

<sup>1</sup>

# Doctoral Thesis

<sup>2</sup>

## Microbiota in Human Diseases

<sup>3</sup>

Jaewoong Lee

<sup>4</sup>

Department of Biomedical Engineering

<sup>5</sup>

Ulsan National Institute of Science and Technology

<sup>6</sup>

2025

<sup>7</sup>

# Microbiota in Human Diseases

<sup>8</sup>

Jaewoong Lee

<sup>9</sup>

Department of Biomedical Engineering

<sup>10</sup>

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 (Colon) Setion 4...

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 . . . . .	8
28	2.1	Introduction . . . . .	8
29	2.2	Materials and methods . . . . .	10
30	2.2.1	Study design and study participants . . . . .	10
31	2.2.2	Clinical data collection and grouping . . . . .	10
32	2.2.3	Salivary microbiome sample collection . . . . .	10
33	2.2.4	16s rRNA gene sequencing . . . . .	10
34	2.2.5	Bioinformatics analysis . . . . .	11
35	2.2.6	Data and code availability . . . . .	11
36	2.3	Results . . . . .	12
37	2.3.1	Overview of clinical information . . . . .	12
38	2.3.2	Comparison of salivary microbiomes composition . . . . .	12
39	2.3.3	Random forest classification to predict PTB risk . . . . .	12
40	2.4	Discussion . . . . .	20
41	3	Random forest prediction model for periodontitis statuses based on the salivary microbiomes	22
42	3.1	Introduction . . . . .	22
43	3.2	Materials and methods . . . . .	24
44	3.2.1	Study participants enrollment . . . . .	24
45	3.2.2	Periodontal clinical parameter diagnosis . . . . .	24
46	3.2.3	Saliva sampling and DNA extraction procedure . . . . .	26
47	3.2.4	Bioinformatics analysis . . . . .	26
48	3.2.5	Data and code availability . . . . .	27
49	3.3	Results . . . . .	29

50	3.3.1	Summary of clinical information and sequencing data . . . . .	29
51	3.3.2	Diversity indices reveal differences among the periodontitis severities .	29
52	3.3.3	DAT among multiple periodontitis severities and their correlation . . .	29
53	3.3.4	Classification of periodontitis severities by random forest models . . .	30
54	3.4	Discussion . . . . .	51
55	4	Metagenomic signature analysis of Korean colorectal cancer . . . . .	55
56	4.1	Introduction . . . . .	55
57	4.2	Materials and methods . . . . .	57
58	4.2.1	Study participants enrollment . . . . .	57
59	4.2.2	DNA extraction procedure . . . . .	57
60	4.2.3	Bioinformatics analysis . . . . .	57
61	4.2.4	Data and code availability . . . . .	58
62	4.3	Results . . . . .	59
63	4.3.1	Summary of clinical characteristics . . . . .	59
64	4.3.2	Gut microbiome compositions . . . . .	59
65	4.3.3	Diversity indices . . . . .	59
66	4.3.4	DAT selection . . . . .	59
67	4.3.5	Pathway prediction . . . . .	59
68	4.4	Discussion . . . . .	61
69	5	Conclusion . . . . .	62
70	References . . . . .		63
71	Acknowledgments . . . . .		78

72

## List of Figures

73	1	DAT volcano plot . . . . .	14
74	2	Salivary microbiome compositions over DAT . . . . .	15
75	3	Random forest-based PTB prediction model . . . . .	16
76	4	Diversity indices . . . . .	17
77	5	PROM-related DAT . . . . .	18
78	6	Validation of random forest-based PTB prediction model . . . . .	19
79	7	Diversity indices . . . . .	37
80	8	Differentially abundant taxa (DAT) . . . . .	38
81	9	Correlation heatmap . . . . .	39
82	10	Random forest classification metrics . . . . .	40
83	11	Random forest classification metrics from external datasets . . . . .	41
84	12	Rarefaction curves for alpha-diversity indices . . . . .	42
85	13	Salivary microbiome compositions in the different periodontal statuses . . . . .	43
86	14	Correlation plots for differentially abundant taxa . . . . .	44
87	15	Clinical measurements by the periodontitis statuses . . . . .	45
88	16	Number of read counts by the periodontitis statuses . . . . .	46
89	17	Proportion of DAT . . . . .	47

90	18	Random forest classification metrics with the full microbiome compositions and ANCOM-selected DAT compositions . . . . .	48
91			
92	19	Alpha-diversity indices account for evenness . . . . .	49
93	20	Gradient Boosting classification metrics . . . . .	50

## List of Tables

95	1	Confusion matrix . . . . .	6
96	2	Standard clinical information of study participants . . . . .	13
97	3	Clinical characteristics of the study participants . . . . .	32
98	4	Feature combinations and their evaluations . . . . .	33
99	5	List of DAT among the periodontally healthy and periodontitis stages . . . . .	34
100	6	Feature the importance of taxa in the classification of different periodontal statuses. . . . .	35
101	7	Beta-diversity pairwise comparisons on the periodontitis statuses . . . . .	36
102	8	Clinical characteristics of the study participants . . . . .	60

103

## List of Abbreviations

- 104 **ACC** Accuracy  
105 **ASV** Amplicon sequence variant  
106 **AUC** Area-under-curve  
107 **BA** Balanced accuracy  
108 **C-section** Cesarean section  
109 **DAT** Differentially abundant taxa  
110 **F1** F1 score  
111 **Faith PD** Faith's phylogenetic diversity  
112 **FTB** Full-term birth  
113 **GA** Gestational age  
114 **MSI** Microsatellite instability  
115 **MSs** Microsatellite stable  
116 **MWU test** Mann-Whitney U-test  
117 **OS** Overall survival  
118 **PRE** Precision  
119 **PROM** Prelabor rupture of membrane  
120 **PTB** Preterm birth  
121 **ROC curve** Receiver-operating characteristics curve  
122 **rRNA** Ribosomal RNA  
123 **SD** Standard deviation  
124 **SEN** Sensitivity  
125 **SPE** Specificity  
126 **t-SNE** t-distributed stochastic neighbor embedding

<sup>127</sup> **1 Introduction**

<sup>128</sup> The microbiome refers to the complex community of microorganisms, including bacteria, viruses, fungi,  
<sup>129</sup> and other microbes, that inhabit various environment within living organisms (Ursell, Metcalf, Parfrey,  
<sup>130</sup> & Knight, 2012; Gilbert et al., 2018). In humans, the microbiome plays a crucial role in maintaining  
<sup>131</sup> health (Lloyd-Price, Abu-Ali, & Huttenhower, 2016), influencing processes such as digestion (Lim, Park,  
<sup>132</sup> Tong, & Yu, 2020), immune response (Thaiss, Zmora, Levy, & Elinav, 2016; Kogut, Lee, & Santin, 2020;  
<sup>133</sup> C. H. Kim, 2018), and even mental health (Mayer, Tillisch, Gupta, et al., 2015; X. Zhu et al., 2017;  
<sup>134</sup> X. Chen, D'Souza, & Hong, 2013). These microbial communities are not static nor constant, but rather  
<sup>135</sup> dynamic ecosystem that interacts with their host and respond to environmental changes. Recent studies  
<sup>136</sup> have revealed that imbalances in the microbiome, known as dysbiosis, can contribute to a wide range of  
<sup>137</sup> diseases, including obesity (John & Mullin, 2016; Tilg, Kaser, et al., 2011; Castaner et al., 2018), diabetes  
<sup>138</sup> (Barlow, Yu, & Mathur, 2015; Hartstra, Bouter, Bäckhed, & Nieuwdorp, 2015; Sharma & Tripathi, 2019),  
<sup>139</sup> infections (Whiteside, Razvi, Dave, Reid, & Burton, 2015; Alverdy, Hyoju, Weigerinck, & Gilbert, 2017),  
<sup>140</sup> inflammatory conditions (Francescone, Hou, & Grivennikov, 2014; Peirce & Alviña, 2019; Honda &  
<sup>141</sup> Littman, 2012), and cancers (Helmink, Khan, Hermann, Gopalakrishnan, & Wargo, 2019; Cullin, Antunes,  
<sup>142</sup> Straussman, Stein-Thoeringer, & Elinav, 2021; Sepich-Poore et al., 2021; Schwabe & Jobin, 2013). Thus,  
<sup>143</sup> understanding the composition of the human microbiomes is essential for developing new therapeutic  
<sup>144</sup> approaches that target these microbial populations to promote health and prevent diseases.

<sup>145</sup> The microbiome participates a crucial role in overall health, influencing not only digestion and immune  
<sup>146</sup> function but also systemic and neurological processes through the brain-gut axis (Martin, Osadchiy,  
<sup>147</sup> Kalani, & Mayer, 2018; Aziz & Thompson, 1998; R. Li et al., 2024). The gut microbiota interact with  
<sup>148</sup> the host through metabolic byproducts, immune signaling, and the production of neurotransmitters, *e.g.*  
<sup>149</sup> serotonin and dopamine, which are essential for brain function and cognition. Disruptions in microbial  
<sup>150</sup> composition, known as dysbiosis, have been linked to various diseases, including inflammatory bowel  
<sup>151</sup> disease (Sultan et al., 2021; Baldelli, Scaldaferrri, Putignani, & Del Chierico, 2021), obesity (Kang et al.,  
<sup>152</sup> 2022; Hamjane, Mechita, Nourouti, & Barakat, 2024; Pezzino et al., 2023), diabetes (Cai et al., 2024;  
<sup>153</sup> X. Li et al., 2021; Y. Li et al., 2023), and cardiovascular diseases (Manolis, Manolis, Melita, & Manolis,  
<sup>154</sup> 2022; Tian et al., 2021). Furthermore, the brain-gut axis, a bidirectional communication system between  
<sup>155</sup> the gut microbiome composition and the central nervous system, has been implicated in mental disorders,  
<sup>156</sup> *e.g.* anxiety disorder, depressive disorder, and neurodegenerative diseases. Emerging evidence suggested  
<sup>157</sup> that alterations in the host microbiome can influence mood, cognitive function, and even behavior through  
<sup>158</sup> immune modulation, vagus nerve signaling, and microbial metabolites. These findings highlight the  
<sup>159</sup> microbiome as a critical factor in maintaining host health and suggest that targeted interventions, namely  
<sup>160</sup> probiotics, antibiotics, dietary modification, and microbiome-based therapies, may hold promise for  
<sup>161</sup> improving both physical and mental comfort. Hence, understanding the microbial effects could lead to  
<sup>162</sup> novel therapeutic strategies for a wide range of health conditions.

<sup>163</sup> 16S ribosomal RNA (rRNA) gene sequencing is one of the most extensively applied methods for  
<sup>164</sup> characterizing microbial communities by targeting the conserved 16S rRNA gene, which contains both

165 highly conserved and variable regions in bacteria (Tringe & Hugenholtz, 2008; Janda & Abbott, 2007).  
166 The conserved regions enable universal primer binding, while the variable regions provide the specificity  
167 needed to differentiate microbial taxa. Among these regions, the V3-V4 region is frequently selected for  
168 sequencing due to its balance between phylogenetic resolution and sequencing efficiency (Johnson et al.,  
169 2019; López-Aladid et al., 2023). Therefore, the V3-V4 region offers sufficient variability to classify a  
170 wide range of bacteria taxa while maintaining compatibility with widely used sequencing platforms.

171 On the other hand, PathSeq is a computational pipeline designed for the identification and analysis  
172 of microbial sequences within short-read human sequencing data, such as next-generation sequencing  
173 (Kostic et al., 2011; Walker et al., 2018). PathSeq's scalable and effective processing of massive amounts  
174 of sequencing data allows large-scale microbial profiling possible. PathSeq workflow consists of two  
175 main phases: a subtractive phase and an analytic phase. The subtractive phase is removing human-derived  
176 reads by aligning them to a human reference genome; and, the analytic phase is mapping remaining reads  
177 to microbial reference databases, not only bacterial reference genome, but also archaeal, fungal, and viral  
178 reference genomes. This approach allows for the comprehensive detection of microbiome compositions,  
179 without a requirement for targeted amplification. PathSeq presents a more comprehensive and objective  
180 evaluation of microbiome compositions than conventional microbiome profiling techniques including 16S  
181 rRNA gene sequencing, capturing an assortment of microbial species beyond bacteria. Therefore, PathSeq  
182 is an effective instrument for metagenomic research, infectious disease study, and microbiome analysis in  
183 environmental and clinical contexts because of its capacity to operate with complex sequencing datasets  
184 (Ojesina et al., 2013; Park et al., 2024; Tejeda et al., 2021).

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

194 Differentially abundant taxa (DAT) detection is a key analytical approach in microbiome study to  
195 identify microbial taxa that significantly differ in abundance between distinct study participant groups.  
196 This DAT detection method is particularly valuable for understanding how microbial communities vary  
197 across different conditions, such as disease states, environmental factors, and/or experimental treatments.  
198 Various statistical and computational techniques, *e.g.* LEfSe (Segata et al., 2011), DESeq2 (Love, Huber,  
199 & Anders, 2014), ANCOM (Lin & Peddada, 2020), and ANCOM-BC (Lin, Eggesbø, & Peddada,  
200 2022; Lin & Peddada, 2024), are commonly used to assess differential abundance while accounting for  
201 compositional and sparsity-related challenges in microbiome composition data (Swift, Cresswell, Johnson,  
202 Stilianoudakis, & Wei, 2023; Cappellato, Baruzzo, & Di Camillo, 2022). Thus, identifying DAT can  
203 provide insights into microbial biomarkers associated with specific health conditions or disease statuses,

enabling potential applications in diagnostics and therapeutics. However, due to the nature of microbiome composition data and the influence of sequencing depth, appropriate normalization and statistically adjustments are necessary to ensure reliable and stable detection of differentially abundant microbes (Xia, 2023; Pan, 2021). Integrating DAT detection analysis with functional profiling further enhances our understanding of the biological significance of microbial shifts or dysbiosis. As microbiome research advances, improving methodologies for DAT selection remains essential for uncovering meaningful microbial association and their potential roles in human diseases.

Classification is one of the supervised machine learning techniques used to categorized data into predefined classes based on features within the data (Kotsiantis, Zaharakis, & Pintelas, 2006; Sen, Hajra, & Ghosh, 2020). In other words, the method learns the relationship between input features and their corresponding output classes through the process of training a classification model using labeled data. Classification models are essential for advising choices in a wide range of applications, including medical diagnostics (Omondiagbe, Veeramani, & Sidhu, 2019). Thus, researchers could uncover sophisticated connections in input features and corresponding classes and produce reliable prediction by utilizing machine learning classification.

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

Evaluating the performance of a machine learning classification model is essential to ensure its reliability and effectiveness in real-world solutions and applications (Novaković, Veljović, Ilić, Papić, & Tomović, 2017; Hossin & Sulaiman, 2015; Hand, 2012). A confusion matrix is a tabular representation of predictions of classification, showing the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (Table 1). From this matrix, evaluations can be derived: accuracy (ACC; Equation 1), balanced accuracy (BA; Equation 2), F1 score (F1; Equation 3), sensitivity (SEN; Equation 4), specificity (SPE; Equation 5), and precision (PRE; Equation 6). These metrics are in [0, 1] range and high metrics are good metrics. The confusion matrix also helps in identifying specific types of errors, such as a tendency to produce false positive or false negatives, offering valuable insights for improving the classification model. By combining the confusion matrix with other evaluation metrics, researchers can comprehensively assess the classification metrics and refine it for real-world solutions and applications.

The receiver-operating characteristics (ROC) curve is a graphical representation used to evaluate the performance of a classification model by plotting the sensitivity against (1-specificity) at multiple threshold setting (Gonçalves, Subtil, Oliveira, & de Zea Bermudez, 2014; Obuchowski & Bullen, 2018; Centor, 1991). The ROC curve illustrates the trade-off between detecting true positives while minimizing false positives, suggesting determining the optimal decision threshold for classification. A key metric

243 derived from the ROC curve is the area-under-curve (AUC), which quantifies overall ability of the  
244 classification model to discriminate between positive and negative predictions. An AUC value of 0.5  
245 indicates a model performing no better than random chance, while value closer to 1.0 suggests high  
246 predictive accuracy. Thus, by analyzing the AUC value of the ROC curve, researchers can compare  
247 different models and select the better classification model that offers the best balance between sensitivity  
248 and specificity for a given application.

249 (Limitation & Novelty)

Table 1: Confusion matrix

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

250

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

251

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

252

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

253

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

254

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

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

255 **2 Predicting preterm birth using random forest classifier in salivary mi-**  
256 **crobiome**

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

258 Hong, Y. M., **Lee, Jaewoong**, Cho, D. H., Jeon, J. H., Kang, J., Kim, M. G., ... & Kim, J. K. (2023).  
259 Predicting preterm birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1),  
260 21105.

261 **2.1 Introduction**

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

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

279 Preterm prelabor rupture of membranes (PROM) brought on by gestational inflammation and infection  
280 contribute to about 70% of PTB cases (Romero, Dey, & Fisher, 2014). Nevertheless, as antibiotics and  
281 anti-inflammatory therapeutic strategies were ineffective to decrease PTB occurrence rates, the pathology  
282 of PTB has not been entirely elucidated by inflammatory and infectious pathways (Romero, Hassan, et al.,  
283 2014). Recent researches on maternal microbiomes were beginning to examine unidentified connections  
284 of PTB as a consequence of developmental processes in molecular biological technology (Fettweis et al.,  
285 2019).

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

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

294 Multiple unfavorable birth outcomes, including PROM and PTB, have been linked to periodontitis  
295 as an independence risk factor, according to numerous epidemiological researches (Offenbacher et al.,  
296 1996). It is expected that the oral microbiome will be able to explain additional hematogenous pathways  
297 in light of these precedents; however, the oral microbiome composition of fetuses is limited understood.

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

301 **2.2 Materials and methods**

302 **2.2.1 Study design and study participants**

303 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  
304 Declaration of Helsinki (Goodyear, Krleza-Jeric, & Lemmens, 2007). The Institutional Review Board  
305 authorized this study (IRB file No. 2019-01-024). Participants who were admitted for elective cesarean  
306 sections (C-sections) or induction births, as well as those who had written informed consent obtained  
307 with premature labor or PROM, were eligible.  
308

309 **2.2.2 Clinical data collection and grouping**

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

- 312 • maternal age at delivery
- 313 • diabetes mellitus
- 314 • hypertension
- 315 • overweight and obesity
- 316 • C-section
- 317 • history PROM or PTB
- 318 • gestational week on delivery
- 319 • birth weight
- 320 • sex

321 **2.2.3 Salivary microbiome sample collection**

322 Salivary microbiome samples were collected 24 hours before to delivery using mouthwash. The standard  
323 methods of sterilizing were performed. Medical experts oversaw each stage of the sample collecting  
324 procedure. Participants received instruction not to eat, drink, or brush their teeth for 30 minutes before  
325 sampling salivary microbiome. Saliva samples were gathered by washing the mouth for 30 seconds with  
326 12 mL of a mouthwash solution (E-zen Gargle, JN Pharm, Pyeongtaek, Gyeonggi, Korea). The samples  
327 were tagged with the anonymous ID for each participant and kept at 4 °C until they underwent further  
328 processing. Genomic DNA was extracted using an ExgeneTM Clinic SV kit (GeneAll Biotechnology,  
329 Seoul, Korea) following with the manufacturer instructions and store at -20 °C.

330 **2.2.4 16s rRNA gene sequencing**

331 Salivary microbiome samples were transported to the Department of Biomedical Engineering of the  
332 Ulsan National Institute of Science and Technology . 16S rRNA sequencing was then carried out using a  
333 commissioned Illumina MiSeq Reagent Kit v3 (Illumina, San Diego, CA, USA). Library methods were  
334 utilized to amplify the V3-V4 areas. 300 base-pair paired-end reads were produced by sequencing the

335 pooled library using a v3  $\times$ 600 cycle chemistry after the samples had been diluted to a final concentration  
336 of 6 pM with a 20% PhiX control.

337 **2.2.5 Bioinformatics analysis**

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

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

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

353 The taxa for predicting PTB using salivary microbiome data were determined using a random forest  
354 classifier (Breiman, 2001). Through stratified *k*-fold cross-validation (*k* = 5) that preserves the existence  
355 rate of PTB and FTB participants, consistency and trustworthy classification were ensured (Wong & Yeh,  
356 2019).

357 **2.2.6 Data and code availability**

358 All sequences from the 59 study participants have been published to the Sequence Read Archives  
359 (project ID PRJNA985119): <https://dataview.ncbi.nlm.nih.gov/object/PRJNA985119>. Docker  
360 image that employed throughout this study is available in the DockerHub: [https://hub.docker.com/r/fumire/helixco\\_premature](https://hub.docker.com/r/fumire/helixco_premature). Every code used in this study can be found on GitHub: [https://github.com/CompbioLabUnist/Helixco\\_Premature](https://github.com/CompbioLabUnist/Helixco_Premature).

363 **2.3 Results**

364 **2.3.1 Overview of clinical information**

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

371 **2.3.2 Comparison of salivary microbiomes composition**

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

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

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

383 **2.3.3 Random forest classification to predict PTB risk**

384 To classify PTB according to DAT, random forest classifiers were constructed. The nine most significant  
385 DAT were used to obtain the best BA ( $0.765 \pm 0.071$ ; Figure 3a). Moreover, random forest classification  
386 model determined each DAT's importance (Figure 3b). We conducted a validation procedure on nine  
387 twin pregnancies that were excluded in the initial study design in order to confirm the reliability and  
388 dependability of our random forest-based PTB prediction model (Figure 6). Comparable to the PTB  
389 prediction model on the 59 initial singleton study participants, the validation classification on PTB risk of  
390 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  $\chi^2$ -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

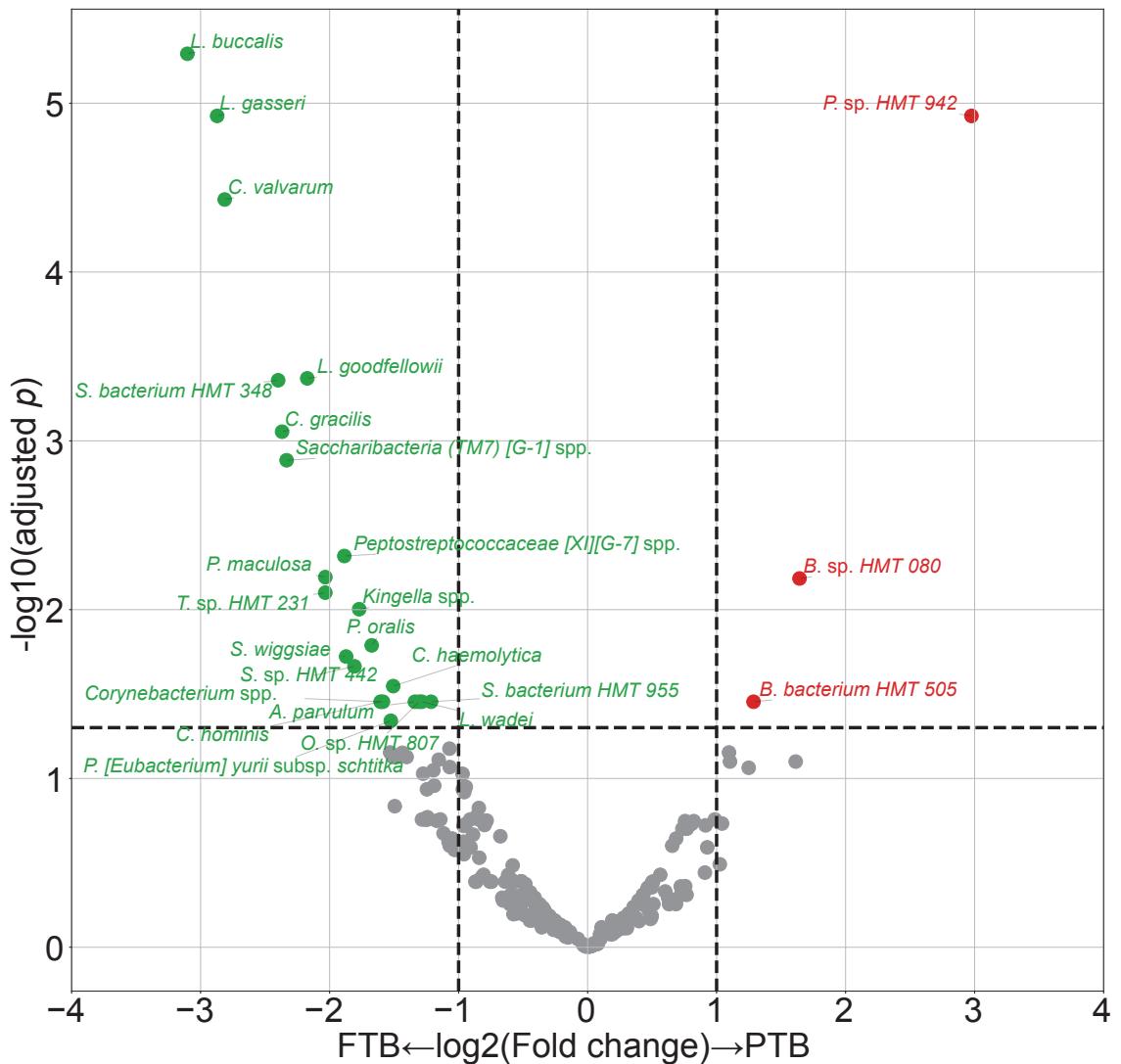


Figure 1: DAT volcano plot.

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

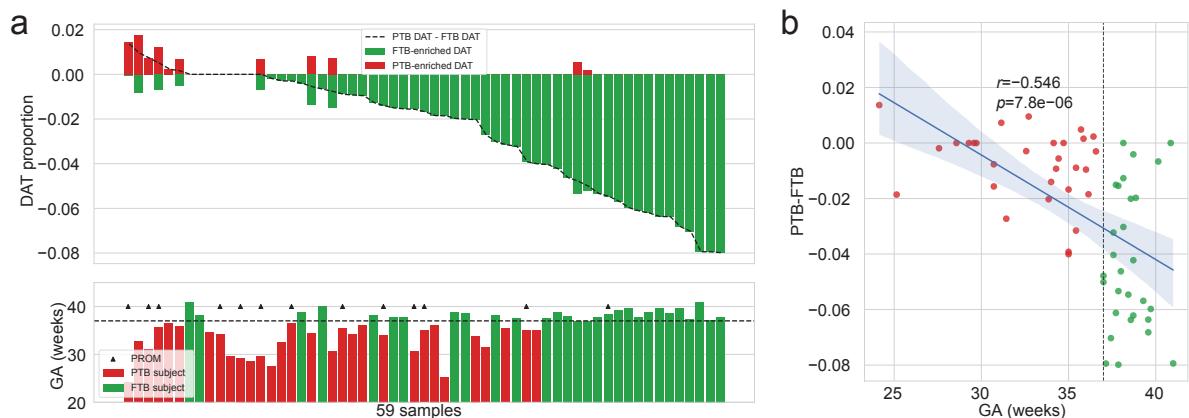


Figure 2: **Salivary microbiome compositions over DAT.**

**(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.

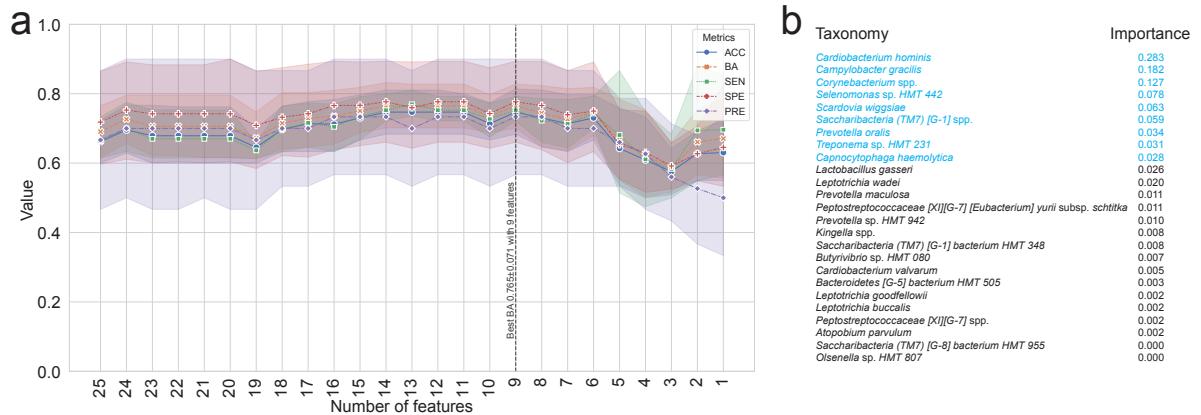


Figure 3: **Random forest-based PTB prediction model.**

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

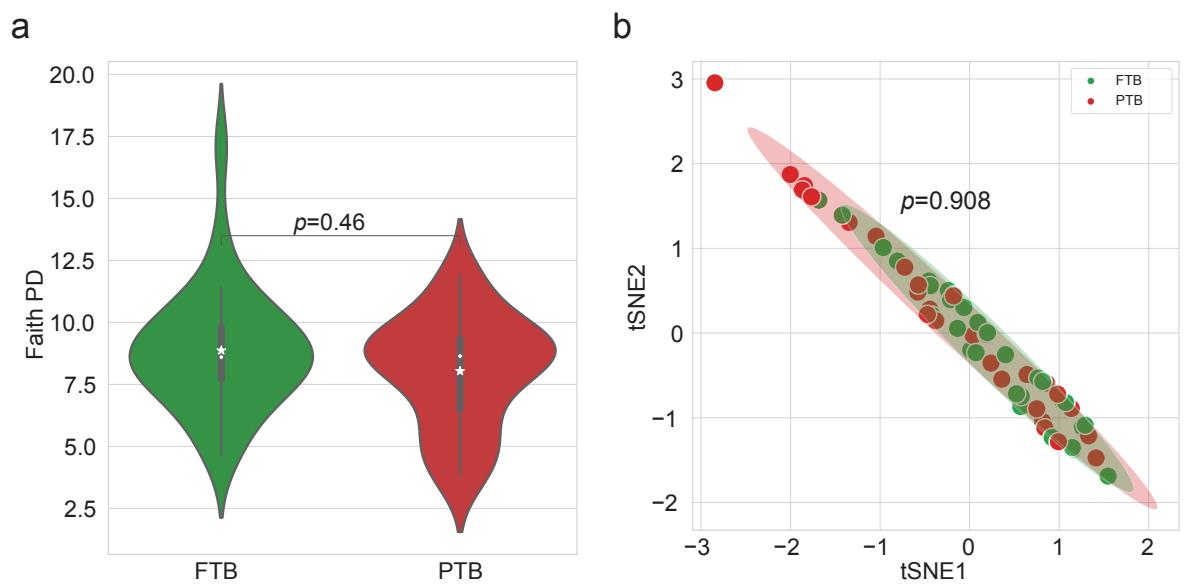


Figure 4: **Diversity indices.**

**(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  $p = 0.908$ )

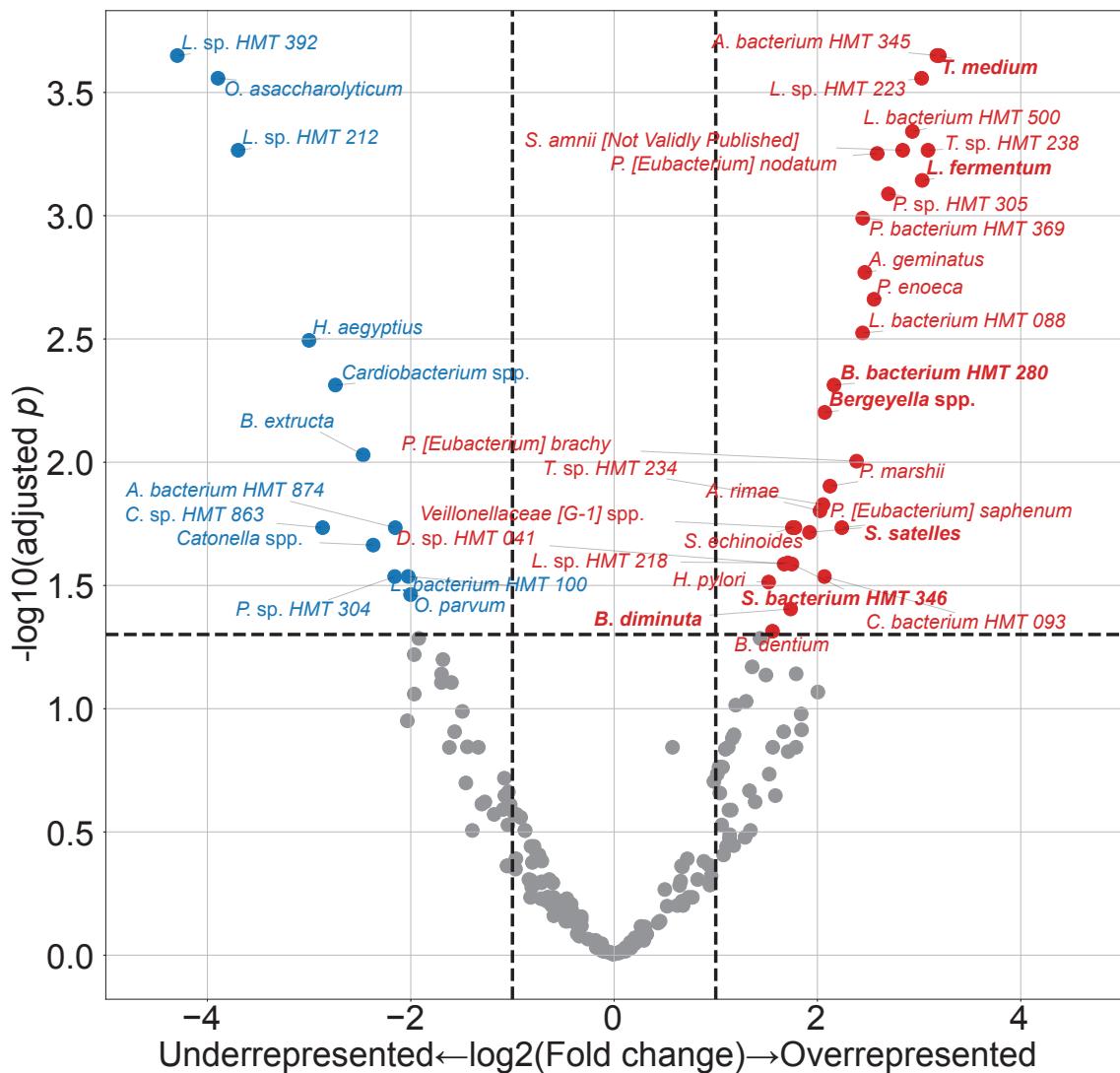
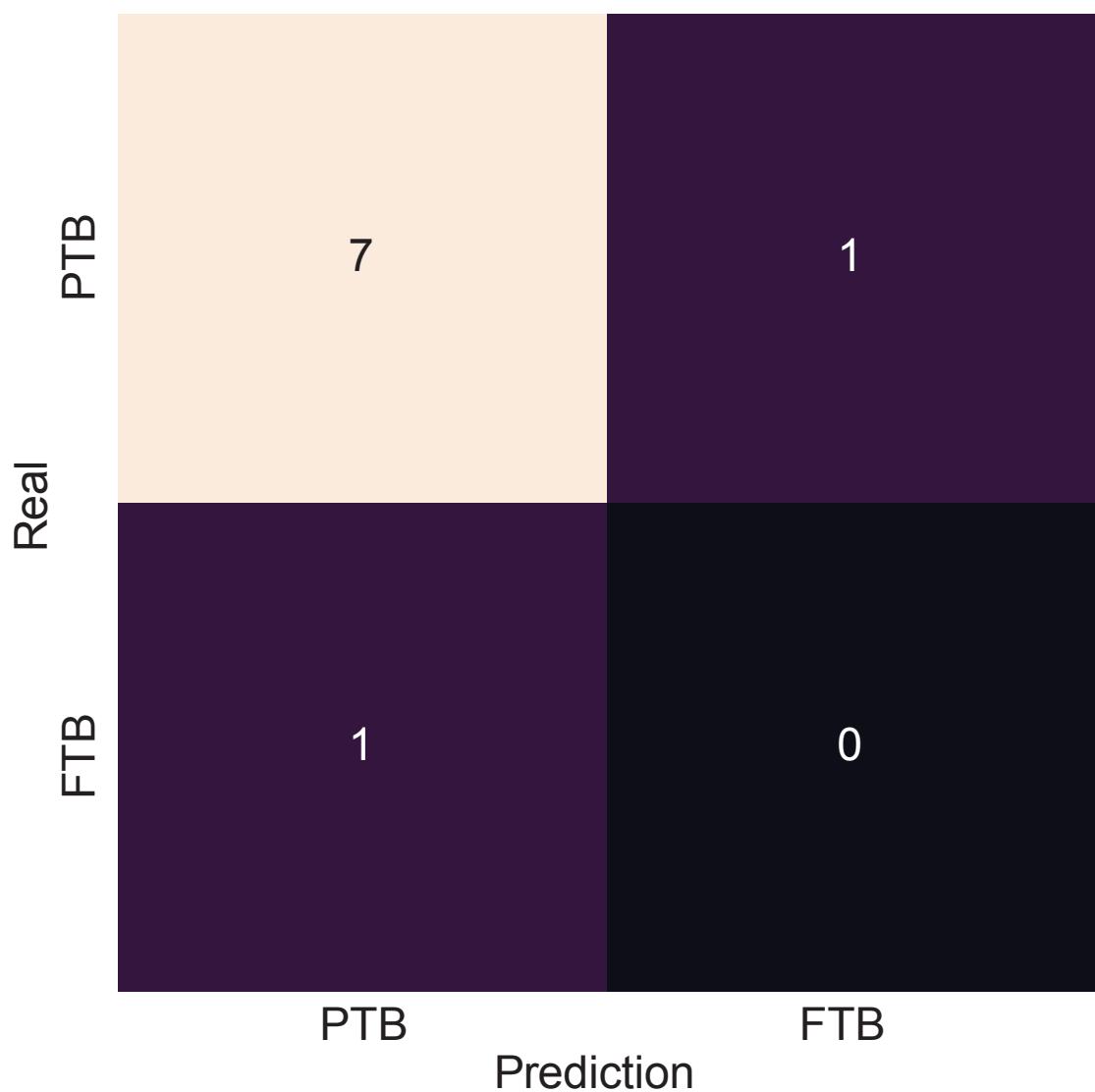


Figure 5: PROM-related DAT.

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 ( $0.714 \pm 0.061$ . Mean  $\pm$  SD)

391 **2.4 Discussion**

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

399 One of the most common salivary taxa that has been examined is *Fusobacterium nucleatum* (Han,  
400 2015; Brennan & Garrett, 2019; Bolstad, Jensen, & Bakken, 1996), that is a Gram-negative, anaerobic, and  
401 filamentous bacteria. *Fusobacterium nucleatum* can be separated from not only the salivary microbiome  
402 but also the vaginal microbiome (Vander Haar, So, Gyamfi-Bannerman, & Han, 2018; Witkin, 2019). In  
403 both animal and human investigation, *Fusobacterium nucleatum* infection has been linked to risk of PTB  
404 (Doyle et al., 2014). According to recent researches, the placenta women who give birth prematurely may  
405 include additional salivary microbiome dysbiosis, such as *Bergeyella* spp. and *Porphyromonas gingivalis*  
406 (León et al., 2007; Katz, Chegini, Shiverick, & Lamont, 2009). Although *Bergeyella* spp. were one of the  
407 PROM-overrepresented DAT (Figure 5), it was excluded in the final 25 PTB-related DAT. Furthermore,  
408 *Porphyromonas gingivalis* and *Campylobacter gracilis* were pathogens of periodontitis in sub-gingival  
409 microbiome (Yang et al., 2022). *Lactobacillus gasseri* was also one of the FTB-enriched DAT (Figure  
410 1), and it is well established that early PTB risk can be reduced by *Lactobacillus gasseri* in the vaginal  
411 microbiome (Basavaprabhu, Sonu, & Prabha, 2020; Payne et al., 2021).

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

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

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

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

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

443 **3 Random forest prediction model for periodontitis statuses based on the**  
444 **salivary microbiomes**

445 This section includes the published contents:

446

447 **3.1 Introduction**

448 Saliva microbial dysbiosis brought on by the accumulation of plaque results in periodontitis, a chronic  
449 inflammatory disease of the tissue that surrounds the tooth (Kinane, Stathopoulou, & Papapanou, 2017).  
450 Loss of periodontal attachment is a consequence of periodontitis, which may lead to irreversible bone loss  
451 and, eventually, permanent tooth loss if left untreated. A new classification criterion of periodontal diseases  
452 was created in 2018, about 20 years after the 1999 statements of the previous one (Papapanou et al.,  
453 2018). Even with this evolution, radiographic and clinical markers of periodontitis progression remain the  
454 primary methods for diagnosing periodontitis (Papapanou et al., 2018). Such tools, nevertheless, frequently  
455 demonstrate the prior damage from periodontitis rather than its present condition. Certain individuals have  
456 a higher risk of periodontitis, a higher chance of developing severe generalized periodontitis, and a worse  
457 response to common salivary bacteria control techniques utilized to prevent and treat periodontitis. As a  
458 result, the 2017 framework for diagnosing periodontitis additionally allows for the potential development  
459 of biomarkers to enhance diagnosis and treatment of periodontitis (Tonetti, Greenwell, & Kornman, 2018).  
460 Instead of only depending on the progression of periodontitis, a new etiological indication based on the  
461 current state must be introduced in order to enable appropriate intervention through early detection of  
462 periodontitis. Thus, the current clinical diagnostic techniques that rely on periodontal probing can be  
463 uncomfortable for patients with periodontitis (Canakci & Canakci, 2007).

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

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

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

496 Recently, we employed multiplex quantitative-PCR and machine learning-based classification model  
497 to predict the severity of periodontitis based on the amount of nine pathogens of periodontitis from  
498 saliva collections (E.-H. Kim et al., 2020). On the other hand, the fact that we focused merely at nine  
499 pathogens for periodontitis and neglected the variety bacterial species associated to the various severities  
500 of periodontitis constrained the breadth of our investigation. By developing a machine learning model  
501 that could classify multiple severities of periodontitis based on the salivary microbiome composition,  
502 this study aims to fill these knowledge gaps and produce more accurate and therapeutically useful  
503 guidance to evaluate progression of periodontitis. Hence, in order to examine the salivary microbiome  
504 composition of both healthy controls and patients with periodontitis in multiple stages, we applied  
505 16S rRNA gene sequencing. Furthermore, employing the 2018 classification criteria, we sought to find  
506 biomarkers (species) for the precise prediction of periodontitis severities (Papapanou et al., 2018; Chapple  
507 et al., 2018).

508 **3.2 Materials and methods**

509 **3.2.1 Study participants enrollment**

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

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

527 **3.2.2 Periodontal clinical parameter diagnosis**

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

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

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

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

545 0. Normal gingiva: without inflammation nor discoloration.

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

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

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

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

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

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

556 • Healthy:

557 1. Bleeding sites < 10%

558 2. Probing depth:  $\leq$  3 mm

559 • Stage I:

560 1. No tooth loss because of periodontitis.

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

562 3. Radiographic bone loss: < 15%

563 • Stage II:

564 1. No tooth loss because of periodontitis.

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

566 3. Radiographic bone loss: 15-33%

567 • Stage III:

568 1. Teeth loss because of periodontitis:  $\leq$  teeth

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

570 3. Radiographic bone loss: > 33%

571 **3.2.3 Saliva sampling and DNA extraction procedure**

572 All study participants received instructions to avoid eating, drinking, brushing, and using mouthwash for  
573 at least an hour prior to the saliva sample collection process. These collections were conducted between  
574 09:00 and 11:00. Mouth rinse was collected by rinsing the mouth for 30 seconds with 12 mL of a solution  
575 (E-zen Gargle, JN Pharm, Korea). All saliva samples were tagged with anonymous ID and stored at -4 °C.

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

- 580 • Forward: 5' -TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNNGCWGCAG-3'  
581 • Reverse: 5' -GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'

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

- 584 1. Heat activation for 30 seconds at 95 °C.  
585 2. 25 cycles for 30 seconds at 95 °C.  
586 3. 30 seconds at 55 °C.  
587 4. 30 seconds at 72 °C.

588 NexteraXT Indexed Primer was applied to amplification 10 µL of the purified initial PCR products for  
589 the final library creation. The second PCR used the same conditions as the first PCR conditions but with  
590 10 cycles. 16S rRNA gene sequencing was performed via 2×300 bp paired-end sequencing at Macrogen  
591 Inc. (Macrogen, Seoul, Korea) using Illumina MiSeq platform (Illumina, San Diego, CA, USA).

592 **3.2.4 Bioinformatics analysis**

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

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

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

604 Aitchison index for a beta-diversity index was calculated using QIIME2 (version 2020.8) (Aitchison,  
605 Barceló-Vidal, Martín-Fernández, & Pawlowsky-Glahn, 2000; Bolyen et al., 2019). We employed the  
606 t-SNE algorithm to illustrate multi-dimensional data from the beta-diversity index computation (Van der  
607 Maaten & Hinton, 2008). The beta-diversity index was compared using the PERMANOVA test (Anderson,  
608 2014; Kelly et al., 2015) and MWU test.

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

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

621 As previously stated (E.-H. Kim et al., 2020), we used stratified  $k$ -fold cross-validation ( $k = 10$ )  
622 by severity of periodontitis to achieve consistent and trustworthy classification results (Wong & Yeh,  
623 2019). Additionally, we utilized various features with confusion matrices and their derivations to evaluate  
624 the classification outcomes in order to identify which features optimize classification evaluations and  
625 decrease sequencing efforts. Using the DAT discovered by ANCOM, we iteratively removed the least  
626 significant taxa from the input features (taxa) of the random forest (Breiman, 2001) and gradient boosting  
627 (Friedman, 2002) classification models using the backward elimination method. Random forest classifier  
628 builds multiple decision trees independently using bootstrapped samples and aggregates their predictions,  
629 enhancing stability and reducing overfitting problems. In contrast, Gradient boosting constructs trees  
630 sequentially, where each new tree improves the errors of the previous ones using gradient descent, leading  
631 to higher classification evaluations.

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

### 636 3.2.5 Data and code availability

637 All sequences from the 250 study participants have been published to the Sequence Read Archives (project  
638 ID PRJNA976179): <https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA976179>. Docker

641 image that employed throughout this study is available in the DockerHub: <https://hub.docker.com/>  
642 repository/docker/fumire/periodontitis\_16s. Every code used in this study can be found on  
643 GitHub: [https://github.com/CompbioLabUnist/Periodontitis\\_16S](https://github.com/CompbioLabUnist/Periodontitis_16S).

644 **3.3 Results**

645 **3.3.1 Summary of clinical information and sequencing data**

646 Among clinical information of the study participants, clinical attachment level, probing depth, plaque  
647 index, and gingival index, were significantly increased with periodontitis severity (Kruskal-Wallis test  
648  $p < 0.001$ ), while sex were observed no significant difference (Table 2). Notably, clinical attachment level  
649 and probing depth have significant differences among the periodontitis severities (MWU test  $p < 0.01$ ;  
650 Figure 15). Additionally,  $71461.00 \pm 11792.30$  and  $45909.78 \pm 11404.65$  reads per sample were obtained  
651 before and after filtering low-quality reads and trimming extra-long tails, respectively (Figure 16). In 250  
652 study subjects, we have found a total of 425 bacterial taxa (Figure 13).

653 **3.3.2 Diversity indices reveal differences among the periodontitis severities**

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

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

664 **3.3.3 DAT among multiple periodontitis severities and their correlation**

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

- 668 • Group 1
  - 669 1. *Treponema* spp.
  - 670 2. *Prevotella* sp. HMT 304
  - 671 3. *Prevotella* sp. HMT 526
  - 672 4. *Peptostreptococcaceae [XI][G-5]* saphenum
  - 673 5. *Treponema* sp. HMT 260
  - 674 6. *Mycoplasma faecium*
  - 675 7. *Peptostreptococcaceae [XI][G-9]* brachy
  - 676 8. *Lachnospiraceae [G-8]* bacterium HMT 500
  - 677 9. *Peptostreptococcaceae [XI][G-6]* nodatum
  - 678 10. *Fretibacterium* spp.

- 679 • Group 2
- 680     1. *Porphyromonas gingivalis*
- 681     2. *Campylobacter showae*
- 682     3. *Filifactor alocis*
- 683     4. *Treponema putidum*
- 684     5. *Tannerella forsythia*
- 685     6. *Prevotella intermedia*
- 686     7. *Porphyromonas* sp. HMT 285

- 687 • Group 3
- 688     1. *Actinomyces* spp.
- 689     2. *Corynebacterium durum*
- 690     3. *Actinomyces graevenitzii*

691 Ten DAT that were significant enriched in stage II and stage III, but deficient in healthy formed Group  
 692 1 (Figure 8). Furthermore, in comparison to the healthy, the seven DAT of Group 2 were significantly  
 693 enriched in each of the stages of periodontitis. On the other hand, three DAT in Group 3 were deficient in  
 694 stage II and stage III, but significantly enriched in healthy. The relative proportions of the DAT further  
 695 supported these findings (Figure 8b), suggesting that the DAT is primarily linked to periodontitis rather  
 696 than other salivary bacteria.

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

### 699 3.3.4 Classification of periodontitis severities by random forest models

700 To confirm that using selected DAT bacterial profiles could have enhanced sequencing expenses without  
 701 losing the classification evaluations, we built the random forest classification models based on DAT and  
 702 full microbiome compositions (Figure 18). DAT based classifier showed non-significant different or better  
 703 evaluations, by removing confounding taxa.

704 Based on the proportion of DAT, random forest classifier were trained to classify the periodontitis  
 705 severities (Table 6). We conducted multi-label classification for the multiple periodontitis severities,  
 706 namely healthy, stage I, stage II, and stage III. In this setting, we classified multiple periodontitis  
 707 severities with the highest BA of  $0.779 \pm 0.029$  (Table 4). AUC ranged between 0.81 and 0.94 (Figure  
 708 10b).

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

713 Based on the findings that the salivary microbiome composition in stage II is more comparable to  
 714 those in stage III than to other severities (Figure 7f and Figure 7j), we combined stage II and stage III to

715 perform a multi-label classification.

716 To examine alternative classification algorithms in comparison to random forest classification, we  
717 selected gradient boost algorithm because it is another algorithm of the few classification algorithms  
718 that can provide feature importances, which is essential for identifying key taxa contributing to the  
719 classification of periodontitis severities. Thus, we assessed gradient boosting algorithms (Figure 20).  
720 However, the classification evaluations obtained from gradient boosting have non-significant differences  
721 compared to random forest classification.

722 Finally, to confirm the reliability and consistency of our random forest classifier, we validated our  
723 classification model using openly accessible 16S rRNA gene sequencing from Spanish participants  
724 (Iniesta et al., 2023) and Portuguese participants (Relvas et al., 2021) (Figure 11). Although some  
725 evaluations, *e.g.* SPE, were low, the other were comparable.

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

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

<b>Index</b>	<b>Healthy</b>	<b>Stage I</b>	<b>Stage II</b>	<b>Stage III</b>	<b>p-value</b>
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 I/II/III		
	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 graevenitzii</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.

<b>Group 1</b>	<b>Group 2</b>	<b>p-value</b>
Healthy	Stage I	0.001
Healthy	Stage II	0.001
Healthy	Stage III	0.001
Stage I	Stage II	0.001
Stage I	Stage III	0.001
Stage II	Stage III	0.737

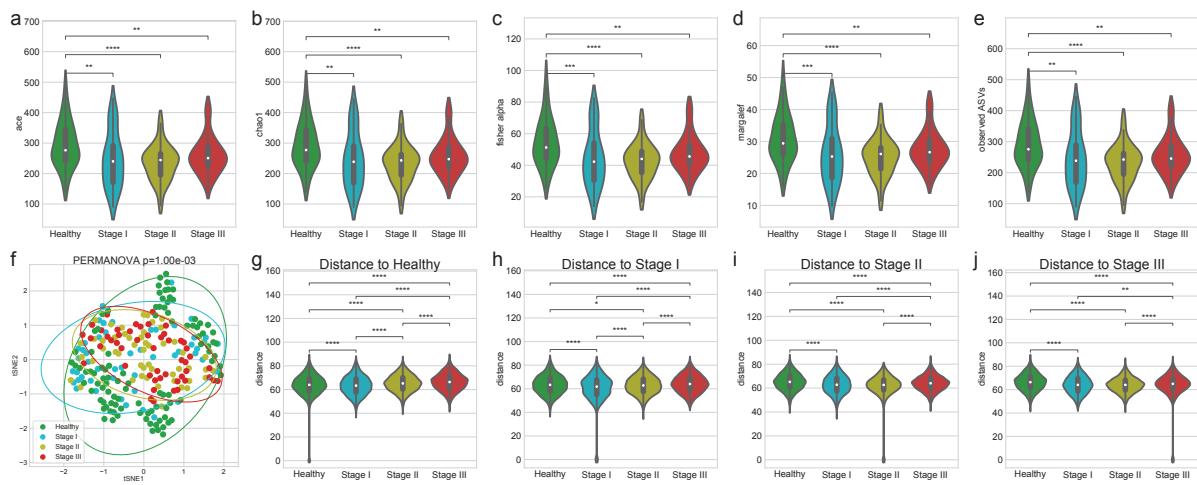


Figure 7: Diversity indices.

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

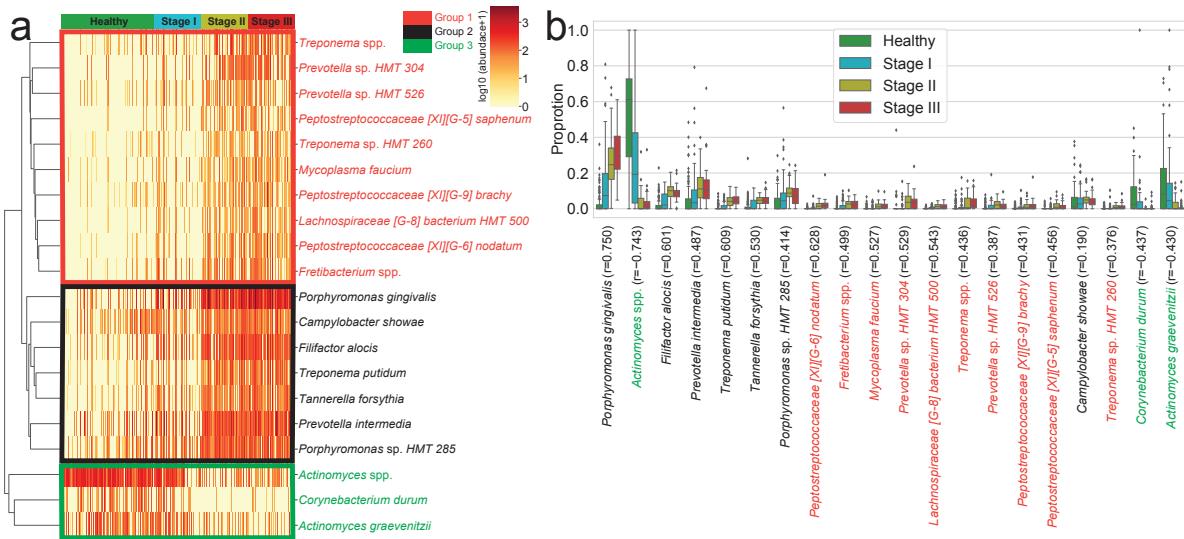
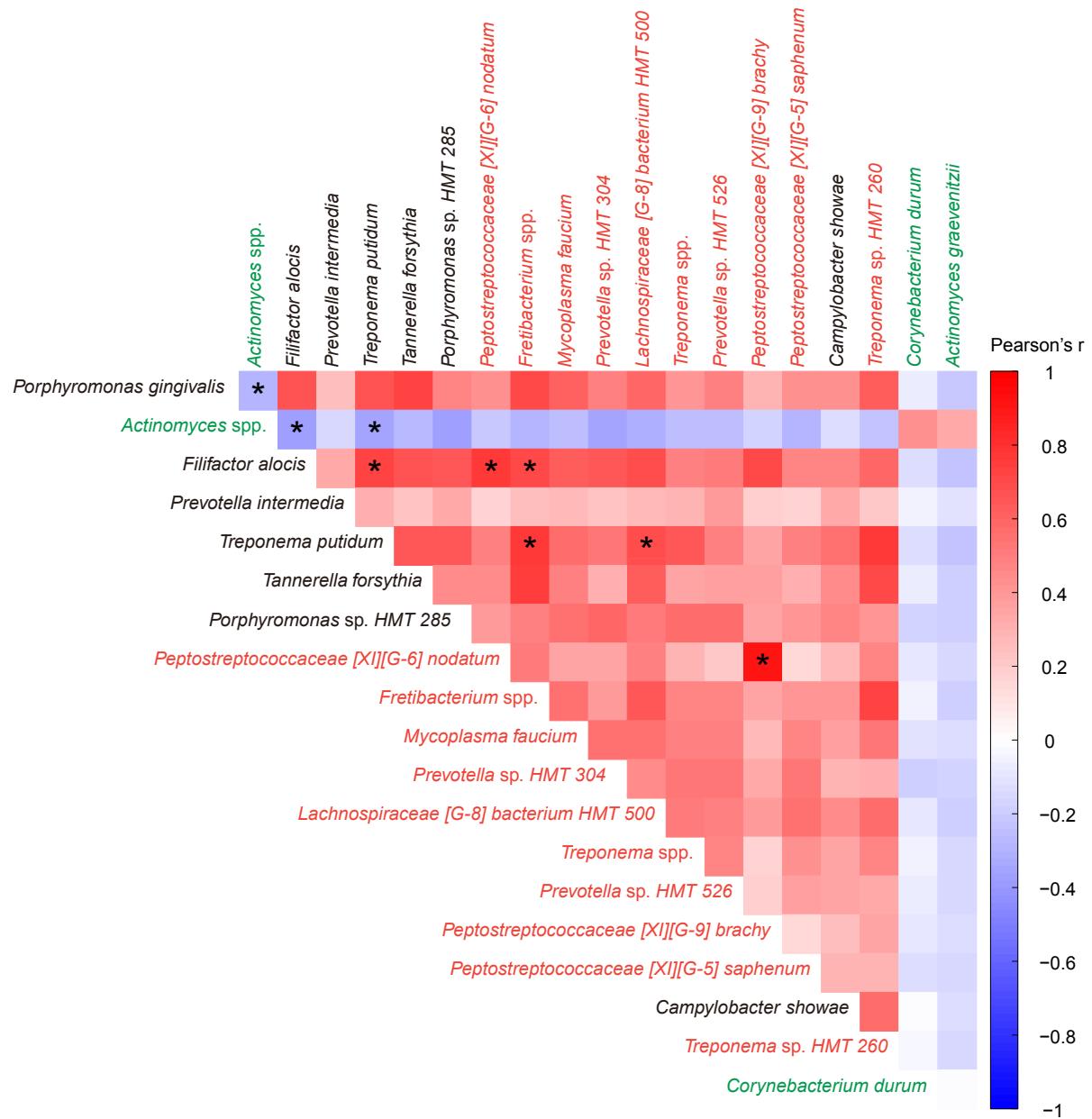


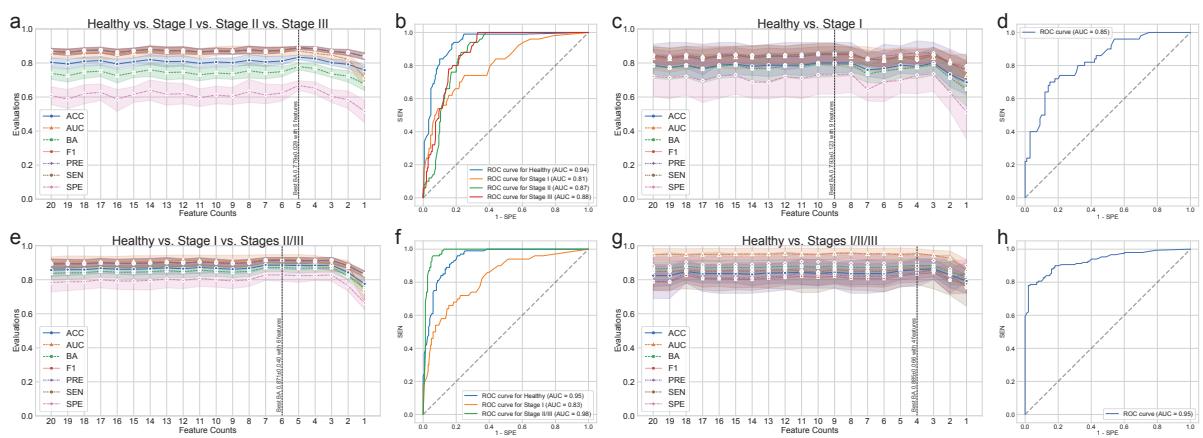
Figure 8: **Differentially abundant taxa (DAT).**

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



**Figure 9: Correlation heatmap.**

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



**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).

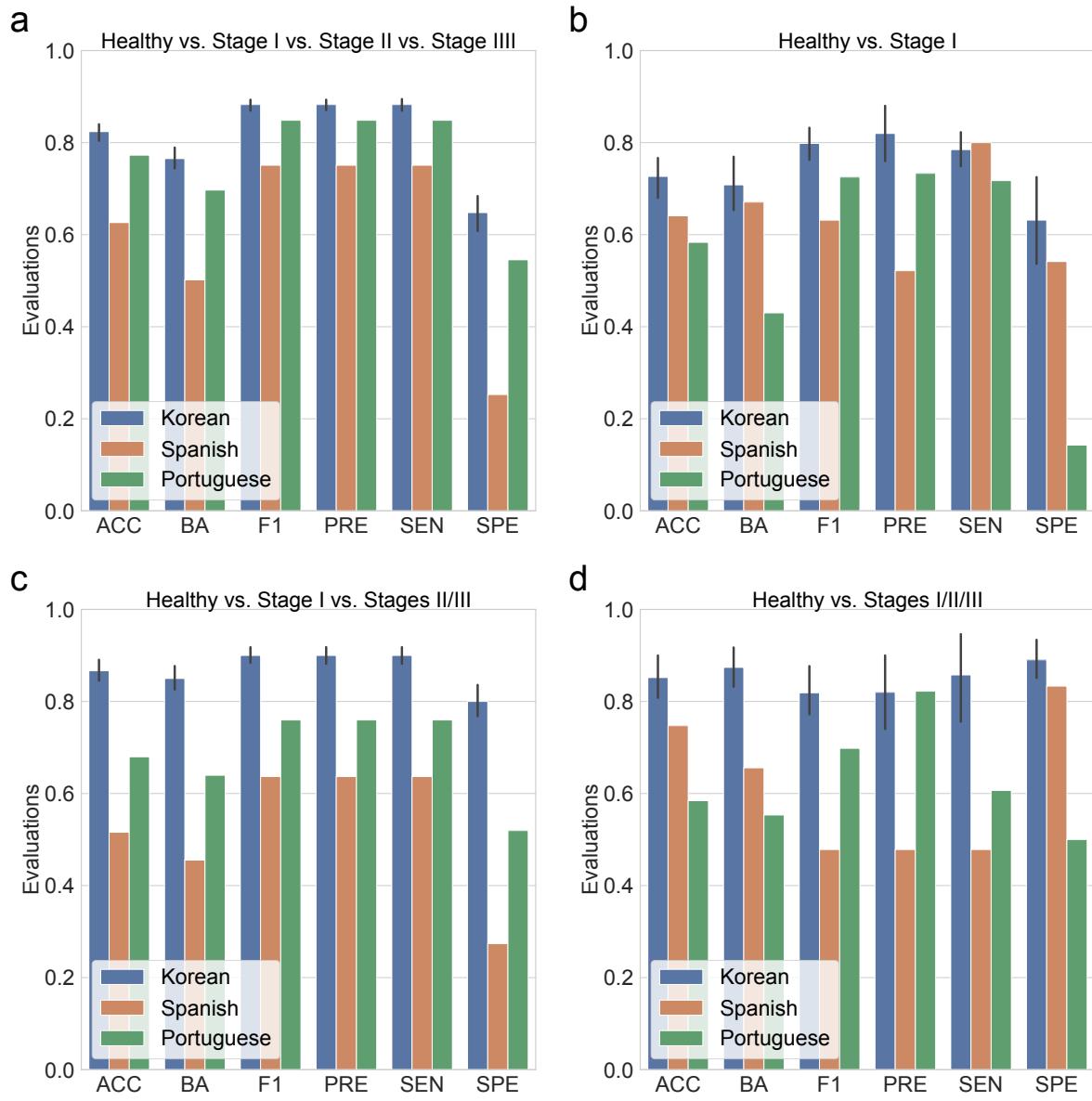


Figure 11: **Random forest classification metrics from external datasets.**

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.

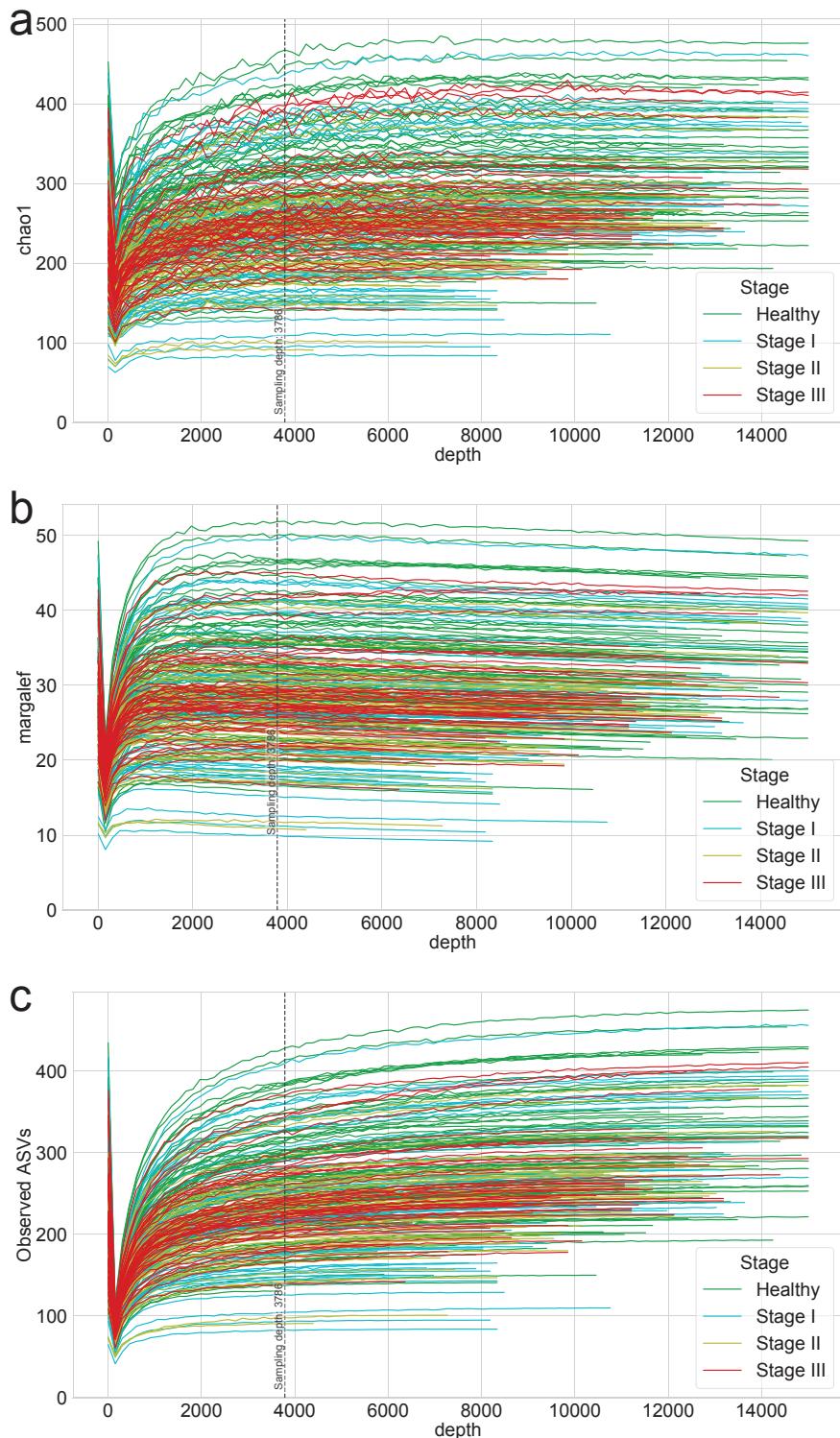
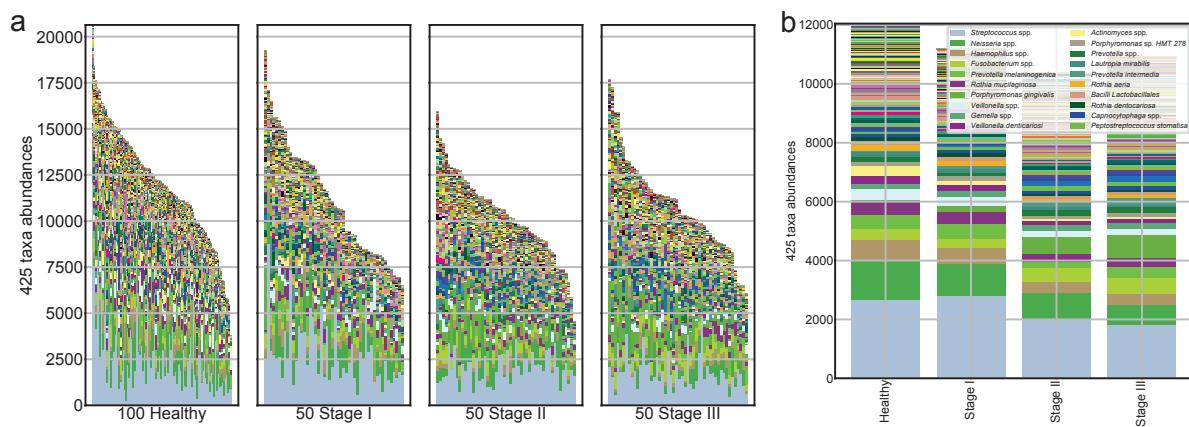


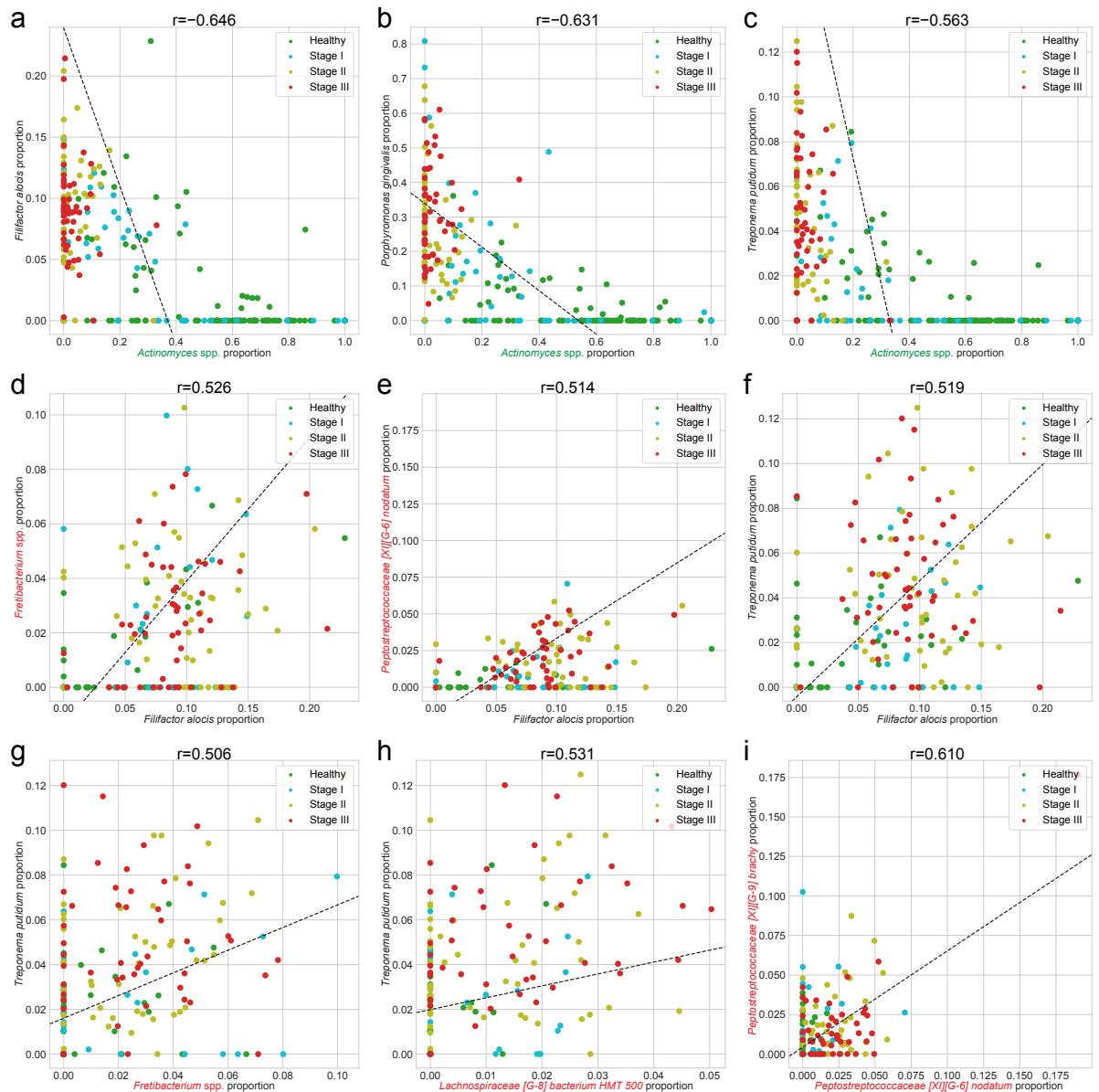
Figure 12: Rarefaction curves for alpha-diversity indices.

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



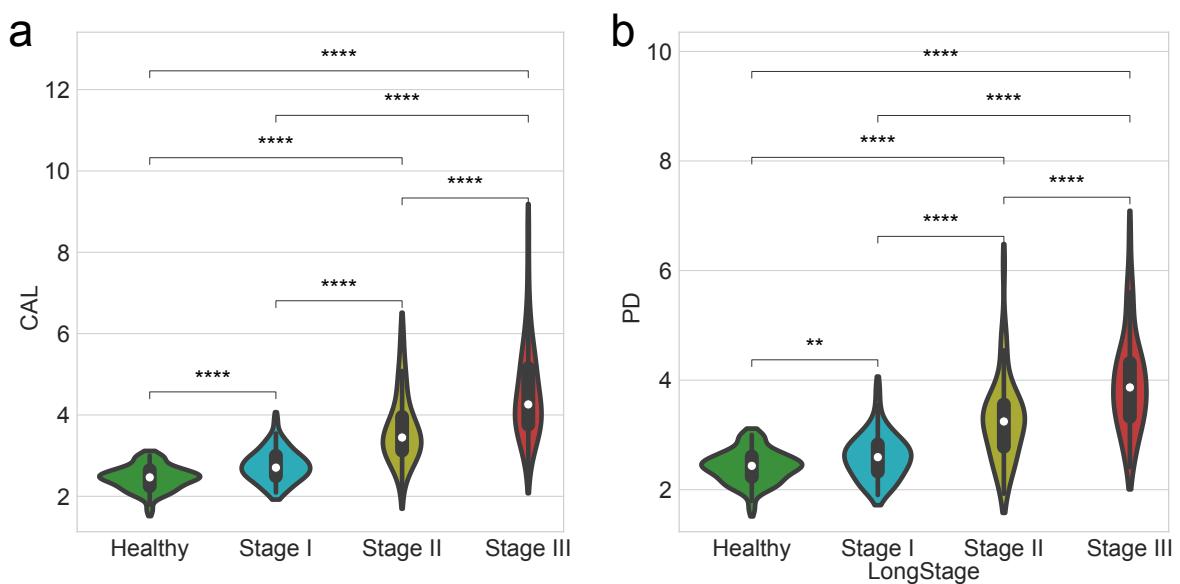
**Figure 13: Salivary microbiome compositions in the different periodontal statuses.**

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 (CAL) **(b)** Probing depth (PD). Statistical significance determined by the MWU test:  $p \leq 0.01$  (\*\*) and  $p \leq 0.0001$  (\*\*\*\*).

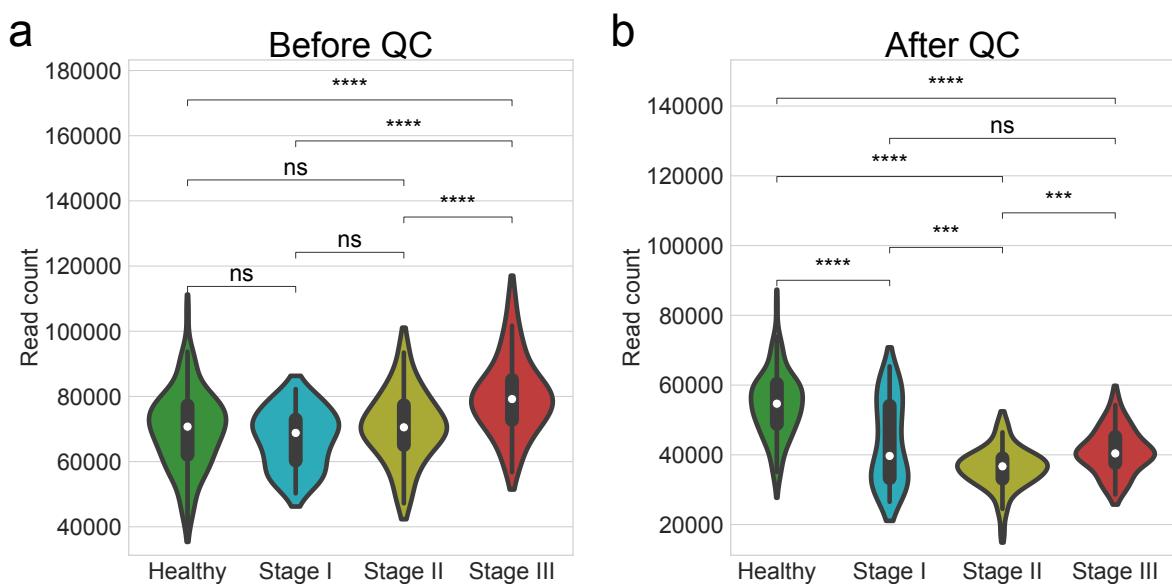


Figure 16: **Number of read counts by the periodontitis statuses.**

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

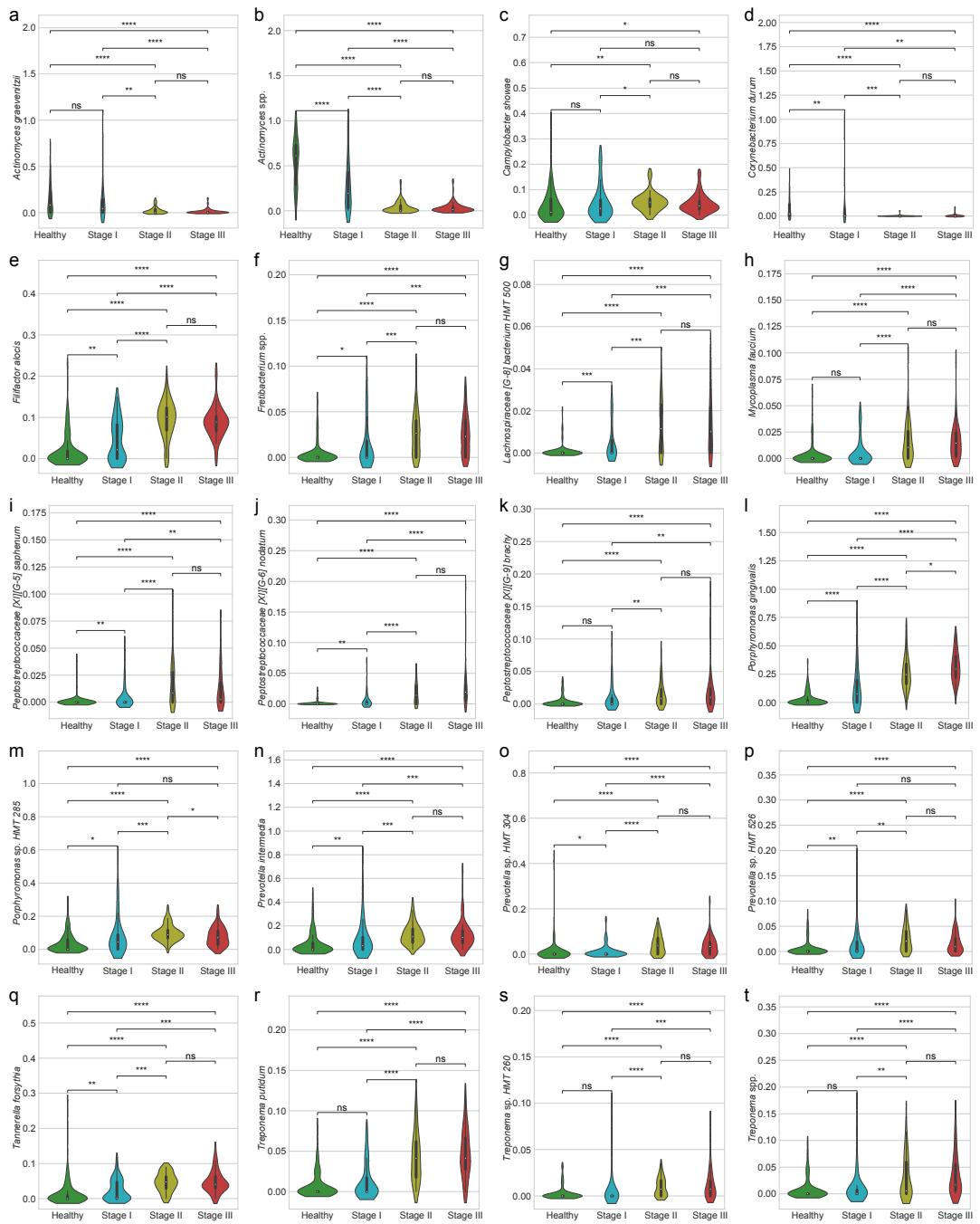
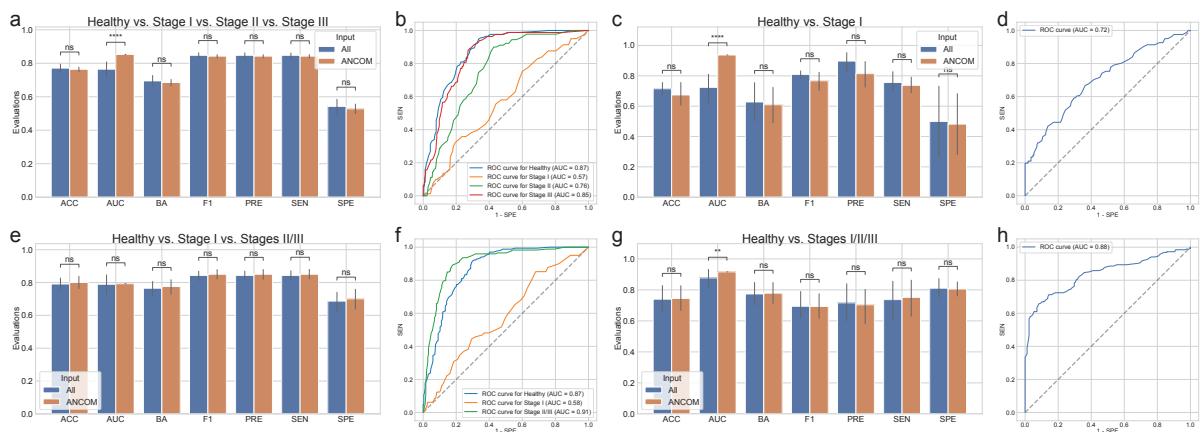


Figure 17: Proportion of DAT.

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

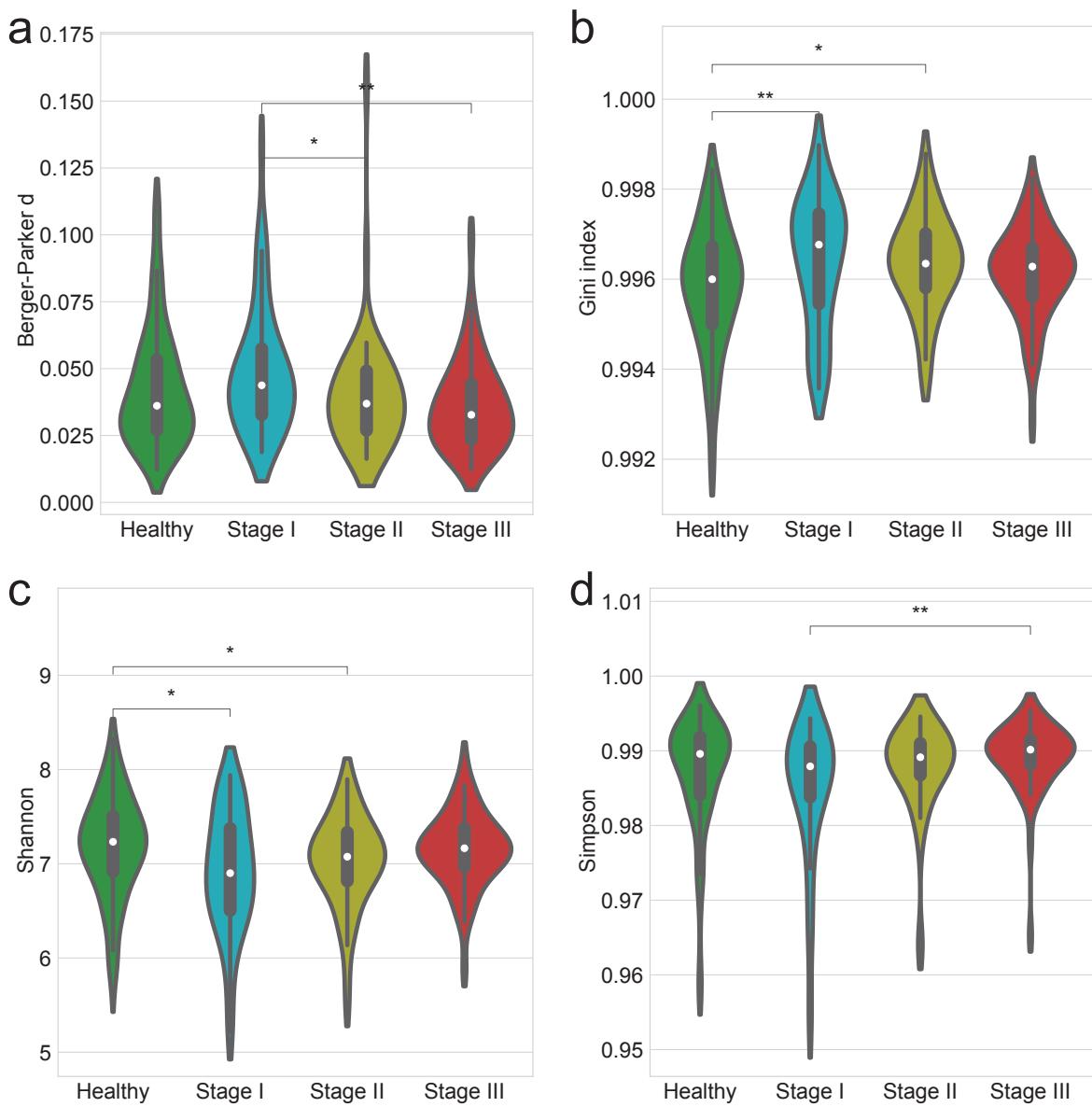
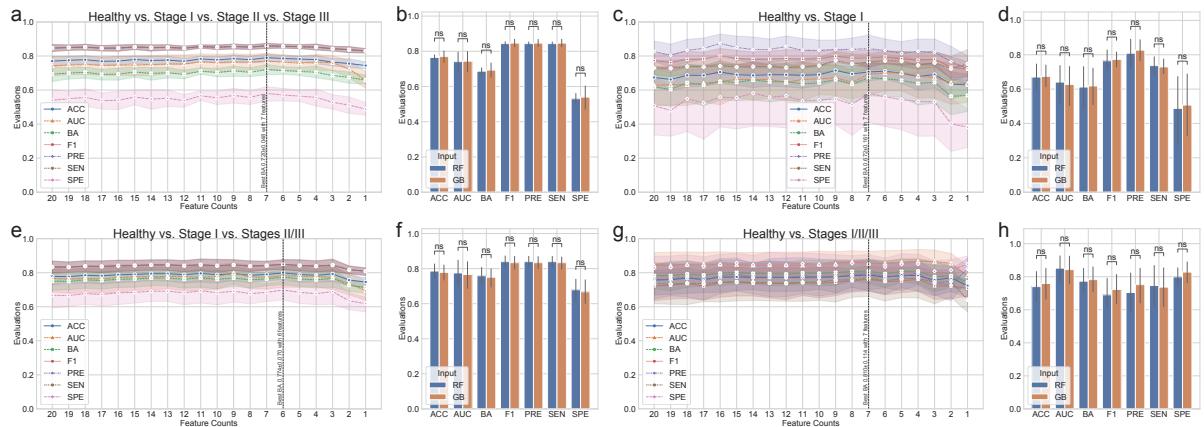


Figure 19: **Alpha-diversity indices account for evenness.**

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.

726 **3.4 Discussion**

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

735 Previous research identified the red complex as the primary pathogens of periodontitis (Listgarten,  
736 1986): *Porphyromonas gingivalis*, *Tannerella forsythia*, and *Treponema denticola*. Other studies, however,  
737 have shown that periodontal pathogens communicate with other bacteria in the salivary microbiome  
738 networks to generate dental plaque prior to the pathogenesis and development of periodontitis (Lamont &  
739 Jenkinson, 2000; Rosan & Lamont, 2000; Yoshimura, Murakami, Nishikawa, Hasegawa, & Kawaminami,  
740 2009).

741 Using subgingival plaque collections, recent researches have suggested a connection between the  
742 periodontitis severity and the salivary microbiome compositions (Altabtbaei et al., 2021; Iniesta et al.,  
743 2023; Nemoto et al., 2021). Therefore, we have examined the salivary microbiome compositions of  
744 patients with multiple severities of periodontitis and periodontally healthy controls, extending on earlier  
745 studies.

746 According to our findings, the salivary microbiome compositions have 425 taxa (Figure 13). We  
747 computed the alpha-diversity indices to determine the variability within each salivary microbiome  
748 composition, including ace (Chao & Lee, 1992), chao1 (Chao, 1984), fisher alpha (Fisher et al., 1943),  
749 margalef (Magurran, 2021), observed ASVs (DeSantis et al., 2006), Berger-Parker *d* (Berger & Parker,  
750 1970), Gini index (Gini, 1912), Shannon (Weaver, 1963), and Simpson (Simpson, 1949) (Figure 7 and  
751 Figure 19). Alpha-diversity indices suggested that the microbial richness of periodontally healthy controls  
752 was higher than that of patients with periodontitis (Figure 7a-e and Figure 19). These results are in line with  
753 findings with that patients with advanced periodontitis, namely stage II and stage III, have less diversified  
754 communities than periodontally healthy controls (Jorth et al., 2014). Recognizing that the periodontitis  
755 severity increases the amount of *Porphyromonas gingivalis*, the salivary microbiome compositions from  
756 periodontally healthy controls conserved microbial networks dominated by *Streptococcus* spp. (Figure  
757 13). *Porphyromonas gingivalis* is one of the known periodontal pathogen that could cause dysbiosys  
758 in the salivary microbiomes, suggesting in the pathophysiology of periodontitis. Despite this finding,  
759 earlier research found that subgingival microbiome of patients with periodontitis had a greater alpha-  
760 diversity index (observed ASVs) than that of healthy controls (Iniesta et al., 2023), might due to the  
761 different sampling sites between saliva and subgingival plaque. On the other hand, another research  
762 has addressed significant discrepancies in alpha-diversity indices from subgingival plaque, saliva, and  
763 tongue biofilms from healthy controls and periodontitis patients, resulting the highest alpha-diversity

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

773 Therefore, we calculated beta-diversity indices to analyze the differences between the study partici-  
774 pants. The distances for the multiple stages of periodontitis, including stage I, stage II, and stage III, as  
775 well as healthy controls (Figure 4g-j and Table 7), suggesting notable differences among the multiple  
776 periodontitis severities. In other words, the composition of the salivary microbiome compositions varies  
777 depending on the periodontitis stages, so that supporting the findings from a previous study (Iniesta et al.,  
778 2023). Taken together that it is nearly impossible to fully restore the attachment level after it has been lost  
779 due to the progression and development of periodontitis, the ability to rapidly screen for periodontitis in  
780 its early phases using saliva collections would be highly beneficial for effective disease management and  
781 treatment.

782 Of the total of 425 taxa in the salivary microbiome composition that have been identified (Figure 13),  
783 ANCOM was applied to select 20 taxa as the DAT that indicated notable abundance variation among  
784 the periodontitis severities (Figure 8 and Table 5). Three sub-groups were formed from the DAT using  
785 hierarchical clustering (Figure 8a). Surprisingly, two of the red complex pathogens (Rôças, Siqueira Jr,  
786 Santos, Coelho, & de Janeiro, 2001), *Porphyromonas gingivalis* and *Tannerella forsythia*, were classified  
787 in Group 2 and were more prevalent in stage II and stage II periodontitis compared to healthy controls.  
788 *Campylobacter showae* was additionally placed in Group 2 of the orange complex pathogens (Gambin et  
789 al., 2021). Furthermoe, some of the DAT in Group 2 have reported their crucial roles in pathogenesis  
790 and development of periodontitis: *Filifactor alocis* (Aruni et al., 2015), *Treponema putidum* (Wyss et  
791 al., 2004), *Tannerella forsythia* (Stafford, Roy, Honma, & Sharma, 2012; W. Zhu & Lee, 2016), and  
792 *Prevotella intermedia* (Karched, Bhardwaj, Qudeimat, Al-Khabbaz, & Ellepol, 2022). Taken together,  
793 this indicates that DAT in Group 2 is essential to periodontitis. The portion of some Group 1 DAT,  
794 including *Peptostreptococcaceae[XI][G-5] saphenum*, *Peptostreptococcaceae[XI][G-6] nodatum*, and  
795 *Peptostreptococcaceae[XI][G-9] brachy*, in healthy controls and patients with periodontitis significantly  
796 differed, according to earlier research (Lafaurie et al., 2022). These outcomes support our research,  
797 implying that Group 1 DAT are also essential to the etiology and progression of periodontitis. However,  
798 in contrast to patients with periodontitis, Group 3 DAT, namely *Corynebacterium durum* and *Actinomyces*  
799 *graevenitzii*, were enriched in healthy controls, which is consistent with earlier research (Redanz et al.,  
800 2021; Nibali et al., 2020).

801 In our correlation analysis (Figure 9), we have discovered strongly negative correlations (coefficient  $\leq$   
802  $-0.5$ ) 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  $\geq 0.5$ ) (Figure 14). Interestingly, there were strongly negative correlations (coefficient  $\leq -0.5$ ) 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  $\geq 0.5$ ) (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  $\geq 0.5$ ), 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. This finding aligns with a recent study that identifies *Actinomyces* spp. as the most prevalent bacteria in both the healthy gingivitis controls, while *Porphyromonas gingivalis* is recognized as the most predominant taxon within the periodontitis subjects, based on analyses of subgingival plaque samples (Nemoto et al., 2021). We have previously developed machine learning models for the classification of periodontitis, with the objective of predicting the severities of chronic periodontitis by analyzing the copy numbers of nine known salivary bacteria species. We classified healthy controls and patients with periodontitis utilizing bacterial combinations in conjunction with a random forest model (E.-H. Kim et al., 2020):

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

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

- AUC: 92%

- 842     • BA: 84%  
843     • SEN: 94%  
844     • SPE: 74%
- 845     By separating patients with periodontitis from healthy controls using only four DAT, *e.g.* *Actinomyces*  
846     *graevenitzii*, *Actinomyces* spp., *Corynebacterium durum*, and *Porphyromonas gingivalis*, our machine  
847     learning model performed better than previously published models (Figure 10, Table 4, and Table 6):  
848     • AUC:  $95.3\% \pm 4.9\%$   
849     • BA:  $88.5\% \pm 6.6\%$   
850     • SEN:  $86.4\% \pm 15.7\%$   
851     • SPE:  $90.5\% \pm 7.0\%$
- 852     This result showed that by detecting Group 3 bacteria that were substantially abundant in health  
853     controls than patients with periodontitis, our study increased BA by at least 5% and SPE by at least 17%.  
854     Furthermore, we have validated our machine-learning prediction model using openly accessible 16S  
855     gene rRNA sequencing data from Portuguese (Iniesta et al., 2023) and Spanish participants (Relvas et  
856     al., 2021) in order to ensure the consistency of our random forest classification model (Figure 11). Our  
857     classification models employed in this study were primarily developed and assessed on Korean study par-  
858     ticipants, which may limit their generalizability to other ethnic groups with different salivary microbiome  
859     compositions (Premaraj et al., 2020; Renson et al., 2019). Therefore, the evaluations of this periodonti-  
860     tis classification models can be affected by ethnic-specific variances and differences, highlighting the  
861     necessity for additional validation and adjustment across a spectrum of ethnic backgrounds.
- 862     Regarding the clinical characteristics and potential confounders influencing the analysis of salivary  
863     microbiome compositions connected with periodontitis severity, this study had a number of limitations  
864     that were pointed out. We did not offer clinical information, such as the percentage of teeth, the percentage  
865     of bleeding on probing, nor dental furcation involvement, even though we did gather information on  
866     attachment level, probing depth, plaque index, and gingival index; this might have it challenging to present  
867     thorough and in-depth data about periodontal health. Moreover, the broad age range may make it tougher  
868     to evaluate the relationship between age and periodontitis statuses, providing the necessity for future  
869     studies to consider into account more comprehensive clinical characteristics associated with periodontitis.  
870     Additionally, potential confounders—*e.g.* body mass index (Bombin, Yan, Bombin, Mosley, & Ferguson,  
871     2022) and e-cigarette use (Suzuki, Nakano, Yoneda, Hirofushi, & Hanioka, 2022)—which might have  
872     affected dental health and salivary microbiome composition were disregarding consideration in addition to  
873     smoking status and systemic diseases. Thus, future research incorporating these components would offer a  
874     more thorough knowledge of how lifestyle factors interact and affect the salivary microbiome composition  
875     and periodontal health. Throughout, resolving these limitations will advance our understanding in  
876     pathogenesis and development of periodontitis, offering significant novel insights on the causal connection  
877     between systemic diseases and the salivary microbiome compositions.

878 **4 Metagenomic signature analysis of Korean colorectal cancer**

879 **4.1 Introduction**

880 Colorectal cancer (CRC) is one of the most prevalent and life-threatening malignancies worldwide  
881 (Kuipers et al., 2015; Center, Jemal, Smith, & Ward, 2009; N. Li et al., 2021), with its incidence  
882 influenced by a combination of genetic (Zhuang et al., 2021; Peltomaki, 2003), environmental (O'Sullivan  
883 et al., 2022; Raut et al., 2021), and lifestyle factors (X. Chen et al., 2021; Bai et al., 2022; Zhou et  
884 al., 2022; X. Chen, Li, Guo, Hoffmeister, & Brenner, 2022). Established risk factors include a often  
885 diet in red and processed meats (Kennedy, Alexander, Taillie, & Jaacks, 2024; Abu-Ghazaleh, Chua,  
886 & Gopalan, 2021), obesity (Mandic, Safizadeh, Niedermaier, Hoffmeister, & Brenner, 2023; Bardou  
887 et al., 2022), cigarette smoking (X. Chen et al., 2021; Bai et al., 2022), alcohol consumption (Zhou et  
888 al., 2022; X. Chen et al., 2022), and a sedentary lifestyle (An & Park, 2022), all of which contribute to  
889 chronic inflammation, mutagenesis, and metabolic regulation. Additionally, underlying conditions, e.g.  
890 Lynch syndrome (Vasen, Mecklin, Khan, & Lynch, 1991; Hampel et al., 2008) and familial adenomatous  
891 polyposis (Inra et al., 2015; Burt et al., 2004), significantly increase risk of CRC due to persistent mucosal  
892 inflammation and somatic mutations that promote tumorigenesis.

893 The gut microbiome plays a fundamental role in maintaining host health by helping digestion  
894 (Joscelyn & Kasper, 2014; Cerqueira, Photenhauer, Pollet, Brown, & Koropatkin, 2020), regulating  
895 metabolism (Dabke, Hendrick, Devkota, et al., 2019; Utzschneider, Kratz, Damman, & Hullarg, 2016;  
896 Magnúsdóttir & Thiele, 2018), adjusting immune function (Kau, Ahern, Griffin, Goodman, & Gordon,  
897 2011; Shi, Li, Duan, & Niu, 2017; Broom & Kogut, 2018), and even coordinating neurological processes  
898 by the brain-gut axis (Martin et al., 2018; Aziz & Thompson, 1998; R. Li et al., 2024). Comprising  
899 these gut microbiota, including, archaea, bacteria, fungi, and viruses, the gut microbiome contributes  
900 to the synthesis of essential vitamins, and production of fatty acids, which influence intestinal integrity  
901 and immune responses. Thus, well-balanced gut microbiome composition modulates systemic immune  
902 function by interacting with gut-associated lymphoid tissue, shaping immune tolerance and response  
903 to infections. Hence, emerging evidence suggests that dysbiosis in the gut microbiome composition are  
904 associated not only a narrow range of diseases, e.g. diarrhea and enteritis (Paganini & Zimmermann,  
905 2017; Gao, Yin, Xu, Li, & Yin, 2019) but also a wide range of diseases, e.g. obesity, diabetes, and cancers  
906 (Barlow et al., 2015; Hartstra et al., 2015; Helmink et al., 2019; Cullin et al., 2021).

907 Recent studies have highlighted the crucial role of the gut microbiome in tumorigenesis and progres-  
908 sion of CRC (Song, Chan, & Sun, 2020; Rebersek, 2021), with dysbiosis emerging as a potential risk  
909 factor. Dysbiosis in gut microbiome compositions can promote tumorigenesis of many cancers, including  
910 CRC, through several signaling cascades, including inflammation, mutagenesis, and altered metabolism  
911 in host. Certain bacteria species, such as *Fusobacterium* genus (Hashemi Goradel et al., 2019; Bullman et  
912 al., 2017; Flanagan et al., 2014), *Bacteroides* genus (Ulger Toprak et al., 2006; Boleij et al., 2015), and  
913 *Escherichia coli* (Swidsinski et al., 1998; Bonnet et al., 2014), have been associated with development  
914 and progression of CRC by producing pro-inflammatory signals, generating toxins including mutagens,

915 and disrupting the intestinal barriers including mucous surface. In contrast, beneficial bacteria, such as  
916 *Lactobacillus* genus (Ghorbani et al., 2022; Ghanavati et al., 2020) and *Bifidobacterium* genus (Le Leu,  
917 Hu, Brown, Woodman, & Young, 2010; Fahmy et al., 2019), are regarded to apply protective roles by  
918 maintaining homeostasis of gut microbiome compositions and regulating immune responses including  
919 inflammation.

920 Furthermore, identifying metagenome biomarkers in Korean CRC patients is essential, as the gut  
921 microbiome compositions significantly vary by ethnicity due to genetic, dietary, and environmental  
922 factor (Fortenberry, 2013; Merrill & Mangano, 2023; Parizadeh & Arrieta, 2023). Additionally, ethnicity-  
923 specific microbiome composition signatures may affect the reliability of previously established biomarkers  
924 derived from predominantly Western CRC cohorts (Network et al., 2012), necessitating population-  
925 specific investigations. By identifying metagenomic biomarkers tailored to Korean CRC patients, we  
926 can improve early detection rate of early-stage CRC, develop more accurate risk of CRC, and explore  
927 microbiome-targeted therapies that consider host-microbiome interactions within the Korean population.

928 Accordingly, this study aims to identify microbiome-based biomarkers specific to CRC within  
929 the Korean population, addressing the critical demand for ethnicity-specific microbiome research. By  
930 leveraging metagenomic sequencing and advanced computational biology analysis, this study seeks to  
931 uncover novel microbial signatures associated with Korean CRC patients. As part of the larger "Multi-  
932 genomic analysis for biomarker development in colon cancer" project (NTIS No. 1711055951), this study  
933 investigates microbial signatures within next-generation sequencing data to enhance precision medicine  
934 approaches for CRC and to develop robust microbiome-based biomarkers for early detection, prognosis,  
935 and therapeutic stratification, complementing genomic and epigenomic markers. Hence, this research  
936 represents a crucial step toward personalized cancer diagnostic and therapeutic strategies tailored to the  
937 Korean population.

938 **4.2 Materials and methods**

939 **4.2.1 Study participants enrollment**

940 To achieve metagenomic observations of CRC, a total of 211 Korean CRC patients were enrolled (Table  
941 8). The tissue samples were collected from both the tumor lesion and its corresponding adjacent normal  
942 lesion to enable comparative metagenomic analyses. Tumor tissue samples were obtained from confirmed  
943 CRC lesions, ensuring adequate representation of CRC-associated microbial alterations. Adjacent normal  
944 tissues were collected from non-cancerous regions away from the tumor margin to serve as a control  
945 for baseline molecular and microbial composition. Moreover, clinical information was collected for all  
946 study participants included in this study to investigate potential associations between gut microbiome  
947 compositions and clinical outcomes. Key clinical characteristics recorded included overall survival (OS),  
948 recurrence, age at diagnosis and sex. Additionally, microsatellite instability (MSI) status, a critical  
949 molecular feature of CRC (Boland & Goel, 2010; Söreide, Janssen, Söiland, Körner, & Baak, 2006),  
950 was evaluated using next-generation sequencing methods to classify CRC as MSI-high, MSI-low, or  
951 microsatellite stable (MSS). These clinical parameters were integrated with metagenomic data to explore  
952 potential microbiome-based biomarkers for CRC prognosis and progression. Ethical approval was  
953 obtained for clinical data collection, and all patient information was anonymized to ensure confidentiality  
954 in accordance with institutional guidelines.

955 **4.2.2 DNA extraction procedure**

956 Tissue samples were immediately processed under sterile conditions to prevent contamination and  
957 preserved in -80 °Cstorage for downstream DNA extraction and whole-genome sequencing. Furthermore,  
958 produced sequencing data were provided by the "Multi-genomic analysis for biomarker development  
959 in colon cancer" project (NTIS No. 1711055951) in mapped BAM format, aligned to the hg38 human  
960 reference genome. The preprocessing pipeline utilized by the main project included high-throughput  
961 whole-genome sequencing using standardized alignment algorithm, BWA (H. Li & Durbin, 2009).  
962 In addition to the mapped human sequences, our whole-genome sequencing data retained unmapped  
963 sequences, which contain potential microbial reads that were not aligned to the human reference genome.

964 **4.2.3 Bioinformatics analysis**

965 To identify microbial signatures associated with CRC, we employed PathSeq (Kostic et al., 2011; Walker  
966 et al., 2018), a computational pipeline designed for metagenomic analysis of high-throughput sequencing  
967 data including the whole-genome sequences. After processing these sequencing data through the PathSeq  
968 pipeline, a comprehensive bioinformatics analyses were conducted to characterize microbial signatures  
969 associated with CRC. Prevalent taxa identification was performed by determining microbial taxa present  
970 in the majority of the study participants, filtering out low-abundance and rare taxa to ensure robust down-  
971 stream analyses. To assess microbial community structure, diversity indices were calculated, including  
972 alpha-diversity to evaluate single-sample diversity and beta-diversity to compare microbial composition

973 between the tumor tissues and their corresponding adjacent normal tissues. Differentially abundant taxa  
974 (DAT) were identified using statistical method, (DESeq2, ANCOM), adjusting for sequencing depth and  
975 potential confounders to highlight taxa significantly associated with CRC. To explore functional implica-  
976 tions, microbial pathway prediction was performed using (PICRUSt3, HUMAnN3), linking microbial  
977 composition to metabolic and functional pathways relevant to carcinogenesis and progression of CRC.  
978 This multi-layered bioinformatics approach enabled a comprehensive investigation of gut microbiome  
979 alteration in CRC, facilitating the identification of potential microbial biomarkers for diagnosis and  
980 prognosis of CRC.

981 **4.2.4 Data and code availability**

982 All sequences from the 211 study participants have been published to the Korea Bioinformation Center  
983 (data ID KGD10008857): <https://kbds.re.kr/KGD10008857>. Docker image that employed through-  
984 out this study is available in the DockerHub: <https://hub.docker.com/repository/docker/fumire/unist-crc-copm/general>. Every code used in this study can be found on GitHub: <https://github.com/CompbioLabUnist/CoPM-ColonCancer>.

987 **4.3 Results**

988 **4.3.1 Summary of clinical characteristics**

989 **4.3.2 Gut microbiome compositions**

990 **4.3.3 Diversity indices**

991 **4.3.4 DAT selection**

992 **4.3.5 Pathway prediction**

Table 8: Clinical characteristics of the study participants

993 **4.4 Discussion**

<sup>994</sup> **5 Conclusion**

<sup>995</sup> In conclusion, the research described in this doctoral dissertation was conducted to identify significant ...

<sup>996</sup> In the section 2, I show that

# <sup>997</sup> References

- <sup>998</sup> Aagaard, K., Ma, J., Antony, K. M., Ganu, R., Petrosino, J., & Versalovic, J. (2014). The placenta harbors  
<sup>999</sup> a unique microbiome. *Science translational medicine*, 6(237), 237ra65–237ra65.
- <sup>1000</sup> Abu-Ghazaleh, N., Chua, W. J., & Gopalan, V. (2021). Intestinal microbiota and its association with  
<sup>1001</sup> colon cancer and red/processed meat consumption. *Journal of gastroenterology and hepatology*,  
<sup>1002</sup> 36(1), 75–88.
- <sup>1003</sup> Abusleme, L., Hoare, A., Hong, B.-Y., & Diaz, P. I. (2021). Microbial signatures of health, gingivitis,  
<sup>1004</sup> and periodontitis. *Periodontology 2000*, 86(1), 57–78.
- <sup>1005</sup> Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., & Pawlowsky-Glahn, V. (2000). Logratio  
<sup>1006</sup> analysis and compositional distance. *Mathematical geology*, 32, 271–275.
- <sup>1007</sup> Aja, E., Mangar, M., Fletcher, H., & Mishra, A. (2021). Filifactor alocis: recent insights and advances.  
<sup>1008</sup> *Journal of dental research*, 100(8), 790–797.
- <sup>1009</sup> Alelyani, S. (2021). Stable bagging feature selection on medical data. *Journal of Big Data*, 8(1), 11.
- <sup>1010</sup> Altabtbaei, K., Maney, P., Ganesan, S. M., Dabdoub, S. M., Nagaraja, H. N., & Kumar, P. S. (2021). Anna  
<sup>1011</sup> karenina and the subgingival microbiome associated with periodontitis. *Microbiome*, 9, 1–15.
- <sup>1012</sup> Altingöz, S. M., Kurgan, Ş., Önder, C., Serdar, M. A., Ünlütürk, U., Uyanık, M., ... Günhan, M.  
<sup>1013</sup> (2021). Salivary and serum oxidative stress biomarkers and advanced glycation end products in  
<sup>1014</sup> periodontitis patients with or without diabetes: A cross-sectional study. *Journal of periodontology*,  
<sup>1015</sup> 92(9), 1274–1285.
- <sup>1016</sup> Alverdy, J., Hyoju, S., Weigerinck, M., & Gilbert, J. (2017). The gut microbiome and the mechanism of  
<sup>1017</sup> surgical infection. *Journal of British Surgery*, 104(2), e14–e23.
- <sup>1018</sup> An, S., & Park, S. (2022). Association of physical activity and sedentary behavior with the risk of  
<sup>1019</sup> colorectal cancer. *Journal of Korean Medical Science*, 37(19).
- <sup>1020</sup> Anderson, M. J. (2014). Permutational multivariate analysis of variance (permanova). *Wiley statsref:  
1021 statistics reference online*, 1–15.
- <sup>1022</sup> Aruni, A. W., Mishra, A., Dou, Y., Chioma, O., Hamilton, B. N., & Fletcher, H. M. (2015). Filifactor  
<sup>1023</sup> alocis—a new emerging periodontal pathogen. *Microbes and infection*, 17(7), 517–530.
- <sup>1024</sup> Aziz, Q., & Thompson, D. G. (1998). Brain-gut axis in health and disease. *Gastroenterology*, 114(3),  
<sup>1025</sup> 559–578.
- <sup>1026</sup> Bai, X., Wei, H., Liu, W., Coker, O. O., Gou, H., Liu, C., ... others (2022). Cigarette smoke promotes  
<sup>1027</sup> colorectal cancer through modulation of gut microbiota and related metabolites. *Gut*, 71(12),

- 1028 2439–2450.
- 1029 Baldelli, V., Scaldaferri, F., Putignani, L., & Del Chierico, F. (2021). The role of enterobacteriaceae in  
1030 gut microbiota dysbiosis in inflammatory bowel diseases. *Microorganisms*, 9(4), 697.
- 1031 Bardou, M., Rouland, A., Martel, M., Loffroy, R., Barkun, A. N., & Chapelle, N. (2022). Obesity and  
1032 colorectal cancer. *Alimentary Pharmacology & Therapeutics*, 56(3), 407–418.
- 1033 Barlow, G. M., Yu, A., & Mathur, R. (2015). Role of the gut microbiome in obesity and diabetes mellitus.  
1034 *Nutrition in clinical practice*, 30(6), 787–797.
- 1035 Basavaprabhu, H., Sonu, K., & Prabha, R. (2020). Mechanistic insights into the action of probiotics  
1036 against bacterial vaginosis and its mediated preterm birth: An overview. *Microbial pathogenesis*,  
1037 141, 104029.
- 1038 Belstrøm, D., Constancias, F., Drautz-Moses, D. I., Schuster, S. C., Veleba, M., Mahé, F., & Givskov, M.  
1039 (2021). Periodontitis associates with species-specific gene expression of the oral microbiota. *npj  
1040 Biofilms and Microbiomes*, 7(1), 76.
- 1041 Berger, W. H., & Parker, F. L. (1970). Diversity of planktonic foraminifera in deep-sea sediments.  
1042 *Science*, 168(3937), 1345–1347.
- 1043 Berghella, V. (2012). Universal cervical length screening for prediction and prevention of preterm birth.  
1044 *Obstetrical & gynecological survey*, 67(10), 653–657.
- 1045 Blencowe, H., Cousens, S., Oestergaard, M. Z., Chou, D., Moller, A.-B., Narwal, R., ... others (2012).  
1046 National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends  
1047 since 1990 for selected countries: a systematic analysis and implications. *The lancet*, 379(9832),  
1048 2162–2172.
- 1049 Boland, C. R., & Goel, A. (2010). Microsatellite instability in colorectal cancer. *Gastroenterology*,  
1050 138(6), 2073–2087.
- 1051 Boleij, A., Hechenbleikner, E. M., Goodwin, A. C., Badani, R., Stein, E. M., Lazarev, M. G., ... others  
1052 (2015). The bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer  
1053 patients. *Clinical Infectious Diseases*, 60(2), 208–215.
- 1054 Bolstad, A., Jensen, H. B., & Bakken, V. (1996). Taxonomy, biology, and periodontal aspects of  
1055 fusobacterium nucleatum. *Clinical microbiology reviews*, 9(1), 55–71.
- 1056 Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... others  
1057 (2019). Reproducible, interactive, scalable and extensible microbiome data science using qiime 2.  
1058 *Nature biotechnology*, 37(8), 852–857.
- 1059 Bombin, A., Yan, S., Bombin, S., Mosley, J. D., & Ferguson, J. F. (2022). Obesity influences composition  
1060 of salivary and fecal microbiota and impacts the interactions between bacterial taxa. *Physiological  
1061 reports*, 10(7), e15254.
- 1062 Bonnet, M., Buc, E., Sauvanet, P., Darcha, C., Dubois, D., Pereira, B., ... Darfeuille-Michaud, A. (2014).  
1063 Colonization of the human gut by e. coli and colorectal cancer risk. *Clinical Cancer Research*,  
1064 20(4), 859–867.
- 1065 Breiman, L. (2001). Random forests. *Machine learning*, 45, 5–32.
- 1066 Brennan, C. A., & Garrett, W. S. (2019). Fusobacterium nucleatum—symbiont, opportunist and

- 1067 oncobacterium. *Nature Reviews Microbiology*, 17(3), 156–166.
- 1068 Broom, L. J., & Kogut, M. H. (2018). The role of the gut microbiome in shaping the immune system of  
1069 chickens. *Veterinary immunology and immunopathology*, 204, 44–51.
- 1070 Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: improving accuracy of classifier  
1071 ensembles by using random feature subsets. *Pattern recognition*, 36(6), 1291–1302.
- 1072 Bullman, S., Pedamallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., ... others (2017). Analysis  
1073 of fusobacterium persistence and antibiotic response in colorectal cancer. *Science*, 358(6369),  
1074 1443–1448.
- 1075 Burt, R. W., Leppert, M. F., Slattery, M. L., Samowitz, W. S., Spirio, L. N., Kerber, R. A., ... others  
1076 (2004). Genetic testing and phenotype in a large kindred with attenuated familial adenomatous  
1077 polyposis. *Gastroenterology*, 127(2), 444–451.
- 1078 Cai, Y., Li, Y., Xiong, Y., Geng, X., Kang, Y., & Yang, Y. (2024). Diabetic foot exacerbates gut  
1079 mycobiome dysbiosis in adult patients with type 2 diabetes mellitus: revealing diagnostic markers.  
1080 *Nutrition & Diabetes*, 14(1), 71.
- 1081 Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016).  
1082 Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7),  
1083 581–583.
- 1084 Canakci, V., & Canakci, C. F. (2007). Pain levels in patients during periodontal probing and mechanical  
1085 non-surgical therapy. *Clinical oral investigations*, 11, 377–383.
- 1086 Cappellato, M., Baruzzo, G., & Di Camillo, B. (2022). Investigating differential abundance methods in  
1087 microbiome data: A benchmark study. *PLoS computational biology*, 18(9), e1010467.
- 1088 Castaner, O., Goday, A., Park, Y.-M., Lee, S.-H., Magkos, F., Shiow, S.-A. T. E., & Schröder, H. (2018).  
1089 The gut microbiome profile in obesity: a systematic review. *International journal of endocrinology*,  
1090 2018(1), 4095789.
- 1091 Center, M. M., Jemal, A., Smith, R. A., & Ward, E. (2009). Worldwide variations in colorectal cancer.  
1092 *CA: a cancer journal for clinicians*, 59(6), 366–378.
- 1093 Centor, R. M. (1991). Signal detectability: the use of roc curves and their analyses. *Medical decision  
1094 making*, 11(2), 102–106.
- 1095 Cerqueira, F. M., Photenhauer, A. L., Pollet, R. M., Brown, H. A., & Koropatkin, N. M. (2020). Starch  
1096 digestion by gut bacteria: crowdsourcing for carbs. *Trends in Microbiology*, 28(2), 95–108.
- 1097 Champagne, C., McNairn, H., Daneshfar, B., & Shang, J. (2014). A bootstrap method for assessing  
1098 classification accuracy and confidence for agricultural land use mapping in canada. *International  
1099 Journal of Applied Earth Observation and Geoinformation*, 29, 44–52.
- 1100 Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian  
1101 Journal of statistics*, 265–270.
- 1102 Chao, A., & Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the  
1103 American statistical Association*, 87(417), 210–217.
- 1104 Chapple, I. L., Mealey, B. L., Van Dyke, T. E., Bartold, P. M., Dommisch, H., Eickholz, P., ... others  
1105 (2018). Periodontal health and gingival diseases and conditions on an intact and a reduced

- 1106 periodontium: Consensus report of workgroup 1 of the 2017 world workshop on the classification  
1107 of periodontal and peri-implant diseases and conditions. *Journal of periodontology*, 89, S74–S84.
- 1108 Chen, T., Marsh, P., & Al-Hebshi, N. (2022). Smdi: an index for measuring subgingival microbial  
1109 dysbiosis. *Journal of dental research*, 101(3), 331–338.
- 1110 Chen, T., Yu, W.-H., Izard, J., Baranova, O. V., Lakshmanan, A., & Dewhirst, F. E. (2010). The human  
1111 oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and  
1112 genomic information. *Database*, 2010.
- 1113 Chen, X., D’Souza, R., & Hong, S.-T. (2013). The role of gut microbiota in the gut-brain axis: current  
1114 challenges and perspectives. *Protein & cell*, 4, 403–414.
- 1115 Chen, X., Jansen, L., Guo, F., Hoffmeister, M., Chang-Claude, J., & Brenner, H. (2021). Smoking,  
1116 genetic predisposition, and colorectal cancer risk. *Clinical and translational gastroenterology*,  
1117 12(3), e00317.
- 1118 Chen, X., Li, H., Guo, F., Hoffmeister, M., & Brenner, H. (2022). Alcohol consumption, polygenic risk  
1119 score, and early-and late-onset colorectal cancer risk. *EClinicalMedicine*, 49.
- 1120 Chew, R. J. J., Tan, K. S., Chen, T., Al-Hebshi, N. N., & Goh, C. E. (2024). Quantifying periodontitis-  
1121 associated oral dysbiosis in tongue and saliva microbiomes—an integrated data analysis. *Journal  
1122 of Periodontology*.
- 1123 Čižmárová, B., Tomečková, V., Hubková, B., Hurajtová, A., Ohlasová, J., & Birková, A. (2022). Salivary  
1124 redox homeostasis in human health and disease. *International Journal of Molecular Sciences*,  
1125 23(17), 10076.
- 1126 Cullin, N., Antunes, C. A., Straussman, R., Stein-Thoeringer, C. K., & Elinav, E. (2021). Microbiome  
1127 and cancer. *Cancer Cell*, 39(10), 1317–1341.
- 1128 Dabke, K., Hendrick, G., Devkota, S., et al. (2019). The gut microbiome and metabolic syndrome. *The  
1129 Journal of clinical investigation*, 129(10), 4050–4057.
- 1130 DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... Andersen, G. L.  
1131 (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with  
1132 arb. *Applied and environmental microbiology*, 72(7), 5069–5072.
- 1133 Doyle, R., Alber, D., Jones, H., Harris, K., Fitzgerald, F., Peebles, D., & Klein, N. (2014). Term and  
1134 preterm labour are associated with distinct microbial community structures in placental membranes  
1135 which are independent of mode of delivery. *Placenta*, 35(12), 1099–1101.
- 1136 Fahmy, C. A., Gamal-Eldeen, A. M., El-Hussieny, E. A., Raafat, B. M., Mehanna, N. S., Talaat, R. M., &  
1137 Shaaban, M. T. (2019). Bifidobacterium longum suppresses murine colorectal cancer through the  
1138 modulation of oncomirs and tumor suppressor mirnas. *Nutrition and cancer*, 71(4), 688–700.
- 1139 Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological conservation*, 61(1),  
1140 1–10.
- 1141 Fettweis, J. M., Serrano, M. G., Brooks, J. P., Edwards, D. J., Girerd, P. H., Parikh, H. I., ... others  
1142 (2019). The vaginal microbiome and preterm birth. *Nature medicine*, 25(6), 1012–1021.
- 1143 Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and  
1144 the number of individuals in a random sample of an animal population. *The Journal of Animal*

- 1145        *Ecology*, 42–58.
- 1146    Flanagan, L., Schmid, J., Ebert, M., Soucek, P., Kunicka, T., Liska, V., ... others (2014). Fusobacterium  
1147        nucleatum associates with stages of colorectal neoplasia development, colorectal cancer and disease  
1148        outcome. *European journal of clinical microbiology & infectious diseases*, 33, 1381–1390.
- 1149    Fortenberry, J. D. (2013). The uses of race and ethnicity in human microbiome research. *Trends in  
1150        microbiology*, 21(4), 165–166.
- 1151    Francescone, R., Hou, V., & Grivennikov, S. I. (2014). Microbiome, inflammation, and cancer. *The  
1152        Cancer Journal*, 20(3), 181–189.
- 1153    Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4),  
1154        367–378.
- 1155    Gambin, D. J., Vitali, F. C., De Carli, J. P., Mazzon, R. R., Gomes, B. P., Duque, T. M., & Trentin, M. S.  
1156        (2021). Prevalence of red and orange microbial complexes in endodontic-periodontal lesions: a  
1157        systematic review and meta-analysis. *Clinical Oral Investigations*, 1–14.
- 1158    Gao, J., Yin, J., Xu, K., Li, T., & Yin, Y. (2019). What is the impact of diet on nutritional diarrhea  
1159        associated with gut microbiota in weaning piglets: a system review. *BioMed research international*,  
1160        2019(1), 6916189.
- 1161    Ghanavati, R., Akbari, A., Mohammadi, F., Asadollahi, P., Javadi, A., Talebi, M., & Rohani, M. (2020).  
1162        Lactobacillus species inhibitory effect on colorectal cancer progression through modulating the  
1163        wnt/β-catenin signaling pathway. *Molecular and Cellular Biochemistry*, 470, 1–13.
- 1164    Ghorbani, E., Avan, A., Ryzhikov, M., Ferns, G., Khazaei, M., & Soleimanpour, S. (2022). Role of  
1165        lactobacillus strains in the management of colorectal cancer: An overview of recent advances.  
1166        *Nutrition*, 103, 111828.
- 1167    Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current  
1168        understanding of the human microbiome. *Nature medicine*, 24(4), 392–400.
- 1169    Gini, C. (1912). Variabilità e mutabilità (variability and mutability). *Tipografia di Paolo Cuppini,  
1170        Bologna, Italy*, 156.
- 1171    Goldenberg, R. L., Culhane, J. F., Iams, J. D., & Romero, R. (2008). Epidemiology and causes of preterm  
1172        birth. *The lancet*, 371(9606), 75–84.
- 1173    Gonçalves, L., Subtil, A., Oliveira, M. R., & de Zea Bermudez, P. (2014). Roc curve estimation: An  
1174        overview. *REVSTAT-Statistical journal*, 12(1), 1–20.
- 1175    Goodyear, M. D., Krleza-Jeric, K., & Lemmens, T. (2007). *The declaration of helsinki* (Vol. 335) (No.  
1176        7621). British Medical Journal Publishing Group.
- 1177    Haffajee, A., Teles, R., & Socransky, S. (2006). Association of eubacterium nodatum and treponema  
1178        denticola with human periodontitis lesions. *Oral microbiology and immunology*, 21(5), 269–282.
- 1179    Hajishengallis, G. (2015). Periodontitis: from microbial immune subversion to systemic inflammation.  
1180        *Nature reviews immunology*, 15(1), 30–44.
- 1181    Hamjane, N., Mechita, M. B., Nourouti, N. G., & Barakat, A. (2024). Gut microbiota dysbiosis-associated  
1182        obesity and its involvement in cardiovascular diseases and type 2 diabetes. a systematic review.  
1183        *Microvascular Research*, 151, 104601.

- 1184 Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*,  
1185 29(2), 147–160.
- 1186 Hampel, H., Frankel, W. L., Martin, E., Arnold, M., Khanduja, K., Kuebler, P., ... others (2008).  
1187 Feasibility of screening for lynch syndrome among patients with colorectal cancer. *Journal of*  
1188 *Clinical Oncology*, 26(35), 5783–5788.
- 1189 Han, Y. W. (2015). Fusobacterium nucleatum: a commensal-turned pathogen. *Current opinion in*  
1190 *microbiology*, 23, 141–147.
- 1191 Han, Y. W., & Wang, X. (2013). Mobile microbiome: oral bacteria in extra-oral infections and  
1192 inflammation. *Journal of dental research*, 92(6), 485–491.
- 1193 Hand, D. J. (2012). Assessing the performance of classification methods. *International Statistical Review*,  
1194 80(3), 400–414.
- 1195 Hartstra, A. V., Bouter, K. E., Bäckhed, F., & Nieuwdorp, M. (2015). Insights into the role of the  
1196 microbiome in obesity and type 2 diabetes. *Diabetes care*, 38(1), 159–165.
- 1197 Hashemi Goradel, N., Heidarzadeh, S., Jahangiri, S., Farhood, B., Mortezaee, K., Khanlarkhani, N., &  
1198 Negahdari, B. (2019). Fusobacterium nucleatum and colorectal cancer: A mechanistic overview.  
1199 *Journal of Cellular Physiology*, 234(3), 2337–2344.
- 1200 Helmink, B. A., Khan, M. W., Hermann, A., Gopalakrishnan, V., & Wargo, J. A. (2019). The microbiome,  
1201 cancer, and cancer therapy. *Nature medicine*, 25(3), 377–388.
- 1202 Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2),  
1203 427–432.
- 1204 Hiranmayi, K. V., Sirisha, K., Rao, M. R., & Sudhakar, P. (2017). Novel pathogens in periodontal  
1205 microbiology. *Journal of Pharmacy and Bioallied Sciences*, 9(3), 155–163.
- 1206 Honda, K., & Littman, D. R. (2012). The microbiome in infectious disease and inflammation. *Annual*  
1207 *review of immunology*, 30(1), 759–795.
- 1208 Honest, H., Forbes, C., Durée, K., Norman, G., Duffy, S., Tsourapas, A., ... others (2009). Screening to  
1209 prevent spontaneous preterm birth: systematic reviews of accuracy and effectiveness literature with  
1210 economic modelling. *Health Technol Assess*, 13(43), 1–627.
- 1211 Hong, Y. M., Lee, J., Cho, D. H., Jeon, J. H., Kang, J., Kim, M.-G., ... J. K. (2023). Predicting preterm  
1212 birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1), 21105.
- 1213 Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations.  
1214 *International journal of data mining & knowledge management process*, 5(2), 1.
- 1215 Huang, R.-Y., Lin, C.-D., Lee, M.-S., Yeh, C.-L., Shen, E.-C., Chiang, C.-Y., ... Fu, E. (2007). Mandibular  
1216 disto-lingual root: a consideration in periodontal therapy. *Journal of periodontology*, 78(8), 1485–  
1217 1490.
- 1218 Iams, J. D., & Berghella, V. (2010). Care for women with prior preterm birth. *American journal of*  
1219 *obstetrics and gynecology*, 203(2), 89–100.
- 1220 Ide, M., & Papapanou, P. N. (2013). Epidemiology of association between maternal periodontal  
1221 disease and adverse pregnancy outcomes—systematic review. *Journal of clinical periodontology*,  
1222 40, S181–S194.

- 1223 Iniesta, M., Chamorro, C., Ambrosio, N., Marín, M. J., Sanz, M., & Herrera, D. (2023). Subgingival  
1224 microbiome in periodontal health, gingivitis and different stages of periodontitis. *Journal of*  
1225 *Clinical Periodontology*, 50(7), 905–920.
- 1226 Inra, J. A., Steyerberg, E. W., Grover, S., McFarland, A., Syngal, S., & Kastrinos, F. (2015). Racial  
1227 variation in frequency and phenotypes of apc and mutyh mutations in 6,169 individuals undergoing  
1228 genetic testing. *Genetics in Medicine*, 17(10), 815–821.
- 1229 Janda, J. M., & Abbott, S. L. (2007). 16s rrna gene sequencing for bacterial identification in the diagnostic  
1230 laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9), 2761–2764.
- 1231 Jiang, W., & Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach  
1232 for estimating the prediction error in microarray classification. *Statistics in medicine*, 26(29),  
1233 5320–5334.
- 1234 John, G. K., & Mullin, G. E. (2016). The gut microbiome and obesity. *Current oncology reports*, 18,  
1235 1–7.
- 1236 Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., . . . others (2019).  
1237 Evaluation of 16s rrna gene sequencing for species and strain-level microbiome analysis. *Nature*  
1238 *communications*, 10(1), 5029.
- 1239 Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N., & Whiteley, M. (2014). Metatranscriptomics  
1240 of the human oral microbiome during health and disease. *MBio*, 5(2), 10–1128.
- 1241 Joscelyn, J., & Kasper, L. H. (2014). Digesting the emerging role for the gut microbiome in central  
1242 nervous system demyelination. *Multiple Sclerosis Journal*, 20(12), 1553–1559.
- 1243 Kang, Y., Kang, X., Yang, H., Liu, H., Yang, X., Liu, Q., . . . others (2022). Lactobacillus acidophilus ame-  
1244 liorates obesity in mice through modulation of gut microbiota dysbiosis and intestinal permeability.  
1245 *Pharmacological research*, 175, 106020.
- 1246 Karched, M., Bhardwaj, R. G., Qudeimat, M., Al-Khabbaz, A., & Ellepol, A. (2022). Proteomic analysis  
1247 of the periodontal pathogen prevotella intermedia secretomes in biofilm and planktonic lifestyles.  
1248 *Scientific Reports*, 12(1), 5636.
- 1249 Katz, J., Chegini, N., Shiverick, K., & Lamont, R. (2009). Localization of p. gingivalis in preterm delivery  
1250 placenta. *Journal of dental research*, 88(6), 575–578.
- 1251 Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., & Gordon, J. I. (2011). Human nutrition, the  
1252 gut microbiome and the immune system. *Nature*, 474(7351), 327–336.
- 1253 Kelly, B. J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J. D., Collman, R. G., . . . Li, H. (2015).  
1254 Power and sample-size estimation for microbiome studies using pairwise distances and permanova.  
1255 *Bioinformatics*, 31(15), 2461–2468.
- 1256 Kennedy, J., Alexander, P., Taillie, L. S., & Jaacks, L. M. (2024). Estimated effects of reductions in  
1257 processed meat consumption and unprocessed red meat consumption on occurrences of type 2  
1258 diabetes, cardiovascular disease, colorectal cancer, and mortality in the usa: a microsimulation  
1259 study. *The Lancet Planetary Health*, 8(7), e441–e451.
- 1260 Kim, B.-R., Shin, J., Guevarra, R. B., Lee, J. H., Kim, D. W., Seol, K.-H., . . . Isaacson, R. E. (2017).  
1261 Deciphering diversity indices for a better understanding of microbial communities. *Journal of*

- 1262        *Microbiology and Biotechnology*, 27(12), 2089–2093.
- 1263    Kim, C. H. (2018). Immune regulation by microbiome metabolites. *Immunology*, 154(2), 220–229.
- 1264    Kim, E.-H., Kim, S., Kim, H.-J., Jeong, H.-o., Lee, J., Jang, J., ... others (2020). Prediction of chronic  
1265        periodontitis severity using machine learning models based on salivary bacterial copy number.  
1266        *Frontiers in Cellular and Infection Microbiology*, 10, 571515.
- 1267    Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and  
1268        bootstrap. *Computational statistics & data analysis*, 53(11), 3735–3745.
- 1269    Kinane, D. F., Stathopoulou, P. G., & Papapanou, P. N. (2017). Periodontal diseases. *Nature reviews  
1270        Disease primers*, 3(1), 1–14.
- 1271    Kindinger, L. M., Bennett, P. R., Lee, Y. S., Marchesi, J. R., Smith, A., Caciato, S., ... MacIntyre,  
1272        D. A. (2017). The interaction between vaginal microbiota, cervical length, and vaginal progesterone  
1273        treatment for preterm birth risk. *Microbiome*, 5, 1–14.
- 1274    Kogut, M. H., Lee, A., & Santin, E. (2020). Microbiome and pathogen interaction with the immune  
1275        system. *Poultry science*, 99(4), 1906–1913.
- 1276    Kostic, A. D., Ojesina, A. I., Pedamallu, C. S., Jung, J., Verhaak, R. G., Getz, G., & Meyerson, M. (2011).  
1277        Pathseq: software to identify or discover microbes by deep sequencing of human tissue. *Nature  
1278        biotechnology*, 29(5), 393–396.
- 1279    Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification  
1280        and combining techniques. *Artificial Intelligence Review*, 26, 159–190.
- 1281    Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., ... Watanabe, T.  
1282        (2015). Colorectal cancer. *Nature reviews. Disease primers*, 1, 15065.
- 1283    Lafaurie, G. I., Neuta, Y., Ríos, R., Pacheco-Montealegre, M., Pianeta, R., Castillo, D. M., ... oth-  
1284        ers (2022). Differences in the subgingival microbiome according to stage of periodontitis: A  
1285        comparison of two geographic regions. *PLoS one*, 17(8), e0273523.
- 1286    Lamont, R. J., & Jenkinson, H. F. (2000). Subgingival colonization by porphyromonas gingivalis. *Oral  
1287        Microbiology and Immunology: Mini-review*, 15(6), 341–349.
- 1288    Lamont, R. J., Koo, H., & Hajishengallis, G. (2018). The oral microbiota: dynamic communities and  
1289        host interactions. *Nature reviews microbiology*, 16(12), 745–759.
- 1290    Leitich, H., & Kaider, A. (2003). Fetal fibronectin—how useful is it in the prediction of preterm birth?  
1291        *BJOG: An International Journal of Obstetrics & Gynaecology*, 110, 66–70.
- 1292    Le Leu, R. K., Hu, Y., Brown, I. L., Woodman, R. J., & Young, G. P. (2010). Synbiotic intervention of  
1293        bifidobacterium lactis and resistant starch protects against colorectal cancer development in rats.  
1294        *Carcinogenesis*, 31(2), 246–251.
- 1295    León, R., Silva, N., Ovalle, A., Chaparro, A., Ahumada, A., Gajardo, M., ... Gamonal, J. (2007).  
1296        Detection of porphyromonas gingivalis in the amniotic fluid in pregnant women with a diagnosis  
1297        of threatened premature labor. *Journal of periodontology*, 78(7), 1249–1255.
- 1298    Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform.  
1299        *bioinformatics*, 25(14), 1754–1760.
- 1300    Li, N., Lu, B., Luo, C., Cai, J., Lu, M., Zhang, Y., ... Dai, M. (2021). Incidence, mortality, survival,

- 1301 risk factor and screening of colorectal cancer: A comparison among china, europe, and northern  
1302 america. *Cancer letters*, 522, 255–268.
- 1303 Li, R., Miao, Z., Liu, Y., Chen, X., Wang, H., Su, J., & Chen, J. (2024). The brain–gut–bone axis in  
1304 neurodegenerative diseases: insights, challenges, and future prospects. *Advanced Science*, 11(38),  
1305 2307971.
- 1306 Li, X., Yu, D., Wang, Y., Yuan, H., Ning, X., Rui, B., ... Li, M. (2021). The intestinal dysbiosis of  
1307 mothers with gestational diabetes mellitus (gdm) and its impact on the gut microbiota of their  
1308 newborns. *Canadian Journal of Infectious Diseases and Medical Microbiology*, 2021(1), 3044534.
- 1309 Li, Y., Qian, F., Cheng, X., Wang, D., Wang, Y., Pan, Y., ... Tian, Y. (2023). Dysbiosis of oral microbiota  
1310 and metabolite profiles associated with type 2 diabetes mellitus. *Microbiology spectrum*, 11(1),  
1311 e03796–22.
- 1312 Lim, J. W., Park, T., Tong, Y. W., & Yu, Z. (2020). The microbiome driving anaerobic digestion and  
1313 microbial analysis. In *Advances in bioenergy* (Vol. 5, pp. 1–61). Elsevier.
- 1314 Lin, H., Eggesbø, M., & Peddada, S. D. (2022). Linear and nonlinear correlation estimators unveil  
1315 undescribed taxa interactions in microbiome data. *Nature communications*, 13(1), 4946.
- 1316 Lin, H., & Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nature  
1317 communications*, 11(1), 3514.
- 1318 Lin, H., & Peddada, S. D. (2024). Multigroup analysis of compositions of microbiomes with covariate  
1319 adjustments and repeated measures. *Nature Methods*, 21(1), 83–91.
- 1320 Listgarten, M. A. (1986). Pathogenesis of periodontitis. *Journal of clinical periodontology*, 13(5),  
1321 418–425.
- 1322 Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome  
1323 medicine*, 8, 1–11.
- 1324 López-Aladid, R., Fernández-Barat, L., Alcaraz-Serrano, V., Bueno-Freire, L., Vázquez, N., Pastor-  
1325 Ibáñez, R., ... Torres, A. (2023). Determining the most accurate 16s rrna hypervariable region for  
1326 taxonomic identification from respiratory samples. *Scientific reports*, 13(1), 3974.
- 1327 Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for  
1328 rna-seq data with deseq2. *Genome biology*, 15, 1–21.
- 1329 Magnúsdóttir, S., & Thiele, I. (2018). Modeling metabolism of the human gut microbiome. *Current  
1330 opinion in biotechnology*, 51, 90–96.
- 1331 Magurran, A. E. (2021). Measuring biological diversity. *Current Biology*, 31(19), R1174–R1177.
- 1332 Mandic, M., Safizadeh, F., Niedermaier, T., Hoffmeister, M., & Brenner, H. (2023). Association of  
1333 overweight, obesity, and recent weight loss with colorectal cancer risk. *JAMA network Open*, 6(4),  
1334 e239556–e239556.
- 1335 Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically  
1336 larger than the other. *The annals of mathematical statistics*, 50–60.
- 1337 Manolis, A. A., Manolis, T. A., Melita, H., & Manolis, A. S. (2022). Gut microbiota and cardiovascular  
1338 disease: symbiosis versus dysbiosis. *Current Medicinal Chemistry*, 29(23), 4050–4077.
- 1339 Martin, C. R., Osadchiy, V., Kalani, A., & Mayer, E. A. (2018). The brain-gut-microbiome axis. *Cellular*

- 1340 and molecular gastroenterology and hepatology, 6(2), 133–148.
- 1341 Mayer, E. A., Tillisch, K., Gupta, A., et al. (2015). Gut/brain axis and the microbiota. *The Journal of*  
1342 *clinical investigation*, 125(3), 926–938.
- 1343 Melguizo-Rodríguez, L., Costela-Ruiz, V. J., Manzano-Moreno, F. J., Ruiz, C., & Illescas-Montes, R.  
1344 (2020). Salivary biomarkers and their application in the diagnosis and monitoring of the most  
1345 common oral pathologies. *International journal of molecular sciences*, 21(14), 5173.
- 1346 Merrill, L. C., & Mangano, K. M. (2023). Racial and ethnic differences in studies of the gut microbiome  
1347 and osteoporosis. *Current Osteoporosis Reports*, 21(5), 578–591.
- 1348 Miller, C. S., Ding, X., Dawson III, D. R., & Ebersole, J. L. (2021). Salivary biomarkers for discriminating  
1349 periodontitis in the presence of diabetes. *Journal of clinical periodontology*, 48(2), 216–225.
- 1350 Morita, T., Yamazaki, Y., Mita, A., Takada, K., Seto, M., Nishinoue, N., ... Maeno, M. (2010). A cohort  
1351 study on the association between periodontal disease and the development of metabolic syndrome.  
1352 *Journal of periodontology*, 81(4), 512–519.
- 1353 Na, H. S., Kim, S. Y., Han, H., Kim, H.-J., Lee, J.-Y., Lee, J.-H., & Chung, J. (2020). Identification of  
1354 potential oral microbial biomarkers for the diagnosis of periodontitis. *Journal of clinical medicine*,  
1355 9(5), 1549.
- 1356 Nemoto, T., Shiba, T., Komatsu, K., Watanabe, T., Shimogishi, M., Shibasaki, M., ... others (2021).  
1357 Discrimination of bacterial community structures among healthy, gingivitis, and periodontitis  
1358 statuses through integrated metatranscriptomic and network analyses. *Msystems*, 6(6), e00886–21.
- 1359 Nesbitt, M. J., Reynolds, M. A., Shiau, H., Choe, K., Simonsick, E. M., & Ferrucci, L. (2010). Association  
1360 of periodontitis and metabolic syndrome in the baltimore longitudinal study of aging. *Aging clinical*  
1361 *and experimental research*, 22, 238–242.
- 1362 Network, C. G. A., et al. (2012). Comprehensive molecular characterization of human colon and rectal  
1363 cancer. *Nature*, 487(7407), 330.
- 1364 Nibali, L., Sousa, V., Davrandi, M., Spratt, D., Alyahya, Q., Dopico, J., & Donos, N. (2020). Differences  
1365 in the periodontal microbiome of successfully treated and persistent aggressive periodontitis.  
1366 *Journal of Clinical Periodontology*, 47(8), 980–990.
- 1367 Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Ž., & Tomović, M. (2017). Evaluation of classification  
1368 models in machine learning. *Theory and Applications of Mathematics & Computer Science*, 7(1),  
1369 39.
- 1370 Obuchowski, N. A., & Bullen, J. A. (2018). Receiver operating characteristic (roc) curves: review of  
1371 methods with applications in diagnostic medicine. *Physics in Medicine & Biology*, 63(7), 07TR01.
- 1372 Offenbacher, S., Katz, V., Fertik, G., Collins, J., Boyd, D., Maynor, G., ... Beck, J. (1996). Periodontal  
1373 infection as a possible risk factor for preterm low birth weight. *Journal of periodontology*, 67,  
1374 1103–1113.
- 1375 Ojesina, A. I., Pedamallu, C. S., Kostic, A., Jung, J., Auclair, D., Lohr, J., ... Meyerson, M. (2013). High  
1376 throughput sequencing-based pathogen discovery in multiple myeloma. *Blood*, 122(21), 5322.
- 1377 Omundiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019). Machine learning classification techniques  
1378 for breast cancer diagnosis. In *Iop conference series: materials science and engineering* (Vol. 495,

- 1379 p. 012033).
- 1380 O'Sullivan, D. E., Sutherland, R. L., Town, S., Chow, K., Fan, J., Forbes, N., ... Brenner, D. R. (2022).  
1381 Risk factors for early-onset colorectal cancer: a systematic review and meta-analysis. *Clinical*  
1382 *gastroenterology and hepatology*, 20(6), 1229–1240.
- 1383 Paganini, D., & Zimmermann, M. B. (2017). The effects of iron fortification and supplementation on the  
1384 gut microbiome and diarrhea in infants and children: a review. *The American journal of clinical*  
1385 *nutrition*, 106, 1688S–1693S.
- 1386 Pan, A. Y. (2021). Statistical analysis of microbiome data: the challenge of sparsity. *Current Opinion in*  
1387 *Endocrine and Metabolic Research*, 19, 35–40.
- 1388 Papapanou, P. N., Sanz, M., Buduneli, N., Dietrich, T., Feres, M., Fine, D. H., ... others (2018).  
1389 Periodontitis: Consensus report of workgroup 2 of the 2017 world workshop on the classification of  
1390 periodontal and peri-implant diseases and conditions. *Journal of periodontology*, 89, S173–S182.
- 1391 Parizadeh, M., & Arrieta, M.-C. (2023). The global human gut microbiome: genes, lifestyles, and diet.  
1392 *Trends in Molecular Medicine*.
- 1393 Park, J., Park, S. H., Lee, D., Lee, J. E., Lee, D., Na, K. J., ... Im, H.-J. (2024). Detecting cancer microbiota  
1394 using unmapped rna reads on spatial transcriptomics. *Cancer Research*, 84(6\_Supplement), 4881–  
1395 4881.
- 1396 Payne, M. S., Newnham, J. P., Doherty, D. A., Furfarro, L. L., Pendal, N. L., Loh, D. E., & Keelan, J. A.  
1397 (2021). A specific bacterial dna signature in the vagina of australian women in midpregnancy  
1398 predicts high risk of spontaneous preterm birth (the predict1000 study). *American journal of*  
1399 *obstetrics and gynecology*, 224(2), 206–e1.
- 1400 Peirce, J. M., & Alviña, K. (2019). The role of inflammation and the gut microbiome in depression and  
1401 anxiety. *Journal of neuroscience research*, 97(10), 1223–1241.
- 1402 Peltomaki, P. (2003). Role of dna mismatch repair defects in the pathogenesis of human cancer. *Journal*  
1403 *of clinical oncology*, 21(6), 1174–1179.
- 1404 Pezzino, S., Sofia, M., Greco, L. P., Litrico, G., Filippello, G., Sarvà, I., ... Latteri, S. (2023). Microbiome  
1405 dysbiosis: a pathological mechanism at the intersection of obesity and glaucoma. *International*  
1406 *Journal of Molecular Sciences*, 24(2), 1166.
- 1407 Premaraj, T. S., Vella, R., Chung, J., Lin, Q., Hunter, P., Underwood, K., ... Zhou, Y. (2020). Ethnic  
1408 variation of oral microbiota in children. *Scientific reports*, 10(1), 14788.
- 1409 Raut, J. R., Schöttker, B., Holleczeck, B., Guo, F., Bhardwaj, M., Miah, K., ... Brenner, H. (2021).  
1410 A microrna panel compared to environmental and polygenic scores for colorectal cancer risk  
1411 prediction. *Nature Communications*, 12(1), 4811.
- 1412 Rebersek, M. (2021). Gut microbiome and its role in colorectal cancer. *BMC cancer*, 21(1), 1325.
- 1413 Redanz, U., Redanz, S., Treerat, P., Prakasam, S., Lin, L.-J., Merritt, J., & Kreth, J. (2021). Differential  
1414 response of oral mucosal and gingival cells to corynebacterium durum, streptococcus sanguinis, and  
1415 porphyromonas gingivalis multispecies biofilms. *Frontiers in cellular and infection microbiology*,  
1416 11, 686479.
- 1417 Relvas, M., Regueira-Iglesias, A., Balsa-Castro, C., Salazar, F., Pacheco, J., Cabral, C., ... Tomás, I.

- 1418 (2021). Relationship between dental and periodontal health status and the salivary microbiome:  
1419 bacterial diversity, co-occurrence networks and predictive models. *Scientific reports*, 11(1), 929.
- 1420 Renson, A., Jones, H. E., Beghini, F., Segata, N., Zolnik, C. P., Usyk, M., ... others (2019). Sociodemographic variation in the oral microbiome. *Annals of epidemiology*, 35, 73–80.
- 1422 Rideout, J. R., Caporaso, G., Bolyen, E., McDonald, D., Baeza, Y. V., Alastuey, J. C., ... Sharma, K.  
1423 (2018, December). *biocore/scikit-bio: scikit-bio 0.5.5: More compositional methods added*. Zenodo.  
1424 Retrieved from <https://doi.org/10.5281/zenodo.2254379> doi: 10.5281/zenodo.2254379
- 1425 Rôcas, I. N., Siqueira Jr, J. F., Santos, K. R., Coelho, A. M., & de Janeiro, R. (2001). “red complex”(*bacteroides forsythus*, *porphyromonas gingivalis*, and *treponema denticola*) in endodontic  
1426 infections: a molecular approach. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology,*  
1427 and *Endodontology*, 91(4), 468–471.
- 1429 Romero, R., Dey, S. K., & Fisher, S. J. (2014). Preterm labor: one syndrome, many causes. *Science*,  
1430 345(6198), 760–765.
- 1431 Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., ... others (2014). The  
1432 composition and stability of the vaginal microbiota of normal pregnant women is different from  
1433 that of non-pregnant women. *Microbiome*, 2, 1–19.
- 1434 Rosan, B., & Lamont, R. J. (2000). Dental plaque formation. *Microbes and infection*, 2(13), 1599–1607.
- 1435 Schwabe, R. F., & Jobin, C. (2013). The microbiome and cancer. *Nature Reviews Cancer*, 13(11),  
1436 800–812.
- 1437 Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011).  
1438 Metagenomic biomarker discovery and explanation. *Genome biology*, 12, 1–18.
- 1439 Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A  
1440 survey and review. In *Emerging technology in modelling and graphics: Proceedings of iem graph*  
1441 2018 (pp. 99–111).
- 1442 Sepich-Poore, G. D., Zitvogel, L., Straussman, R., Hasty, J., Wargo, J. A., & Knight, R. (2021). The  
1443 microbiome and human cancer. *Science*, 371(6536), eabc4552.
- 1444 Sharma, S., & Tripathi, P. (2019). Gut microbiome and type 2 diabetes: where we are and where to go?  
1445 *The Journal of nutritional biochemistry*, 63, 101–108.
- 1446 Shi, N., Li, N., Duan, X., & Niu, H. (2017). Interaction between the gut microbiome and mucosal  
1447 immune system. *Military Medical Research*, 4, 1–7.
- 1448 Simpson, E. (1949). Measurement of diversity. *Nature*, 163.
- 1449 Song, M., Chan, A. T., & Sun, J. (2020). Influence of the gut microbiome, diet, and environment on risk  
1450 of colorectal cancer. *Gastroenterology*, 158(2), 322–340.
- 1451 Söreide, K., Janssen, E., Söiland, H., Körner, H., & Baak, J. (2006). Microsatellite instability in colorectal  
1452 cancer. *Journal of British Surgery*, 93(4), 395–406.
- 1453 Sotiriadis, A., Papatheodorou, S., Kavvadias, A., & Makrydimas, G. (2010). Transvaginal cervical  
1454 length measurement for prediction of preterm birth in women with threatened preterm labor: a  
1455 meta-analysis. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International*  
1456 *Society of Ultrasound in Obstetrics and Gynecology*, 35(1), 54–64.

- 1457 Spss, I., et al. (2011). Ibm spss statistics for windows, version 20.0. *New York: IBM Corp*, 440, 394.
- 1458 Stafford, G., Roy, S., Honma, K., & Sharma, A. (2012). Sialic acid, periodontal pathogens and tannerella  
1459 forsythia: stick around and enjoy the feast! *Molecular Oral Microbiology*, 27(1), 11–22.
- 1460 Stout, M. J., Conlon, B., Landeau, M., Lee, I., Bower, C., Zhao, Q., ... Mysorekar, I. U. (2013).  
1461 Identification of intracellular bacteria in the basal plate of the human placenta in term and preterm  
1462 gestations. *American journal of obstetrics and gynecology*, 208(3), 226–e1.
- 1463 Sultan, S., El-Mowafy, M., Elgaml, A., Ahmed, T. A., Hassan, H., & Mottawea, W. (2021). Metabolic  
1464 influences of gut microbiota dysbiosis on inflammatory bowel disease. *Frontiers in physiology*, 12,  
1465 715506.
- 1466 Suzuki, N., Nakano, Y., Yoneda, M., Hirofumi, T., & Hanioka, T. (2022). The effects of cigarette  
1467 smoking on the salivary and tongue microbiome. *Clinical and Experimental Dental Research*, 8(1),  
1468 449–456.
- 1469 Swidsinski, A., Khilkin, M., Kerjaschki, D., Schreiber, S., Ortner, M., Weber, J., & Lochs, H. (1998).  
1470 Association between intraepithelial escherichia coli and colorectal cancer. *Gastroenterology*,  
1471 115(2), 281–286.
- 1472 Swift, D., Cresswell, K., Johnson, R., Stilianoudakis, S., & Wei, X. (2023). A review of normalization  
1473 and differential abundance methods for microbiome counts data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1), e1586.
- 1474 Tanner, A. C., Kent Jr, R., Kanasi, E., Lu, S. C., Paster, B. J., Sonis, S. T., ... Van Dyke, T. E. (2007).  
1475 Clinical characteristics and microbiota of progressing slight chronic periodontitis in adults. *Journal of clinical periodontology*, 34(11), 917–930.
- 1476 Tanner, A. C., Paster, B. J., Lu, S. C., Kanasi, E., Kent Jr, R., Van Dyke, T., & Sonis, S. T. (2006).  
1477 Subgingival and tongue microbiota during early periodontitis. *Journal of dental research*, 85(4),  
1478 318–323.
- 1479 Tejeda, M., Farrell, J., Zhu, C., Haines, J. L., Wang, L.-S., Schellenberg, G. D., ... others (2021). Multiple  
1480 viruses detected in human dna are associated with alzheimer disease risk. *Alzheimer's & Dementia*,  
1481 17, e054585.
- 1482 Teles, F., Wang, Y., Hajishengallis, G., Hasturk, H., & Marchesan, J. T. (2021). Impact of systemic  
1483 factors in shaping the periodontal microbiome. *Periodontology 2000*, 85(1), 126–160.
- 1484 Thaiss, C. A., Zmora, N., Levy, M., & Elinav, E. (2016). The microbiome and innate immunity. *Nature*,  
1485 535(7610), 65–74.
- 1486 Tian, R., Liu, H., Feng, S., Wang, H., Wang, Y., Wang, Y., ... Zhang, S. (2021). Gut microbiota dysbiosis  
1487 in stable coronary artery disease combined with type 2 diabetes mellitus influences cardiovascular  
1488 prognosis. *Nutrition, Metabolism and Cardiovascular Diseases*, 31(5), 1454–1466.
- 1489 Tilg, H., Kaser, A., et al. (2011). Gut microbiome, obesity, and metabolic dysfunction. *The Journal of  
1490 clinical investigation*, 121(6), 2126–2132.
- 1491 Tonetti, M. S., Greenwell, H., & Kornman, K. S. (2018). Staging and grading of periodontitis: Framework  
1492 and proposal of a new classification and case definition. *Journal of periodontology*, 89, S159–S172.
- 1493 Tringe, S. G., & Hugenholtz, P. (2008). A renaissance for the pioneering 16s rrna gene. *Current opinion*

- 1496        *in microbiology*, 11(5), 442–446.
- 1497    Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., ... others (2017). A  
1498        guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological  
1499        Reviews*, 92(2), 698–715.
- 1500    Ulger Toprak, N., Yagci, A., Gulluoglu, B., Akin, M., Demirkalem, P., Celenk, T., & Soyletir, G. (2006).  
1501        A possible role of bacteroides fragilis enterotoxin in the aetiology of colorectal cancer. *Clinical  
1502        microbiology and infection*, 12(8), 782–786.
- 1503    Ursell, L. K., Metcalf, J. L., Parfrey, L. W., & Knight, R. (2012). Defining the human microbiome.  
1504        *Nutrition reviews*, 70(suppl\_1), S38–S44.
- 1505    Utzschneider, K. M., Kratz, M., Damman, C. J., & Hullarg, M. (2016). Mechanisms linking the gut  
1506        microbiome and glucose metabolism. *The Journal of Clinical Endocrinology & Metabolism*,  
1507        101(4), 1445–1454.
- 1508    Vander Haar, E. L., So, J., Gyamfi-Bannerman, C., & Han, Y. W. (2018). Fusobacterium nucleatum and  
1509        adverse pregnancy outcomes: epidemiological and mechanistic evidence. *Anaerobe*, 50, 55–59.
- 1510    Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning  
1511        research*, 9(11).
- 1512    Vasen, H. F., Mecklin, J.-P., Khan, P. M., & Lynch, H. T. (1991). The international collaborative group  
1513        on hereditary non-polyposis colorectal cancer (icg-hnpcc). *Diseases of the Colon & Rectum*, 34(5),  
1514        424–425.
- 1515    Walker, M. A., Pedamallu, C. S., Ojesina, A. I., Bullman, S., Sharpe, T., Whelan, C. W., & Meyerson, M.  
1516        (2018). Gatk pathseq: a customizable computational tool for the discovery and identification of  
1517        microbial sequences in libraries from eukaryotic hosts. *Bioinformatics*, 34(24), 4287–4289.
- 1518    Weaver, W. (1963). *The mathematical theory of communication*. University of Illinois Press.
- 1519    Whiteside, S. A., Razvi, H., Dave, S., Reid, G., & Burton, J. P. (2015). The microbiome of the urinary  
1520        tract—a role beyond infection. *Nature Reviews Urology*, 12(2), 81–90.
- 1521    Witkin, S. (2019). Vaginal microbiome studies in pregnancy must also analyse host factors. *BJOG: An  
1522        International Journal of Obstetrics & Gynaecology*, 126(3), 359–359.
- 1523    Wong, T.-T., & Yeh, P.-Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE  
1524        Transactions on Knowledge and Data Engineering*, 32(8), 1586–1594.
- 1525    Wyss, C., Moter, A., Choi, B.-K., Dewhirst, F., Xue, Y., Schüpbach, P., ... Guggenheim, B. (2004).  
1526        *Treponema putidum* sp. nov., a medium-sized proteolytic spirochaete isolated from lesions of  
1527        human periodontitis and acute necrotizing ulcerative gingivitis. *International journal of systematic  
1528        and evolutionary microbiology*, 54(4), 1117–1122.
- 1529    Xia, Y. (2023). Statistical normalization methods in microbiome data with application to microbiome  
1530        cancer research. *Gut Microbes*, 15(2), 2244139.
- 1531    Yaman, E., & Subasi, A. (2019). Comparison of bagging and boosting ensemble machine learning methods  
1532        for automated emg signal classification. *BioMed research international*, 2019(1), 9152506.
- 1533    Yang, I., Claussen, H., Arthur, R. A., Hertzberg, V. S., Geurs, N., Corwin, E. J., & Dunlop, A. L. (2022).  
1534        Subgingival microbiome in pregnancy and a potential relationship to early term birth. *Frontiers in*

- 1535            *cellular and infection microbiology*, 12, 873683.
- 1536    Yoshimura, F., Murakami, Y., Nishikawa, K., Hasegawa, Y., & Kawaminami, S. (2009). Surface  
1537       components of *porphyromonas gingivalis*. *Journal of periodontal research*, 44(1), 1–12.
- 1538    Zhang, C.-Z., Cheng, X.-Q., Li, J.-Y., Zhang, P., Yi, P., Xu, X., & Zhou, X.-D. (2016). Saliva in the  
1539       diagnosis of diseases. *International journal of oral science*, 8(3), 133–137.
- 1540    Zhou, X., Wang, L., Xiao, J., Sun, J., Yu, L., Zhang, H., ... others (2022). Alcohol consumption,  
1541       dna methylation and colorectal cancer risk: Results from pooled cohort studies and mendelian  
1542       randomization analysis. *International journal of cancer*, 151(1), 83–94.
- 1543    Zhu, W., & Lee, S.-W. (2016). Surface interactions between two of the main periodontal pathogens:  
1544       *Porphyromonas gingivalis* and *tannerella forsythia*. *Journal of periodontal & implant science*,  
1545       46(1), 2–9.
- 1546    Zhu, X., Han, Y., Du, J., Liu, R., Jin, K., & Yi, W. (2017). Microbiota-gut-brain axis and the central  
1547       nervous system. *Oncotarget*, 8(32), 53829.
- 1548    Zhuang, Y., Wang, H., Jiang, D., Li, Y., Feng, L., Tian, C., ... others (2021). Multi gene mutation  
1549       signatures in colorectal cancer patients: predict for the diagnosis, pathological classification, staging  
1550       and prognosis. *BMC cancer*, 21, 1–16.

## Acknowledgments

1552 I would like to disclose my earnest appreciation for my advisor, Professor Semin Lee, who provided  
 1553 solicitous supervision and cherished opportunities throughout the course of my research. His advice and  
 1554 consultation encouraged me to become as a researcher and to receive all humility and gentleness. I am  
 1555 also grateful to all of my committee members, Professor AAA, Professor BBB, Professor CCC, and  
 1556 Professor DDD, for their critical and meaningful mentions and suggestions.

1557 I extend my deepest gratitude to my Lord, *the Flying Spaghetti Monster*, His Noodly Appendage  
 1558 has guided me through the twist and turns of this academic journey. His presence, ever comforting and  
 1559 mysterious, has been a source of strength and humor during both highs and lows. In moments of doubt, I  
 1560 found solace in the belief that you were there, gently reminding me to keep faith in the process. His Holy  
 1561 Noodle has nourished my mind, and for that, I am truly overwhelmed. May His Holy Noodle continue to  
 1562 guide me in all my future endeavors. *R’Amen.*

1563 (Professors)

1564 I would like to extend my heartfelt gratitude to my colleagues of the Computational Biology Lab @  
 1565 UNIST, whose collaboration, friendship, brotherhood, and support have been an invaluable part of my  
 1566 journey. Your willingness to share insights, engage in thoughtful discussions, and offer encouragement  
 1567 during the challenging moments of research has significantly shaped my academic experience. The  
 1568 camaraderie in Computational Biology Lab made even the most demanding days more enjoyable, and I  
 1569 am deeply grateful for the collaborative environment we created together. I appreciate you for standing  
 1570 by my side throughout this Ph.D. journey.

1571 I would like to express my heartfelt gratitude to my family, whose unwavering support has been the  
 1572 foundation of everything I have achieved. Your love, encouragement, and belief in me have sustained me  
 1573 through every challenge, and I could not have come this far without you. From your words of wisdom to  
 1574 your patience and understanding, each of you has played a vital role in helping me navigate this journey.  
 1575 The strength and comfort I have drawn from our family bond have been my greatest source of resilience.  
 1576 Your presence, both near and far, has filled my life with warmth and motivation. I am deeply grateful for  
 1577 your unconditional love and for always being there when I needed you the most. Thank you for being my  
 1578 constant source of strength and inspiration.

1579 I am incredibly pleased to my friends, especially my GSHS alumni (○망특), for their unwavering  
 1580 support and encouragement throughout this journey. The bonds we formed back in our school days have  
 1581 only grown stronger over the years, and I am fortunate to have had such loyal and understanding friends  
 1582 by my side. Your constant words of motivation, and even moments of levity during stressful times have  
 1583 helped keep me grounded. Whether it was a late-night conversations, a shared laugh, or a simple message  
 1584 of reassurance, you all have played a vital role in keeping me focused and motivated. I am relieved for the  
 1585 ways you celebrated each small achievement with me and how you patiently listened to my worries. The  
 1586 memories of our shared past provided me with comfort and a sense of stability when the road ahead felt  
 1587 uncertain. I could not have reached this point without the love and friendship that you all have generously  
 1588 given. Each of your, in your unique way, has contributed to this dissertation, even if indirectly, and for

1589 that, I am forever beholden. I look forward to continuing our friendship as we all grow in our individual  
1590 paths, knowing that the support we share is something truly special.

1591 (Girlfriend)

1592 I would like to express my sincere gratitude to the amazing members of my animal protection groups,  
1593 DRDR (두루두루) and UNIMALS (유니멀스), whose dedication and compassion have been a constant  
1594 source of motivation. Your unwavering commitment to improving the lives of animals has inspired me  
1595 throughout this journey. I am also thankful for the beautiful cats we have cared for, whose presence  
1596 brought both joy and purpose to our allegiance. Their playful spirits and gentle companionship served as  
1597 daily reminders of why we continue to fight for animal rights. The bond we share, both with each other  
1598 and with the animals we protect, has enriched my life in countless ways. I appreciate you all again for  
1599 your support, dedication, and for being part of this meaningful cause.

1600 I would like to express my deepest gratitude to everyone I have had the honor of meeting throughout  
1601 this journey. Your kindness, encouragement, and support have carried me through both the challenging  
1602 and rewarding moments of my life. Whether through a kind word, thoughtful advice, or simply being  
1603 there when I needed it most, your presence has made all the difference. I am incredibly fortunate to have  
1604 received such generosity and warmth from those around me, and I do not take it for granted. Every act  
1605 of kindness, no matter how big or small, has been a source of strength and motivation for me. To all  
1606 my friends, colleagues, mentors, and beloved ones, thank you for your unwavering support. I am truly  
1607 grateful for each of you, and your kindness has left an indelible mark on my journey.

1608 My Lord, *the Flying Spaghetti Monster*,  
1609 give us grace to accept with serenity the things that cannot be changed,  
1610 courage to change the things that should be changed,  
1611 and the wisdom to distinguish the one from the other.

1612  
1613 Glory be to *the Meatball*, to *the Sauce*, and to *the Holy Noodle*.  
1614 As it was in the beginning, is now, and ever shall be.

1615 *R'Amen.*



*May your progress be evident to all*

