

¹

Doctoral Thesis

²

Microbiota in Human Diseases

³

Jaewoong Lee

⁴

Department of Biomedical Engineering

⁵

Ulsan National Institute of Science and Technology

⁶

2025

7

Microbiota in Human Diseases

8

Jaewoong Lee

9

Department of Biomedical Engineering

10

Ulsan National Institute of Science and Technology



CHURCH OF THE FLYING SPAGHETTI MONSTER

February 09, 2021

Letter of Good Standing

Dear Sir or Madam:

I am pleased to verify that _____

JAEWOONG LEE

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

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

Any questions can be directed to the undersigned.

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

Representative,
Church of the Flying Spaghetti Monster
Bobby Henderson



CHURCH OF THE FLYING SPAGHETTI MONSTER

February 09, 2021

Letter of Good Standing

Dear Sir or Madam:

I am pleased to verify that _____

JAEWOONG LEE

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

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

Any questions can be directed to the undersigned.

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

Representative,
Church of the Flying Spaghetti Monster
Bobby Henderson

13

Abstract

14 (Microbiome)

15 (PTB) Section 2 introduces...

16 (Periodontitis) Section 3 describes...

17 (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	67
69	5	Conclusion	68
70	References		69
71	Acknowledgments		84

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
94	21	Gut microbiome compositions in genus level	61
95	22	Gut microbiome compositions in species level	62
96	23	Alpha-diversity indices in genus level	63
97	24	Alpha-diversity indices in species level	64
98	25	Alpha-diversity indices with recurrence in genus level	65
99	26	Alpha-diversity indices with recurrence in species level	66

List of Tables

101	1	Confusion matrix	6
102	2	Standard clinical information of study participants	13
103	3	Clinical characteristics of the study participants	32
104	4	Feature combinations and their evaluations	33
105	5	List of DAT among the periodontally healthy and periodontitis stages	34
106	6	Feature the importance of taxa in the classification of different periodontal statuses.	35
107	7	Beta-diversity pairwise comparisons on the periodontitis statuses	36
108	8	Clinical characteristics of CRC study participants	60

109

List of Abbreviations

110 **ACC** Accuracy

111 **ACE** Abundance-based coverage estimator

112 **ASV** Amplicon sequence variant

113 **AUC** Area-under-curve

114 **BA** Balanced accuracy

115 **C-section** Cesarean section

116 **DAT** Differentially abundant taxa

117 **F1** F1 score

118 **Faith PD** Faith's phylogenetic diversity

119 **FTB** Full-term birth

120 **GA** Gestational age

121 **MSI** Microsatellite instability

122 **MSS** Microsatellite stable

123 **MWU test** Mann-Whitney U-test

124 **OS** Overall survival

125 **PRE** Precision

126 **PROM** Prelabor rupture of membrane

127 **PTB** Preterm birth

128 **ROC curve** Receiver-operating characteristics curve

129 **rRNA** Ribosomal RNA

130 **SD** Standard deviation

131 **SEN** Sensitivity

132 **SPE** Specificity

133 **t-SNE** t-distributed stochastic neighbor embedding

134 **1 Introduction**

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

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

170 16S ribosomal RNA (rRNA) gene sequencing is one of the most extensively applied methods for
171 characterizing microbial communities by targeting the conserved 16S rRNA gene, which contains both

172 highly conserved and variable regions in bacteria (Tringe & Hugenholtz, 2008; Janda & Abbott, 2007).
173 The conserved regions enable universal primer binding, while the variable regions provide the specificity
174 needed to differentiate microbial taxa. Among these regions, the V3-V4 region is frequently selected for
175 sequencing due to its balance between phylogenetic resolution and sequencing efficiency (Johnson et al.,
176 2019; López-Aladid et al., 2023). Therefore, the V3-V4 region offers sufficient variability to classify a
177 wide range of bacteria taxa while maintaining compatibility with widely used sequencing platforms.

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

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

201 Differentially abundant taxa (DAT) detection is a key analytical approach in microbiome study to
202 identify microbial taxa that significantly differ in abundance between distinct study participant groups.
203 This DAT detection method is particularly valuable for understanding how microbial communities vary
204 across different conditions, such as disease states, environmental factors, and/or experimental treatments.
205 Various statistical and computational techniques, *e.g.* LEfSe (Segata et al., 2011), DESeq2 (Love, Huber,
206 & Anders, 2014), ANCOM (Lin & Peddada, 2020), and ANCOM-BC (Lin, Eggesbø, & Peddada,
207 2022; Lin & Peddada, 2024), are commonly used to assess differential abundance while accounting for
208 compositional and sparsity-related challenges in microbiome composition data (Swift, Cresswell, Johnson,
209 Stilianoudakis, & Wei, 2023; Cappellato, Baruzzo, & Di Camillo, 2022). Thus, identifying DAT can
210 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

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

256 This dissertation present a comprehensive, multi-disease human microbiome analysis, bridging the
257 association between preterm birth (PTB) (Section 2), periodontitis (Section 3), and colorectal cancer
258 (CRC) (Section 4) through a unified metagenomic approach. While previous studies have examined the
259 role and characteristics of human microbiome in these diseases individually, this dissertation uniquely
260 integrates human microbiome-driven insights across these diseases to identify shared and disease-specific
261 microbial signatures. By applying high-throughput metagenomic sequencing, microbial diversity analysis,
262 and advanced bioinformatics techniques, this dissertation aims to uncover novel microbiome-based
263 biomarkers and mechanistic insights into how microbial communities influence these conditions. These
264 findings contribute to a broader understanding of microbiome-mediated disease interactions and pave the
265 way for personalized medicine strategies, including microbiome-targeted diagnostics and therapeutics.

Table 1: Confusion matrix

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

266

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

267

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

268

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

269

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

270

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

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

271 **2 Predicting preterm birth using random forest classifier in salivary mi-**
272 **crobiome**

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

274 Hong, Y. M., **Lee, Jaewoong**, Cho, D. H., Jeon, J. H., Kang, J., Kim, M. G., ... & Kim, J. K. (2023).
275 Predicting preterm birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1),
276 21105.

277 **2.1 Introduction**

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

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

295 Preterm prelabor rupture of membranes (PROM) brought on by gestational inflammation and infection
296 contribute to about 70% of PTB cases (Romero, Dey, & Fisher, 2014). Nevertheless, as antibiotics and
297 anti-inflammatory therapeutic strategies were ineffective to decrease PTB occurrence rates, the pathology
298 of PTB has not been entirely elucidated by inflammatory and infectious pathways (Romero, Hassan, et al.,
299 2014). Recent researches on maternal microbiomes were beginning to examine unidentified connections
300 of PTB as a consequence of developmental processes in molecular biological technology (Fettweis et al.,
301 2019).

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

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

310 Multiple unfavorable birth outcomes, including PROM and PTB, have been linked to periodontitis
311 as an independence risk factor, according to numerous epidemiological researches (Offenbacher et al.,
312 1996). It is expected that the oral microbiome will be able to explain additional hematogenous pathways
313 in light of these precedents; however, the oral microbiome composition of fetuses is limited understood.

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

317 **2.2 Materials and methods**

318 **2.2.1 Study design and study participants**

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

325 **2.2.2 Clinical data collection and grouping**

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

- 328 • maternal age at delivery
- 329 • diabetes mellitus
- 330 • hypertension
- 331 • overweight and obesity
- 332 • C-section
- 333 • history PROM or PTB
- 334 • gestational week on delivery
- 335 • birth weight
- 336 • sex

337 **2.2.3 Salivary microbiome sample collection**

338 Salivary microbiome samples were collected 24 hours before to delivery using mouthwash. The standard
339 methods of sterilizing were performed. Medical experts oversaw each stage of the sample collecting
340 procedure. Participants received instruction not to eat, drink, or brush their teeth for 30 minutes before
341 sampling salivary microbiome. Saliva samples were gathered by washing the mouth for 30 seconds with
342 12 mL of a mouthwash solution (E-zен Gargle, JN Pharm, Pyeongtaek, Gyeonggi, Korea). The samples
343 were tagged with the anonymous ID for each participant and kept at 4 °C until they underwent further
344 processing. Genomic DNA was extracted using an ExgeneTM Clinic SV kit (GeneAll Biotechnology,
345 Seoul, Korea) following with the manufacturer instructions and store at -20 °C.

346 **2.2.4 16s rRNA gene sequencing**

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

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

353 **2.2.5 Bioinformatics analysis**

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

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

365 Taxonomic assignment were implemented with HOMD (version 15.22) (T. Chen et al., 2010).
366 Afterward, DESeq2 was implemented to identify differentially abundant taxa (DAT) that could dis-
367 tinguish between salivary microbiome from PTB and FTB participants (Love et al., 2014). Taxa with
368 $|\log_2 \text{FoldChange}| > 1$ and $p < 0.05$ were considered as statistically significant.

369 The taxa for predicting PTB using salivary microbiome data were determined using a random forest
370 classifier (Breiman, 2001). Through stratified *k*-fold cross-validation (*k* = 5) that preserves the existence
371 rate of PTB and FTB participants, consistency and trustworthy classification were ensured (Wong & Yeh,
372 2019).

373 **2.2.6 Data and code availability**

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

379 **2.3 Results**

380 **2.3.1 Overview of clinical information**

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

387 **2.3.2 Comparison of salivary microbiomes composition**

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

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

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

399 **2.3.3 Random forest classification to predict PTB risk**

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

Table 2: Standard clinical information of study participants.

Continuous variable for independent *t*-test. Categorical variable for Pearson's χ^2 -square test. Continuous variable: mean \pm SD. Categorical variable: count (proportion)

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

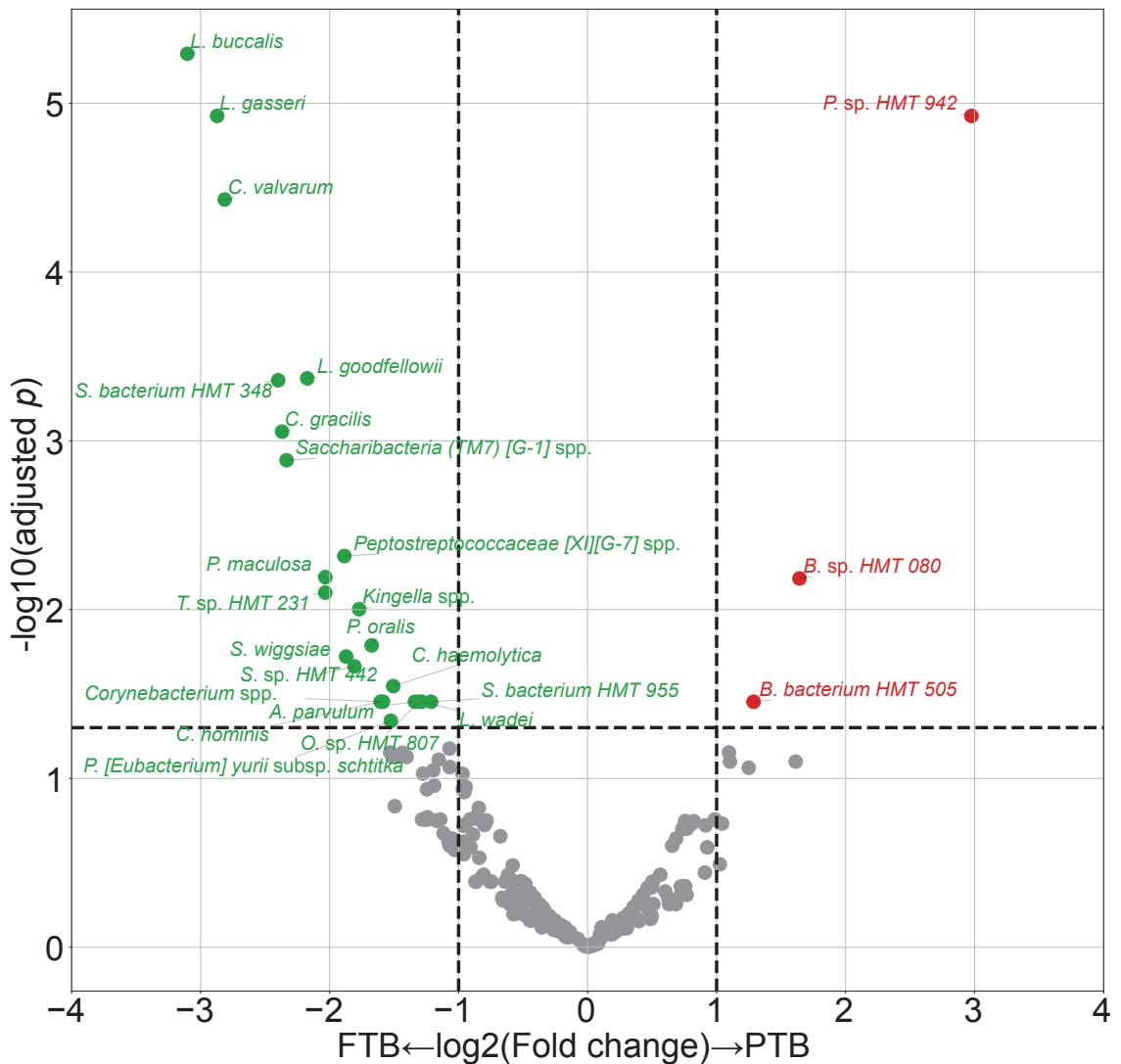


Figure 1: DAT volcano plot.

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

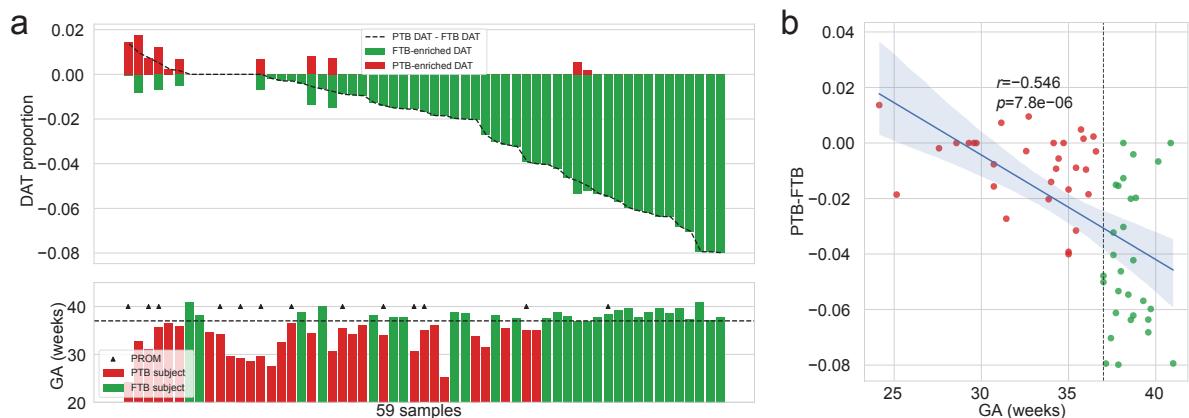


Figure 2: **Salivary microbiome compositions over DAT.**

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

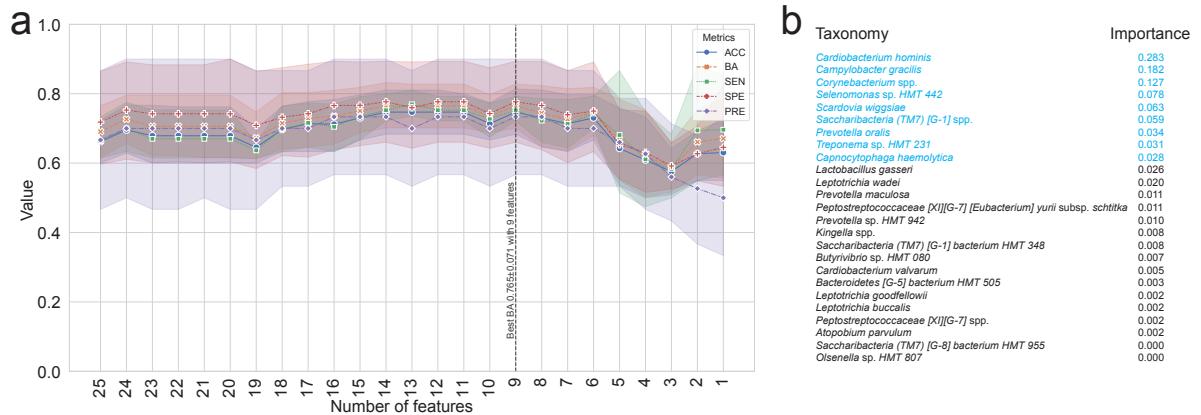


Figure 3: **Random forest-based PTB prediction model.**

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

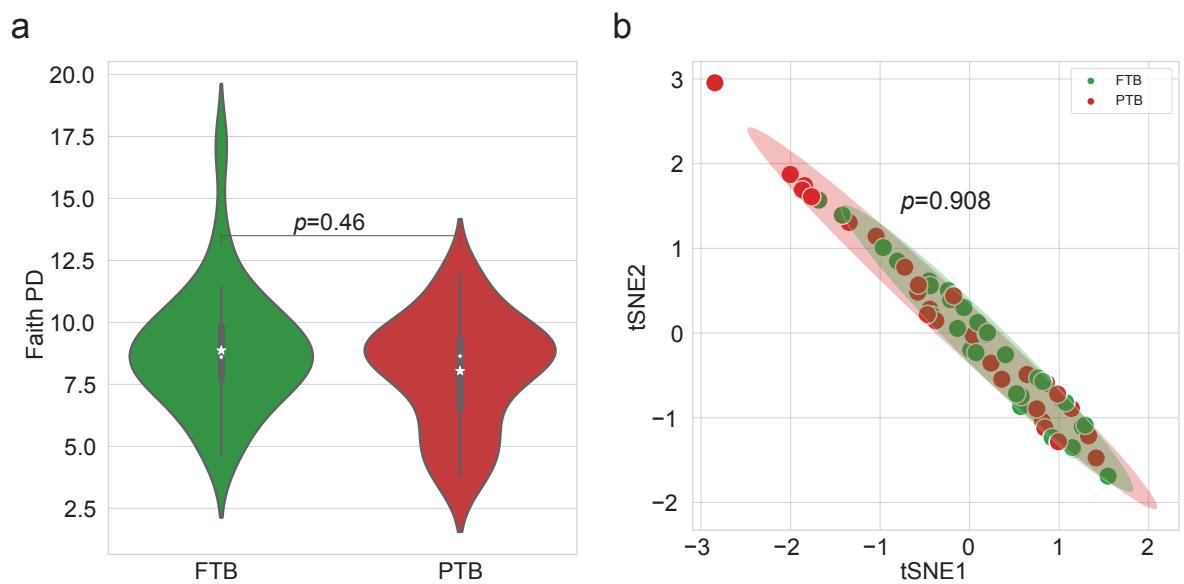


Figure 4: **Diversity indices.**

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

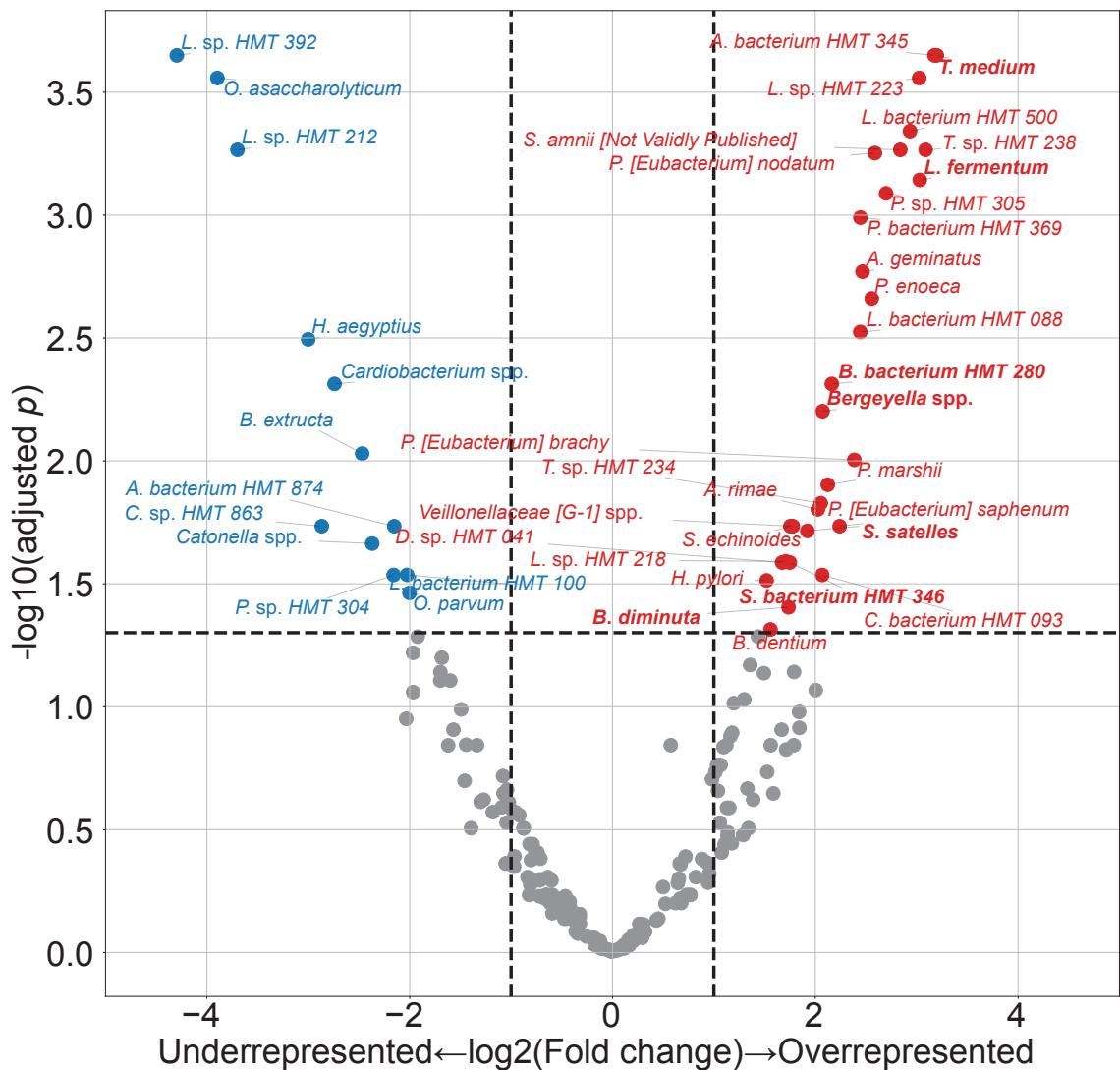


Figure 5: **PROM-related DAT**.

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

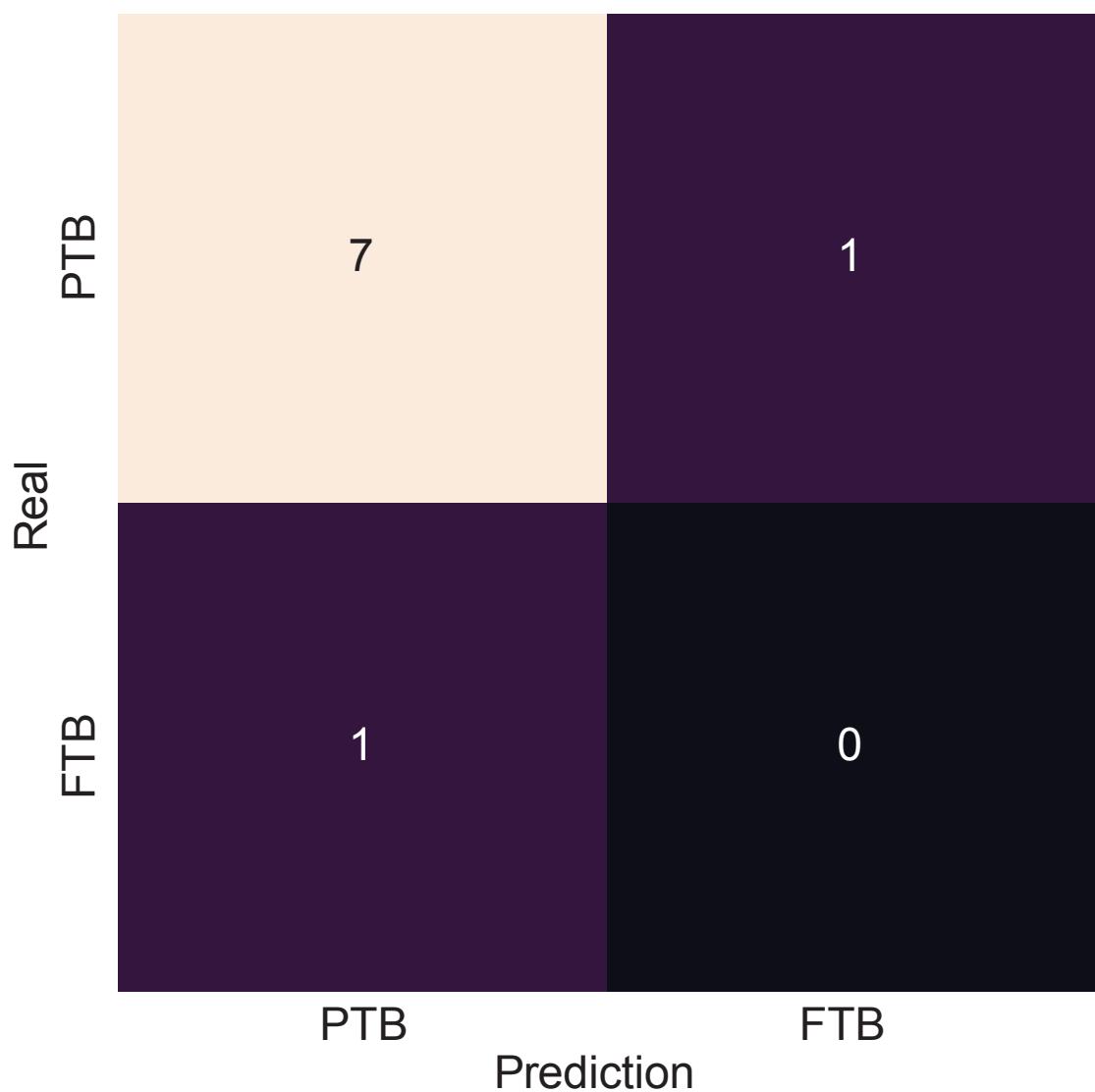


Figure 6: Validation of random forest-based PTB prediction model.

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

407 **2.4 Discussion**

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

415 One of the most common salivary taxa that has been examined is *Fusobacterium nucleatum* (Han,
416 2015; Brennan & Garrett, 2019; Bolstad, Jensen, & Bakken, 1996), that is a Gram-negative, anaerobic, and
417 filamentous bacteria. *Fusobacterium nucleatum* can be separated from not only the salivary microbiome
418 but also the vaginal microbiome (Vander Haar, So, Gyamfi-Bannerman, & Han, 2018; Witkin, 2019). In
419 both animal and human investigation, *Fusobacterium nucleatum* infection has been linked to risk of PTB
420 (Doyle et al., 2014). According to recent researches, the placenta women who give birth prematurely may
421 include additional salivary microbiome dysbiosis, such as *Bergeyella* spp. and *Porphyromonas gingivalis*
422 (León et al., 2007; Katz, Chegini, Shiverick, & Lamont, 2009). Although *Bergeyella* spp. were one of the
423 PROM-overrepresented DAT (Figure 5), it was excluded in the final 25 PTB-related DAT. Furthermore,
424 *Porphyromonas gingivalis* and *Campylobacter gracilis* were pathogens of periodontitis in sub-gingival
425 microbiome (Yang et al., 2022). *Lactobacillus gasseri* was also one of the FTB-enriched DAT (Figure
426 1), and it is well established that early PTB risk can be reduced by *Lactobacillus gasseri* in the vaginal
427 microbiome (Basavaprabhu, Sonu, & Prabha, 2020; Payne et al., 2021).

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

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

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

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

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

459 **3 Random forest prediction model for periodontitis statuses based on the**
460 **salivary microbiomes**

461 **3.1 Introduction**

462 Saliva microbial dysbiosis brought on by the accumulation of plaque results in periodontitis, a chronic
463 inflammatory disease of the tissue that surrounds the tooth (Kinane, Stathopoulou, & Papapanou, 2017).
464 Loss of periodontal attachment is a consequence of periodontitis, which may lead to irreversible bone loss
465 and, eventually, permanent tooth loss if left untreated. A new classification criterion of periodontal diseases
466 was created in 2018, about 20 years after the 1999 statements of the previous one (Papapanou et al.,
467 2018). Even with this evolution, radiographic and clinical markers of periodontitis progression remain the
468 primary methods for diagnosing periodontitis (Papapanou et al., 2018). Such tools, nevertheless, frequently
469 demonstrate the prior damage from periodontitis rather than its present condition. Certain individuals have
470 a higher risk of periodontitis, a higher chance of developing severe generalized periodontitis, and a worse
471 response to common salivary bacteria control techniques utilized to prevent and treat periodontitis. As a
472 result, the 2017 framework for diagnosing periodontitis additionally allows for the potential development
473 of biomarkers to enhance diagnosis and treatment of periodontitis (Tonetti, Greenwell, & Kornman, 2018).
474 Instead of only depending on the progression of periodontitis, a new etiological indication based on the
475 current state must be introduced in order to enable appropriate intervention through early detection of
476 periodontitis. Thus, the current clinical diagnostic techniques that rely on periodontal probing can be
477 uncomfortable for patients with periodontitis (Canakci & Canakci, 2007).

478 Due to the development of salivaomics, in this manner, the examination of saliva has emerged as
479 a significant alternative to the conventional ways of identifying periodontitis (Altingöz et al., 2021;
480 Melguizo-Rodríguez, Costela-Ruiz, Manzano-Moreno, Ruiz, & Illescas-Montes, 2020). Given that saliva
481 sampling is non-invasive, painless, and accessible to non-specialists, it may be a valuable instrument for
482 diagnosing periodontitis (Zhang et al., 2016). Furthermore, much research has suggested that periodontitis
483 could be a trigger in the development and exacerbation of metabolic syndrome (Morita et al., 2010; Nesbitt
484 et al., 2010). Consequently, alteration in these levels of salivary microbiome markers may serve as high
485 effective diagnostic, prognostic, and therapeutic indicators for periodontitis and other systemic diseases
486 (Miller, Ding, Dawson III, & Ebersole, 2021; Čižmárová et al., 2022). The pathogenesis of periodontitis
487 typically comprises qualitative as well as quantitative alterations in the salivary microbial community,
488 despite that it is a complex disease impacted by a number of contributing factors including age, smoking
489 status, stress, and nourishment (Abusleme, Hoare, Hong, & Diaz, 2021; Lafaurie et al., 2022). Depending
490 on the severity of periodontitis, the salivary microbial community's diversity and characteristics vary
491 (Abusleme et al., 2021), indicating that a new etiological diagnostic standards might be microbial
492 community profiling based on clinical diagnostic criteria. As a consequence, salivary microbiome
493 compositions have been characterized in numerous research in connection with periodontitis. High-
494 throughput sequencing, including 16S rRNA gene sequencing, has recently used in multiple studies to
495 identify variations in the bacterial composition of sub-gingival plaque collections from periodontal healthy

496 individuals and patients with periodontitis (Altabtbaei et al., 2021; Iniesta et al., 2023; Nemoto et al., 2021).
497 This realization has rendered clear that alterations in the salivary microbial community—especially, shifts to
498 dysbiosis—are significant contributors to the pathogenesis and development of periodontitis (Lamont, Koo,
499 & Hajishengallis, 2018). Yet most of these research either focused only on the microbiome alterations in
500 sub-gingival plaque collection, comprised a limited number of periodontitis study participants, or did not
501 account for the impact of multiple severities of periodontitis.

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

510 Recently, we employed multiplex quantitative-PCR and machine learning-based classification model
511 to predict the severity of periodontitis based on the amount of nine pathogens of periodontitis from
512 saliva collections (E.-H. Kim et al., 2020). On the other hand, the fact that we focused merely at nine
513 pathogens for periodontitis and neglected the variety bacterial species associated to the various severities
514 of periodontitis constrained the breadth of our investigation. By developing a machine learning model
515 that could classify multiple severities of periodontitis based on the salivary microbiome composition,
516 this study aims to fill these knowledge gaps and produce more accurate and therapeutically useful
517 guidance to evaluate progression of periodontitis. Hence, in order to examine the salivary microbiome
518 composition of both healthy controls and patients with periodontitis in multiple stages, we applied
519 16S rRNA gene sequencing. Furthermore, employing the 2018 classification criteria, we sought to find
520 biomarkers (species) for the precise prediction of periodontitis severities (Papapanou et al., 2018; Chapple
521 et al., 2018).

522 **3.2 Materials and methods**

523 **3.2.1 Study participants enrollment**

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

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

541 **3.2.2 Periodontal clinical parameter diagnosis**

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

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

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

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

559 0. Normal gingiva: without inflammation nor discoloration.

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

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

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

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

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

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

570 • Healthy:

571 1. Bleeding sites < 10%

572 2. Probing depth: \leq 3 mm

573 • Stage I:

574 1. No tooth loss because of periodontitis.

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

576 3. Radiographic bone loss: < 15%

577 • Stage II:

578 1. No tooth loss because of periodontitis.

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

580 3. Radiographic bone loss: 15-33%

581 • Stage III:

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

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

584 3. Radiographic bone loss: > 33%

585 **3.2.3 Saliva sampling and DNA extraction procedure**

586 All study participants received instructions to avoid eating, drinking, brushing, and using mouthwash for
587 at least an hour prior to the saliva sample collection process. These collections were conducted between
588 09:00 and 11:00. Mouth rinse was collected by rinsing the mouth for 30 seconds with 12 mL of a solution
589 (E-zen Gargle, JN Pharm, Korea). All saliva samples were tagged with anonymous ID and stored at -4 °C.

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

- 594 • Forward: 5' -TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNNGCWGCAG-3'
595 • Reverse: 5' -GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'

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

- 598 1. Heat activation for 30 seconds at 95 °C.
599 2. 25 cycles for 30 seconds at 95 °C.
600 3. 30 seconds at 55 °C.
601 4. 30 seconds at 72 °C.

602 NexteraXT Indexed Primer was applied to amplification 10 µL of the purified initial PCR products for
603 the final library creation. The second PCR used the same conditions as the first PCR conditions but with
604 10 cycles. 16S rRNA gene sequencing was performed via 2×300 bp paired-end sequencing at Macrogen
605 Inc. (Macrogen, Seoul, Korea) using Illumina MiSeq platform (Illumina, San Diego, CA, USA).

606 **3.2.4 Bioinformatics analysis**

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

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

618 • Shannon (Weaver, 1963)
619 • Simpson (Simpson, 1949)
620 Aitchison index for a beta-diversity index was calculated using QIIME2 (version 2020.8) (Aitchison,
621 Barceló-Vidal, Martín-Fernández, & Pawlowsky-Glahn, 2000; Bolyen et al., 2019). We employed the
622 t-SNE algorithm to illustrate multi-dimensional data from the beta-diversity index computation (Van der
623 Maaten & Hinton, 2008). The beta-diversity index was compared using the PERMANOVA test (Anderson,
624 2014; Kelly et al., 2015) and MWU test.

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

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

637 As previously stated (E.-H. Kim et al., 2020), we used stratified k -fold cross-validation ($k = 10$)
638 by severity of periodontitis to achieve consistent and trustworthy classification results (Wong & Yeh,
639 2019). Additionally, we utilized various features with confusion matrices and their derivations to evaluate
640 the classification outcomes in order to identify which features optimize classification evaluations and
641 decrease sequencing efforts. Using the DAT discovered by ANCOM, we iteratively removed the least
642 significant taxa from the input features (taxa) of the random forest (Breiman, 2001) and gradient boosting
643 (Friedman, 2002) classification models using the backward elimination method. Random forest classifier
644 builds multiple decision trees independently using bootstrapped samples and aggregates their predictions,
645 enhancing stability and reducing overfitting problems. In contrast, Gradient boosting constructs trees
646 sequentially, where each new tree improves the errors of the previous ones using gradient descent, leading
647 to higher classification evaluations.

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

652 3.2.5 Data and code availability

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

655 image that employed throughout this study is available in the DockerHub: <https://hub.docker.com/>
656 repository/docker/fumire/periodontitis_16s. Every code used in this study can be found on
657 GitHub: https://github.com/CompbioLabUnist/Periodontitis_16S.

658 **3.3 Results**

659 **3.3.1 Summary of clinical information and sequencing data**

660 Among clinical information of the study participants, clinical attachment level, probing depth, plaque
661 index, and gingival index, were significantly increased with periodontitis severity (Kruskal-Wallis test
662 $p < 0.001$), while sex were observed no significant difference (Table 2). Notably, clinical attachment level
663 and probing depth have significant differences among the periodontitis severities (MWU test $p < 0.01$;
664 Figure 15). Additionally, 71461.00 ± 11792.30 and 45909.78 ± 11404.65 reads per sample were obtained
665 before and after filtering low-quality reads and trimming extra-long tails, respectively (Figure 16). In 250
666 study subjects, we have found a total of 425 bacterial taxa (Figure 13).

667 **3.3.2 Diversity indices reveal differences among the periodontitis severities**

668 Rarefaction curves showed that the sequencing depth was sufficient (Figure 12). Alpha-diversity in-
669 dices indicated significant differences between the healthy and the periodontitis stages (MWU test
670 $p < 0.01$; Figure 7a-e); however, there were no significant differences between the periodontitis stages.
671 This emphasizes how essential it is to classify the salivary microbiome compositions and distinguish
672 between the stages of periodontitis using machine learning approaches.

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

678 **3.3.3 DAT among multiple periodontitis severities and their correlation**

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

- 682 • Group 1
 - 683 1. *Treponema* spp.
 - 684 2. *Prevotella* sp. HMT 304
 - 685 3. *Prevotella* sp. HMT 526
 - 686 4. *Peptostreptococcaceae [XI][G-5]* saphenum
 - 687 5. *Treponema* sp. HMT 260
 - 688 6. *Mycoplasma faecium*
 - 689 7. *Peptostreptococcaceae [XI][G-9]* brachy
 - 690 8. *Lachnospiraceae [G-8]* bacterium HMT 500
 - 691 9. *Peptostreptococcaceae [XI][G-6]* nodatum
 - 692 10. *Fretibacterium* spp.

- 693 • Group 2
- 694 1. *Porphyromonas gingivalis*
- 695 2. *Campylobacter showae*
- 696 3. *Filifactor alocis*
- 697 4. *Treponema putidum*
- 698 5. *Tannerella forsythia*
- 699 6. *Prevotella intermedia*
- 700 7. *Porphyromonas* sp. HMT 285

- 701 • Group 3
- 702 1. *Actinomyces* spp.
- 703 2. *Corynebacterium durum*
- 704 3. *Actinomyces graevenitzii*

705 Ten DAT that were significant enriched in stage II and stage III, but deficient in healthy formed Group
706 1 (Figure 8). Furthermore, in comparison to the healthy, the seven DAT of Group 2 were significantly
707 enriched in each of the stages of periodontitis. On the other hand, three DAT in Group 3 were deficient in
708 stage II and stage III, but significantly enriched in healthy. The relative proportions of the DAT further
709 supported these findings (Figure 8b), suggesting that the DAT is primarily linked to periodontitis rather
710 than other salivary bacteria.

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

713 3.3.4 Classification of periodontitis severities by random forest models

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

718 Based on the proportion of DAT, random forest classifier were trained to classify the periodontitis
719 severities (Table 6). We conducted multi-label classification for the multiple periodontitis severities,
720 namely healthy, stage I, stage II, and stage III. In this setting, we classified multiple periodontitis
721 severities with the highest BA of 0.779 ± 0.029 (Table 4). AUC ranged between 0.81 and 0.94 (Figure
722 10b).

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

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

729 perform a multi-label classification.

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

736 Finally, to confirm the reliability and consistency of our random forest classifier, we validated our
737 classification model using openly accessible 16S rRNA gene sequencing from Spanish participants
738 (Iniesta et al., 2023) and Portuguese participants (Relvas et al., 2021) (Figure 11). Although some
739 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.

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

Table 4: Feature combinations and their evaluations

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

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

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

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

Table 6: Feature the importance of taxa in the classification of different periodontal statuses
 Taxa are ranked in descending order of importance; from most important to least important.

Condition	Healthy vs. Stage I vs. Stage II vs. Stage III			Healthy vs. Stage I			Healthy vs. Stage I vs. Stage II/III			Healthy vs. Stage 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>	0.050	<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>	0.032	<i>Filifactor alocis</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>	0.037	<i>Treponema spp.</i>	0.025	<i>Treponema spp.</i>	0.022		
8	<i>Corynebacterium durum</i>	0.032	<i>Tannerella forsythia</i>	0.026	<i>Tannerella forsythia</i>	0.029	<i>Treponema spp.</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>	0.026	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.018	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.015		
10	<i>Tannerella forsythia</i>	0.026	<i>Prevotella intermedia</i>	0.023	<i>Freibacterium spp.</i>	0.018	<i>Peptostreptococcaceae (G-8) bacterium HMT 500</i>	0.014	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.010		
11	<i>Treponema putidum</i>	0.025	<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		
12	<i>Freibacterium spp.</i>	0.023	<i>Prevotella sp. HMT 260</i>	0.019	<i>Treponema putidum</i>	0.014	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.010	<i>Freibacterium spp.</i>	0.009		
13	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.021	<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>Treponema putidum</i>	0.006		
14	<i>Treponema sp. HMT 260</i>	0.019	<i>Prevotella sp. HMT 526</i>	0.018	<i>Treponema sp. HMT 260</i>	0.008	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.008	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.004		
15	<i>Prevotella sp. HMT 526</i>	0.018	<i>Prevotella sp. HMT 304</i>	0.017	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.008	<i>Freibacterium spp.</i>	0.008	<i>Treponema sp. HMT 260</i>	0.004		
16	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.018	<i>Mycoplasma faecium</i>	0.014	<i>Mycoplasma faecium</i>	0.004	<i>Prevotella sp. HMT 304</i>	0.005	<i>Mycoplasma faecium</i>	0.004		
17	<i>Prevotella sp. HMT 304</i>	0.017	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.014	<i>Prevotella sp. HMT 304</i>	0.003	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.003	<i>Prevotella sp. HMT 304</i>	0.002		
18	<i>Mycoplasma faecium</i>	0.014	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.013	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.003	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.004	<i>Prevotella sp. HMT 304</i>	0.001		

Table 7: Beta-diversity pairwise comparisons on the periodontitis statuses

Statistically significant (p-value) was determined by the PERMANOVA test.

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

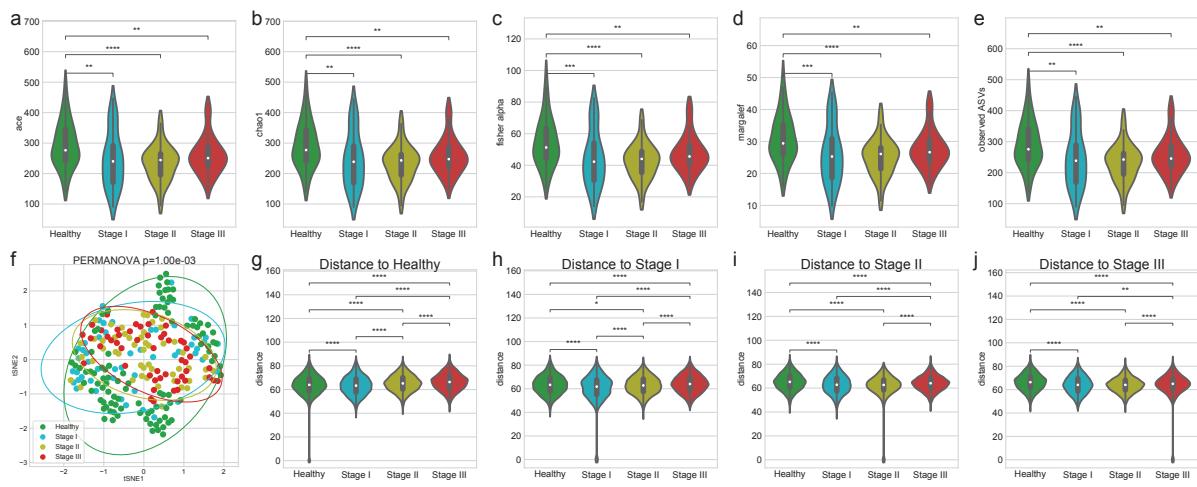


Figure 7: Diversity indices.

Alpha-diversity indices (a-e) indicate that healthy controls have increased heterogeneity than periodontitis stages as measured by: (a) ACE (b) Chao1 (c) Fisher alpha (d) Margalef, and (e) observed ASVs. (f) The beta-diversity index (weighted UniFrac) was visualized using a tSNE-transformed plot. The confidence ellipses are shown to display the distribution of each periodontitis stage. The distance to each stage demonstrated that each periodontitis stage was distinguished from the other periodontitis stages: (g) distance to Healthy (h) distance to Stage I (i) distance to Stage II, and (j) distance to Stage III. Statistical significance determined by the MWU test and the PERMANOVA test: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***) , and $p \leq 0.0001$ (****).

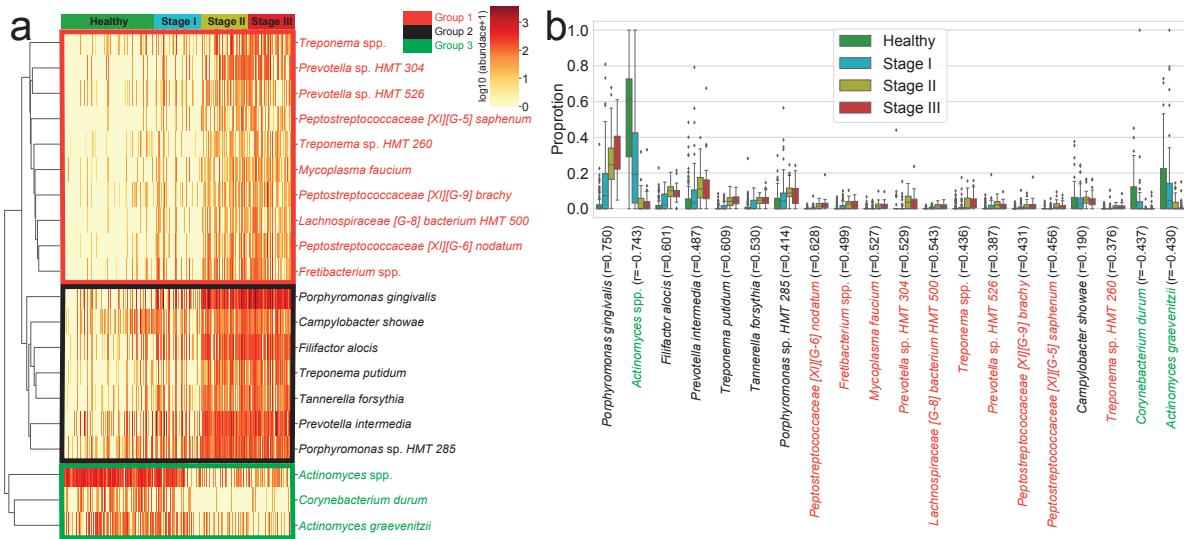


Figure 8: **Differentially abundant taxa (DAT).**

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.

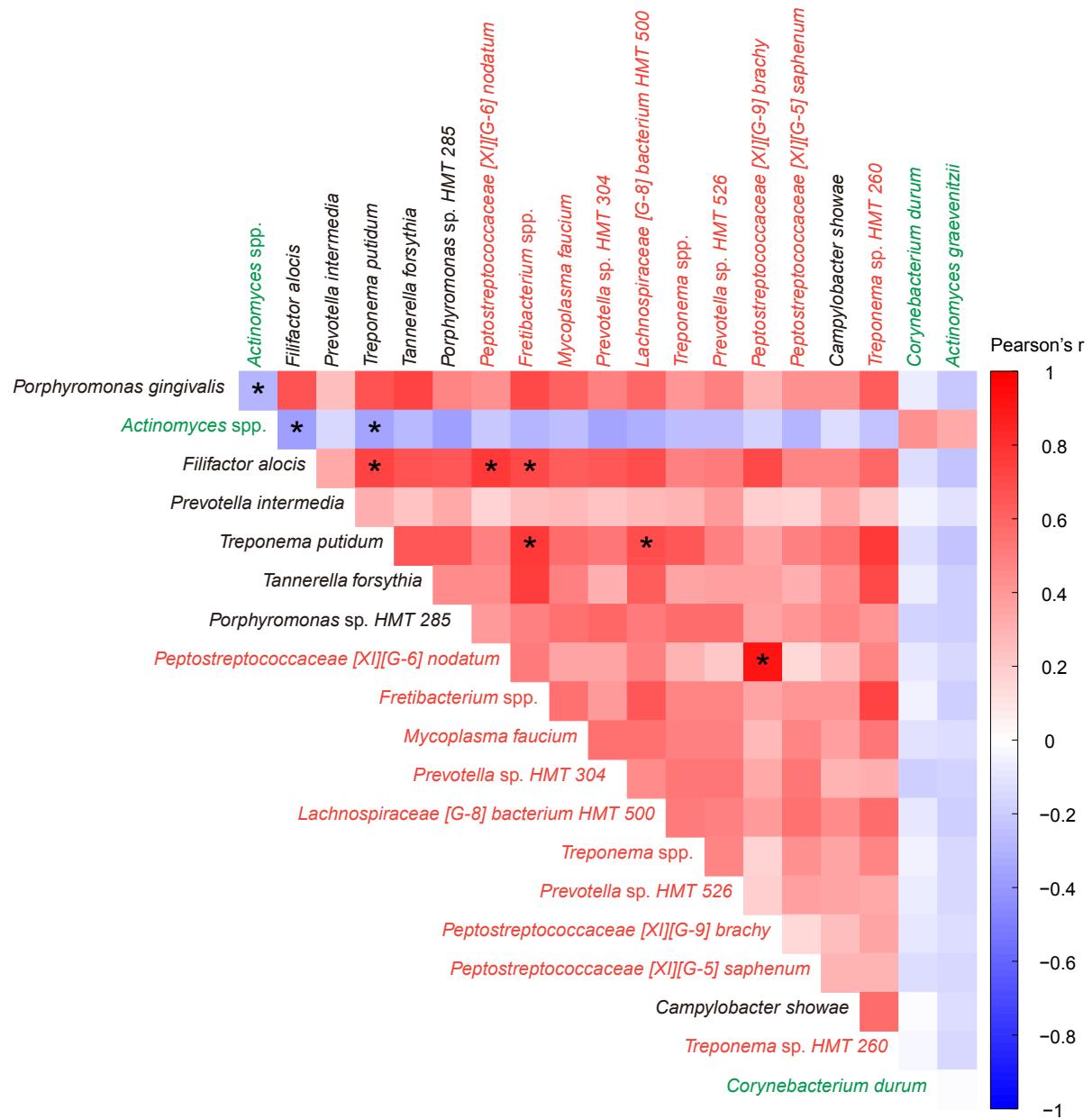


Figure 9: Correlation heatmap.

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

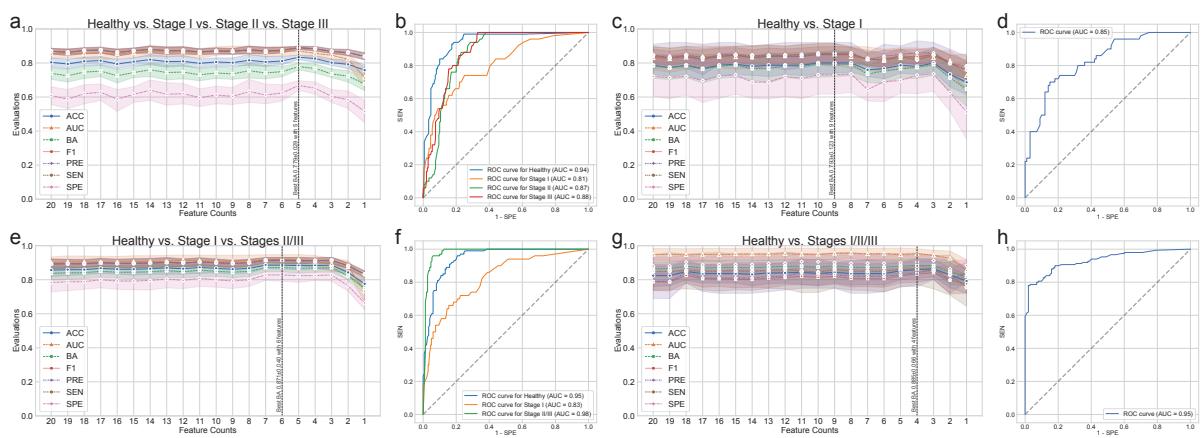


Figure 10: Random forest classification metrics.

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).

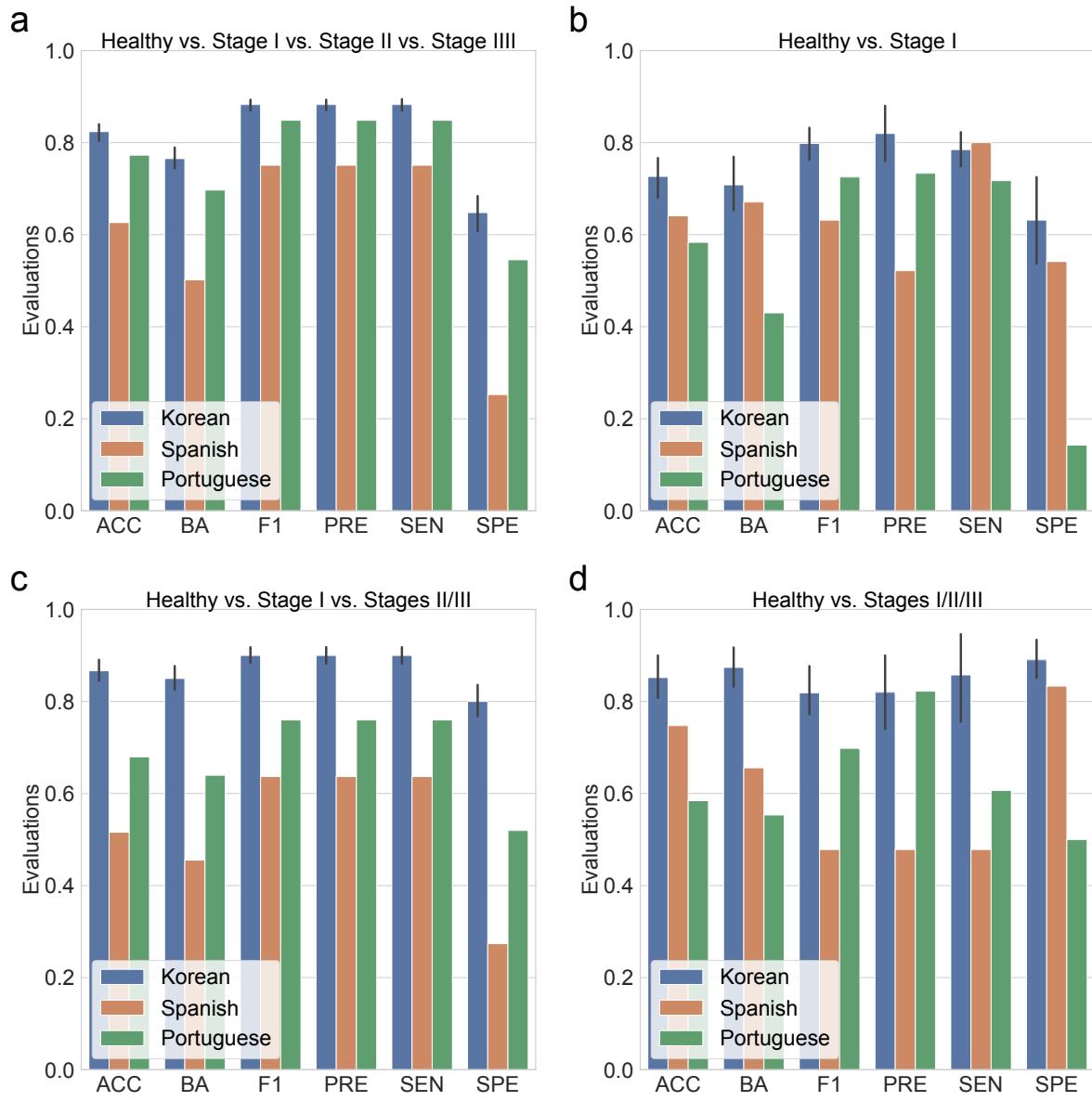


Figure 11: **Random forest classification metrics from external datasets.**

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.

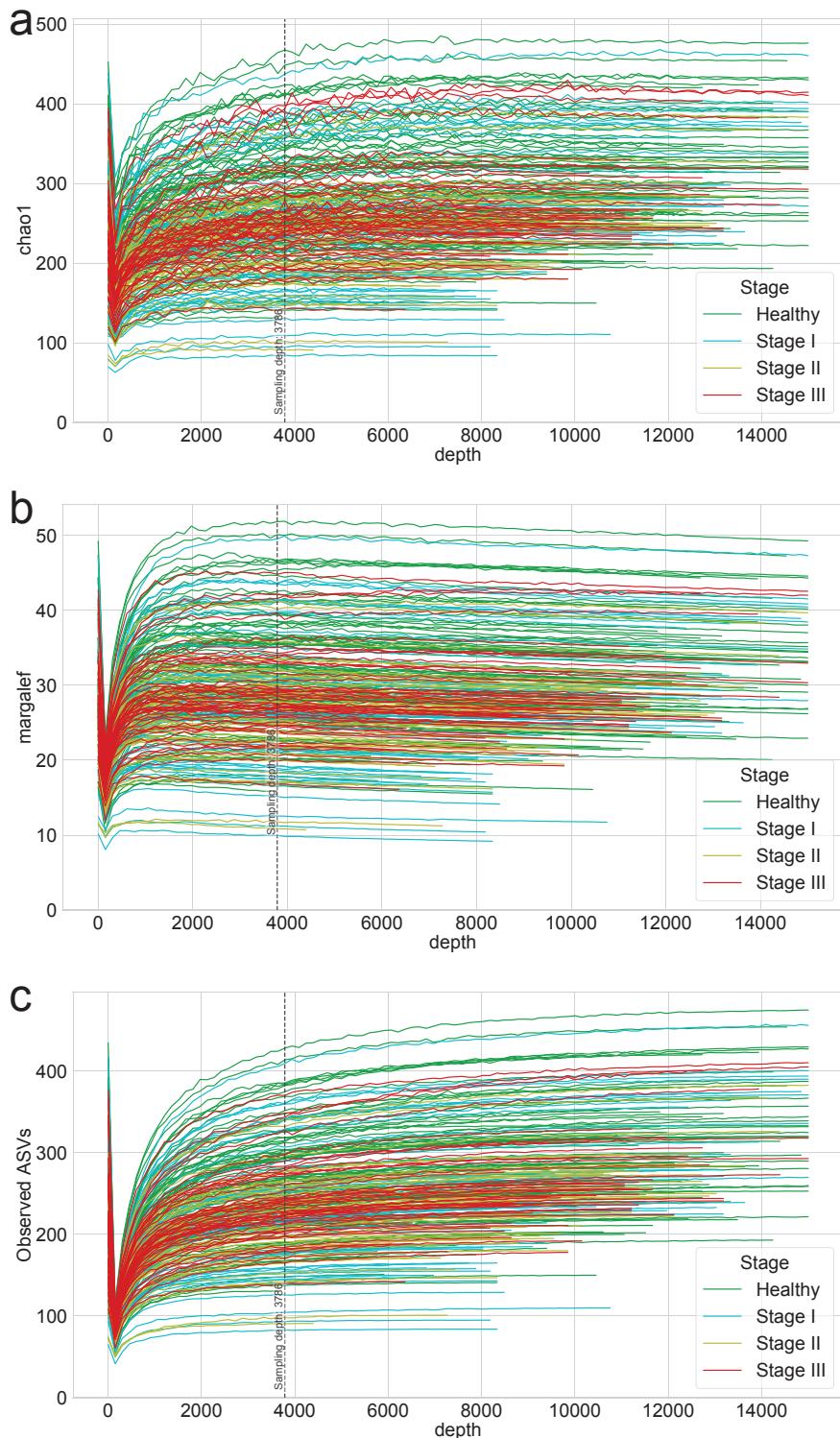


Figure 12: Rarefaction curves for alpha-diversity indices.

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

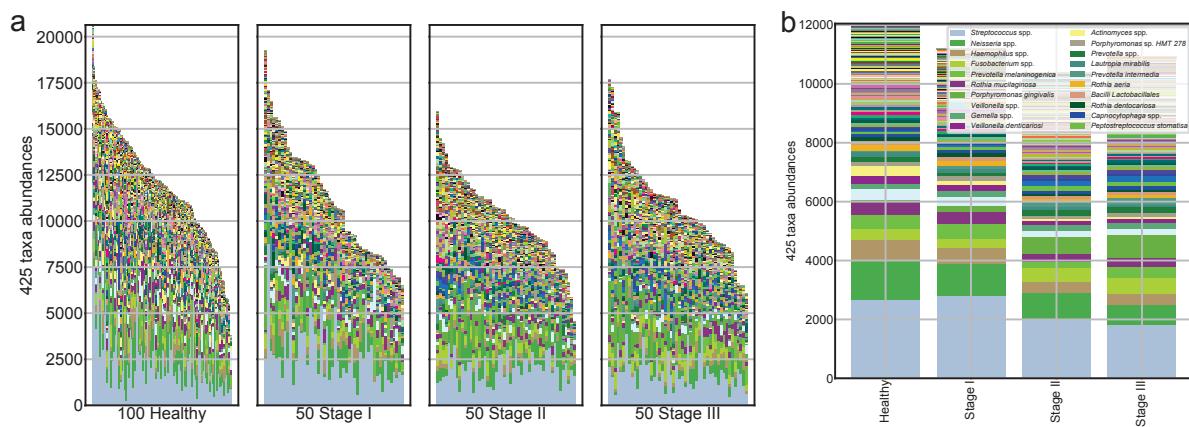


Figure 13: Salivary microbiome compositions in the different periodontal statuses.

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**).

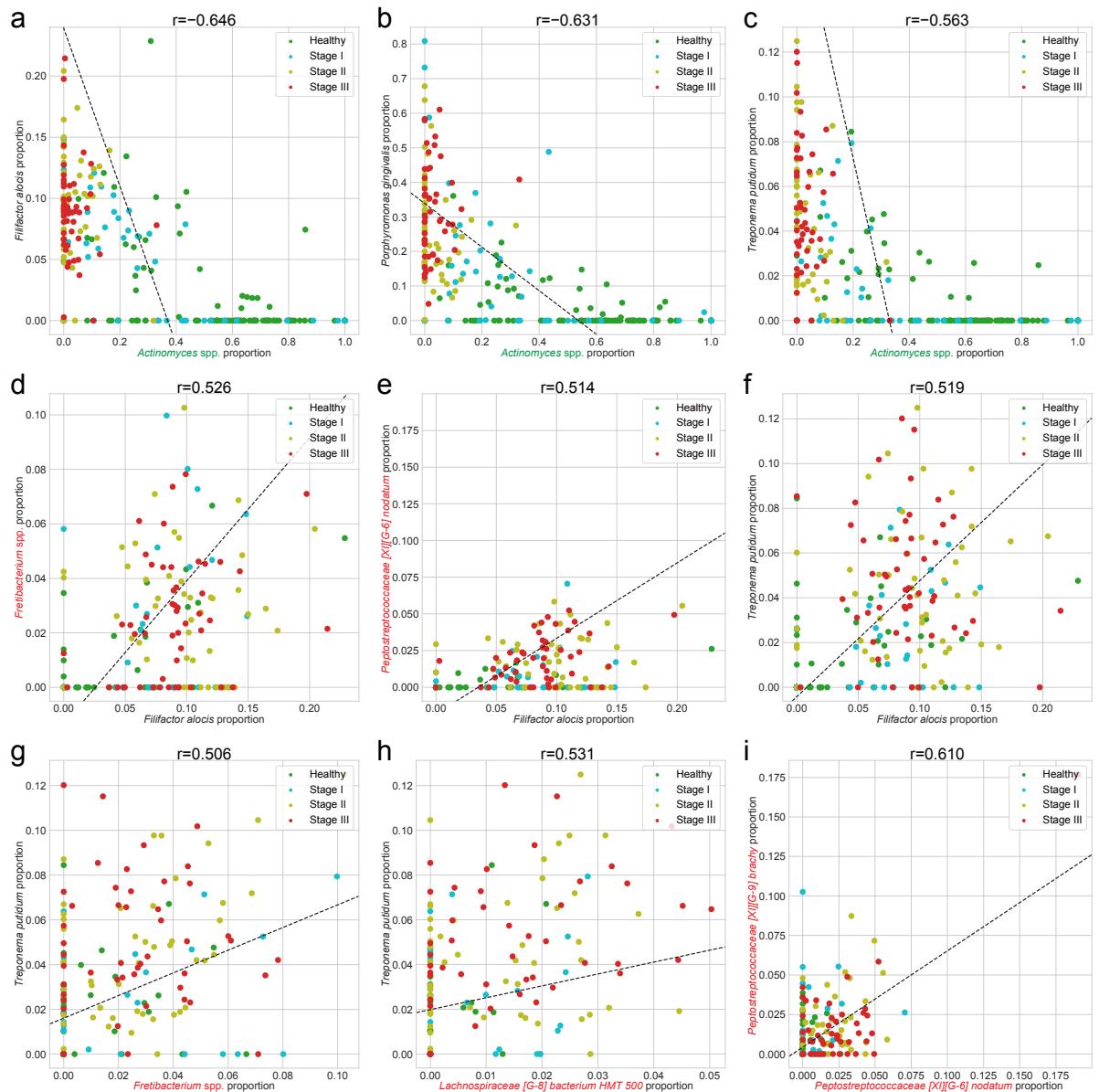


Figure 14: Correlation plots for differentially abundant taxa.

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).

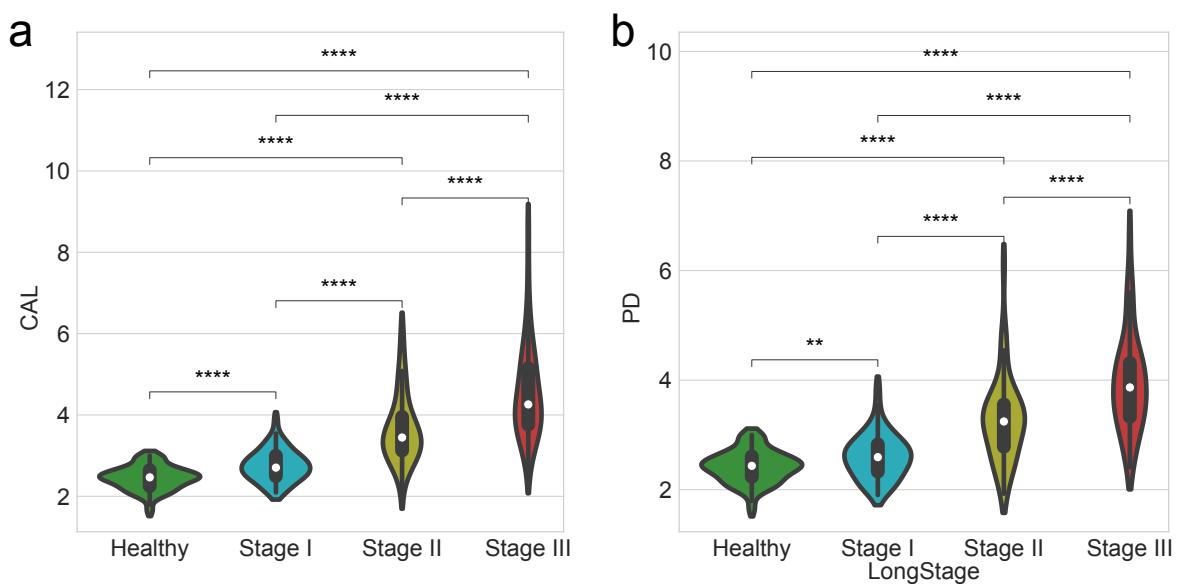


Figure 15: **Clinical measurements by the periodontitis statuses.**

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: $p < 0.01$ (**) and $p < 0.0001$ (****).

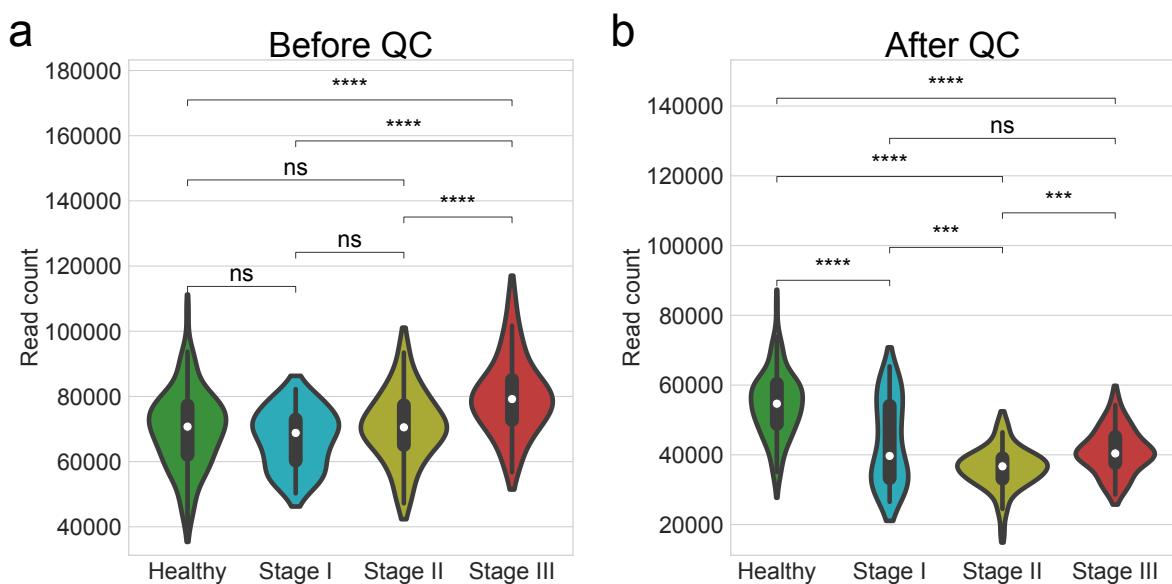


Figure 16: **Number of read counts by the periodontitis statuses.**

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: $p \geq 0.05$ (ns), $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***) $,$ and $p < 0.0001$ (****).

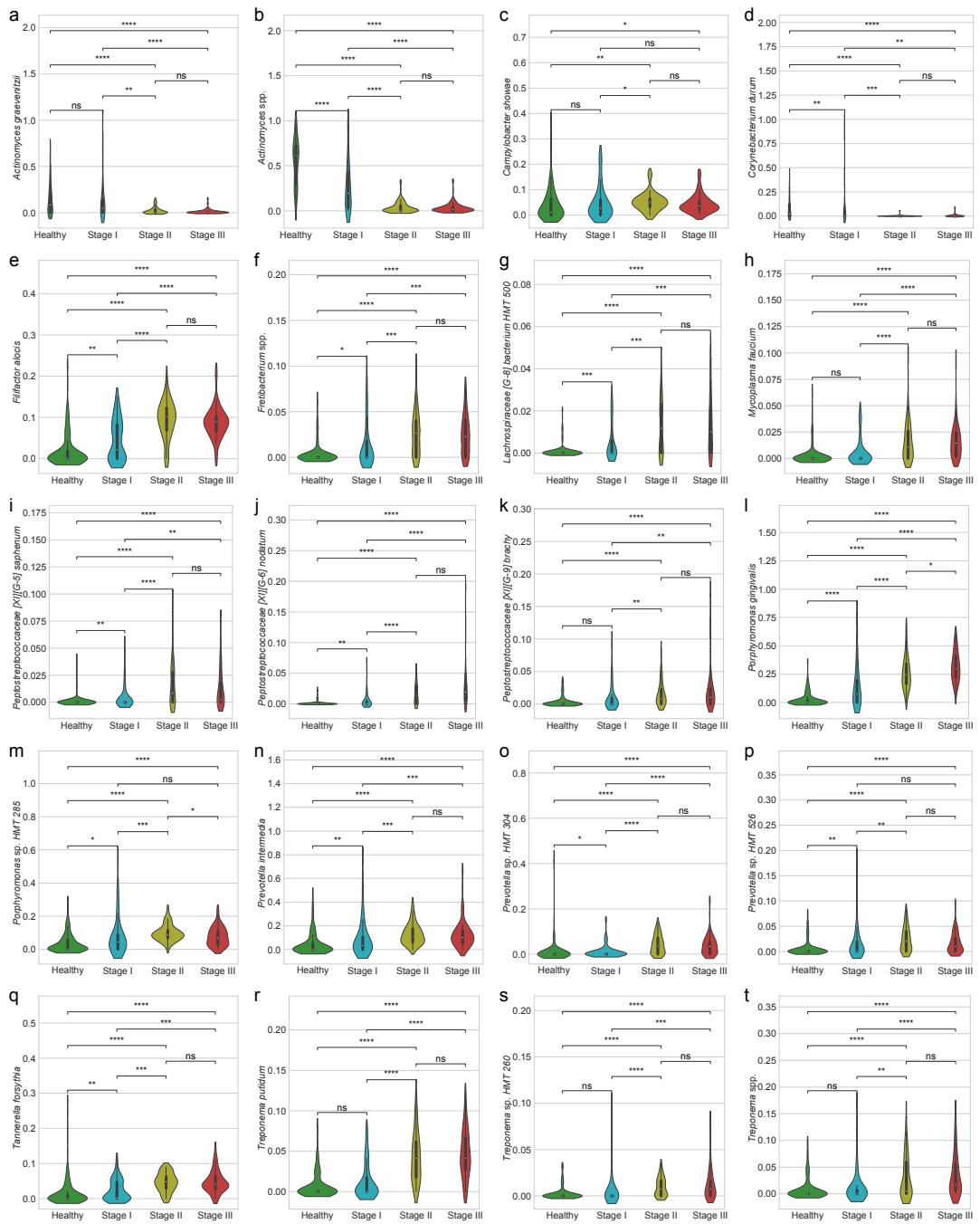


Figure 17: Proportion of DAT.

(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: $p \geq 0.05$ (ns), $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), and $p < 0.0001$ (****).

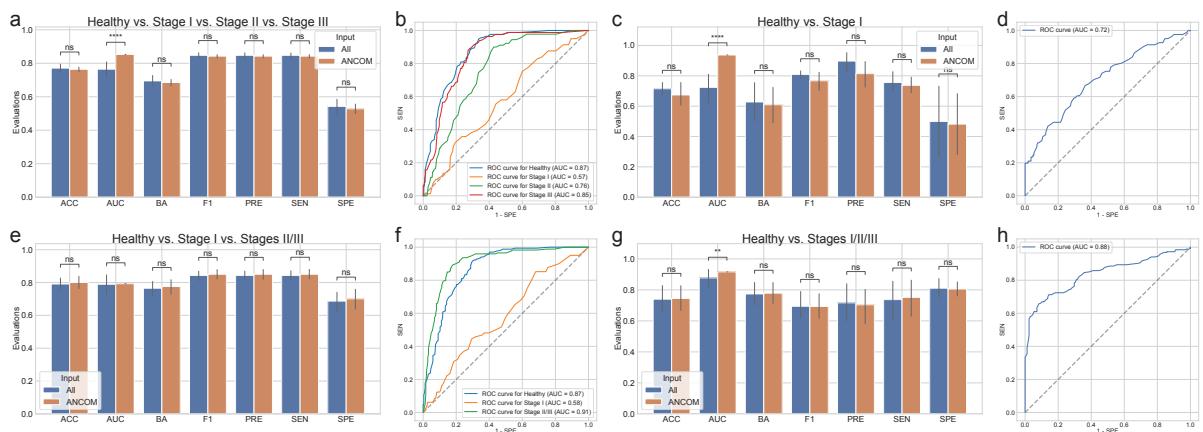


Figure 18: Random forest classification metrics with the full microbiome compositions and ANCOM-selected DAT compositions.

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: $p \geq 0.05$ (ns), $p < 0.01$ (**), and $p < 0.0001$ (****).

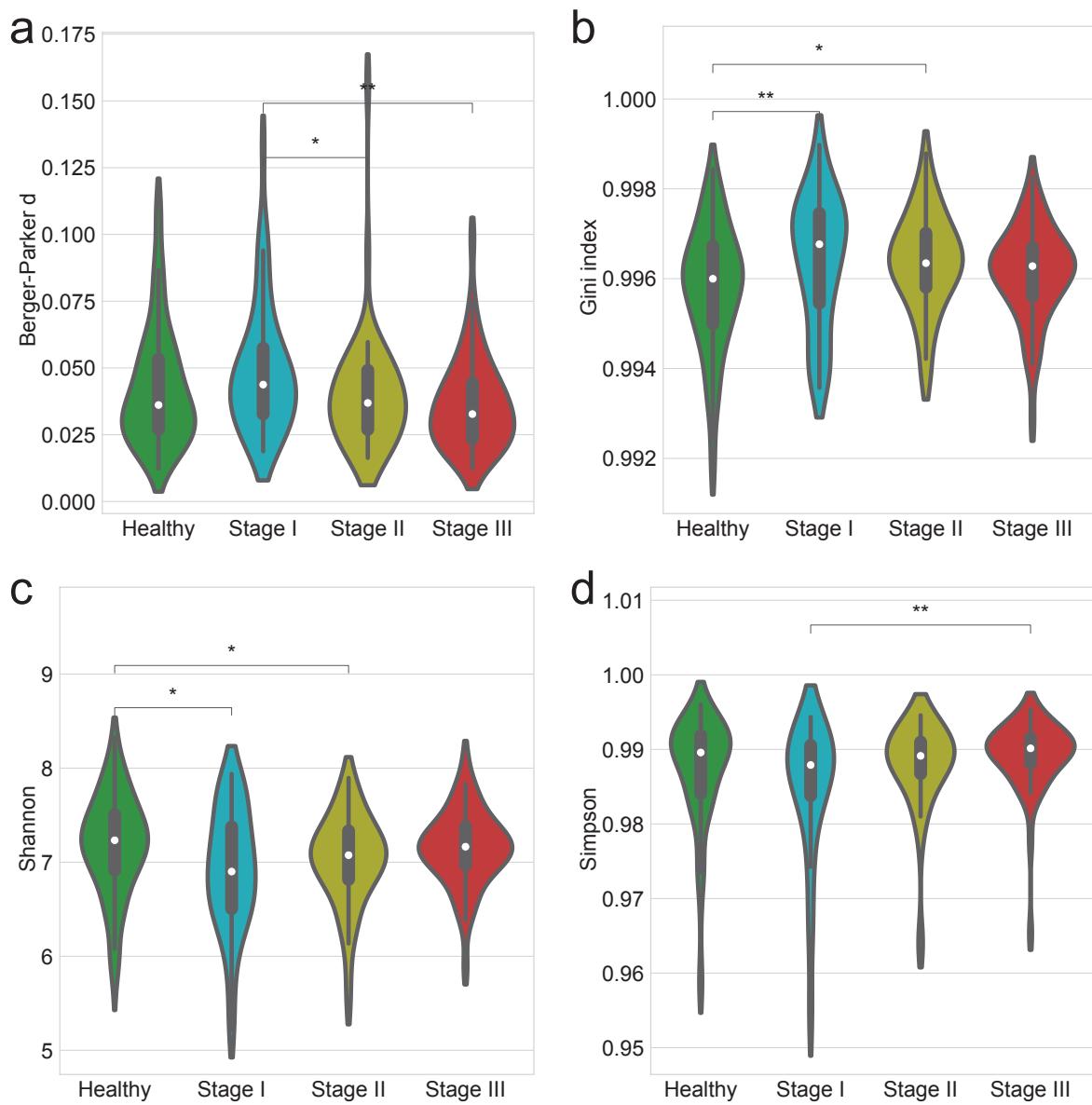


Figure 19: **Alpha-diversity indices account for evenness.**

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: $p < 0.05$ (*) and $p < 0.01$ (**)

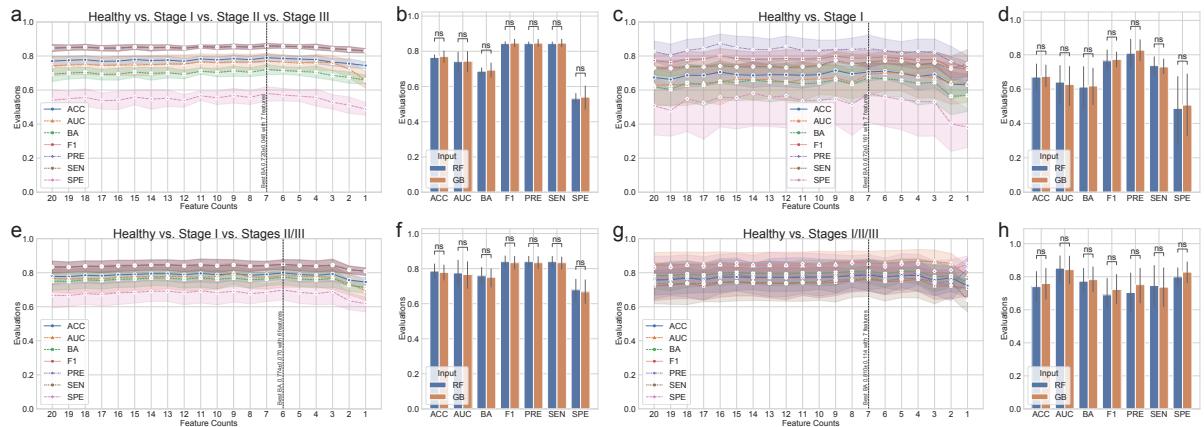


Figure 20: Gradient Boosting classification metrics.

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. MWU test: $p \geq 0.05$ (ns)

740 **3.4 Discussion**

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

749 Previous research identified the red complex as the primary pathogens of periodontitis (Listgarten,
750 1986): *Porphyromonas gingivalis*, *Tannerella forsythia*, and *Treponema denticola*. Other studies, however,
751 have shown that periodontal pathogens communicate with other bacteria in the salivary microbiome
752 networks to generate dental plaque prior to the pathogenesis and development of periodontitis (Lamont &
753 Jenkinson, 2000; Rosan & Lamont, 2000; Yoshimura, Murakami, Nishikawa, Hasegawa, & Kawaminami,
754 2009).

755 Using subgingival plaque collections, recent researches have suggested a connection between the
756 periodontitis severity and the salivary microbiome compositions (Altabtbaei et al., 2021; Iniesta et al.,
757 2023; Nemoto et al., 2021). Therefore, we have examined the salivary microbiome compositions of
758 patients with multiple severities of periodontitis and periodontally healthy controls, extending on earlier
759 studies.

760 According to our findings, the salivary microbiome compositions have 425 taxa (Figure 13). We
761 computed the alpha-diversity indices to determine the variability within each salivary microbiome
762 composition, including ace (Chao & Lee, 1992), chao1 (Chao, 1984), fisher alpha (Fisher et al., 1943),
763 margalef (Magurran, 2021), observed ASVs (DeSantis et al., 2006), Berger-Parker *d* (Berger & Parker,
764 1970), Gini index (Gini, 1912), Shannon (Weaver, 1963), and Simpson (Simpson, 1949) (Figure 7 and
765 Figure 19). Alpha-diversity indices suggested that the microbial richness of periodontally healthy controls
766 was higher than that of patients with periodontitis (Figure 7a-e and Figure 19). These results are in line with
767 findings with that patients with advanced periodontitis, namely stage II and stage III, have less diversified
768 communities than periodontally healthy controls (Jorth et al., 2014). Recognizing that the periodontitis
769 severity increases the amount of *Porphyromonas gingivalis*, the salivary microbiome compositions from
770 periodontally healthy controls conserved microbial networks dominated by *Streptococcus* spp. (Figure
771 13). *Porphyromonas gingivalis* is one of the known periodontal pathogen that could cause dysbiosys
772 in the salivary microbiomes, suggesting in the pathophysiology of periodontitis. Despite this finding,
773 earlier research found that subgingival microbiome of patients with periodontitis had a greater alpha-
774 diversity index (observed ASVs) than that of healthy controls (Iniesta et al., 2023), might due to the
775 different sampling sites between saliva and subgingival plaque. On the other hand, another research
776 has addressed significant discrepancies in alpha-diversity indices from subgingival plaque, saliva, and
777 tongue biofilms from healthy controls and periodontitis patients, resulting the highest alpha-diversity

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

787 Therefore, we calculated beta-diversity indices to analyze the differences between the study partici-
788 pants. The distances for the multiple stages of periodontitis, including stage I, stage II, and stage III, as
789 well as healthy controls (Figure 4g-j and Table 7), suggesting notable differences among the multiple
790 periodontitis severities. In other words, the composition of the salivary microbiome compositions varies
791 depending on the periodontitis stages, so that supporting the findings from a previous study (Iniesta et al.,
792 2023). Taken together that it is nearly impossible to fully restore the attachment level after it has been lost
793 due to the progression and development of periodontitis, the ability to rapidly screen for periodontitis in
794 its early phases using saliva collections would be highly beneficial for effective disease management and
795 treatment.

796 Of the total of 425 taxa in the salivary microbiome composition that have been identified (Figure 13),
797 ANCOM was applied to select 20 taxa as the DAT that indicated notable abundance variation among
798 the periodontitis severities (Figure 8 and Table 5). Three sub-groups were formed from the DAT using
799 hierarchical clustering (Figure 8a). Surprisingly, two of the red complex pathogens (Rôças, Siqueira Jr,
800 Santos, Coelho, & de Janeiro, 2001), *Porphyromonas gingivalis* and *Tannerella forsythia*, were classified
801 in Group 2 and were more prevalent in stage II and stage II periodontitis compared to healthy controls.
802 *Campylobacter showae* was additionally placed in Group 2 of the orange complex pathogens (Gambin et
803 al., 2021). Furthermore, some of the DAT in Group 2 have reported their crucial roles in pathogenesis
804 and development of periodontitis: *Filifactor alocis* (Aruni et al., 2015), *Treponema putidum* (Wyss et
805 al., 2004), *Tannerella forsythia* (Stafford, Roy, Honma, & Sharma, 2012; W. Zhu & Lee, 2016), and
806 *Prevotella intermedia* (Karched, Bhardwaj, Qudeimat, Al-Khabbaz, & Ellepolo, 2022). Taken together,
807 this indicates that DAT in Group 2 is essential to periodontitis. The portion of some Group 1 DAT,
808 including *Peptostreptococcaceae[XI][G-5] saphenum*, *Peptostreptococcaceae[XI][G-6] nodatum*, and
809 *Peptostreptococcaceae[XI][G-9] brachy*, in healthy controls and patients with periodontitis significantly
810 differed, according to earlier research (Lafaurie et al., 2022). These outcomes support our research,
811 implying that Group 1 DAT are also essential to the etiology and progression of periodontitis. However,
812 in contrast to patients with periodontitis, Group 3 DAT, namely *Corynebacterium durum* and *Actinomyces*
813 *graevenitzii*, were enriched in healthy controls, which is consistent with earlier research (Redanz et al.,
814 2021; Nibali et al., 2020).

815 In our correlation analysis (Figure 9), we have discovered strongly negative correlations (coefficient \leq
816 -0.5) between DAT of Group 3 and these of Group 1 and Group 2; we have also identified nine DAT

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

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

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

Porphyromonas gingivalis and *Actinomyces* spp. were the two most important taxa (feature) in all classification settings. 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%

- 856 • BA: 84%
857 • SEN: 94%
858 • SPE: 74%
- 859 By separating patients with periodontitis from healthy controls using only four DAT, *e.g.* *Actinomyces*
860 *graevenitzii*, *Actinomyces* spp., *Corynebacterium durum*, and *Porphyromonas gingivalis*, our machine
861 learning model performed better than previously published models (Figure 10, Table 4, and Table 6):
- 862 • AUC: $95.3\% \pm 4.9\%$
863 • BA: $88.5\% \pm 6.6\%$
864 • SEN: $86.4\% \pm 15.7\%$
865 • SPE: $90.5\% \pm 7.0\%$
- 866 This result showed that by detecting Group 3 bacteria that were substantially abundant in health
867 controls than patients with periodontitis, our study increased BA by at least 5% and SPE by at least 17%.
- 868 Furthermore, we have validated our machine-learning prediction model using openly accessible 16S
869 gene rRNA sequencing data from Portuguese (Iniesta et al., 2023) and Spanish participants (Relvas et
870 al., 2021) in order to ensure the consistency of our random forest classification model (Figure 11). Our
871 classification models employed in this study were primarily developed and assessed on Korean study par-
872 ticipants, which may limit their generalizability to other ethnic groups with different salivary microbiome
873 compositions (Premaraj et al., 2020; Renson et al., 2019). Therefore, the evaluations of this periodonti-
874 tis classification models can be affected by ethnic-specific variances and differences, highlighting the
875 necessity for additional validation and adjustment across a spectrum of ethnic backgrounds.
- 876 Regarding the clinical characteristics and potential confounders influencing the analysis of salivary
877 microbiome compositions connected with periodontitis severity, this study had a number of limitations
878 that were pointed out. We did not offer clinical information, such as the percentage of teeth, the percentage
879 of bleeding on probing, nor dental furcation involvement, even though we did gather information on
880 attachment level, probing depth, plaque index, and gingival index (Renvert & Persson, 2002); this might
881 have it challenging to present thorough and in-depth data about periodontal health. Moreover, the broad age
882 range may make it tougher to evaluate the relationship between age and periodontitis statuses, providing
883 the necessity for future studies to consider into account more comprehensive clinical characteristics
884 associated with periodontitis. Additionally, potential confounders—*e.g.* body mass index (Bombin, Yan,
885 Bombin, Mosley, & Ferguson, 2022) and e-cigarette use (Suzuki, Nakano, Yoneda, Hirofushi, & Hanioka,
886 2022)—which might have affected dental health and salivary microbiome composition were disregarding
887 consideration in addition to smoking status and systemic diseases. Thus, future research incorporating
888 these components would offer a more thorough knowledge of how lifestyle factors interact and affect the
889 salivary microbiome composition and periodontal health. Throughout, resolving these limitations will
890 advance our understanding in pathogenesis and development of periodontitis, offering significant novel
891 insights on the causal connection between systemic diseases and the salivary microbiome compositions.

892 **4 Metagenomic signature analysis of Korean colorectal cancer**

893 **4.1 Introduction**

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

907 The gut microbiome plays a fundamental role in maintaining host health by helping digestion
908 (Joscelyn & Kasper, 2014; Cerqueira, Photenhauer, Pollet, Brown, & Koropatkin, 2020), regulating
909 metabolism (Dabke, Hendrick, Devkota, et al., 2019; Utzschneider, Kratz, Damman, & Hullarg, 2016;
910 Magnúsdóttir & Thiele, 2018), adjusting immune function (Kau, Ahern, Griffin, Goodman, & Gordon,
911 2011; Shi, Li, Duan, & Niu, 2017; Broom & Kogut, 2018), and even coordinating neurological processes
912 by the brain-gut axis (Martin et al., 2018; Aziz & Thompson, 1998; R. Li et al., 2024). Comprising
913 these gut microbiota, including, archaea, bacteria, fungi, and viruses, the gut microbiome contributes
914 to the synthesis of essential vitamins, and production of fatty acids, which influence intestinal integrity
915 and immune responses. Thus, well-balanced gut microbiome composition modulates systemic immune
916 function by interacting with gut-associated lymphoid tissue, shaping immune tolerance and response
917 to infections. Hence, emerging evidence suggests that dysbiosis in the gut microbiome composition are
918 associated not only a narrow range of diseases, e.g. diarrhea and enteritis (Paganini & Zimmermann,
919 2017; Gao, Yin, Xu, Li, & Yin, 2019) but also a wide range of diseases, e.g. obesity, diabetes, and cancers
920 (Barlow et al., 2015; Hartstra et al., 2015; Helmink et al., 2019; Cullin et al., 2021).

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

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

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

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

952 **4.2 Materials and methods**

953 **4.2.1 Study participants enrollment**

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

969 **4.2.2 DNA extraction procedure**

970 Tissue samples were immediately processed under sterile conditions to prevent contamination and
971 preserved in low temperature (-80°C) storage for downstream DNA extraction and whole-genome
972 sequencing. Furthermore, produced sequencing data were provided by the "Multi-genomic analysis
973 for biomarker development in colon cancer" project (NTIS No. 1711055951) in mapped BAM format,
974 aligned to the hg38 human reference genome. The preprocessing pipeline utilized by the main project
975 included high-throughput whole-genome sequencing using standardized alignment algorithm, BWA
976 (H. Li & Durbin, 2009). In addition to the mapped human sequences, our whole-genome sequencing
977 data retained unmapped sequences, which contain potential microbial reads that were not aligned to the
978 human reference genome.

979 **4.2.3 Bioinformatics analysis**

980 To identify microbial signatures associated with CRC, we employed PathSeq (Kostic et al., 2011;
981 Walker et al., 2018), a computational pipeline designed for metagenomic analysis of high-throughput
982 sequencing data including the whole-genome sequences. After processing these sequencing data through
983 the PathSeq pipeline, a comprehensive bioinformatics analyses were conducted to characterize microbial
984 signatures associated with CRC. Prevalent taxa identification was performed by determining microbial
985 taxa present in the majority of the study participants, filtering out low-abundance and rare taxa to ensure
986 robust downstream analyses. To assess microbial community structure, diversity indices were calculated,

987 including alpha-diversity to evaluate single-sample diversity and beta-diversity to compare microbial
988 composition between the tumor tissues and their corresponding adjacent normal tissues. Differentially
989 abundant taxa (DAT) were identified using statistical method, (DESeq2 (Love et al., 2014), ANCOM
990 (Lin & Peddada, 2020)), adjusting for sequencing depth and potential confounders to highlight taxa
991 significantly associated with CRC. To explore functional implications, microbial pathway prediction was
992 performed using (PICRUSt3, HUMAnN3), linking microbial composition to metabolic and functional
993 pathways relevant to carcinogenesis and progression of CRC. This multi-layered bioinformatics approach
994 enabled a comprehensive investigation of gut microbiome alteration in CRC, facilitating the identification
995 of potential microbial biomarkers for diagnosis and prognosis of CRC.

996 **4.2.4 Data and code availability**

997 All sequences from the 211 study participants have been published to the Korea Bioinformation Center
998 (data ID KGD10008857): <https://kbds.re.kr/KGD10008857>. Docker image that employed through-
999 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>.

₁₀₀₂ **4.3 Results**

₁₀₀₃ **4.3.1 Summary of clinical characteristics**

₁₀₀₄ **4.3.2 Gut microbiome compositions**

₁₀₀₅ **4.3.3 Diversity indices**

₁₀₀₆ **4.3.4 DAT selection**

₁₀₀₇ **4.3.5 Pathway prediction**

Table 8: Clinical characteristics of CRC study participants

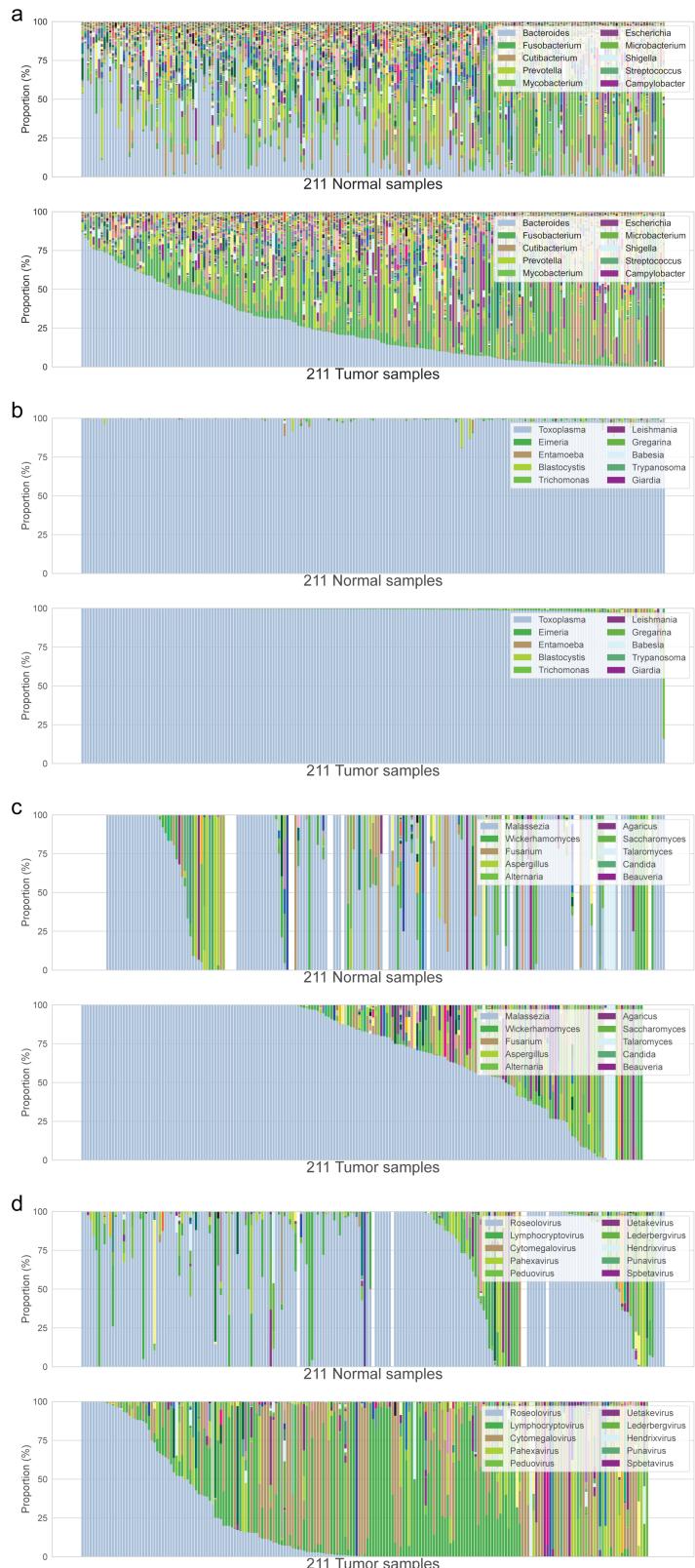


Figure 21: Gut microbiome compositions in genus level.

Taxa were sorted from the most prevalent taxon to the least prevalent taxon. CRC patients were sorted by the most prevalent taxon in descending order. **(a)** Bacteria kingdom **(b)** Eukaryota kingdom **(c)** Fungi kingdom **(d)** Viruses kingdom

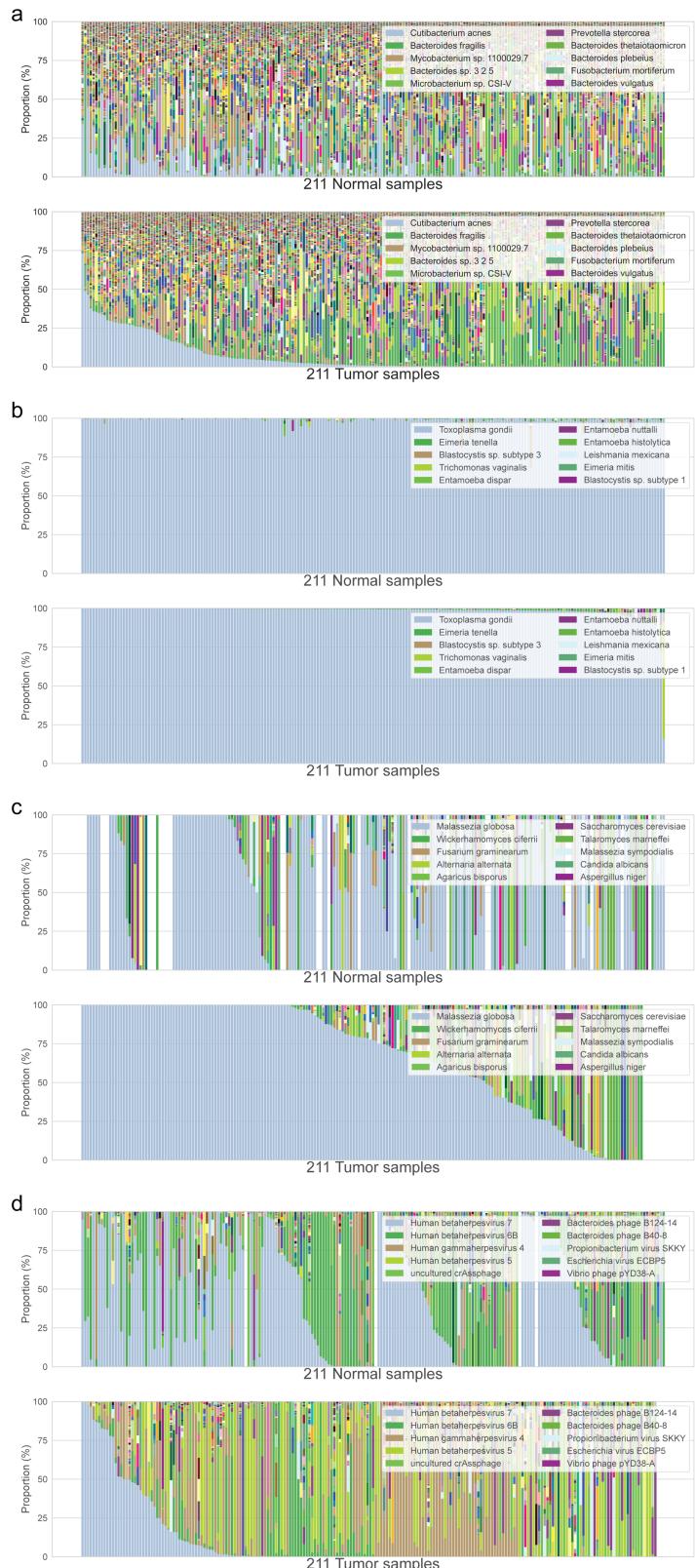


Figure 22: Gut microbiome compositions in species level.

Taxa were sorted from the most prevalent taxon to the least prevalent taxon. CRC patients were sorted by the most prevalent taxon in descending order. **(a)** Bacteria kingdom **(b)** Eukaryota kingdom **(c)** Fungi kingdom **(d)** Viruses kingdom

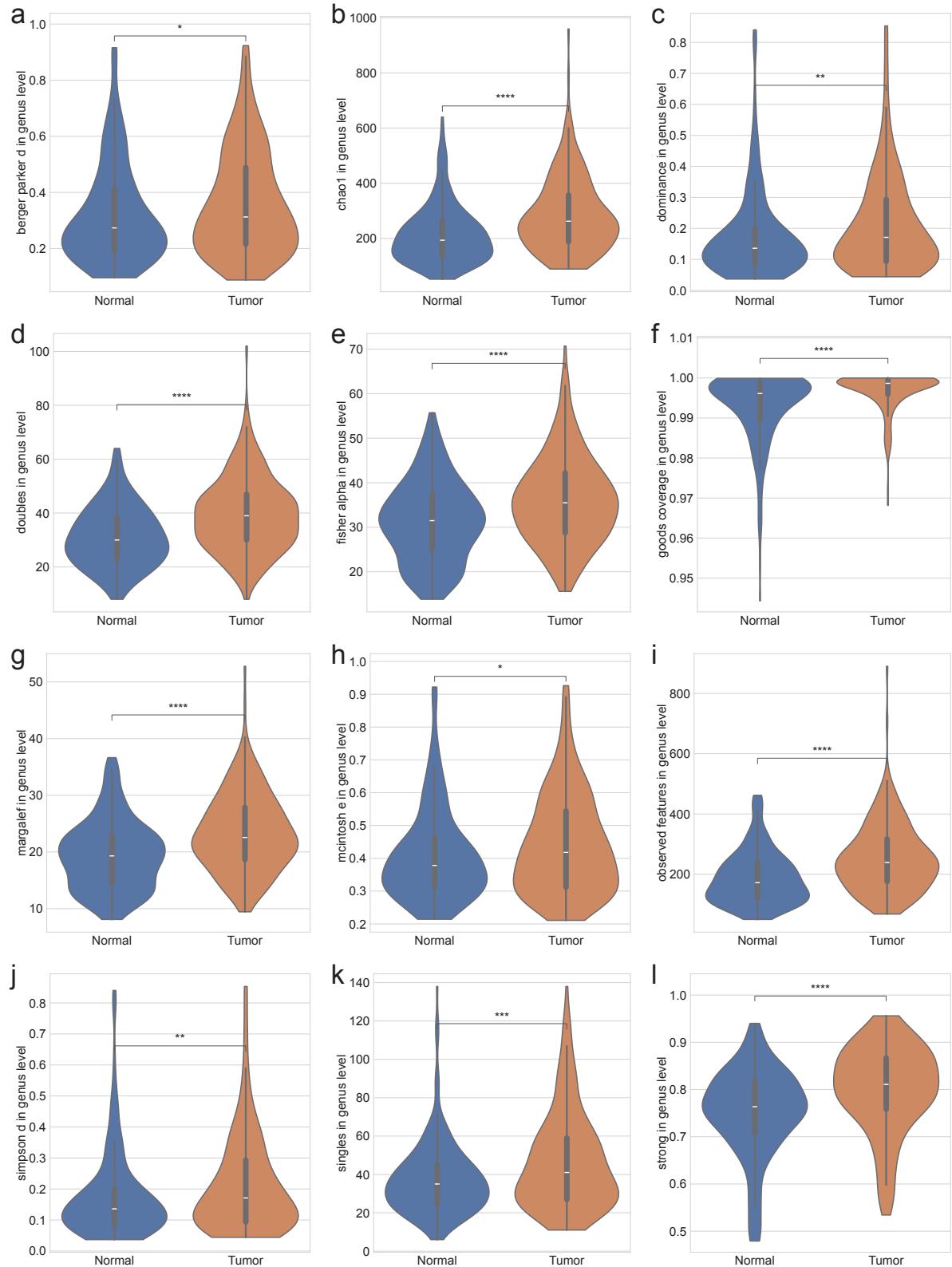


Figure 23: Alpha-diversity indices in genus level.

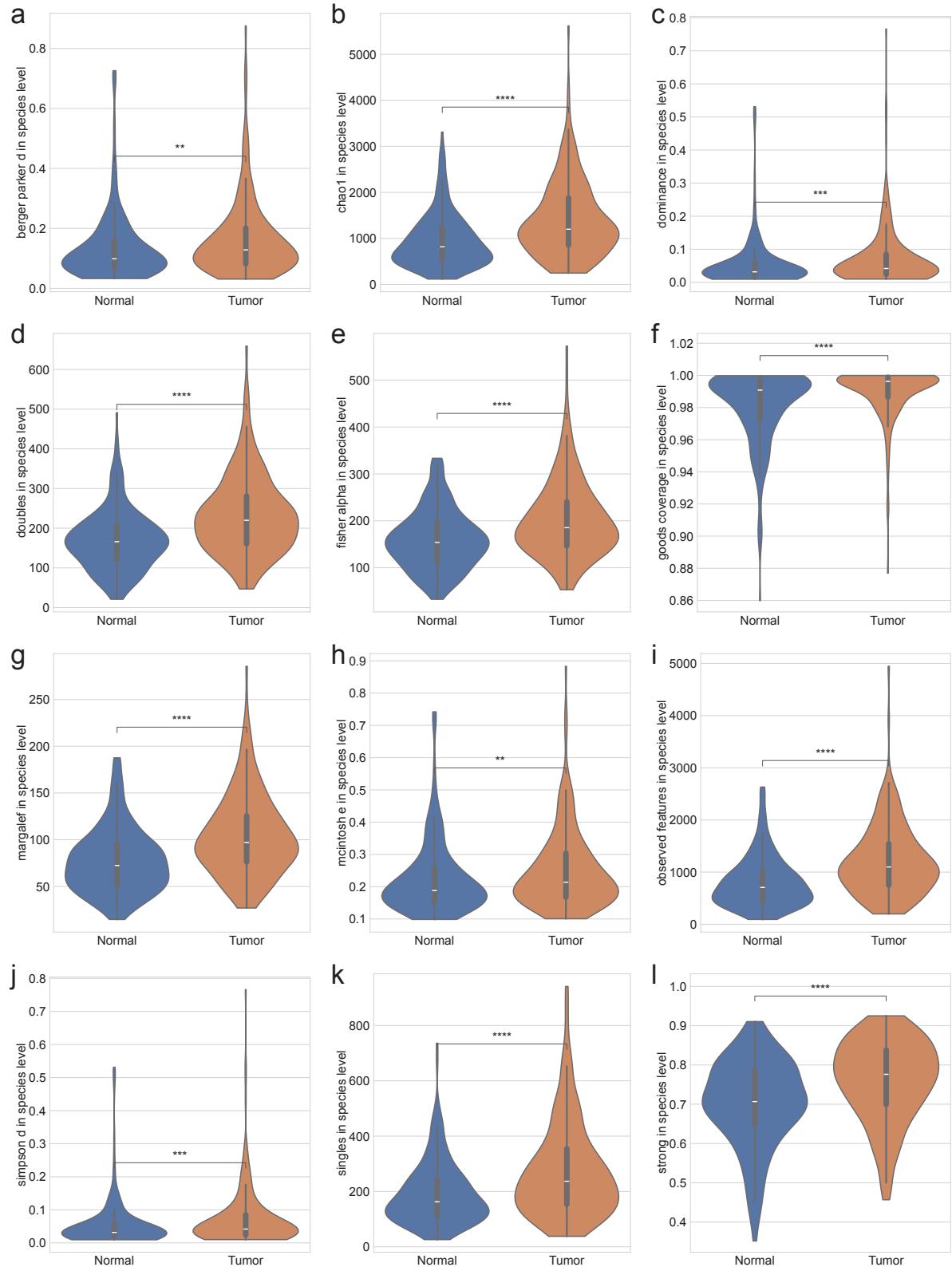


Figure 24: **Alpha-diversity indices in species level.**

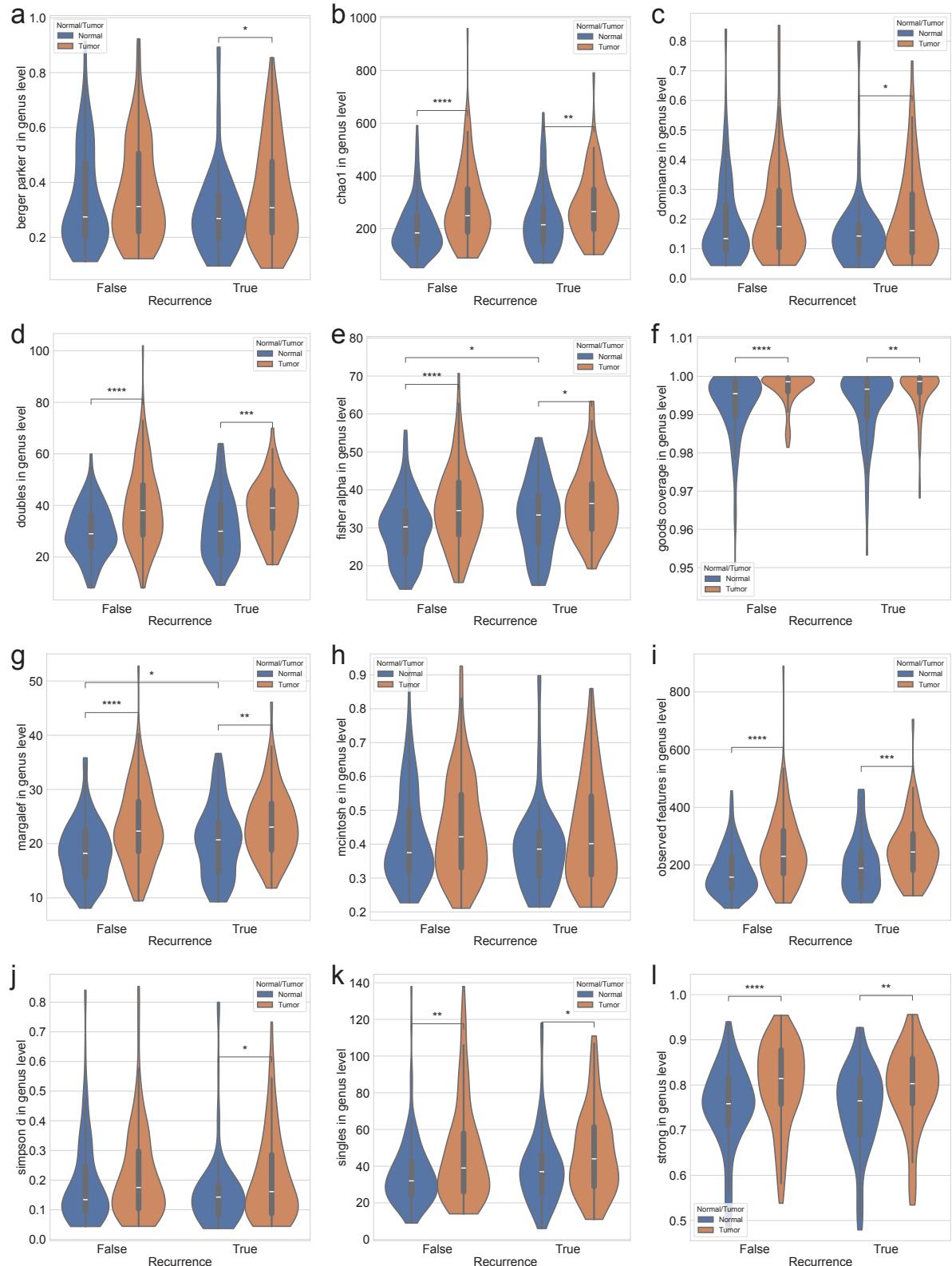


Figure 25: Alpha-diversity indices with recurrence in genus level.

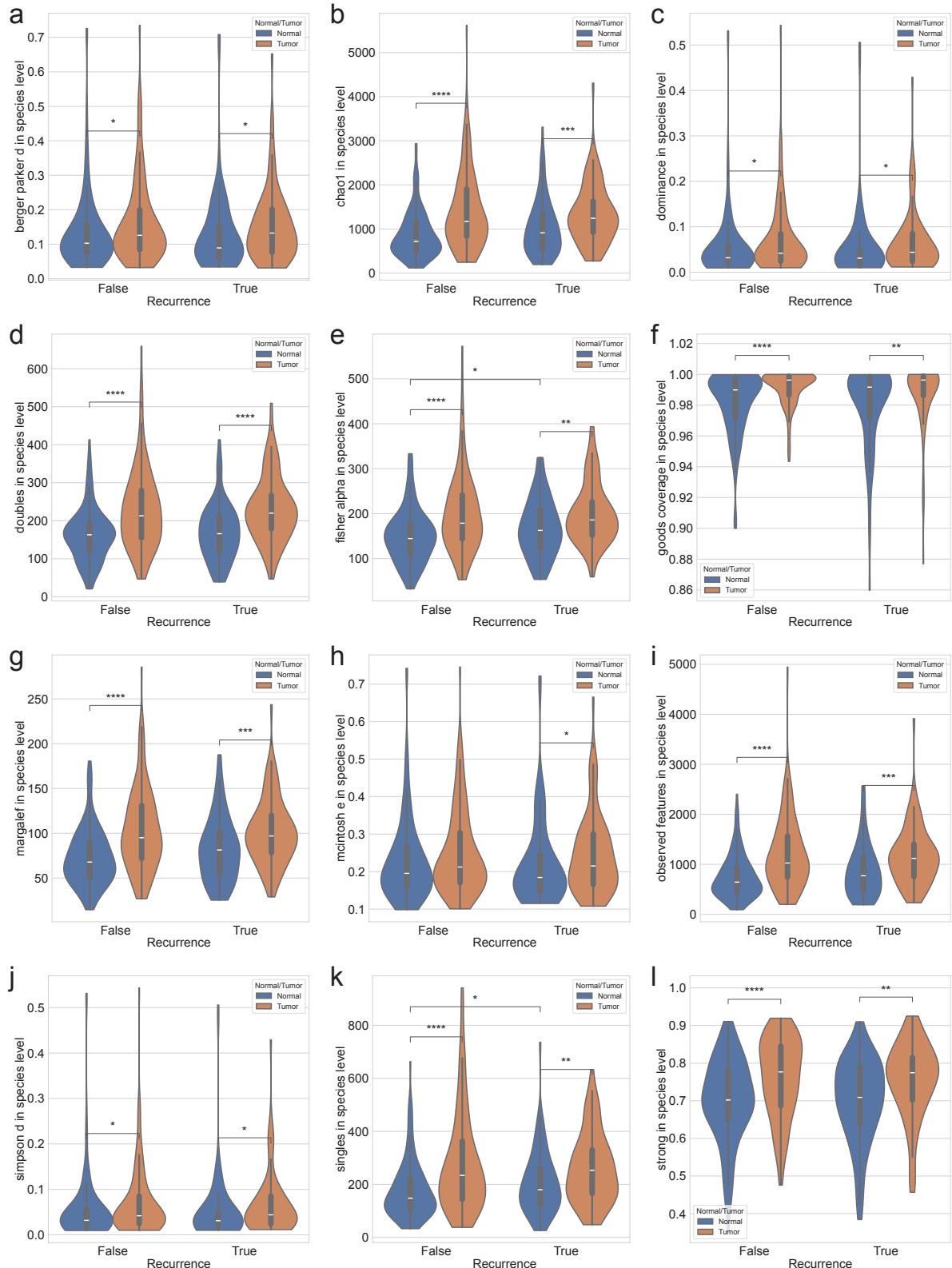


Figure 26: Alpha-diversity indices with recurrence in species level.

1008 **4.4 Discussion**

₁₀₀₉ **5 Conclusion**

₁₀₁₀ In conclusion, the research described in this doctoral dissertation was conducted to identify significant ...

₁₀₁₁ In the section 2, I show that

¹⁰¹² References

- ¹⁰¹³ Aagaard, K., Ma, J., Antony, K. M., Ganu, R., Petrosino, J., & Versalovic, J. (2014). The placenta harbors
¹⁰¹⁴ a unique microbiome. *Science translational medicine*, 6(237), 237ra65–237ra65.
- ¹⁰¹⁵ Abu-Ghazaleh, N., Chua, W. J., & Gopalan, V. (2021). Intestinal microbiota and its association with
¹⁰¹⁶ colon cancer and red/processed meat consumption. *Journal of gastroenterology and hepatology*,
¹⁰¹⁷ 36(1), 75–88.
- ¹⁰¹⁸ Abusleme, L., Hoare, A., Hong, B.-Y., & Diaz, P. I. (2021). Microbial signatures of health, gingivitis,
¹⁰¹⁹ and periodontitis. *Periodontology 2000*, 86(1), 57–78.
- ¹⁰²⁰ Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., & Pawlowsky-Glahn, V. (2000). Logratio
¹⁰²¹ analysis and compositional distance. *Mathematical geology*, 32, 271–275.
- ¹⁰²² Aja, E., Mangar, M., Fletcher, H., & Mishra, A. (2021). Filifactor alocis: recent insights and advances.
¹⁰²³ *Journal of dental research*, 100(8), 790–797.
- ¹⁰²⁴ Alelyani, S. (2021). Stable bagging feature selection on medical data. *Journal of Big Data*, 8(1), 11.
- ¹⁰²⁵ Altabtbaei, K., Maney, P., Ganesan, S. M., Dabdoub, S. M., Nagaraja, H. N., & Kumar, P. S. (2021). Anna
¹⁰²⁶ karenina and the subgingival microbiome associated with periodontitis. *Microbiome*, 9, 1–15.
- ¹⁰²⁷ Altingöz, S. M., Kurgan, Ş., Önder, C., Serdar, M. A., Ünlütürk, U., Uyanık, M., ... Günhan, M.
¹⁰²⁸ (2021). Salivary and serum oxidative stress biomarkers and advanced glycation end products in
¹⁰²⁹ periodontitis patients with or without diabetes: A cross-sectional study. *Journal of periodontology*,
¹⁰³⁰ 92(9), 1274–1285.
- ¹⁰³¹ Alverdy, J., Hyoju, S., Weigerinck, M., & Gilbert, J. (2017). The gut microbiome and the mechanism of
¹⁰³² surgical infection. *Journal of British Surgery*, 104(2), e14–e23.
- ¹⁰³³ An, S., & Park, S. (2022). Association of physical activity and sedentary behavior with the risk of
¹⁰³⁴ colorectal cancer. *Journal of Korean Medical Science*, 37(19).
- ¹⁰³⁵ Anderson, M. J. (2014). Permutational multivariate analysis of variance (permanova). *Wiley statsref:
1036* statistics reference online, 1–15.
- ¹⁰³⁷ Aruni, A. W., Mishra, A., Dou, Y., Chioma, O., Hamilton, B. N., & Fletcher, H. M. (2015). Filifactor
¹⁰³⁸ alocis—a new emerging periodontal pathogen. *Microbes and infection*, 17(7), 517–530.
- ¹⁰³⁹ Aziz, Q., & Thompson, D. G. (1998). Brain-gut axis in health and disease. *Gastroenterology*, 114(3),
¹⁰⁴⁰ 559–578.
- ¹⁰⁴¹ Bai, X., Wei, H., Liu, W., Coker, O. O., Gou, H., Liu, C., ... others (2022). Cigarette smoke promotes
¹⁰⁴² colorectal cancer through modulation of gut microbiota and related metabolites. *Gut*, 71(12),

- 1043 2439–2450.
- 1044 Baldelli, V., Scaldaferri, F., Putignani, L., & Del Chierico, F. (2021). The role of enterobacteriaceae in
1045 gut microbiota dysbiosis in inflammatory bowel diseases. *Microorganisms*, 9(4), 697.
- 1046 Bardou, M., Rouland, A., Martel, M., Loffroy, R., Barkun, A. N., & Chapelle, N. (2022). Obesity and
1047 colorectal cancer. *Alimentary Pharmacology & Therapeutics*, 56(3), 407–418.
- 1048 Barlow, G. M., Yu, A., & Mathur, R. (2015). Role of the gut microbiome in obesity and diabetes mellitus.
1049 *Nutrition in clinical practice*, 30(6), 787–797.
- 1050 Basavaprabhu, H., Sonu, K., & Prabha, R. (2020). Mechanistic insights into the action of probiotics
1051 against bacterial vaginosis and its mediated preterm birth: An overview. *Microbial pathogenesis*,
1052 141, 104029.
- 1053 Belstrøm, D., Constancias, F., Drautz-Moses, D. I., Schuster, S. C., Veleba, M., Mahé, F., & Givskov, M.
1054 (2021). Periodontitis associates with species-specific gene expression of the oral microbiota. *npj
1055 Biofilms and Microbiomes*, 7(1), 76.
- 1056 Berger, W. H., & Parker, F. L. (1970). Diversity of planktonic foraminifera in deep-sea sediments.
1057 *Science*, 168(3937), 1345–1347.
- 1058 Berghella, V. (2012). Universal cervical length screening for prediction and prevention of preterm birth.
1059 *Obstetrical & gynecological survey*, 67(10), 653–657.
- 1060 Blencowe, H., Cousens, S., Oestergaard, M. Z., Chou, D., Moller, A.-B., Narwal, R., ... others (2012).
1061 National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends
1062 since 1990 for selected countries: a systematic analysis and implications. *The lancet*, 379(9832),
1063 2162–2172.
- 1064 Boland, C. R., & Goel, A. (2010). Microsatellite instability in colorectal cancer. *Gastroenterology*,
1065 138(6), 2073–2087.
- 1066 Boleij, A., Hechenbleikner, E. M., Goodwin, A. C., Badani, R., Stein, E. M., Lazarev, M. G., ... others
1067 (2015). The bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer
1068 patients. *Clinical Infectious Diseases*, 60(2), 208–215.
- 1069 Bolstad, A., Jensen, H. B., & Bakken, V. (1996). Taxonomy, biology, and periodontal aspects of
1070 fusobacterium nucleatum. *Clinical microbiology reviews*, 9(1), 55–71.
- 1071 Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... others
1072 (2019). Reproducible, interactive, scalable and extensible microbiome data science using qiime 2.
1073 *Nature biotechnology*, 37(8), 852–857.
- 1074 Bombin, A., Yan, S., Bombin, S., Mosley, J. D., & Ferguson, J. F. (2022). Obesity influences composition
1075 of salivary and fecal microbiota and impacts the interactions between bacterial taxa. *Physiological
1076 reports*, 10(7), e15254.
- 1077 Bonnet, M., Buc, E., Sauvanet, P., Darcha, C., Dubois, D., Pereira, B., ... Darfeuille-Michaud, A. (2014).
1078 Colonization of the human gut by e. coli and colorectal cancer risk. *Clinical Cancer Research*,
1079 20(4), 859–867.
- 1080 Breiman, L. (2001). Random forests. *Machine learning*, 45, 5–32.
- 1081 Brennan, C. A., & Garrett, W. S. (2019). Fusobacterium nucleatum—symbiont, opportunist and

- 1082 oncobacterium. *Nature Reviews Microbiology*, 17(3), 156–166.
- 1083 Broom, L. J., & Kogut, M. H. (2018). The role of the gut microbiome in shaping the immune system of
1084 chickens. *Veterinary immunology and immunopathology*, 204, 44–51.
- 1085 Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: improving accuracy of classifier
1086 ensembles by using random feature subsets. *Pattern recognition*, 36(6), 1291–1302.
- 1087 Bullman, S., Pedamallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., ... others (2017). Analysis
1088 of fusobacterium persistence and antibiotic response in colorectal cancer. *Science*, 358(6369),
1089 1443–1448.
- 1090 Burt, R. W., Leppert, M. F., Slattery, M. L., Samowitz, W. S., Spirio, L. N., Kerber, R. A., ... others
1091 (2004). Genetic testing and phenotype in a large kindred with attenuated familial adenomatous
1092 polyposis. *Gastroenterology*, 127(2), 444–451.
- 1093 Cai, Y., Li, Y., Xiong, Y., Geng, X., Kang, Y., & Yang, Y. (2024). Diabetic foot exacerbates gut
1094 mycobiome dysbiosis in adult patients with type 2 diabetes mellitus: revealing diagnostic markers.
1095 *Nutrition & Diabetes*, 14(1), 71.
- 1096 Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016).
1097 Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7),
1098 581–583.
- 1099 Canakci, V., & Canakci, C. F. (2007). Pain levels in patients during periodontal probing and mechanical
1100 non-surgical therapy. *Clinical oral investigations*, 11, 377–383.
- 1101 Cappellato, M., Baruzzo, G., & Di Camillo, B. (2022). Investigating differential abundance methods in
1102 microbiome data: A benchmark study. *PLoS computational biology*, 18(9), e1010467.
- 1103 Castaner, O., Goday, A., Park, Y.-M., Lee, S.-H., Magkos, F., Shiow, S.-A. T. E., & Schröder, H. (2018).
1104 The gut microbiome profile in obesity: a systematic review. *International journal of endocrinology*,
1105 2018(1), 4095789.
- 1106 Center, M. M., Jemal, A., Smith, R. A., & Ward, E. (2009). Worldwide variations in colorectal cancer.
1107 *CA: a cancer journal for clinicians*, 59(6), 366–378.
- 1108 Centor, R. M. (1991). Signal detectability: the use of roc curves and their analyses. *Medical decision
1109 making*, 11(2), 102–106.
- 1110 Cerqueira, F. M., Photenhauer, A. L., Pollet, R. M., Brown, H. A., & Koropatkin, N. M. (2020). Starch
1111 digestion by gut bacteria: crowdsourcing for carbs. *Trends in Microbiology*, 28(2), 95–108.
- 1112 Champagne, C., McNairn, H., Daneshfar, B., & Shang, J. (2014). A bootstrap method for assessing
1113 classification accuracy and confidence for agricultural land use mapping in canada. *International
1114 Journal of Applied Earth Observation and Geoinformation*, 29, 44–52.
- 1115 Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian
1116 Journal of statistics*, 265–270.
- 1117 Chao, A., & Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the
1118 American statistical Association*, 87(417), 210–217.
- 1119 Chapple, I. L., Mealey, B. L., Van Dyke, T. E., Bartold, P. M., Dommisch, H., Eickholz, P., ... others
1120 (2018). Periodontal health and gingival diseases and conditions on an intact and a reduced

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

- 1160 *Ecology*, 42–58.
- 1161 Flanagan, L., Schmid, J., Ebert, M., Soucek, P., Kunicka, T., Liska, V., ... others (2014). Fusobacterium
1162 nucleatum associates with stages of colorectal neoplasia development, colorectal cancer and disease
1163 outcome. *European journal of clinical microbiology & infectious diseases*, 33, 1381–1390.
- 1164 Fortenberry, J. D. (2013). The uses of race and ethnicity in human microbiome research. *Trends in
1165 microbiology*, 21(4), 165–166.
- 1166 Francescone, R., Hou, V., & Grivennikov, S. I. (2014). Microbiome, inflammation, and cancer. *The
1167 Cancer Journal*, 20(3), 181–189.
- 1168 Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4),
1169 367–378.
- 1170 Gamin, D. J., Vitali, F. C., De Carli, J. P., Mazzon, R. R., Gomes, B. P., Duque, T. M., & Trentin, M. S.
1171 (2021). Prevalence of red and orange microbial complexes in endodontic-periodontal lesions: a
1172 systematic review and meta-analysis. *Clinical Oral Investigations*, 1–14.
- 1173 Gao, J., Yin, J., Xu, K., Li, T., & Yin, Y. (2019). What is the impact of diet on nutritional diarrhea
1174 associated with gut microbiota in weaning piglets: a system review. *BioMed research international*,
1175 2019(1), 6916189.
- 1176 Ghanavati, R., Akbari, A., Mohammadi, F., Asadollahi, P., Javadi, A., Talebi, M., & Rohani, M. (2020).
1177 Lactobacillus species inhibitory effect on colorectal cancer progression through modulating the
1178 wnt/β-catenin signaling pathway. *Molecular and Cellular Biochemistry*, 470, 1–13.
- 1179 Ghorbani, E., Avan, A., Ryzhikov, M., Ferns, G., Khazaei, M., & Soleimanpour, S. (2022). Role of
1180 lactobacillus strains in the management of colorectal cancer: An overview of recent advances.
1181 *Nutrition*, 103, 111828.
- 1182 Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current
1183 understanding of the human microbiome. *Nature medicine*, 24(4), 392–400.
- 1184 Gini, C. (1912). Variabilità e mutabilità (variability and mutability). *Tipografia di Paolo Cuppini,
1185 Bologna, Italy*, 156.
- 1186 Goldenberg, R. L., Culhane, J. F., Iams, J. D., & Romero, R. (2008). Epidemiology and causes of preterm
1187 birth. *The lancet*, 371(9606), 75–84.
- 1188 Gonçalves, L., Subtil, A., Oliveira, M. R., & de Zea Bermudez, P. (2014). Roc curve estimation: An
1189 overview. *REVSTAT-Statistical journal*, 12(1), 1–20.
- 1190 Goodyear, M. D., Krleza-Jeric, K., & Lemmens, T. (2007). *The declaration of helsinki* (Vol. 335) (No.
1191 7621). British Medical Journal Publishing Group.
- 1192 Haffajee, A., Teles, R., & Socransky, S. (2006). Association of eubacterium nodatum and treponema
1193 denticola with human periodontitis lesions. *Oral microbiology and immunology*, 21(5), 269–282.
- 1194 Hajishengallis, G. (2015). Periodontitis: from microbial immune subversion to systemic inflammation.
1195 *Nature reviews immunology*, 15(1), 30–44.
- 1196 Hamjane, N., Mechita, M. B., Nourouti, N. G., & Barakat, A. (2024). Gut microbiota dysbiosis-associated
1197 obesity and its involvement in cardiovascular diseases and type 2 diabetes. a systematic review.
1198 *Microvascular Research*, 151, 104601.

- 1199 Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*,
1200 29(2), 147–160.
- 1201 Hampel, H., Frankel, W. L., Martin, E., Arnold, M., Khanduja, K., Kuebler, P., ... others (2008).
1202 Feasibility of screening for lynch syndrome among patients with colorectal cancer. *Journal of
1203 Clinical Oncology*, 26(35), 5783–5788.
- 1204 Han, Y. W. (2015). Fusobacterium nucleatum: a commensal-turned pathogen. *Current opinion in
1205 microbiology*, 23, 141–147.
- 1206 Han, Y. W., & Wang, X. (2013). Mobile microbiome: oral bacteria in extra-oral infections and
1207 inflammation. *Journal of dental research*, 92(6), 485–491.
- 1208 Hand, D. J. (2012). Assessing the performance of classification methods. *International Statistical Review*,
1209 80(3), 400–414.
- 1210 Hartstra, A. V., Bouter, K. E., Bäckhed, F., & Nieuwdorp, M. (2015). Insights into the role of the
1211 microbiome in obesity and type 2 diabetes. *Diabetes care*, 38(1), 159–165.
- 1212 Hashemi Goradel, N., Heidarzadeh, S., Jahangiri, S., Farhood, B., Mortezaee, K., Khanlarkhani, N., &
1213 Negahdari, B. (2019). Fusobacterium nucleatum and colorectal cancer: A mechanistic overview.
1214 *Journal of Cellular Physiology*, 234(3), 2337–2344.
- 1215 Helmink, B. A., Khan, M. W., Hermann, A., Gopalakrishnan, V., & Wargo, J. A. (2019). The microbiome,
1216 cancer, and cancer therapy. *Nature medicine*, 25(3), 377–388.
- 1217 Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2),
1218 427–432.
- 1219 Hiranmayi, K. V., Sirisha, K., Rao, M. R., & Sudhakar, P. (2017). Novel pathogens in periodontal
1220 microbiology. *Journal of Pharmacy and Bioallied Sciences*, 9(3), 155–163.
- 1221 Honda, K., & Littman, D. R. (2012). The microbiome in infectious disease and inflammation. *Annual
1222 review of immunology*, 30(1), 759–795.
- 1223 Honest, H., Forbes, C., Durée, K., Norman, G., Duffy, S., Tsourapas, A., ... others (2009). Screening to
1224 prevent spontaneous preterm birth: systematic reviews of accuracy and effectiveness literature with
1225 economic modelling. *Health Technol Assess*, 13(43), 1–627.
- 1226 Hong, Y. M., Lee, J., Cho, D. H., Jeon, J. H., Kang, J., Kim, M.-G., ... J. K. (2023). Predicting preterm
1227 birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1), 21105.
- 1228 Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations.
1229 *International journal of data mining & knowledge management process*, 5(2), 1.
- 1230 Huang, R.-Y., Lin, C.-D., Lee, M.-S., Yeh, C.-L., Shen, E.-C., Chiang, C.-Y., ... Fu, E. (2007). Mandibular
1231 disto-lingual root: a consideration in periodontal therapy. *Journal of periodontology*, 78(8), 1485–
1232 1490.
- 1233 Iams, J. D., & Berghella, V. (2010). Care for women with prior preterm birth. *American journal of
1234 obstetrics and gynecology*, 203(2), 89–100.
- 1235 Ide, M., & Papapanou, P. N. (2013). Epidemiology of association between maternal periodontal
1236 disease and adverse pregnancy outcomes—systematic review. *Journal of clinical periodontology*,
1237 40, S181–S194.

- 1238 Iniesta, M., Chamorro, C., Ambrosio, N., Marín, M. J., Sanz, M., & Herrera, D. (2023). Subgingival
1239 microbiome in periodontal health, gingivitis and different stages of periodontitis. *Journal of*
1240 *Clinical Periodontology*, 50(7), 905–920.
- 1241 Inra, J. A., Steyerberg, E. W., Grover, S., McFarland, A., Syngal, S., & Kastrinos, F. (2015). Racial
1242 variation in frequency and phenotypes of apc and mutyh mutations in 6,169 individuals undergoing
1243 genetic testing. *Genetics in Medicine*, 17(10), 815–821.
- 1244 Janda, J. M., & Abbott, S. L. (2007). 16s rrna gene sequencing for bacterial identification in the diagnostic
1245 laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9), 2761–2764.
- 1246 Jiang, W., & Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach
1247 for estimating the prediction error in microarray classification. *Statistics in medicine*, 26(29),
1248 5320–5334.
- 1249 John, G. K., & Mullin, G. E. (2016). The gut microbiome and obesity. *Current oncology reports*, 18,
1250 1–7.
- 1251 Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., . . . others (2019).
1252 Evaluation of 16s rrna gene sequencing for species and strain-level microbiome analysis. *Nature*
1253 *communications*, 10(1), 5029.
- 1254 Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N., & Whiteley, M. (2014). Metatranscriptomics
1255 of the human oral microbiome during health and disease. *MBio*, 5(2), 10–1128.
- 1256 Joscelyn, J., & Kasper, L. H. (2014). Digesting the emerging role for the gut microbiome in central
1257 nervous system demyelination. *Multiple Sclerosis Journal*, 20(12), 1553–1559.
- 1258 Kang, Y., Kang, X., Yang, H., Liu, H., Yang, X., Liu, Q., . . . others (2022). Lactobacillus acidophilus ame-
1259 liorates obesity in mice through modulation of gut microbiota dysbiosis and intestinal permeability.
1260 *Pharmacological research*, 175, 106020.
- 1261 Karched, M., Bhardwaj, R. G., Qudeimat, M., Al-Khabbaz, A., & Ellepol, A. (2022). Proteomic analysis
1262 of the periodontal pathogen prevotella intermedia secretomes in biofilm and planktonic lifestyles.
1263 *Scientific Reports*, 12(1), 5636.
- 1264 Katz, J., Chegini, N., Shiverick, K., & Lamont, R. (2009). Localization of p. gingivalis in preterm delivery
1265 placenta. *Journal of dental research*, 88(6), 575–578.
- 1266 Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., & Gordon, J. I. (2011). Human nutrition, the
1267 gut microbiome and the immune system. *Nature*, 474(7351), 327–336.
- 1268 Kelly, B. J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J. D., Collman, R. G., . . . Li, H. (2015).
1269 Power and sample-size estimation for microbiome studies using pairwise distances and permanova.
1270 *Bioinformatics*, 31(15), 2461–2468.
- 1271 Kennedy, J., Alexander, P., Taillie, L. S., & Jaacks, L. M. (2024). Estimated effects of reductions in
1272 processed meat consumption and unprocessed red meat consumption on occurrences of type 2
1273 diabetes, cardiovascular disease, colorectal cancer, and mortality in the usa: a microsimulation
1274 study. *The Lancet Planetary Health*, 8(7), e441–e451.
- 1275 Kim, B.-R., Shin, J., Guevarra, R. B., Lee, J. H., Kim, D. W., Seol, K.-H., . . . Isaacson, R. E. (2017).
1276 Deciphering diversity indices for a better understanding of microbial communities. *Journal of*

- 1277 *Microbiology and Biotechnology*, 27(12), 2089–2093.
- 1278 Kim, C. H. (2018). Immune regulation by microbiome metabolites. *Immunology*, 154(2), 220–229.
- 1279 Kim, E.-H., Kim, S., Kim, H.-J., Jeong, H.-o., Lee, J., Jang, J., ... others (2020). Prediction of chronic
1280 periodontitis severity using machine learning models based on salivary bacterial copy number.
1281 *Frontiers in Cellular and Infection Microbiology*, 10, 571515.
- 1282 Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and
1283 bootstrap. *Computational statistics & data analysis*, 53(11), 3735–3745.
- 1284 Kinane, D. F., Stathopoulou, P. G., & Papapanou, P. N. (2017). Periodontal diseases. *Nature reviews
1285 Disease primers*, 3(1), 1–14.
- 1286 Kindinger, L. M., Bennett, P. R., Lee, Y. S., Marchesi, J. R., Smith, A., Cacciato, S., ... MacIntyre,
1287 D. A. (2017). The interaction between vaginal microbiota, cervical length, and vaginal progesterone
1288 treatment for preterm birth risk. *Microbiome*, 5, 1–14.
- 1289 Kogut, M. H., Lee, A., & Santin, E. (2020). Microbiome and pathogen interaction with the immune
1290 system. *Poultry science*, 99(4), 1906–1913.
- 1291 Kostic, A. D., Ojesina, A. I., Pedamallu, C. S., Jung, J., Verhaak, R. G., Getz, G., & Meyerson, M. (2011).
1292 Pathseq: software to identify or discover microbes by deep sequencing of human tissue. *Nature
1293 biotechnology*, 29(5), 393–396.
- 1294 Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification
1295 and combining techniques. *Artificial Intelligence Review*, 26, 159–190.
- 1296 Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., ... Watanabe, T.
1297 (2015). Colorectal cancer. *Nature reviews. Disease primers*, 1, 15065.
- 1298 Lafaurie, G. I., Neuta, Y., Ríos, R., Pacheco-Montealegre, M., Pianeta, R., Castillo, D. M., ... oth-
1299 ers (2022). Differences in the subgingival microbiome according to stage of periodontitis: A
1300 comparison of two geographic regions. *PLoS one*, 17(8), e0273523.
- 1301 Lamont, R. J., & Jenkinson, H. F. (2000). Subgingival colonization by porphyromonas gingivalis. *Oral
1302 Microbiology and Immunology: Mini-review*, 15(6), 341–349.
- 1303 Lamont, R. J., Koo, H., & Hajishengallis, G. (2018). The oral microbiota: dynamic communities and
1304 host interactions. *Nature reviews microbiology*, 16(12), 745–759.
- 1305 Leitich, H., & Kaider, A. (2003). Fetal fibronectin—how useful is it in the prediction of preterm birth?
1306 *BJOG: An International Journal of Obstetrics & Gynaecology*, 110, 66–70.
- 1307 Le Leu, R. K., Hu, Y., Brown, I. L., Woodman, R. J., & Young, G. P. (2010). Synbiotic intervention of
1308 bifidobacterium lactis and resistant starch protects against colorectal cancer development in rats.
1309 *Carcinogenesis*, 31(2), 246–251.
- 1310 León, R., Silva, N., Ovalle, A., Chaparro, A., Ahumada, A., Gajardo, M., ... Gamonal, J. (2007).
1311 Detection of porphyromonas gingivalis in the amniotic fluid in pregnant women with a diagnosis
1312 of threatened premature labor. *Journal of periodontology*, 78(7), 1249–1255.
- 1313 Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform.
1314 *bioinformatics*, 25(14), 1754–1760.
- 1315 Li, N., Lu, B., Luo, C., Cai, J., Lu, M., Zhang, Y., ... Dai, M. (2021). Incidence, mortality, survival,

- 1316 risk factor and screening of colorectal cancer: A comparison among china, europe, and northern
1317 america. *Cancer letters*, 522, 255–268.
- 1318 Li, R., Miao, Z., Liu, Y., Chen, X., Wang, H., Su, J., & Chen, J. (2024). The brain–gut–bone axis in
1319 neurodegenerative diseases: insights, challenges, and future prospects. *Advanced Science*, 11(38),
1320 2307971.
- 1321 Li, X., Yu, D., Wang, Y., Yuan, H., Ning, X., Rui, B., ... Li, M. (2021). The intestinal dysbiosis of
1322 mothers with gestational diabetes mellitus (gdm) and its impact on the gut microbiota of their
1323 newborns. *Canadian Journal of Infectious Diseases and Medical Microbiology*, 2021(1), 3044534.
- 1324 Li, Y., Qian, F., Cheng, X., Wang, D., Wang, Y., Pan, Y., ... Tian, Y. (2023). Dysbiosis of oral microbiota
1325 and metabolite profiles associated with type 2 diabetes mellitus. *Microbiology spectrum*, 11(1),
1326 e03796–22.
- 1327 Lim, J. W., Park, T., Tong, Y. W., & Yu, Z. (2020). The microbiome driving anaerobic digestion and
1328 microbial analysis. In *Advances in bioenergy* (Vol. 5, pp. 1–61). Elsevier.
- 1329 Lin, H., Eggesbø, M., & Peddada, S. D. (2022). Linear and nonlinear correlation estimators unveil
1330 undescribed taxa interactions in microbiome data. *Nature communications*, 13(1), 4946.
- 1331 Lin, H., & Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nature
1332 communications*, 11(1), 3514.
- 1333 Lin, H., & Peddada, S. D. (2024). Multigroup analysis of compositions of microbiomes with covariate
1334 adjustments and repeated measures. *Nature Methods*, 21(1), 83–91.
- 1335 Listgarten, M. A. (1986). Pathogenesis of periodontitis. *Journal of clinical periodontology*, 13(5),
1336 418–425.
- 1337 Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome
1338 medicine*, 8, 1–11.
- 1339 López-Aladid, R., Fernández-Barat, L., Alcaraz-Serrano, V., Bueno-Freire, L., Vázquez, N., Pastor-
1340 Ibáñez, R., ... Torres, A. (2023). Determining the most accurate 16s rrna hypervariable region for
1341 taxonomic identification from respiratory samples. *Scientific reports*, 13(1), 3974.
- 1342 Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for
1343 rna-seq data with deseq2. *Genome biology*, 15, 1–21.
- 1344 Magnúsdóttir, S., & Thiele, I. (2018). Modeling metabolism of the human gut microbiome. *Current
1345 opinion in biotechnology*, 51, 90–96.
- 1346 Magurran, A. E. (2021). Measuring biological diversity. *Current Biology*, 31(19), R1174–R1177.
- 1347 Mandic, M., Safizadeh, F., Niedermaier, T., Hoffmeister, M., & Brenner, H. (2023). Association of
1348 overweight, obesity, and recent weight loss with colorectal cancer risk. *JAMA network Open*, 6(4),
1349 e239556–e239556.
- 1350 Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically
1351 larger than the other. *The annals of mathematical statistics*, 50–60.
- 1352 Manolis, A. A., Manolis, T. A., Melita, H., & Manolis, A. S. (2022). Gut microbiota and cardiovascular
1353 disease: symbiosis versus dysbiosis. *Current Medicinal Chemistry*, 29(23), 4050–4077.
- 1354 Martin, C. R., Osadchiy, V., Kalani, A., & Mayer, E. A. (2018). The brain-gut-microbiome axis. *Cellular*

- 1355 and molecular gastroenterology and hepatology, 6(2), 133–148.
- 1356 Mayer, E. A., Tillisch, K., Gupta, A., et al. (2015). Gut/brain axis and the microbiota. *The Journal of*
1357 *clinical investigation*, 125(3), 926–938.
- 1358 Melguizo-Rodríguez, L., Costela-Ruiz, V. J., Manzano-Moreno, F. J., Ruiz, C., & Illescas-Montes, R.
1359 (2020). Salivary biomarkers and their application in the diagnosis and monitoring of the most
1360 common oral pathologies. *International journal of molecular sciences*, 21(14), 5173.
- 1361 Merrill, L. C., & Mangano, K. M. (2023). Racial and ethnic differences in studies of the gut microbiome
1362 and osteoporosis. *Current Osteoporosis Reports*, 21(5), 578–591.
- 1363 Miller, C. S., Ding, X., Dawson III, D. R., & Ebersole, J. L. (2021). Salivary biomarkers for discriminating
1364 periodontitis in the presence of diabetes. *Journal of clinical periodontology*, 48(2), 216–225.
- 1365 Morita, T., Yamazaki, Y., Mita, A., Takada, K., Seto, M., Nishinoue, N., ... Maeno, M. (2010). A cohort
1366 study on the association between periodontal disease and the development of metabolic syndrome.
1367 *Journal of periodontology*, 81(4), 512–519.
- 1368 Na, H. S., Kim, S. Y., Han, H., Kim, H.-J., Lee, J.-Y., Lee, J.-H., & Chung, J. (2020). Identification of
1369 potential oral microbial biomarkers for the diagnosis of periodontitis. *Journal of clinical medicine*,
1370 9(5), 1549.
- 1371 Nemoto, T., Shiba, T., Komatsu, K., Watanabe, T., Shimogishi, M., Shibasaki, M., ... others (2021).
1372 Discrimination of bacterial community structures among healthy, gingivitis, and periodontitis
1373 statuses through integrated metatranscriptomic and network analyses. *Msystems*, 6(6), e00886–21.
- 1374 Nesbitt, M. J., Reynolds, M. A., Shiau, H., Choe, K., Simonsick, E. M., & Ferrucci, L. (2010). Association
1375 of periodontitis and metabolic syndrome in the baltimore longitudinal study of aging. *Aging clinical*
1376 *and experimental research*, 22, 238–242.
- 1377 Network, C. G. A., et al. (2012). Comprehensive molecular characterization of human colon and rectal
1378 cancer. *Nature*, 487(7407), 330.
- 1379 Nibali, L., Sousa, V., Davrandi, M., Spratt, D., Alyahya, Q., Dopico, J., & Donos, N. (2020). Differences
1380 in the periodontal microbiome of successfully treated and persistent aggressive periodontitis.
1381 *Journal of Clinical Periodontology*, 47(8), 980–990.
- 1382 Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Ž., & Tomović, M. (2017). Evaluation of classification
1383 models in machine learning. *Theory and Applications of Mathematics & Computer Science*, 7(1),
1384 39.
- 1385 Obuchowski, N. A., & Bullen, J. A. (2018). Receiver operating characteristic (roc) curves: review of
1386 methods with applications in diagnostic medicine. *Physics in Medicine & Biology*, 63(7), 07TR01.
- 1387 Offenbacher, S., Katz, V., Fertik, G., Collins, J., Boyd, D., Maynor, G., ... Beck, J. (1996). Periodontal
1388 infection as a possible risk factor for preterm low birth weight. *Journal of periodontology*, 67,
1389 1103–1113.
- 1390 Ojesina, A. I., Pedamallu, C. S., Kostic, A., Jung, J., Auclair, D., Lohr, J., ... Meyerson, M. (2013). High
1391 throughput sequencing-based pathogen discovery in multiple myeloma. *Blood*, 122(21), 5322.
- 1392 Omundiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019). Machine learning classification techniques
1393 for breast cancer diagnosis. In *Iop conference series: materials science and engineering* (Vol. 495,

- 1394 p. 012033).
- 1395 O'Sullivan, D. E., Sutherland, R. L., Town, S., Chow, K., Fan, J., Forbes, N., ... Brenner, D. R. (2022).
1396 Risk factors for early-onset colorectal cancer: a systematic review and meta-analysis. *Clinical*
1397 *gastroenterology and hepatology*, 20(6), 1229–1240.
- 1398 Paganini, D., & Zimmermann, M. B. (2017). The effects of iron fortification and supplementation on the
1399 gut microbiome and diarrhea in infants and children: a review. *The American journal of clinical*
1400 *nutrition*, 106, 1688S–1693S.
- 1401 Pan, A. Y. (2021). Statistical analysis of microbiome data: the challenge of sparsity. *Current Opinion in*
1402 *Endocrine and Metabolic Research*, 19, 35–40.
- 1403 Papapanou, P. N., Sanz, M., Buduneli, N., Dietrich, T., Feres, M., Fine, D. H., ... others (2018).
1404 Periodontitis: Consensus report of workgroup 2 of the 2017 world workshop on the classification of
1405 periodontal and peri-implant diseases and conditions. *Journal of periodontology*, 89, S173–S182.
- 1406 Parizadeh, M., & Arrieta, M.-C. (2023). The global human gut microbiome: genes, lifestyles, and diet.
1407 *Trends in Molecular Medicine*.
- 1408 Park, J., Park, S. H., Lee, D., Lee, J. E., Lee, D., Na, K. J., ... Im, H.-J. (2024). Detecting cancer microbiota
1409 using unmapped rna reads on spatial transcriptomics. *Cancer Research*, 84(6_Supplement), 4881–
1410 4881.
- 1411 Payne, M. S., Newnham, J. P., Doherty, D. A., Furfarro, L. L., Pendal, N. L., Loh, D. E., & Keelan, J. A.
1412 (2021). A specific bacterial dna signature in the vagina of australian women in midpregnancy
1413 predicts high risk of spontaneous preterm birth (the predict1000 study). *American journal of*
1414 *obstetrics and gynecology*, 224(2), 206–e1.
- 1415 Peirce, J. M., & Alviña, K. (2019). The role of inflammation and the gut microbiome in depression and
1416 anxiety. *Journal of neuroscience research*, 97(10), 1223–1241.
- 1417 Peltomaki, P. (2003). Role of dna mismatch repair defects in the pathogenesis of human cancer. *Journal*
1418 *of clinical oncology*, 21(6), 1174–1179.
- 1419 Pezzino, S., Sofia, M., Greco, L. P., Litrico, G., Filippello, G., Sarvà, I., ... Latteri, S. (2023). Microbiome
1420 dysbiosis: a pathological mechanism at the intersection of obesity and glaucoma. *International*
1421 *Journal of Molecular Sciences*, 24(2), 1166.
- 1422 Premaraj, T. S., Vella, R., Chung, J., Lin, Q., Hunter, P., Underwood, K., ... Zhou, Y. (2020). Ethnic
1423 variation of oral microbiota in children. *Scientific reports*, 10(1), 14788.
- 1424 Raut, J. R., Schöttker, B., Holleczeck, B., Guo, F., Bhardwaj, M., Miah, K., ... Brenner, H. (2021).
1425 A microrna panel compared to environmental and polygenic scores for colorectal cancer risk
1426 prediction. *Nature Communications*, 12(1), 4811.
- 1427 Rebersek, M. (2021). Gut microbiome and its role in colorectal cancer. *BMC cancer*, 21(1), 1325.
- 1428 Redanz, U., Redanz, S., Treerat, P., Prakasam, S., Lin, L.-J., Merritt, J., & Kreth, J. (2021). Differential
1429 response of oral mucosal and gingival cells to corynebacterium durum, streptococcus sanguinis, and
1430 porphyromonas gingivalis multispecies biofilms. *Frontiers in cellular and infection microbiology*,
1431 11, 686479.
- 1432 Relvas, M., Regueira-Iglesias, A., Balsa-Castro, C., Salazar, F., Pacheco, J., Cabral, C., ... Tomás, I.

- 1433 (2021). Relationship between dental and periodontal health status and the salivary microbiome:
1434 bacterial diversity, co-occurrence networks and predictive models. *Scientific reports*, 11(1), 929.
- 1435 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.
- 1437 Renvert, S., & Persson, G. (2002). A systematic review on the use of residual probing depth, bleeding on
1438 probing and furcation status following initial periodontal therapy to predict further attachment and
1439 tooth loss. *Journal of clinical periodontology*, 29, 82–89.
- 1440 Rideout, J. R., Caporaso, G., Bolyen, E., McDonald, D., Baeza, Y. V., Alastuey, J. C., ... Sharma, K.
1441 (2018, December). *biocore/scikit-bio: scikit-bio 0.5.5: More compositional methods added*. Zenodo.
1442 Retrieved from <https://doi.org/10.5281/zenodo.2254379> doi: 10.5281/zenodo.2254379
- 1443 Rôças, I. N., Siqueira Jr, J. F., Santos, K. R., Coelho, A. M., & de Janeiro, R. (2001). “red complex”
1444 (bacteroides forsythus, porphyromonas gingivalis, and treponema denticola) in endodontic
1445 infections: a molecular approach. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology,*
1446 *and Endodontology*, 91(4), 468–471.
- 1447 Romero, R., Dey, S. K., & Fisher, S. J. (2014). Preterm labor: one syndrome, many causes. *Science*,
1448 345(6198), 760–765.
- 1449 Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., ... others (2014). The
1450 composition and stability of the vaginal microbiota of normal pregnant women is different from
1451 that of non-pregnant women. *Microbiome*, 2, 1–19.
- 1452 Rosan, B., & Lamont, R. J. (2000). Dental plaque formation. *Microbes and infection*, 2(13), 1599–1607.
- 1453 Schwabe, R. F., & Jobin, C. (2013). The microbiome and cancer. *Nature Reviews Cancer*, 13(11),
1454 800–812.
- 1455 Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011).
1456 Metagenomic biomarker discovery and explanation. *Genome biology*, 12, 1–18.
- 1457 Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A
1458 survey and review. In *Emerging technology in modelling and graphics: Proceedings of iem graph*
1459 2018 (pp. 99–111).
- 1460 Sepich-Poore, G. D., Zitvogel, L., Straussman, R., Hasty, J., Wargo, J. A., & Knight, R. (2021). The
1461 microbiome and human cancer. *Science*, 371(6536), eabc4552.
- 1462 Sharma, S., & Tripathi, P. (2019). Gut microbiome and type 2 diabetes: where we are and where to go?
1463 *The Journal of nutritional biochemistry*, 63, 101–108.
- 1464 Shi, N., Li, N., Duan, X., & Niu, H. (2017). Interaction between the gut microbiome and mucosal
1465 immune system. *Military Medical Research*, 4, 1–7.
- 1466 Simpson, E. (1949). Measurement of diversity. *Nature*, 163.
- 1467 Song, M., Chan, A. T., & Sun, J. (2020). Influence of the gut microbiome, diet, and environment on risk
1468 of colorectal cancer. *Gastroenterology*, 158(2), 322–340.
- 1469 Söreide, K., Janssen, E., Söiland, H., Körner, H., & Baak, J. (2006). Microsatellite instability in colorectal
1470 cancer. *Journal of British Surgery*, 93(4), 395–406.
- 1471 Sotiriadis, A., Papatheodorou, S., Kavvadias, A., & Makrydimas, G. (2010). Transvaginal cervical

- length measurement for prediction of preterm birth in women with threatened preterm labor: a meta-analysis. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 35(1), 54–64.
- Spss, I., et al. (2011). Ibm spss statistics for windows, version 20.0. *New York: IBM Corp*, 440, 394.
- Stafford, G., Roy, S., Honma, K., & Sharma, A. (2012). Sialic acid, periodontal pathogens and tannerella forsythia: stick around and enjoy the feast! *Molecular Oral Microbiology*, 27(1), 11–22.
- Stout, M. J., Conlon, B., Landeau, M., Lee, I., Bower, C., Zhao, Q., ... Mysorekar, I. U. (2013). Identification of intracellular bacteria in the basal plate of the human placenta in term and preterm gestations. *American journal of obstetrics and gynecology*, 208(3), 226–e1.
- Sultan, S., El-Mowafy, M., Elgaml, A., Ahmed, T. A., Hassan, H., & Mottawea, W. (2021). Metabolic influences of gut microbiota dysbiosis on inflammatory bowel disease. *Frontiers in physiology*, 12, 715506.
- Suzuki, N., Nakano, Y., Yoneda, M., Hirofumi, T., & Hanioka, T. (2022). The effects of cigarette smoking on the salivary and tongue microbiome. *Clinical and Experimental Dental Research*, 8(1), 449–456.
- Swidsinski, A., Khilkin, M., Kerjaschki, D., Schreiber, S., Ortner, M., Weber, J., & Lochs, H. (1998). Association between intraepithelial escherichia coli and colorectal cancer. *Gastroenterology*, 115(2), 281–286.
- Swift, D., Cresswell, K., Johnson, R., Stilianoudakis, S., & Wei, X. (2023). A review of normalization and differential abundance methods for microbiome counts data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1), e1586.
- Tanner, A. C., Kent Jr, R., Kanasi, E., Lu, S. C., Paster, B. J., Sonis, S. T., ... Van Dyke, T. E. (2007). Clinical characteristics and microbiota of progressing slight chronic periodontitis in adults. *Journal of clinical periodontology*, 34(11), 917–930.
- Tanner, A. C., Paster, B. J., Lu, S. C., Kanasi, E., Kent Jr, R., Van Dyke, T., & Sonis, S. T. (2006). Subgingival and tongue microbiota during early periodontitis. *Journal of dental research*, 85(4), 318–323.
- Tejeda, M., Farrell, J., Zhu, C., Haines, J. L., Wang, L.-S., Schellenberg, G. D., ... others (2021). Multiple viruses detected in human dna are associated with alzheimer disease risk. *Alzheimer's & Dementia*, 17, e054585.
- Teles, F., Wang, Y., Hajishengallis, G., Hasturk, H., & Marchesan, J. T. (2021). Impact of systemic factors in shaping the periodontal microbiome. *Periodontology 2000*, 85(1), 126–160.
- Thaiss, C. A., Zmora, N., Levy, M., & Elinav, E. (2016). The microbiome and innate immunity. *Nature*, 535(7610), 65–74.
- Tian, R., Liu, H., Feng, S., Wang, H., Wang, Y., Wang, Y., ... Zhang, S. (2021). Gut microbiota dysbiosis in stable coronary artery disease combined with type 2 diabetes mellitus influences cardiovascular prognosis. *Nutrition, Metabolism and Cardiovascular Diseases*, 31(5), 1454–1466.
- Tilg, H., Kaser, A., et al. (2011). Gut microbiome, obesity, and metabolic dysfunction. *The Journal of clinical investigation*, 121(6), 2126–2132.

- 1511 Tonetti, M. S., Greenwell, H., & Kornman, K. S. (2018). Staging and grading of periodontitis: Framework
1512 and proposal of a new classification and case definition. *Journal of periodontology*, 89, S159–S172.
- 1513 Tringe, S. G., & Hugenholtz, P. (2008). A renaissance for the pioneering 16s rRNA gene. *Current opinion*
1514 in *microbiology*, 11(5), 442–446.
- 1515 Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., ... others (2017). A
1516 guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological*
1517 *Reviews*, 92(2), 698–715.
- 1518 Ulger Toprak, N., Yagci, A., Gulluoglu, B., Akin, M., Demirkalem, P., Celenk, T., & Soyletir, G. (2006).
1519 A possible role of *Bacteroides fragilis* enterotoxin in the aetiology of colorectal cancer. *Clinical*
1520 *microbiology and infection*, 12(8), 782–786.
- 1521 Ursell, L. K., Metcalf, J. L., Parfrey, L. W., & Knight, R. (2012). Defining the human microbiome.
1522 *Nutrition reviews*, 70(suppl_1), S38–S44.
- 1523 Utzschneider, K. M., Kratz, M., Damman, C. J., & Hullarg, M. (2016). Mechanisms linking the gut
1524 microbiome and glucose metabolism. *The Journal of Clinical Endocrinology & Metabolism*,
1525 101(4), 1445–1454.
- 1526 Vander Haar, E. L., So, J., Gyamfi-Bannerman, C., & Han, Y. W. (2018). *Fusobacterium nucleatum* and
1527 adverse pregnancy outcomes: epidemiological and mechanistic evidence. *Anaerobe*, 50, 55–59.
- 1528 Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning*
1529 *research*, 9(11).
- 1530 Vasen, H. F., Mecklin, J.-P., Khan, P. M., & Lynch, H. T. (1991). The international collaborative group
1531 on hereditary non-polyposis colorectal cancer (icg-hnppcc). *Diseases of the Colon & Rectum*, 34(5),
1532 424–425.
- 1533 Vilar, E., & Gruber, S. B. (2010). Microsatellite instability in colorectal cancer—the stable evidence.
1534 *Nature reviews Clinical oncology*, 7(3), 153–162.
- 1535 Walker, M. A., Pedamallu, C. S., Ojesina, A. I., Bullman, S., Sharpe, T., Whelan, C. W., & Meyerson, M.
1536 (2018). Gatk pathseq: a customizable computational tool for the discovery and identification of
1537 microbial sequences in libraries from eukaryotic hosts. *Bioinformatics*, 34(24), 4287–4289.
- 1538 Weaver, W. (1963). *The mathematical theory of communication*. University of Illinois Press.
- 1539 Whiteside, S. A., Razvi, H., Dave, S., Reid, G., & Burton, J. P. (2015). The microbiome of the urinary
1540 tract—a role beyond infection. *Nature Reviews Urology*, 12(2), 81–90.
- 1541 Witkin, S. (2019). Vaginal microbiome studies in pregnancy must also analyse host factors. *BJOG: An*
1542 *International Journal of Obstetrics & Gynaecology*, 126(3), 359–359.
- 1543 Wong, T.-T., & Yeh, P.-Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE*
1544 *Transactions on Knowledge and Data Engineering*, 32(8), 1586–1594.
- 1545 Wyss, C., Moter, A., Choi, B.-K., Dewhirst, F., Xue, Y., Schüpbach, P., ... Guggenheim, B. (2004).
1546 *Treponema putidum* sp. nov., a medium-sized proteolytic spirochaete isolated from lesions of
1547 human periodontitis and acute necrotizing ulcerative gingivitis. *International journal of systematic*
1548 and *evolutionary microbiology*, 54(4), 1117–1122.
- 1549 Xia, Y. (2023). Statistical normalization methods in microbiome data with application to microbiome

- 1550 cancer research. *Gut Microbes*, 15(2), 2244139.
- 1551 Yaman, E., & Subasi, A. (2019). Comparison of bagging and boosting ensemble machine learning methods
1552 for automated emg signal classification. *BioMed research international*, 2019(1), 9152506.
- 1553 Yang, I., Claussen, H., Arthur, R. A., Hertzberg, V. S., Geurs, N., Corwin, E. J., & Dunlop, A. L. (2022).
1554 Subgingival microbiome in pregnancy and a potential relationship to early term birth. *Frontiers in*
1555 *cellular and infection microbiology*, 12, 873683.
- 1556 Yoshimura, F., Murakami, Y., Nishikawa, K., Hasegawa, Y., & Kawaminami, S. (2009). Surface
1557 components of porphyromonas gingivalis. *Journal of periodontal research*, 44(1), 1–12.
- 1558 Zhang, C.-Z., Cheng, X.-Q., Li, J.-Y., Zhang, P., Yi, P., Xu, X., & Zhou, X.-D. (2016). Saliva in the
1559 diagnosis of diseases. *International journal of oral science*, 8(3), 133–137.
- 1560 Zhou, X., Wang, L., Xiao, J., Sun, J., Yu, L., Zhang, H., ... others (2022). Alcohol consumption,
1561 dna methylation and colorectal cancer risk: Results from pooled cohort studies and mendelian
1562 randomization analysis. *International journal of cancer*, 151(1), 83–94.
- 1563 Zhu, W., & Lee, S.-W. (2016). Surface interactions between two of the main periodontal pathogens:
1564 Porphyromonas gingivalis and tannerella forsythia. *Journal of periodontal & implant science*,
1565 46(1), 2–9.
- 1566 Zhu, X., Han, Y., Du, J., Liu, R., Jin, K., & Yi, W. (2017). Microbiota-gut-brain axis and the central
1567 nervous system. *Oncotarget*, 8(32), 53829.
- 1568 Zhuang, Y., Wang, H., Jiang, D., Li, Y., Feng, L., Tian, C., ... others (2021). Multi gene mutation
1569 signatures in colorectal cancer patients: predict for the diagnosis, pathological classification, staging
1570 and prognosis. *BMC cancer*, 21, 1–16.

Acknowledgments

1572 I would like to disclose my earnest appreciation for my advisor, Professor **Semin Lee**, who provided
 1573 solicitous supervision and cherished opportunities throughout the course of my research. His advice and
 1574 consultation encouraged me to become as a researcher and to receive all humility and gentleness. I am also
 1575 grateful to all of my committee members, Professor **Taejoon Kwon**, Professor **Eunhee Kim**, Professor
 1576 **Kyemyung Park**, and Professor **Min Hyuk Lim**, for their meaningful mentions and suggestions.

1577 I extend my deepest gratitude to my Lord, *the Flying Spaghetti Monster*, His Noodly Appendage
 1578 has guided me through the twist and turns of this academic journey. His presence, ever comforting and
 1579 mysterious, has been a source of strength and humor during both highs and lows. In moments of doubt, I
 1580 found solace in the belief that you were there, gently reminding me to keep faith in the process. His Holy
 1581 Noodle has nourished my mind, and for that, I am truly overwhelmed. May His Holy Noodle continue to
 1582 guide me in all my future endeavors. *R’Amen.*

1583 I would like to extend my heartfelt gratitude to Professor **You Mi Hong** for her invaluable guidance
 1584 and insightful advice on PTB study. Her expertise in maternal and fetal health, along with her deep under-
 1585 standing of statistical and clinical interpretations, greatly contributed to refining the analytical framework
 1586 of this study. Her constructive feedback and thoughtful discussions provided critical perspectives that
 1587 enhanced the robustness and relevance of the research findings. I sincerely appreciate her generosity
 1588 in sharing her knowledge and effort, as well as her encouragement throughout my Ph.D. journey. Her
 1589 support has been instrumental in strengthening this work, and I am truly grateful for her contributions.

1590 I also would like to express my sincere gratitude for Professor **Jun Hyeok Lim** for his invaluable
 1591 guidance and insightful advice on lung cancer study. His expertise in cancer genomics and data interpreta-
 1592 tion provided essential perspectives that greatly enriched the analytical approach of my Ph.D. journey. His
 1593 constructive feedback and thoughtful discussion helped refine methodologies and enhance the scientific
 1594 rigor of the research. I deeply appreciate his willingness to share his knowledge and expertise, which has
 1595 been instrumental in shaping key aspects of this work. His support and encouragement have been truly
 1596 inspiring, and I am grateful for the opportunity to have benefited from his mentorship.

1597 I would like to extend my heartfelt gratitude to my colleagues of the **Computational Biology Lab @**
 1598 **UNIST**, whose collaboration, friendship, brotherhood, and support have been an invaluable part of my
 1599 journey. Your willingness to share insights, engage in thoughtful discussions, and offer encouragement
 1600 during the challenging moments of research has significantly shaped my academic experience. The
 1601 camaraderie in Computational Biology Lab made even the most demanding days more enjoyable, and I
 1602 am deeply grateful for the collaborative environment we created together. I appreciate you for standing
 1603 by my side throughout this Ph.D. journey.

1604 I would like to express my heartfelt gratitude to **my family**, whose unwavering support has been the
 1605 foundation of everything I have achieved. Your love, encouragement, and belief in me have sustained me
 1606 through every challenge, and I could not have come this far without you. From your words of wisdom to
 1607 your patience and understanding, each of you has played a vital role in helping me navigate this journey.
 1608 The strength and comfort I have drawn from our family bond have been my greatest source of resilience.

1609 Your presence, both near and far, has filled my life with warmth and motivation. I am deeply grateful for
1610 your unconditional love and for always being there when I needed you the most. Thank you for being my
1611 constant source of strength and inspiration.

1612 I am incredibly pleased to my friends, especially my GSHS alumni (**이망특**), for their unwavering
1613 support and encouragement throughout this journey. The bonds we formed back in our school days have
1614 only grown stronger over the years, and I am fortunate to have had such loyal and understanding friends
1615 by my side. Your constant words of motivation, and even moments of levity during stressful times have
1616 helped keep me grounded. Whether it was a late-night conversations, a shared laugh, or a simple message
1617 of reassurance, you all have played a vital role in keeping me focused and motivated. I am relieved for the
1618 ways you celebrated each small achievement with me and how you patiently listened to my worries. The
1619 memories of our shared past provided me with comfort and a sense of stability when the road ahead felt
1620 uncertain. I could not have reached this point without the love and friendship that you all have generously
1621 given. Each of your, in your unique way, has contributed to this dissertation, even if indirectly, and for
1622 that, I am forever beholden. I look forward to continuing our friendship as we all grow in our individual
1623 paths, knowing that the support we share is something truly special.

1624 I would like to express my deepest recognition to **my girlfriend (expected)** for her unwavering
1625 support, patience, and companionship throughout my Ph.D. journey. Her presence has been a constant
1626 source of comfort and motivation, helping me navigate the challenges of research and writing with
1627 renewed energy. Through moments of frustration and accomplishment alike, her encouragement has
1628 reminded me of the importance of balance and perseverance. Her kindness, understanding, and belief
1629 in me have been invaluable, making even the most difficult days feel lighter. I am truly grateful for her
1630 support and for sharing this journey with me, and I look forward to all the moments we will continue to
1631 experience together.

1632 I would like to express my sincere gratitude to the amazing members of my animal protection groups,
1633 DRDR (**두루두루**) and UNIMALS (**유니멀스**), whose dedication and compassion have been a constant
1634 source of motivation. Your unwavering commitment to improving the lives of animals has inspired me
1635 throughout this journey. I am also thankful for the beautiful cats we have cared for, whose presence
1636 brought both joy and purpose to our allegiance. Their playful spirits and gentle companionship served as
1637 daily reminders of why we continue to fight for animal rights. The bond we share, both with each other
1638 and with the animals we protect, has enriched my life in countless ways. I appreciate you all again for
1639 your support, dedication, and for being part of this meaningful cause.

1640 I would like to express my deepest gratitude to **everyone** I have had the honor of meeting throughout
1641 this journey. Your kindness, encouragement, and support have carried me through both the challenging
1642 and rewarding moments of my life. Whether through a kind word, thoughtful advice, or simply being
1643 there when I needed it most, your presence has made all the difference. I am incredibly fortunate to have
1644 received such generosity and warmth from those around me, and I do not take it for granted. Every act
1645 of kindness, no matter how big or small, has been a source of strength and motivation for me. To all
1646 my friends, colleagues, mentors, and beloved ones, thank you for your unwavering support. I am truly
1647 grateful for each of you, and your kindness has left an indelible mark on my journey.

1648 My Lord, *the Flying Spaghetti Monster*,
1649 give us grace to accept with serenity the things that cannot be changed,
1650 courage to change the things that should be changed,
1651 and the wisdom to distinguish the one from the other.

1652
1653 Glory be to *the Meatball*, to *the Sauce*, and to *the Holy Noodle*.
1654 As it was in the beginning, is now, and ever shall be.

1655 *R'Amen.*



May your progress be evident to all

