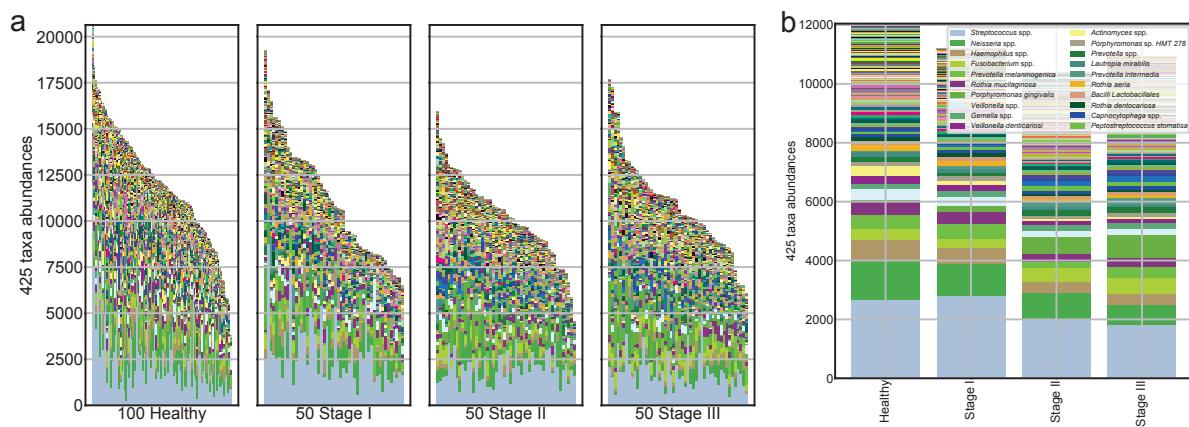


Figure 13: Salivary microbiome compositions in the different periodontal statuses.

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**).

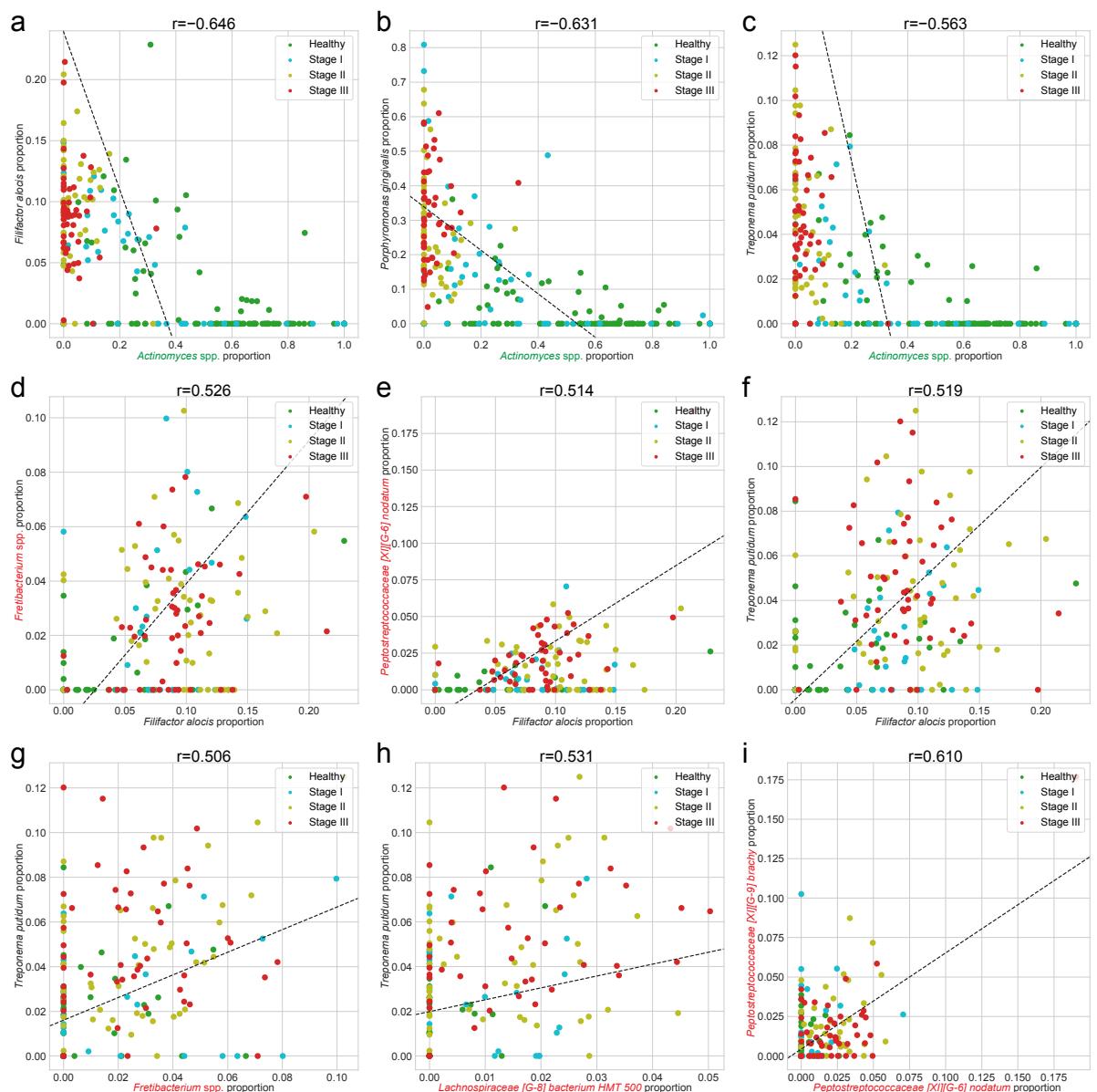


Figure 14: Correlation plots for differentially abundant taxa.

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).

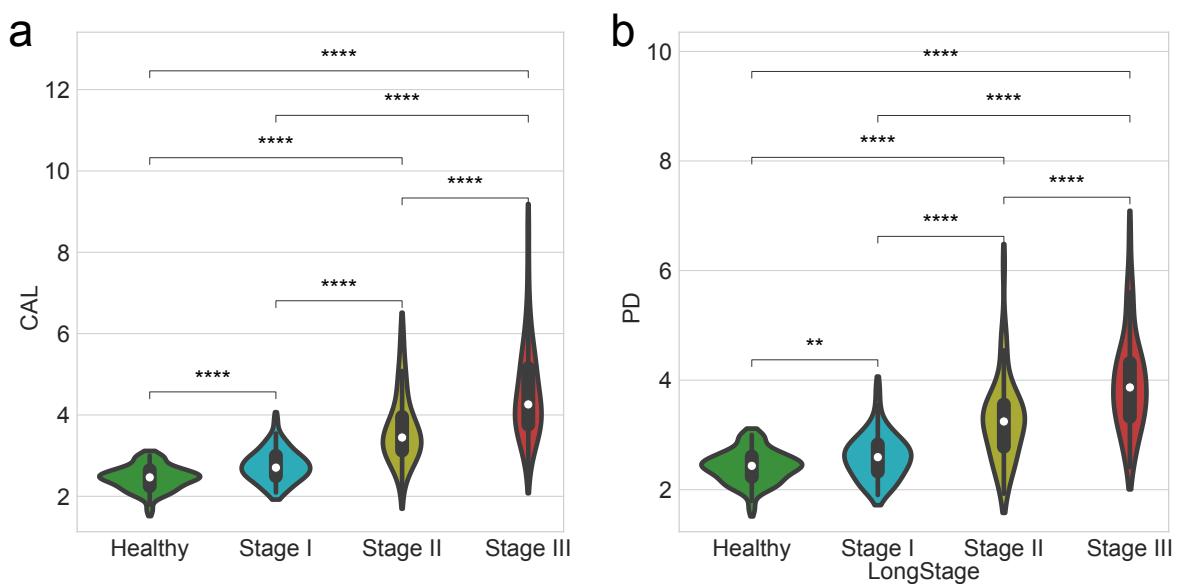


Figure 15: **Clinical measurements by the periodontitis statuses.**

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 $p \leq 0.0001$ (****).

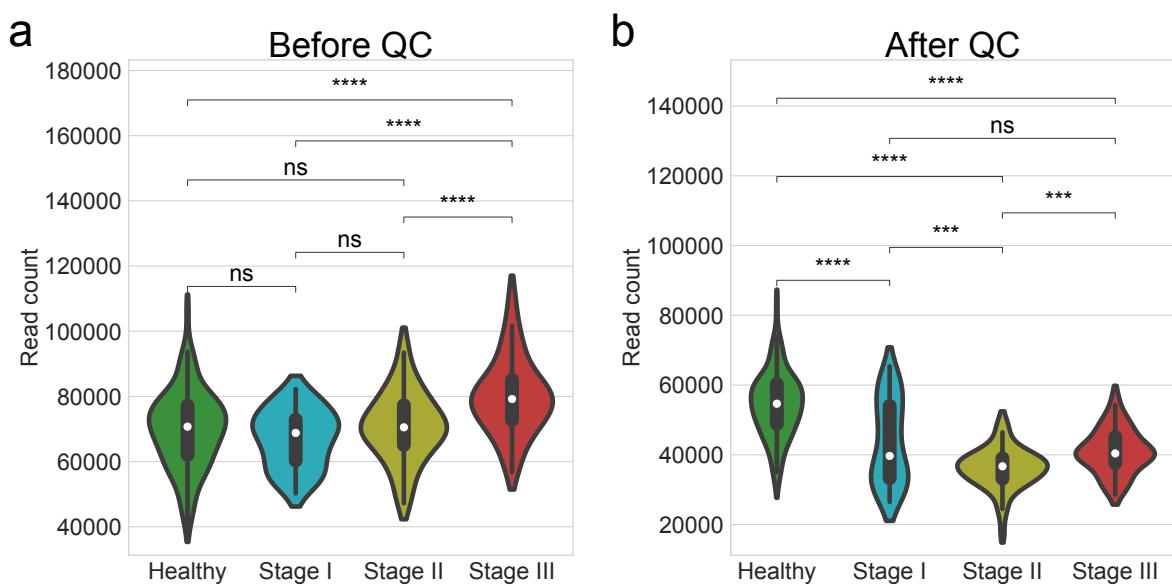


Figure 16: **Number of read counts by the periodontitis statuses.**

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 $p \leq 0.0001$ (****).

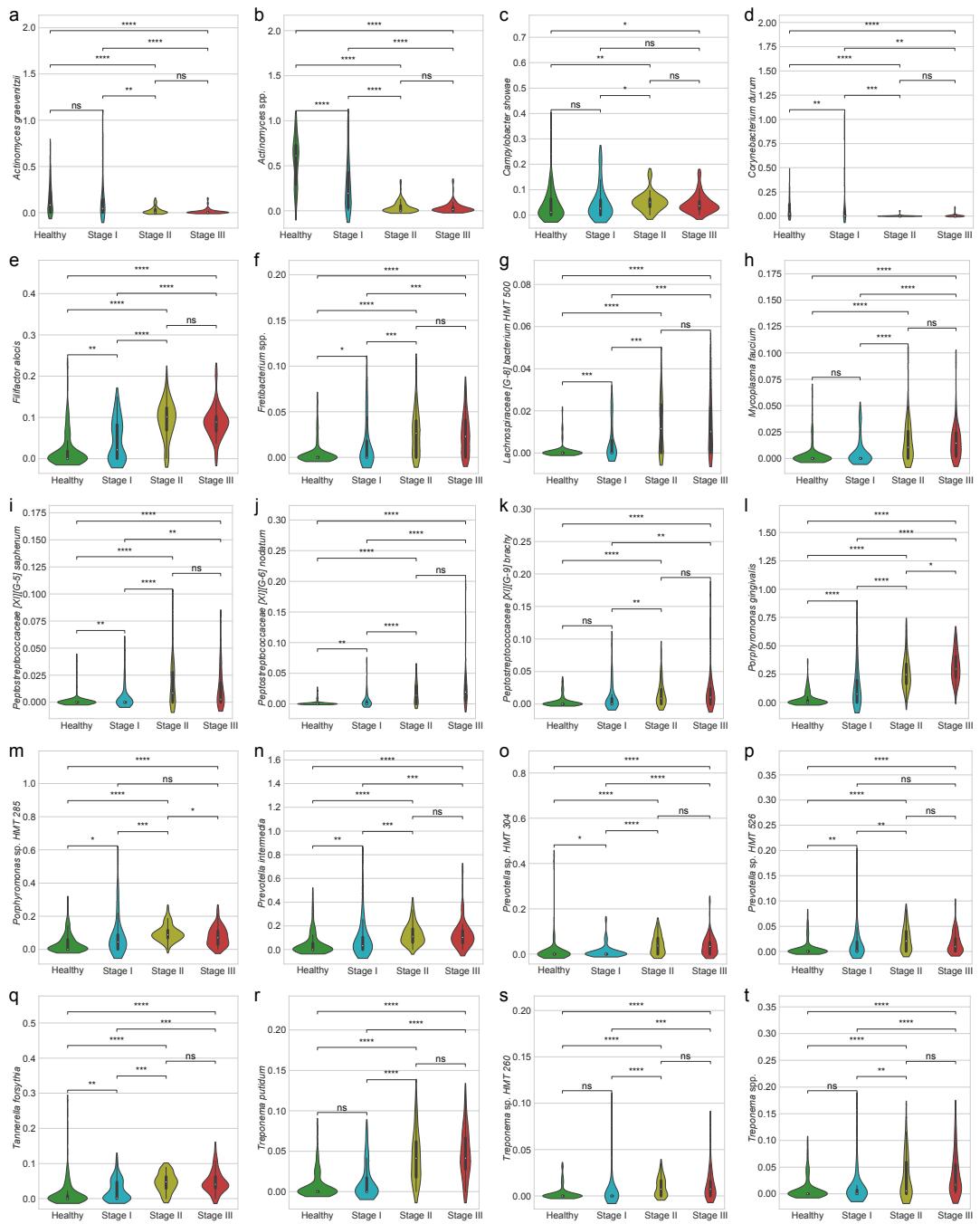


Figure 17: Proportion of DAT.

(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 faecium* (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 $p \leq 0.0001$ (****).

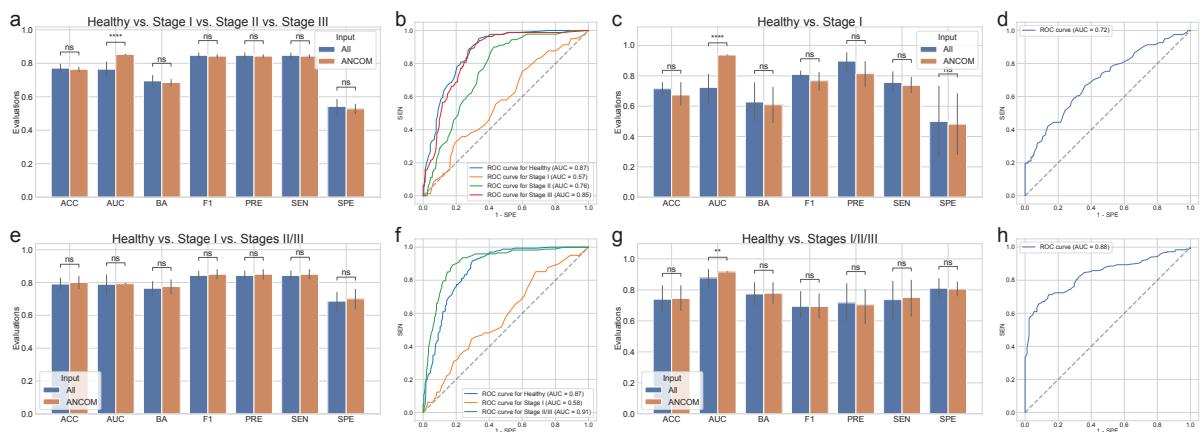


Figure 18: Random forest classification metrics with the full microbiome compositions and ANCOM-selected DAT compositions.

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 $p \leq 0.0001$ (****).

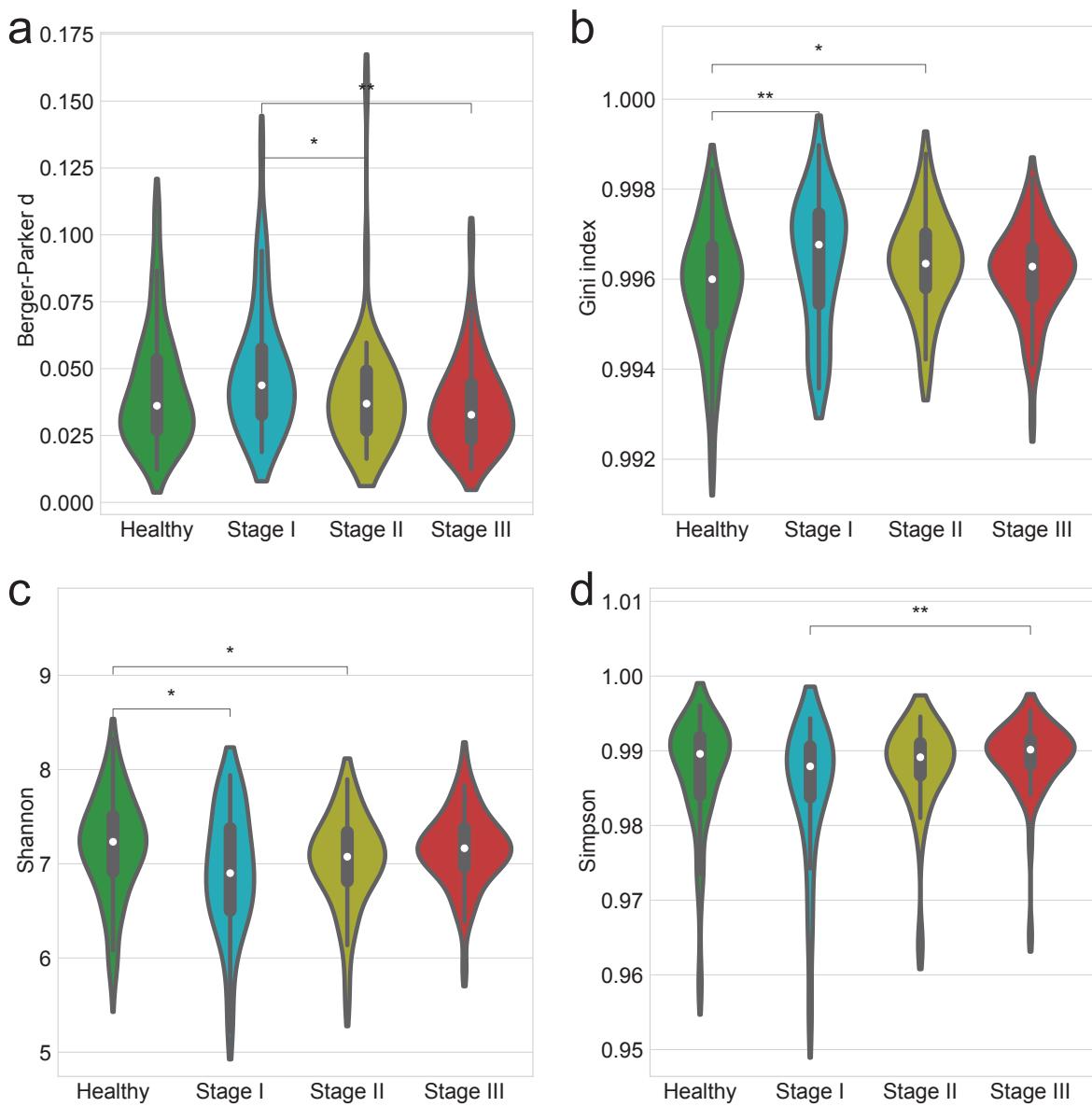


Figure 19: **Alpha-diversity indices account for evenness.**

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 $p \leq 0.01$ (**)

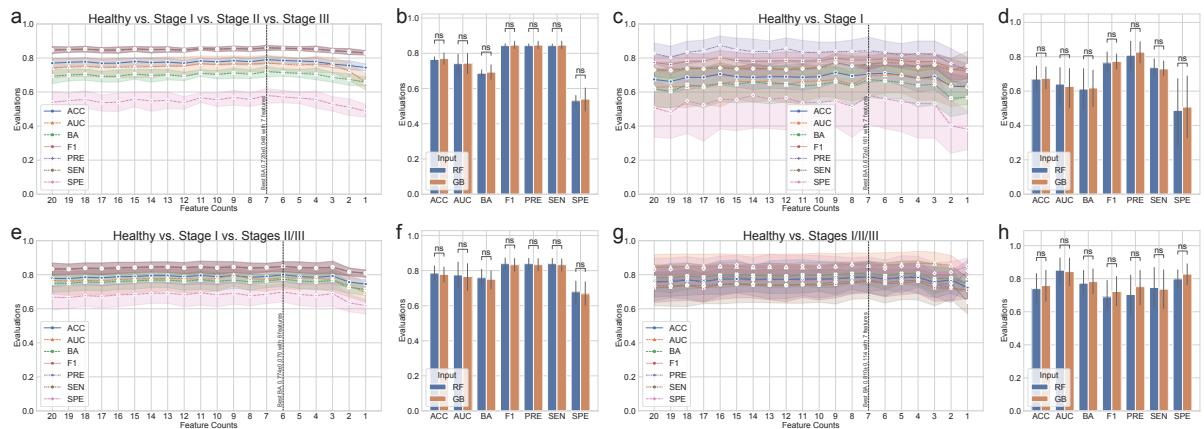


Figure 20: Gradient Boosting classification metrics.

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.

635 **3.4 Discussion**

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

644 Previous research identified the red complex as the primary pathogens of periodontitis (Listgarten,
645 1986): *Porphyromonas gingivalis*, *Tannerella forsythia*, and *Treponema denticola*. Other studies, however,
646 have shown that periodontal pathogens communicate with other bacteria in the salivary microbiome
647 networks to generate dental plaque prior to the pathogenesis and development of periodontitis (Lamont &
648 Jenkinson, 2000; Rosan & Lamont, 2000; Yoshimura, Murakami, Nishikawa, Hasegawa, & Kawaminami,
649 2009).

650 Using subgingival plaque collections, recent researches have suggested a connection between the
651 periodontitis severity and the salivary microbiome compositions (Altabtbaei et al., 2021; Iniesta et al.,
652 2023; Nemoto et al., 2021). Therefore, we have examined the salivary microbiome compositions of
653 patients with multiple severities of periodontitis and periodontally healthy controls, extending on earlier
654 studies.

655 According to our findings, the salivary microbiome compositions have 425 taxa (Figure 13). We
656 computed the alpha-diversity indices to determine the variability within each salivary microbiome
657 composition, including ace (Chao & Lee, 1992), chao1 (Chao, 1984), fisher alpha (Fisher et al., 1943),
658 margalef (Magurran, 2021), observed ASVs (DeSantis et al., 2006), Berger-Parker *d* (Berger & Parker,
659 1970), Gini index (Gini, 1912), Shannon (Weaver, 1963), and Simpson (Simpson, 1949) (Figure 7 and
660 Figure 19). Alpha-diversity indices suggested that the microbial richness of periodontally healthy controls
661 was higher than that of patients with periodontitis (Figure 7a-e and Figure 19). These results are in line with
662 findings with that patients with advanced periodontitis, namely stage II and stage III, have less diversified
663 communities than periodontally healthy controls (Jorth et al., 2014). Recognizing that the periodontitis
664 severity increases the amount of *Porphyromonas gingivalis*, the salivary microbiome compositions from
665 periodontally healthy controls conserved microbial networks dominated by *Streptococcus* spp. (Figure
666 13). *Porphyromonas gingivalis* is one of the known periodontal pathogen that could cause dysbiosys
667 in the salivary microbiomes, suggesting in the pathophysiology of periodontitis. Despite this finding,
668 earlier research found that subgingival microbiome of patients with periodontitis had a greater alpha-
669 diversity index (observed ASVs) than that of healthy controls (Iniesta et al., 2023), might due to the
670 different sampling sites between saliva and subgingival plaque. On the other hand, another research
671 has addressed significant discrepancies in alpha-diversity indices from subgingival plaque, saliva, and
672 tongue biofilms from healthy controls and periodontitis patients, resulting the highest alpha-diversity

673 index in saliva collections (Belstrøm et al., 2021). Moreover, early-stage periodontitis, namely stage I,
674 did not determine statisticall ysiginificant differences in alpha-diversity indices compared to advanced
675 periodontitis, including stage II and stage III (Figure 7a-e). Accordingly, saliva collection of stage I
676 periodontitis may exhibit heterogeneity, indicating a midpoint condition between a healthy state and
677 advanced periodontitis (stage II and stage III). Likewise, gingivitis is often associated with low abundances
678 of the majority of periodontal pathogens, including *Porphyromonas gingivalis*, *Tannerella forsythia*, and
679 *Treponema denticola* (Abusleme et al., 2021). Compared to healthy controls, patients with stage I
680 periodontitis have higher detection rates of *Porphyromonas gingivalis* and *Tannerella forsythia* (Tanner et
681 al., 2006, 2007).

682 Therefore, we calculated beta-diversity indices to analyze the differences between the study partici-
683 pants. The distances for the multiple stages of periodontitis, including stage I, stage II, and stage III, as
684 well as healthy controls (Figure 4g-j and Table 7), suggesting notable differences among the multiple
685 periodontitis severities. In other words, the composition of the salivary microbiome compositions varies
686 depending on the periodontitis stages, so that supporting the findings from a previous study (Iniesta et al.,
687 2023). Taken together that it is nearly impossible to fully restore the attachment level after it has been lost
688 due to the progression and development of periodontitis, the ability to rapidly screen for periodontitis in
689 its early phases using saliva collections would be highly beneficial for effective disease management and
690 treatment.

691 Of the total of 425 taxa in the salivary microbiome composition that have been identified (Figure 13),
692 ANCOM was applied to select 20 taxa as the DAT that indicated notable abundance variation among
693 the periodontitis severities (Figure 8 and Table 5). Three sub-groups were formed from the DAT using
694 hierarchical clustering (Figure 8a). Surprisingly, two of the red complex pathogens (Rôças, Siqueira Jr,
695 Santos, Coelho, & de Janeiro, 2001), *Porphyromonas gingivalis* and *Tannerella forsythia*, were classified
696 in Group 2 and were more prevalent in stage II and stage II periodontitis compared to healthy controls.
697 *Campylobacter showae* was additionally placed in Group 2 of the orange complex pathogens (Gambin et
698 al., 2021). Furthermoe, some of the DAT in Group 2 have reported their crucial roles in pathogenesis
699 and development of periodontitis: *Filifactor alocis* (Aruni et al., 2015), *Treponema putidum* (Wyss et
700 al., 2004), *Tannerella forsythia* (Stafford, Roy, Honma, & Sharma, 2012; W. Zhu & Lee, 2016), and
701 *Prevotella intermedia* (Karched, Bhardwaj, Qudeimat, Al-Khabbaz, & Ellepol, 2022). Taken together,
702 this indicates that DAT in Group 2 is essential to periodontitis. The portion of some Group 1 DAT,
703 including *Peptostreptococcaceae[XI][G-5] saphenum*, *Peptostreptococcaceae[XI][G-6] nodatum*, and
704 *Peptostreptococcaceae[XI][G-9] brachy*, in healthy controls and patients with periodontitis significantly
705 differed, according to earlier research (Lafaurie et al., 2022). These outcomes support our research,
706 implying that Group 1 DAT are also essential to the etiology and progression of periodontitis. However,
707 in contrast to patients with periodontitis, Group 3 DAT, namely *Corynebacterium durum* and *Actinomyces*
708 *graevenitzii*, were enriched in healthy controls, which is consistent with earlier research (Redanz et al.,
709 2021; Nibali et al., 2020).

710 In our correlation analysis (Figure 9), we have discovered strongly negative correlations (coefficient \leq
711 -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 periodontitis 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.

737 **4 Lung microbiome**

738 **4.1 Introduction**

⁷³⁹ **4.2 Materials and methods**

740 **4.3 Results**

741 **4.4 Discussion**

5 Conclusion

⁷⁴² In conclusion, the research described in this doctoral dissertation was conducted to identify significant ...

⁷⁴⁴ In the section 2, I show that

745 References

- 746 Aagaard, K., Ma, J., Antony, K. M., Ganu, R., Petrosino, J., & Versalovic, J. (2014). The placenta harbors
747 a unique microbiome. *Science translational medicine*, 6(237), 237ra65–237ra65.
- 748 Abusleme, L., Hoare, A., Hong, B.-Y., & Diaz, P. I. (2021). Microbial signatures of health, gingivitis,
749 and periodontitis. *Periodontology 2000*, 86(1), 57–78.
- 750 Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., & Pawlowsky-Glahn, V. (2000). Logratio
751 analysis and compositional distance. *Mathematical geology*, 32, 271–275.
- 752 Aja, E., Mangar, M., Fletcher, H., & Mishra, A. (2021). Filifactor alocis: recent insights and advances.
753 *Journal of dental research*, 100(8), 790–797.
- 754 Alelyani, S. (2021). Stable bagging feature selection on medical data. *Journal of Big Data*, 8(1), 11.
- 755 Altabtbaei, K., Maney, P., Ganesan, S. M., Dabdoub, S. M., Nagaraja, H. N., & Kumar, P. S. (2021). Anna
756 karenina and the subgingival microbiome associated with periodontitis. *Microbiome*, 9, 1–15.
- 757 Altingöz, S. M., Kurgan, Ş., Önder, C., Serdar, M. A., Ünlütürk, U., Uyanık, M., ... Günhan, M.
758 (2021). Salivary and serum oxidative stress biomarkers and advanced glycation end products in
759 periodontitis patients with or without diabetes: A cross-sectional study. *Journal of periodontology*,
760 92(9), 1274–1285.
- 761 Alverdy, J., Hyoju, S., Weigerinck, M., & Gilbert, J. (2017). The gut microbiome and the mechanism of
762 surgical infection. *Journal of British Surgery*, 104(2), e14–e23.
- 763 Anderson, M. J. (2014). Permutational multivariate analysis of variance (permanova). *Wiley statsref:
764 statistics reference online*, 1–15.
- 765 Aruni, A. W., Mishra, A., Dou, Y., Chioma, O., Hamilton, B. N., & Fletcher, H. M. (2015). Filifactor
766 alocis—a new emerging periodontal pathogen. *Microbes and infection*, 17(7), 517–530.
- 767 Barlow, G. M., Yu, A., & Mathur, R. (2015). Role of the gut microbiome in obesity and diabetes mellitus.
768 *Nutrition in clinical practice*, 30(6), 787–797.
- 769 Basavaprabhu, H., Sonu, K., & Prabha, R. (2020). Mechanistic insights into the action of probiotics
770 against bacterial vaginosis and its mediated preterm birth: An overview. *Microbial pathogenesis*,
771 141, 104029.
- 772 Belstrøm, D., Constancias, F., Drautz-Moses, D. I., Schuster, S. C., Veleba, M., Mahé, F., & Givskov, M.
773 (2021). Periodontitis associates with species-specific gene expression of the oral microbiota. *npj
774 Biofilms and Microbiomes*, 7(1), 76.
- 775 Berger, W. H., & Parker, F. L. (1970). Diversity of planktonic foraminifera in deep-sea sediments.

- 776 *Science*, 168(3937), 1345–1347.
- 777 Berghella, V. (2012). Universal cervical length screening for prediction and prevention of preterm birth.
778 *Obstetrical & gynecological survey*, 67(10), 653–657.
- 779 Blencowe, H., Cousens, S., Oestergaard, M. Z., Chou, D., Moller, A.-B., Narwal, R., ... others (2012).
780 National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends
781 since 1990 for selected countries: a systematic analysis and implications. *The lancet*, 379(9832),
782 2162–2172.
- 783 Bolstad, A., Jensen, H. B., & Bakken, V. (1996). Taxonomy, biology, and periodontal aspects of
784 *fusobacterium nucleatum*. *Clinical microbiology reviews*, 9(1), 55–71.
- 785 Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... others
786 (2019). Reproducible, interactive, scalable and extensible microbiome data science using qiime 2.
787 *Nature biotechnology*, 37(8), 852–857.
- 788 Breiman, L. (2001). Random forests. *Machine learning*, 45, 5–32.
- 789 Brennan, C. A., & Garrett, W. S. (2019). *Fusobacterium nucleatum*—symbiont, opportunist and
790 *oncobacterium*. *Nature Reviews Microbiology*, 17(3), 156–166.
- 791 Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: improving accuracy of classifier
792 ensembles by using random feature subsets. *Pattern recognition*, 36(6), 1291–1302.
- 793 Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016).
794 Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7),
795 581–583.
- 796 Canakci, V., & Canakci, C. F. (2007). Pain levels in patients during periodontal probing and mechanical
797 non-surgical therapy. *Clinical oral investigations*, 11, 377–383.
- 798 Castaner, O., Goday, A., Park, Y.-M., Lee, S.-H., Magkos, F., Shiow, S.-A. T. E., & Schröder, H. (2018).
799 The gut microbiome profile in obesity: a systematic review. *International journal of endocrinology*,
800 2018(1), 4095789.
- 801 Champagne, C., McNairn, H., Daneshfar, B., & Shang, J. (2014). A bootstrap method for assessing
802 classification accuracy and confidence for agricultural land use mapping in canada. *International
803 Journal of Applied Earth Observation and Geoinformation*, 29, 44–52.
- 804 Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian
805 Journal of statistics*, 265–270.
- 806 Chao, A., & Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the
807 American statistical Association*, 87(417), 210–217.
- 808 Chapple, I. L., Mealey, B. L., Van Dyke, T. E., Bartold, P. M., Dommisch, H., Eickholz, P., ... others
809 (2018). Periodontal health and gingival diseases and conditions on an intact and a reduced
810 periodontium: Consensus report of workgroup 1 of the 2017 world workshop on the classification
811 of periodontal and peri-implant diseases and conditions. *Journal of periodontology*, 89, S74–S84.
- 812 Chen, T., Marsh, P., & Al-Hebshi, N. (2022). Smdi: an index for measuring subgingival microbial
813 dysbiosis. *Journal of dental research*, 101(3), 331–338.
- 814 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

- 854 denticola with human periodontitis lesions. *Oral microbiology and immunology*, 21(5), 269–282.
- 855 Hajishengallis, G. (2015). Periodontitis: from microbial immune subversion to systemic inflammation.
856 *Nature reviews immunology*, 15(1), 30–44.
- 857 Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*,
858 29(2), 147–160.
- 859 Han, Y. W. (2015). Fusobacterium nucleatum: a commensal-turned pathogen. *Current opinion in
860 microbiology*, 23, 141–147.
- 861 Han, Y. W., & Wang, X. (2013). Mobile microbiome: oral bacteria in extra-oral infections and
862 inflammation. *Journal of dental research*, 92(6), 485–491.
- 863 Hartstra, A. V., Bouter, K. E., Bäckhed, F., & Nieuwdorp, M. (2015). Insights into the role of the
864 microbiome in obesity and type 2 diabetes. *Diabetes care*, 38(1), 159–165.
- 865 Helmink, B. A., Khan, M. W., Hermann, A., Gopalakrishnan, V., & Wargo, J. A. (2019). The microbiome,
866 cancer, and cancer therapy. *Nature medicine*, 25(3), 377–388.
- 867 Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2),
868 427–432.
- 869 Hiranmayi, K. V., Sirisha, K., Rao, M. R., & Sudhakar, P. (2017). Novel pathogens in periodontal
870 microbiology. *Journal of Pharmacy and Bioallied Sciences*, 9(3), 155–163.
- 871 Honda, K., & Littman, D. R. (2012). The microbiome in infectious disease and inflammation. *Annual
872 review of immunology*, 30(1), 759–795.
- 873 Honest, H., Forbes, C., Durée, K., Norman, G., Duffy, S., Tsourapas, A., ... others (2009). Screening to
874 prevent spontaneous preterm birth: systematic reviews of accuracy and effectiveness literature with
875 economic modelling. *Health Technol Assess*, 13(43), 1–627.
- 876 Hong, Y. M., Lee, J., Cho, D. H., Jeon, J. H., Kang, J., Kim, M.-G., ... J. K. (2023). Predicting preterm
877 birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1), 21105.
- 878 Huang, R.-Y., Lin, C.-D., Lee, M.-S., Yeh, C.-L., Shen, E.-C., Chiang, C.-Y., ... Fu, E. (2007). Mandibular
879 disto-lingual root: a consideration in periodontal therapy. *Journal of periodontology*, 78(8), 1485–
880 1490.
- 881 Iams, J. D., & Berghella, V. (2010). Care for women with prior preterm birth. *American journal of
882 obstetrics and gynecology*, 203(2), 89–100.
- 883 Ide, M., & Papapanou, P. N. (2013). Epidemiology of association between maternal periodontal
884 disease and adverse pregnancy outcomes—systematic review. *Journal of clinical periodontology*,
885 40, S181–S194.
- 886 Iniesta, M., Chamorro, C., Ambrosio, N., Marín, M. J., Sanz, M., & Herrera, D. (2023). Subgingival
887 microbiome in periodontal health, gingivitis and different stages of periodontitis. *Journal of
888 Clinical Periodontology*, 50(7), 905–920.
- 889 Janda, J. M., & Abbott, S. L. (2007). 16s rRNA gene sequencing for bacterial identification in the diagnostic
890 laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9), 2761–2764.
- 891 Jiang, W., & Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach
892 for estimating the prediction error in microarray classification. *Statistics in medicine*, 26(29),

- 893 5320–5334.
- 894 John, G. K., & Mullin, G. E. (2016). The gut microbiome and obesity. *Current oncology reports*, 18, 895 1–7.
- 896 Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., ... others (2019). 897 Evaluation of 16s rrna gene sequencing for species and strain-level microbiome analysis. *Nature 898 communications*, 10(1), 5029.
- 899 Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N., & Whiteley, M. (2014). Metatranscriptomics 900 of the human oral microbiome during health and disease. *MBio*, 5(2), 10–1128.
- 901 Karched, M., Bhardwaj, R. G., Qudeimat, M., Al-Khabbaz, A., & Ellepola, A. (2022). Proteomic analysis 902 of the periodontal pathogen prevotella intermedia secretomes in biofilm and planktonic lifestyles. 903 *Scientific Reports*, 12(1), 5636.
- 904 Katz, J., Chegini, N., Shiverick, K., & Lamont, R. (2009). Localization of p. gingivalis in preterm delivery 905 placenta. *Journal of dental research*, 88(6), 575–578.
- 906 Kelly, B. J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J. D., Collman, R. G., ... Li, H. (2015). 907 Power and sample-size estimation for microbiome studies using pairwise distances and permanova. 908 *Bioinformatics*, 31(15), 2461–2468.
- 909 Kim, C. H. (2018). Immune regulation by microbiome metabolites. *Immunology*, 154(2), 220–229.
- 910 Kim, E.-H., Kim, S., Kim, H.-J., Jeong, H.-o., Lee, J., Jang, J., ... others (2020). Prediction of chronic 911 periodontitis severity using machine learning models based on salivary bacterial copy number. 912 *Frontiers in Cellular and Infection Microbiology*, 10, 571515.
- 913 Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and 914 bootstrap. *Computational statistics & data analysis*, 53(11), 3735–3745.
- 915 Kinane, D. F., Stathopoulou, P. G., & Papapanou, P. N. (2017). Periodontal diseases. *Nature reviews 916 Disease primers*, 3(1), 1–14.
- 917 Kindinger, L. M., Bennett, P. R., Lee, Y. S., Marchesi, J. R., Smith, A., Caciato, S., ... MacIntyre, 918 D. A. (2017). The interaction between vaginal microbiota, cervical length, and vaginal progesterone 919 treatment for preterm birth risk. *Microbiome*, 5, 1–14.
- 920 Kogut, M. H., Lee, A., & Santin, E. (2020). Microbiome and pathogen interaction with the immune 921 system. *Poultry science*, 99(4), 1906–1913.
- 922 Lafaurie, G. I., Neuta, Y., Ríos, R., Pacheco-Montealegre, M., Pianeta, R., Castillo, D. M., ... oth- 923 ers (2022). Differences in the subgingival microbiome according to stage of periodontitis: A 924 comparison of two geographic regions. *PloS one*, 17(8), e0273523.
- 925 Lamont, R. J., & Jenkinson, H. F. (2000). Subgingival colonization by porphyromonas gingivalis. *Oral 926 Microbiology and Immunology: Mini-review*, 15(6), 341–349.
- 927 Lamont, R. J., Koo, H., & Hajishengallis, G. (2018). The oral microbiota: dynamic communities and 928 host interactions. *Nature reviews microbiology*, 16(12), 745–759.
- 929 Leitich, H., & Kaider, A. (2003). Fetal fibronectin—how useful is it in the prediction of preterm birth? 930 *BJOG: An International Journal of Obstetrics & Gynaecology*, 110, 66–70.
- 931 León, R., Silva, N., Ovalle, A., Chaparro, A., Ahumada, A., Gajardo, M., ... Gamonal, J. (2007).

- 932 Detection of porphyromonas gingivalis in the amniotic fluid in pregnant women with a diagnosis
933 of threatened premature labor. *Journal of periodontology*, 78(7), 1249–1255.
- 934 Lim, J. W., Park, T., Tong, Y. W., & Yu, Z. (2020). The microbiome driving anaerobic digestion and
935 microbial analysis. In *Advances in bioenergy* (Vol. 5, pp. 1–61). Elsevier.
- 936 Lin, H., & Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nature
937 communications*, 11(1), 3514.
- 938 Listgarten, M. A. (1986). Pathogenesis of periodontitis. *Journal of clinical periodontology*, 13(5),
939 418–425.
- 940 Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome
941 medicine*, 8, 1–11.
- 942 Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for
943 rna-seq data with deseq2. *Genome biology*, 15, 1–21.
- 944 Magurran, A. E. (2021). Measuring biological diversity. *Current Biology*, 31(19), R1174–R1177.
- 945 Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically
946 larger than the other. *The annals of mathematical statistics*, 50–60.
- 947 Mayer, E. A., Tillisch, K., Gupta, A., et al. (2015). Gut/brain axis and the microbiota. *The Journal of
948 clinical investigation*, 125(3), 926–938.
- 949 Melguizo-Rodríguez, L., Costela-Ruiz, V. J., Manzano-Moreno, F. J., Ruiz, C., & Illescas-Montes, R.
950 (2020). Salivary biomarkers and their application in the diagnosis and monitoring of the most
951 common oral pathologies. *International journal of molecular sciences*, 21(14), 5173.
- 952 Miller, C. S., Ding, X., Dawson III, D. R., & Ebersole, J. L. (2021). Salivary biomarkers for discriminating
953 periodontitis in the presence of diabetes. *Journal of clinical periodontology*, 48(2), 216–225.
- 954 Morita, T., Yamazaki, Y., Mita, A., Takada, K., Seto, M., Nishinoue, N., ... Maeno, M. (2010). A cohort
955 study on the association between periodontal disease and the development of metabolic syndrome.
956 *Journal of periodontology*, 81(4), 512–519.
- 957 Nemoto, T., Shiba, T., Komatsu, K., Watanabe, T., Shimogishi, M., Shibasaki, M., ... others (2021).
958 Discrimination of bacterial community structures among healthy, gingivitis, and periodontitis
959 statuses through integrated metatranscriptomic and network analyses. *Msystems*, 6(6), e00886–21.
- 960 Nesbitt, M. J., Reynolds, M. A., Shiau, H., Choe, K., Simonsick, E. M., & Ferrucci, L. (2010). Association
961 of periodontitis and metabolic syndrome in the baltimore longitudinal study of aging. *Aging clinical
962 and experimental research*, 22, 238–242.
- 963 Nibali, L., Sousa, V., Davrandi, M., Spratt, D., Alyahya, Q., Dopico, J., & Donos, N. (2020). Differences
964 in the periodontal microbiome of successfully treated and persistent aggressive periodontitis.
965 *Journal of Clinical Periodontology*, 47(8), 980–990.
- 966 Offenbacher, S., Katz, V., Fertik, G., Collins, J., Boyd, D., Maynor, G., ... Beck, J. (1996). Periodontal
967 infection as a possible risk factor for preterm low birth weight. *Journal of periodontology*, 67,
968 1103–1113.
- 969 Papapanou, P. N., Sanz, M., Buduneli, N., Dietrich, T., Feres, M., Fine, D. H., ... others (2018).
970 Periodontitis: Consensus report of workgroup 2 of the 2017 world workshop on the classification of

- 971 periodontal and peri-implant diseases and conditions. *Journal of periodontology*, 89, S173–S182.
- 972 Payne, M. S., Newnham, J. P., Doherty, D. A., Furfaro, L. L., Pendal, N. L., Loh, D. E., & Keelan, J. A.
973 (2021). A specific bacterial dna signature in the vagina of australian women in midpregnancy
974 predicts high risk of spontaneous preterm birth (the predict1000 study). *American journal of
975 obstetrics and gynecology*, 224(2), 206–e1.
- 976 Peirce, J. M., & Alviña, K. (2019). The role of inflammation and the gut microbiome in depression and
977 anxiety. *Journal of neuroscience research*, 97(10), 1223–1241.
- 978 Redanz, U., Redanz, S., Treerat, P., Prakasam, S., Lin, L.-J., Merritt, J., & Kreth, J. (2021). Differential
979 response of oral mucosal and gingival cells to corynebacterium durum, streptococcus sanguinis, and
980 porphyromonas gingivalis multispecies biofilms. *Frontiers in cellular and infection microbiology*,
981 11, 686479.
- 982 Relvas, M., Regueira-Iglesias, A., Balsa-Castro, C., Salazar, F., Pacheco, J., Cabral, C., ... Tomás, I.
983 (2021). Relationship between dental and periodontal health status and the salivary microbiome:
984 bacterial diversity, co-occurrence networks and predictive models. *Scientific reports*, 11(1), 929.
- 985 Rideout, J. R., Caporaso, G., Bolyen, E., McDonald, D., Baeza, Y. V., Alastuey, J. C., ... Sharma, K.
986 (2018, December). *biocore/scikit-bio: scikit-bio 0.5.5: More compositional methods added*. Zenodo.
987 Retrieved from <https://doi.org/10.5281/zenodo.2254379> doi: 10.5281/zenodo.2254379
- 988 Rôças, I. N., Siqueira Jr, J. F., Santos, K. R., Coelho, A. M., & de Janeiro, R. (2001). “red com-
989 plex”(bacteroides forsythus, porphyromonas gingivalis, and treponema denticola) in endodontic
990 infections: a molecular approach. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology,
991 and Endodontology*, 91(4), 468–471.
- 992 Romero, R., Dey, S. K., & Fisher, S. J. (2014). Preterm labor: one syndrome, many causes. *Science*,
993 345(6198), 760–765.
- 994 Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., ... others (2014). The
995 composition and stability of the vaginal microbiota of normal pregnant women is different from
996 that of non-pregnant women. *Microbiome*, 2, 1–19.
- 997 Rosan, B., & Lamont, R. J. (2000). Dental plaque formation. *Microbes and infection*, 2(13), 1599–1607.
- 998 Schwabe, R. F., & Jobin, C. (2013). The microbiome and cancer. *Nature Reviews Cancer*, 13(11),
999 800–812.
- 1000 Sepich-Poore, G. D., Zitvogel, L., Straussman, R., Hasty, J., Wargo, J. A., & Knight, R. (2021). The
1001 microbiome and human cancer. *Science*, 371(6536), eabc4552.
- 1002 Sharma, S., & Tripathi, P. (2019). Gut microbiome and type 2 diabetes: where we are and where to go?
1003 *The Journal of nutritional biochemistry*, 63, 101–108.
- 1004 Simpson, E. (1949). Measurement of diversity. *Nature*, 163.
- 1005 Sotiriadis, A., Papatheodorou, S., Kavvadias, A., & Makrydimas, G. (2010). Transvaginal cervical
1006 length measurement for prediction of preterm birth in women with threatened preterm labor: a
1007 meta-analysis. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International
1008 Society of Ultrasound in Obstetrics and Gynecology*, 35(1), 54–64.
- 1009 Spss, I., et al. (2011). Ibm spss statistics for windows, version 20.0. *New York: IBM Corp*, 440, 394.

- 1010 Stafford, G., Roy, S., Honma, K., & Sharma, A. (2012). Sialic acid, periodontal pathogens and tannerella
1011 forsythia: stick around and enjoy the feast! *Molecular Oral Microbiology*, 27(1), 11–22.
- 1012 Stout, M. J., Conlon, B., Landeau, M., Lee, I., Bower, C., Zhao, Q., ... Mysorekar, I. U. (2013).
1013 Identification of intracellular bacteria in the basal plate of the human placenta in term and preterm
1014 gestations. *American journal of obstetrics and gynecology*, 208(3), 226–e1.
- 1015 Tanner, A. C., Kent Jr, R., Kanasi, E., Lu, S. C., Paster, B. J., Sonis, S. T., ... Van Dyke, T. E. (2007).
1016 Clinical characteristics and microbiota of progressing slight chronic periodontitis in adults. *Journal
1017 of clinical periodontology*, 34(11), 917–930.
- 1018 Tanner, A. C., Paster, B. J., Lu, S. C., Kanasi, E., Kent Jr, R., Van Dyke, T., & Sonis, S. T. (2006).
1019 Subgingival and tongue microbiota during early periodontitis. *Journal of dental research*, 85(4),
1020 318–323.
- 1021 Teles, F., Wang, Y., Hajishengallis, G., Hasturk, H., & Marchesan, J. T. (2021). Impact of systemic
1022 factors in shaping the periodontal microbiome. *Periodontology 2000*, 85(1), 126–160.
- 1023 Thaiss, C. A., Zmora, N., Levy, M., & Elinav, E. (2016). The microbiome and innate immunity. *Nature*,
1024 535(7610), 65–74.
- 1025 Tilg, H., Kaser, A., et al. (2011). Gut microbiome, obesity, and metabolic dysfunction. *The Journal of
1026 clinical investigation*, 121(6), 2126–2132.
- 1027 Tonetti, M. S., Greenwell, H., & Kornman, K. S. (2018). Staging and grading of periodontitis: Framework
1028 and proposal of a new classification and case definition. *Journal of periodontology*, 89, S159–S172.
- 1029 Tringe, S. G., & Hugenholtz, P. (2008). A renaissance for the pioneering 16s rRNA gene. *Current opinion
1030 in microbiology*, 11(5), 442–446.
- 1031 Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., ... others (2017). A
1032 guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological
1033 Reviews*, 92(2), 698–715.
- 1034 Ursell, L. K., Metcalf, J. L., Parfrey, L. W., & Knight, R. (2012). Defining the human microbiome.
1035 *Nutrition reviews*, 70(suppl_1), S38–S44.
- 1036 Vander Haar, E. L., So, J., Gyamfi-Bannerman, C., & Han, Y. W. (2018). Fusobacterium nucleatum and
1037 adverse pregnancy outcomes: epidemiological and mechanistic evidence. *Anaerobe*, 50, 55–59.
- 1038 Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning
1039 research*, 9(11).
- 1040 Weaver, W. (1963). *The mathematical theory of communication*. University of Illinois Press.
- 1041 Whiteside, S. A., Razvi, H., Dave, S., Reid, G., & Burton, J. P. (2015). The microbiome of the urinary
1042 tract—a role beyond infection. *Nature Reviews Urology*, 12(2), 81–90.
- 1043 Witkin, S. (2019). Vaginal microbiome studies in pregnancy must also analyse host factors. *BJOG: An
1044 International Journal of Obstetrics & Gynaecology*, 126(3), 359–359.
- 1045 Wong, T.-T., & Yeh, P.-Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE
1046 Transactions on Knowledge and Data Engineering*, 32(8), 1586–1594.
- 1047 Wyss, C., Moter, A., Choi, B.-K., Dewhurst, F., Xue, Y., Schüpbach, P., ... Guggenheim, B. (2004).
1048 *Treponema putidum* sp. nov., a medium-sized proteolytic spirochaete isolated from lesions of

- 1049 human periodontitis and acute necrotizing ulcerative gingivitis. *International journal of systematic*
1050 *and evolutionary microbiology*, 54(4), 1117–1122.
- 1051 Yaman, E., & Subasi, A. (2019). Comparison of bagging and boosting ensemble machine learning methods
1052 for automated emg signal classification. *BioMed research international*, 2019(1), 9152506.
- 1053 Yang, I., Claussen, H., Arthur, R. A., Hertzberg, V. S., Geurs, N., Corwin, E. J., & Dunlop, A. L. (2022).
1054 Subgingival microbiome in pregnancy and a potential relationship to early term birth. *Frontiers in*
1055 *cellular and infection microbiology*, 12, 873683.
- 1056 Yoshimura, F., Murakami, Y., Nishikawa, K., Hasegawa, Y., & Kawaminami, S. (2009). Surface
1057 components of *porphyromonas gingivalis*. *Journal of periodontal research*, 44(1), 1–12.
- 1058 Zhang, C.-Z., Cheng, X.-Q., Li, J.-Y., Zhang, P., Yi, P., Xu, X., & Zhou, X.-D. (2016). Saliva in the
1059 diagnosis of diseases. *International journal of oral science*, 8(3), 133–137.
- 1060 Zhu, W., & Lee, S.-W. (2016). Surface interactions between two of the main periodontal pathogens:
1061 *Porphyromonas gingivalis* and *tannerella forsythia*. *Journal of periodontal & implant science*,
1062 46(1), 2–9.
- 1063 Zhu, X., Han, Y., Du, J., Liu, R., Jin, K., & Yi, W. (2017). Microbiota-gut-brain axis and the central
1064 nervous system. *Oncotarget*, 8(32), 53829.

1065

Acknowledgments

1066 I would like to disclose my earnest appreciation for my advisor, Professor Semin Lee, who provided
 1067 solicitous supervision and cherished opportunities throughout the course of my research. His advice and
 1068 consultation encouraged me to become as a researcher and to receive all humility and gentleness. I am
 1069 also grateful to all of my committee members, Professor AAA, Professor BBB, Professor CCC, and
 1070 Professor DDD, for their critical and meaningful mentions and suggestions.

1071 I extend my deepest gratitude to my Lord, *the Flying Spaghetti Monster*, His Noodly Appendage
 1072 has guided me through the twist and turns of this academic journey. His presence, ever comforting and
 1073 mysterious, has been a source of strength and humor during both highs and lows. In moments of doubt, I
 1074 found solace in the belief that you were there, gently reminding me to keep faith in the process. His Holy
 1075 Noodle has nourished my mind, and for that, I am truly overwhelmed. May His Holy Noodle continue to
 1076 guide me in all my future endeavors. R’Amen.

1077 (Professors)

1078 I would like to extend my heartfelt gratitude to my colleagues of the Computational Biology Lab @
 1079 UNIST, whose collaboration, friendship, brotherhood, and support have been an invaluable part of my
 1080 journey. Your willingness to share insights, engage in thoughtful discussions, and offer encouragement
 1081 during the challenging moments of research has significantly shaped my academic experience. The
 1082 camaraderie in Computational Biology Lab made even the most demanding days more enjoyable, and I
 1083 am deeply grateful for the collaborative environment we created together. I appreciate you for standing
 1084 by my side throughout this Ph.D. journey.

1085 I would like to express my heartfelt gratitude to my family, whose unwavering support has been the
 1086 foundation of everything I have achieved. Your love, encouragement, and belief in me have sustained me
 1087 through every challenge, and I could not have come this far without you. From your words of wisdom to
 1088 your patience and understanding, each of you has played a vital role in helping me navigate this journey.
 1089 The strength and comfort I have drawn from our family bond have been my greatest source of resilience.
 1090 Your presence, both near and far, has filled my life with warmth and motivation. I am deeply grateful for
 1091 your unconditional love and for always being there when I needed you the most. Thank you for being my
 1092 constant source of strength and inspiration.

1093 I am incredibly pleased to my friends, especially my GSHS alumni (○망특), for their unwavering
 1094 support and encouragement throughout this journey. The bonds we formed back in our school days have
 1095 only grown stronger over the years, and I am fortunate to have had such loyal and understanding friends
 1096 by my side. Your constant words of motivation, and even moments of levity during stressful times have
 1097 helped keep me grounded. Whether it was a late-night conversations, a shared laugh, or a simple message
 1098 of reassurance, you all have played a vital role in keeping me focused and motivated. I am relieved for the
 1099 ways you celebrated each small achievement with me and how you patiently listened to my worries. The
 1100 memories of our shared past provided me with comfort and a sense of stability when the road ahead felt
 1101 uncertain. I could not have reached this point without the love and friendship that you all have generously
 1102 given. Each of your, in your unique way, has contributed to this dissertation, even if indirectly, and for

1103 that, I am forever beholden. I look forward to continuing our friendship as we all grow in our individual
1104 paths, knowing that the support we share is something truly special.

1105 (Girlfriend)

1106 I would like to express my sincere gratitude to the amazing members of my animal protection groups,
1107 DRDR (두루두루) and UNIMALS (유니멀스), whose dedication and compassion have been a constant
1108 source of motivation. Your unwavering commitment to improving the lives of animals has inspired me
1109 throughout this journey. I am also thankful for the beautiful cats we have cared for, whose presence
1110 brought both joy and purpose to our allegiance. Their playful spirits and gentle companionship served as
1111 daily reminders of why we continue to fight for animal rights. The bond we share, both with each other
1112 and with the animals we protect, has enriched my life in countless ways. I appreciate you all again for
1113 your support, dedication, and for being part of this meaningful cause.

1114 I would like to express my deepest gratitude to everyone I have had the honor of meeting throughout
1115 this journey. Your kindness, encouragement, and support have carried me through both the challenging
1116 and rewarding moments of my life. Whether through a kind word, thoughtful advice, or simply being
1117 there when I needed it most, your presence has made all the difference. I am incredibly fortunate to have
1118 received such generosity and warmth from those around me, and I do not take it for granted. Every act
1119 of kindness, no matter how big or small, has been a source of strength and motivation for me. To all
1120 my friends, colleagues, mentors, and beloved ones, thank you for your unwavering support. I am truly
1121 grateful for each of you, and your kindness has left an indelible mark on my journey.

1122 My Lord, *the Flying Spaghetti Monster*,
1123 give us grace to accept with serenity the things that cannot be changed,
1124 courage to change the things that should be changed,
1125 and the wisdom to distinguish the one from the other.

1126
1127 Glory be to *the Meatball*, to *the Sauce*, and to *the Holy Noodle*.
1128 As it was in the beginning, is now, and ever shall be.

1129 R'Amen.

