

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	1
27	2	Predicting preterm birth using random forest classifier in salivary microbiome	7
28	2.1	Introduction	7
29	2.2	Materials and methods	9
30	2.2.1	Study design and study participants	9
31	2.2.2	Clinical data collection and grouping	9
32	2.2.3	Salivary microbiome sample collection	9
33	2.2.4	16s rRNA gene sequencing	9
34	2.2.5	Bioinformatics analysis	10
35	2.2.6	Data and code availability	10
36	2.3	Results	11
37	2.3.1	Overview of clinical information	11
38	2.3.2	Comparison of salivary microbiomes composition	11
39	2.3.3	Random forest classification to predict PTB risk	11
40	2.4	Discussion	19
41	3	Random forest prediction model for periodontitis statuses based on the salivary microbiomes	21
42	3.1	Introduction	21
43	3.2	Materials and methods	23
44	3.2.1	Study participants enrollment	23
45	3.2.2	Periodontal clinical parameter diagnosis	23
46	3.2.3	Saliva sampling and DNA extraction procedure	25
47	3.2.4	Bioinformatics analysis	25
48	3.2.5	Data and code availability	26
49	3.3	Results	28

50	3.3.1	Summary of clinical information and sequencing data	28
51	3.3.2	Diversity indices reveal differences among the periodontitis severities .	28
52	3.3.3	DAT among multiple periodontitis severities and their correlation . . .	28
53	3.3.4	Classification of periodontitis severities by random forest models . . .	29
54	3.4	Discussion	50
55	4	Metagenomic signature analysis of Korean colorectal cancer	54
56	4.1	Introduction	54
57	4.2	Materials and methods	56
58	4.2.1	Study participants enrollment	56
59	4.2.2	DNA extraction procedure	56
60	4.2.3	Bioinformatics analysis	56
61	4.2.4	Data and code availability	57
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	ML prediction	59
68	4.4	Discussion	72
69	5	Conclusion	73
70	References		74
71	Acknowledgments		90

72

List of Figures

73	1	DAT volcano plot	13
74	2	Salivary microbiome compositions over DAT	14
75	3	Random forest-based PTB prediction model	15
76	4	Diversity indices	16
77	5	PROM-related DAT	17
78	6	Validation of random forest-based PTB prediction model	18
79	7	Diversity indices	36
80	8	Differentially abundant taxa (DAT)	37
81	9	Correlation heatmap	38
82	10	Random forest classification metrics	39
83	11	Random forest classification metrics from external datasets	40
84	12	Rarefaction curves for alpha-diversity indices	41
85	13	Salivary microbiome compositions in the different periodontal statuses	42
86	14	Correlation plots for differentially abundant taxa	43
87	15	Clinical measurements by the periodontitis statuses	44
88	16	Number of read counts by the periodontitis statuses	45
89	17	Proportion of DAT	46

90	18	Random forest classification metrics with the full microbiome compositions and ANCOM-selected DAT compositions	47
91			
92	19	Alpha-diversity indices account for evenness	48
93	20	Gradient Boosting classification metrics	49
94	21	Gut microbiome compositions in genus level	61
95	22	Alpha-diversity indices in genus level	62
96	23	Alpha-diversity indices with recurrence in genus level	63
97	24	Alpha-diversity indices with OS in genus level	64
98	25	Beta-diversity indices in genus level	65
99	26	Beta-diversity indices with recurrence in genus level	66
100	27	Beta-diversity indices with recurrence in genus level	67
101	28	DAT with recurrence in genus level	68
102	29	DAT with OS in genus level	69
103	30	ML classification.	70
104	31	ML regression.	71

List of Tables

106	1	Confusion matrix	5
107	2	Standard clinical information of study participants	12
108	3	Clinical characteristics of the study participants	31
109	4	Feature combinations and their evaluations	32
110	5	List of DAT among the periodontally healthy and periodontitis stages	33
111	6	Feature the importance of taxa in the classification of different periodontal statuses.	34
112	7	Beta-diversity pairwise comparisons on the periodontitis statuses	35
113	8	Clinical characteristics of CRC study participants	60

List of Abbreviations

- ¹¹⁵ **ACC** Accuracy
- ¹¹⁶ **ACE** Abundance-based coverage estimator
- ¹¹⁷ **ASV** Amplicon sequence variant
- ¹¹⁸ **AUC** Area-under-curve
- ¹¹⁹ **BA** Balanced accuracy
- ¹²⁰ **C-section** Cesarean section
- ¹²¹ **DAT** Differentially abundant taxa
- ¹²² **F1** F1 score
- ¹²³ **Faith PD** Faith's phylogenetic diversity
- ¹²⁴ **FTB** Full-term birth
- ¹²⁵ **GA** Gestational age
- ¹²⁶ **MSI** Microsatellite instability
- ¹²⁷ **MSS** Microsatellite stable
- ¹²⁸ **MWU test** Mann-Whitney U-test
- ¹²⁹ **OS** Overall survival
- ¹³⁰ **PRE** Precision
- ¹³¹ **PROM** Prelabor rupture of membrane
- ¹³² **PTB** Preterm birth
- ¹³³ **ROC curve** Receiver-operating characteristics curve
- ¹³⁴ **rRNA** Ribosomal RNA
- ¹³⁵ **SD** Standard deviation
- ¹³⁶ **SEN** Sensitivity
- ¹³⁷ **SPE** Specificity
- ¹³⁸ **t-SNE** t-distributed stochastic neighbor embedding

139 **1 Introduction**

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

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

175 16S ribosomal RNA (rRNA) gene sequencing is one of the most extensively applied methods for
176 characterizing microbial communities by targeting the conserved 16S rRNA gene, which contains both

177 highly conserved and variable regions in bacteria (Tringe & Hugenholtz, 2008; Janda & Abbott, 2007).
178 The conserved regions enable universal primer binding, while the variable regions provide the specificity
179 needed to differentiate microbial taxa. Among these regions, the V3-V4 region is frequently selected for
180 sequencing due to its balance between phylogenetic resolution and sequencing efficiency (Johnson et al.,
181 2019; López-Aladid et al., 2023). Therefore, the V3-V4 region offers sufficient variability to classify a
182 wide range of bacteria taxa while maintaining compatibility with widely used sequencing platforms.

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

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

206 Differentially abundant taxa (DAT) detection is a key analytical approach in microbiome study to
207 identify microbial taxa that significantly differ in abundance between distinct study participant groups.
208 This DAT detection method is particularly valuable for understanding how microbial communities vary
209 across different conditions, such as disease states, environmental factors, and/or experimental treatments.
210 Various statistical and computational techniques, *e.g.* LEfSe (Segata et al., 2011), DESeq2 (Love, Huber,
211 & Anders, 2014), ANCOM (Lin & Peddada, 2020), and ANCOM-BC (Lin, Eggesbø, & Peddada,
212 2022; Lin & Peddada, 2024), are commonly used to assess differential abundance while accounting for
213 compositional and sparsity-related challenges in microbiome composition data (Swift, Cresswell, Johnson,
214 Stilianoudakis, & Wei, 2023; Cappellato, Baruzzo, & Di Camillo, 2022). Thus, identifying DAT can
215 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

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

261 Regression is a powerful predictive machine learning approach used to analyze complex relationships
262 between variables and make continuous value predictions (Maulud & Abdulazeez, 2020; Yildiz, Bilbao, &
263 Sproul, 2017). Beside classification, which assigns discrete labels, regression models estimate numerical
264 outcomes based on input features, making them particularly useful in biological research and clinical
265 applications for predicting disease risk, patient outcomes, and biomarker selection. By leveraging high-
266 throughput biological techniques and clinical information, regression model enables the discovery of
267 hidden patterns and the development of precision medicine strategies. As computational methods advance,
268 integrating regression models with metagenomic data can improve predictive accuracy and facilitate
269 data-driven therapeutic guide in healthcare.

270 This dissertation present a comprehensive, multi-disease human microbiome analysis, bridging the
271 association between preterm birth (PTB) (Section 2), periodontitis (Section 3), and colorectal cancer
272 (CRC) (Section 4) through a unified metagenomic approach. While previous studies have examined the
273 role and characteristics of human microbiome in these diseases individually, this dissertation uniquely
274 integrates human microbiome-driven insights across these diseases to identify shared and disease-specific
275 microbial signatures. By applying high-throughput metagenomic sequencing, microbial diversity analysis,
276 and advanced bioinformatics techniques, this dissertation aims to uncover novel microbiome-based
277 biomarkers and mechanistic insights into how microbial communities influence these conditions. These
278 findings contribute to a broader understanding of microbiome-mediated disease interactions and pave the
279 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)

280

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

281

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

282

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

283

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

284

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

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

285 **2 Predicting preterm birth using random forest classifier in salivary mi-
286 crobiome**

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

288 Hong, Y. M., **Lee, Jaewoong**, Cho, D. H., Jeon, J. H., Kang, J., Kim, M. G., ... & Kim, J. K. (2023).
289 Predicting preterm birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1),
290 21105.

291 **2.1 Introduction**

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

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

309 Preterm prelabor rupture of membranes (PROM) brought on by gestational inflammation and infection
310 contribute to about 70% of PTB cases (Romero, Dey, & Fisher, 2014). Nevertheless, as antibiotics and
311 anti-inflammatory therapeutic strategies were ineffective to decrease PTB occurrence rates, the pathology
312 of PTB has not been entirely elucidated by inflammatory and infectious pathways (Romero, Hassan, et al.,
313 2014). Recent researches on maternal microbiomes were beginning to examine unidentified connections
314 of PTB as a consequence of developmental processes in molecular biological technology (Fettweis et al.,
315 2019).

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

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

324 Multiple unfavorable birth outcomes, including PROM and PTB, have been linked to periodontitis
325 as an independence risk factor, according to numerous epidemiological researches (Offenbacher et al.,
326 1996). It is expected that the oral microbiome will be able to explain additional hematogenous pathways
327 in light of these precedents; however, the oral microbiome composition of fetuses is limited understood.

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

331 **2.2 Materials and methods**

332 **2.2.1 Study design and study participants**

333 Between 2019 and 2021, singleton pregnant women who received treatment to Jeonbuk National University Hospital for childbirth were the participants of this study. This study was conducted according to the
334 Declaration of Helsinki (Goodyear, Krleza-Jeric, & Lemmens, 2007). The Institutional Review Board
335 authorized this study (IRB file No. 2019-01-024). Participants who were admitted for elective cesarean
336 sections (C-sections) or induction births, as well as those who had written informed consent obtained
337 with premature labor or PROM, were eligible.
338

339 **2.2.2 Clinical data collection and grouping**

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

- 342 • maternal age at delivery
- 343 • diabetes mellitus
- 344 • hypertension
- 345 • overweight and obesity
- 346 • C-section
- 347 • history PROM or PTB
- 348 • gestational week on delivery
- 349 • birth weight
- 350 • sex

351 **2.2.3 Salivary microbiome sample collection**

352 Salivary microbiome samples were collected 24 hours before to delivery using mouthwash. The standard
353 methods of sterilizing were performed. Medical experts oversaw each stage of the sample collecting
354 procedure. Participants received instruction not to eat, drink, or brush their teeth for 30 minutes before
355 sampling salivary microbiome. Saliva samples were gathered by washing the mouth for 30 seconds with
356 12 mL of a mouthwash solution (E-zен Gargle, JN Pharm, Pyeongtaek, Gyeonggi, Korea). The samples
357 were tagged with the anonymous ID for each participant and kept at 4 °C until they underwent further
358 processing. Genomic DNA was extracted using an ExgeneTM Clinic SV kit (GeneAll Biotechnology,
359 Seoul, Korea) following with the manufacturer instructions and store at -20 °C.

360 **2.2.4 16s rRNA gene sequencing**

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

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

367 **2.2.5 Bioinformatics analysis**

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

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

379 Taxonomic assignment were implemented with HOMD (version 15.22) (T. Chen et al., 2010).
380 Afterward, DESeq2 was implemented to identify differentially abundant taxa (DAT) that could dis-
381 tinguish between salivary microbiome from PTB and FTB participants (Love et al., 2014). Taxa with
382 $|\log_2 \text{FoldChange}| > 1$ and $p < 0.05$ were considered as statistically significant.

383 The taxa for predicting PTB using salivary microbiome data were determined using a random forest
384 classifier (Breiman, 2001). Through stratified *k*-fold cross-validation (*k* = 5) that preserves the existence
385 rate of PTB and FTB participants, consistency and trustworthy classification were ensured (Wong & Yeh,
386 2019).

387 **2.2.6 Data and code availability**

388 All sequences from the 59 study participants have been published to the Sequence Read Archives
389 (project ID PRJNA985119): <https://dataview.ncbi.nlm.nih.gov/object/PRJNA985119>. Docker
390 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.

393 **2.3 Results**

394 **2.3.1 Overview of clinical information**

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

401 **2.3.2 Comparison of salivary microbiomes composition**

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

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

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

413 **2.3.3 Random forest classification to predict PTB risk**

414 To classify PTB according to DAT, random forest classifiers were constructed. The nine most significant
415 DAT were used to obtain the best BA (0.765 ± 0.071 ; Figure 3a). Moreover, random forest classification
416 model determined each DAT's importance (Figure 3b). We conducted a validation procedure on nine
417 twin pregnancies that were excluded in the initial study design in order to confirm the reliability and
418 dependability of our random forest-based PTB prediction model (Figure 6). Comparable to the PTB
419 prediction model on the 59 initial singleton study participants, the validation classification on PTB risk of
420 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

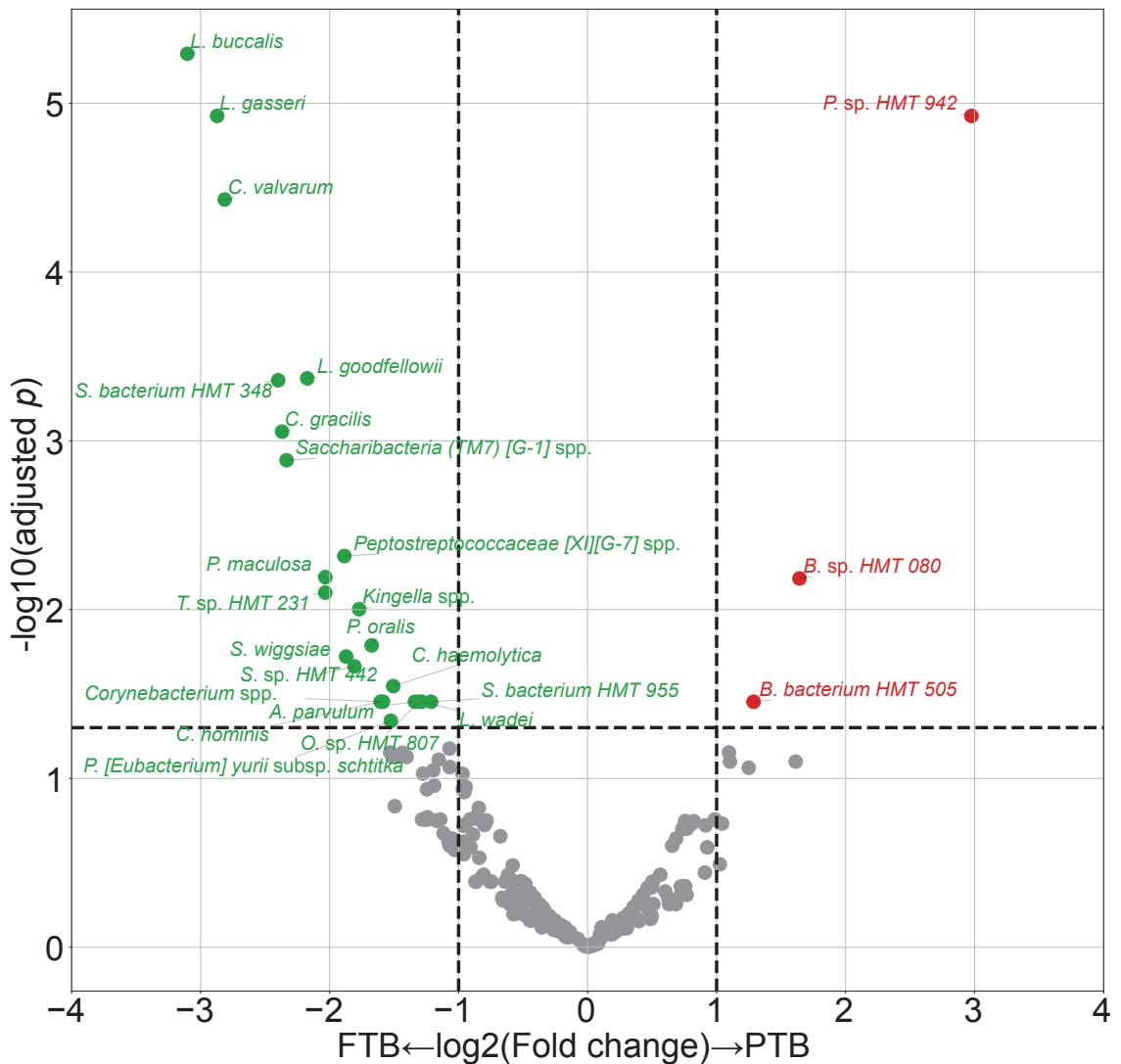


Figure 1: DAT volcano plot.

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

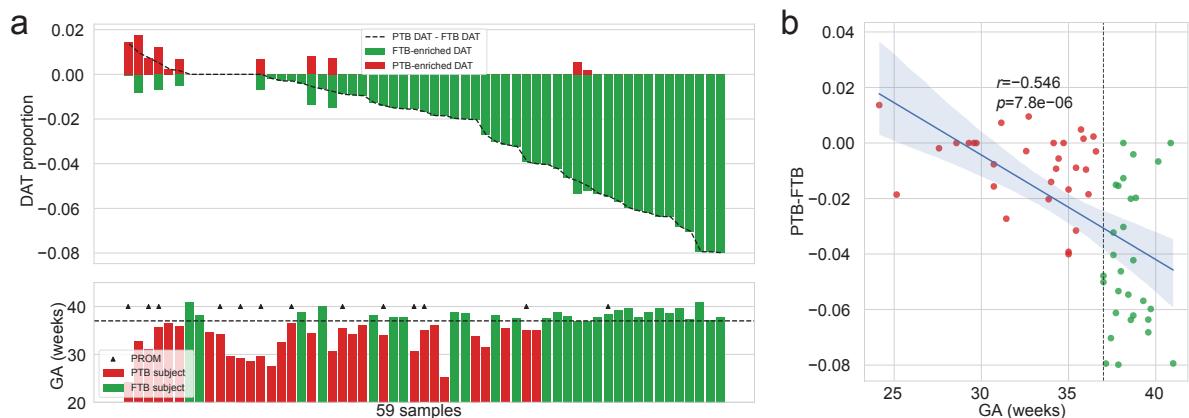


Figure 2: **Salivary microbiome compositions over DAT.**

(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.

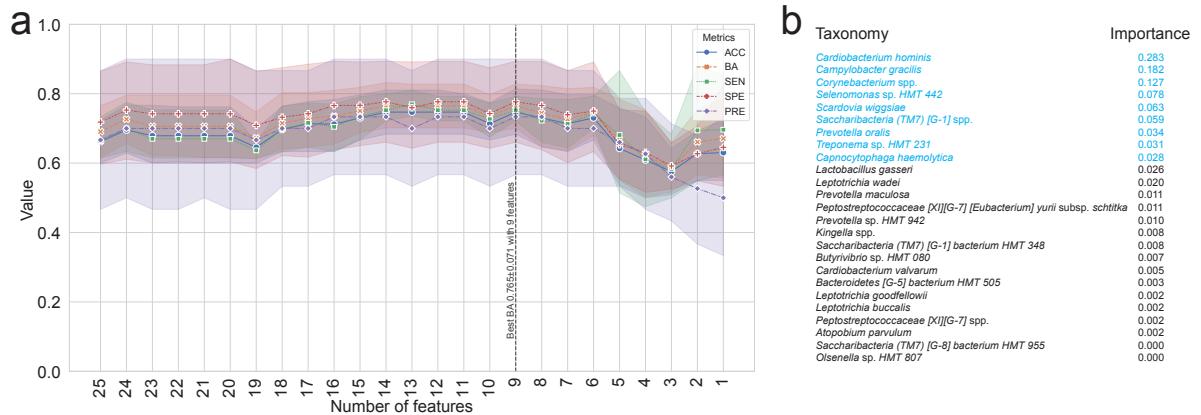


Figure 3: **Random forest-based PTB prediction model.**

(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.

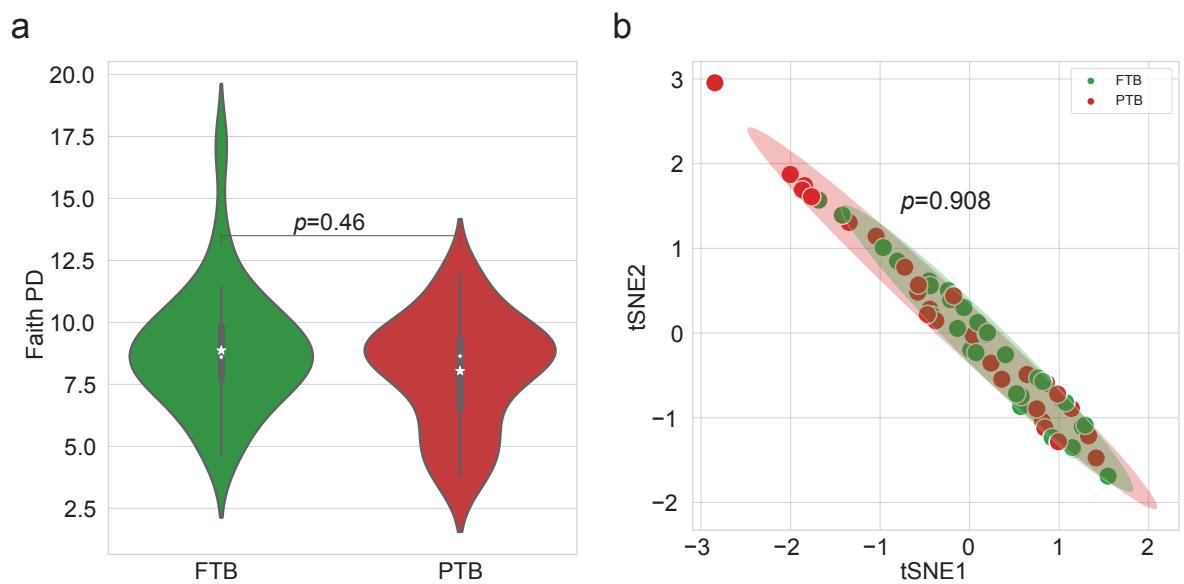


Figure 4: **Diversity indices.**

(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 $p = 0.908$)

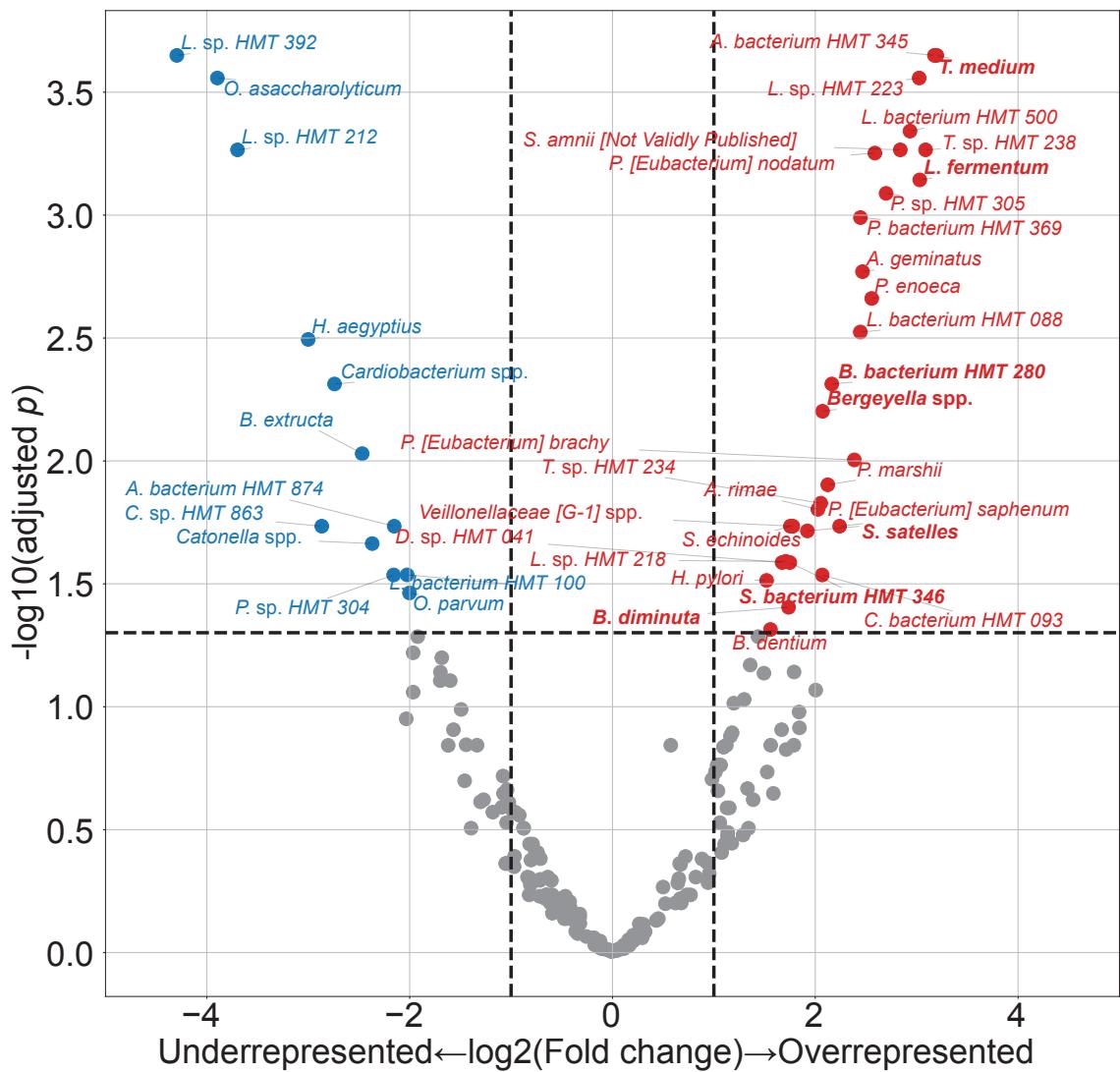


Figure 5: **PROM-related DAT**.

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.

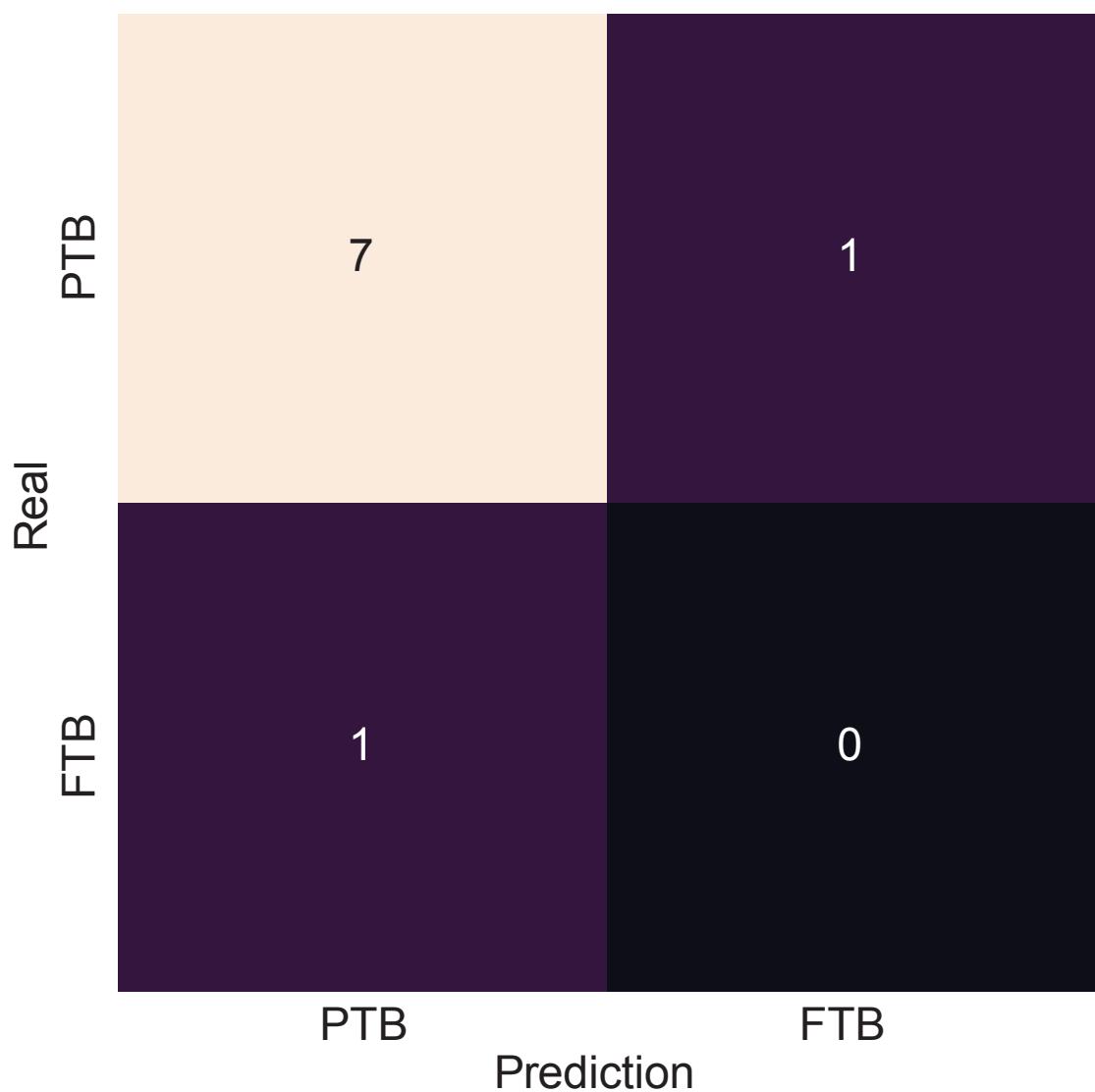


Figure 6: Validation of random forest-based PTB prediction model.

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)

421 **2.4 Discussion**

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

429 One of the most common salivary taxa that has been examined is *Fusobacterium nucleatum* (Han,
430 2015; Brennan & Garrett, 2019; Bolstad, Jensen, & Bakken, 1996), that is a Gram-negative, anaerobic, and
431 filamentous bacteria. *Fusobacterium nucleatum* can be separated from not only the salivary microbiome
432 but also the vaginal microbiome (Vander Haar, So, Gyamfi-Bannerman, & Han, 2018; Witkin, 2019). In
433 both animal and human investigation, *Fusobacterium nucleatum* infection has been linked to risk of PTB
434 (Doyle et al., 2014). According to recent researches, the placenta women who give birth prematurely may
435 include additional salivary microbiome dysbiosis, such as *Bergeyella* spp. and *Porphyromonas gingivalis*
436 (León et al., 2007; Katz, Chegini, Shiverick, & Lamont, 2009). Although *Bergeyella* spp. were one of the
437 PROM-overrepresented DAT (Figure 5), it was excluded in the final 25 PTB-related DAT. Furthermore,
438 *Porphyromonas gingivalis* and *Campylobacter gracilis* were pathogens of periodontitis in sub-gingival
439 microbiome (Yang et al., 2022). *Lactobacillus gasseri* was also one of the FTB-enriched DAT (Figure
440 1), and it is well established that early PTB risk can be reduced by *Lactobacillus gasseri* in the vaginal
441 microbiome (Basavaprabhu, Sonu, & Prabha, 2020; Payne et al., 2021).

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

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

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

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

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

473 **3 Random forest prediction model for periodontitis statuses based on the**
474 **salivary microbiomes**

475 **3.1 Introduction**

476 Saliva microbial dysbiosis brought on by the accumulation of plaque results in periodontitis, a chronic
477 inflammatory disease of the tissue that surrounds the tooth (Kinane, Stathopoulou, & Papapanou, 2017).
478 Loss of periodontal attachment is a consequence of periodontitis, which may lead to irreversible bone loss
479 and, eventually, permanent tooth loss if left untreated. A new classification criterion of periodontal diseases
480 was created in 2018, about 20 years after the 1999 statements of the previous one (Papapanou et al.,
481 2018). Even with this evolution, radiographic and clinical markers of periodontitis progression remain the
482 primary methods for diagnosing periodontitis (Papapanou et al., 2018). Such tools, nevertheless, frequently
483 demonstrate the prior damage from periodontitis rather than its present condition. Certain individuals have
484 a higher risk of periodontitis, a higher chance of developing severe generalized periodontitis, and a worse
485 response to common salivary bacteria control techniques utilized to prevent and treat periodontitis. As a
486 result, the 2017 framework for diagnosing periodontitis additionally allows for the potential development
487 of biomarkers to enhance diagnosis and treatment of periodontitis (Tonetti, Greenwell, & Kornman, 2018).
488 Instead of only depending on the progression of periodontitis, a new etiological indication based on the
489 current state must be introduced in order to enable appropriate intervention through early detection of
490 periodontitis. Thus, the current clinical diagnostic techniques that rely on periodontal probing can be
491 uncomfortable for patients with periodontitis (Canakci & Canakci, 2007).

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

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

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

524 Recently, we employed multiplex quantitative-PCR and machine learning-based classification model
525 to predict the severity of periodontitis based on the amount of nine pathogens of periodontitis from
526 saliva collections (E.-H. Kim et al., 2020). On the other hand, the fact that we focused merely at nine
527 pathogens for periodontitis and neglected the variety bacterial species associated to the various severities
528 of periodontitis constrained the breadth of our investigation. By developing a machine learning model
529 that could classify multiple severities of periodontitis based on the salivary microbiome composition,
530 this study aims to fill these knowledge gaps and produce more accurate and therapeutically useful
531 guidance to evaluate progression of periodontitis. Hence, in order to examine the salivary microbiome
532 composition of both healthy controls and patients with periodontitis in multiple stages, we applied
533 16S rRNA gene sequencing. Furthermore, employing the 2018 classification criteria, we sought to find
534 biomarkers (species) for the precise prediction of periodontitis severities (Papapanou et al., 2018; Chapple
535 et al., 2018).

536 **3.2 Materials and methods**

537 **3.2.1 Study participants enrollment**

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

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

555 **3.2.2 Periodontal clinical parameter diagnosis**

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

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

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

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

573 0. Normal gingiva: without inflammation nor discoloration.

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

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

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

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

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

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

584 • Healthy:

585 1. Bleeding sites < 10%

586 2. Probing depth: \leq 3 mm

587 • Stage I:

588 1. No tooth loss because of periodontitis.

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

590 3. Radiographic bone loss: < 15%

591 • Stage II:

592 1. No tooth loss because of periodontitis.

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

594 3. Radiographic bone loss: 15-33%

595 • Stage III:

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

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

598 3. Radiographic bone loss: > 33%

599 **3.2.3 Saliva sampling and DNA extraction procedure**

600 All study participants received instructions to avoid eating, drinking, brushing, and using mouthwash for
601 at least an hour prior to the saliva sample collection process. These collections were conducted between
602 09:00 and 11:00. Mouth rinse was collected by rinsing the mouth for 30 seconds with 12 mL of a solution
603 (E-zen Gargle, JN Pharm, Korea). All saliva samples were tagged with anonymous ID and stored at -4 °C.

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

- 608 • Forward: 5' -TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNNGCWGCAG-3'
609 • Reverse: 5' -GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'

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

- 612 1. Heat activation for 30 seconds at 95 °C.
613 2. 25 cycles for 30 seconds at 95 °C.
614 3. 30 seconds at 55 °C.
615 4. 30 seconds at 72 °C.

616 NexteraXT Indexed Primer was applied to amplification 10 µL of the purified initial PCR products for
617 the final library creation. The second PCR used the same conditions as the first PCR conditions but with
618 10 cycles. 16S rRNA gene sequencing was performed via 2×300 bp paired-end sequencing at Macrogen
619 Inc. (Macrogen, Seoul, Korea) using Illumina MiSeq platform (Illumina, San Diego, CA, USA).

620 **3.2.4 Bioinformatics analysis**

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

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

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

632 Aitchison index for a beta-diversity index was calculated using QIIME2 (version 2020.8) (Aitchison,
633 Barceló-Vidal, Martín-Fernández, & Pawlowsky-Glahn, 2000; Bolyen et al., 2019). We employed the
634 t-SNE algorithm to illustrate multi-dimensional data from the beta-diversity index computation (Van der
635 Maaten & Hinton, 2008). The beta-diversity index was compared using the PERMANOVA test (Anderson,
636 2014; Kelly et al., 2015) and MWU test.

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

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

649 As previously stated (E.-H. Kim et al., 2020), we used stratified k -fold cross-validation ($k = 10$)
650 by severity of periodontitis to achieve consistent and trustworthy classification results (Wong & Yeh,
651 2019). Additionally, we utilized various features with confusion matrices and their derivations to evaluate
652 the classification outcomes in order to identify which features optimize classification evaluations and
653 decrease sequencing efforts. Using the DAT discovered by ANCOM, we iteratively removed the least
654 significant taxa from the input features (taxa) of the random forest (Breiman, 2001) and gradient boosting
655 (Friedman, 2002) classification models using the backward elimination method. Random forest classifier
656 builds multiple decision trees independently using bootstrapped samples and aggregates their predictions,
657 enhancing stability and reducing overfitting problems. In contrast, Gradient boosting constructs trees
658 sequentially, where each new tree improves the errors of the previous ones using gradient descent, leading
659 to higher classification evaluations.

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

664 3.2.5 Data and code availability

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

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

672 **3.3 Results**

673 **3.3.1 Summary of clinical information and sequencing data**

674 Among clinical information of the study participants, clinical attachment level, probing depth, plaque
675 index, and gingival index, were significantly increased with periodontitis severity (Kruskal-Wallis test
676 $p < 0.001$), while sex were observed no significant difference (Table 2). Notably, clinical attachment level
677 and probing depth have significant differences among the periodontitis severities (MWU test $p < 0.01$;
678 Figure 15). Additionally, 71461.00 ± 11792.30 and 45909.78 ± 11404.65 reads per sample were obtained
679 before and after filtering low-quality reads and trimming extra-long tails, respectively (Figure 16). In 250
680 study subjects, we have found a total of 425 bacterial taxa (Figure 13).

681 **3.3.2 Diversity indices reveal differences among the periodontitis severities**

682 Rarefaction curves showed that the sequencing depth was sufficient (Figure 12). Alpha-diversity in-
683 dices indicated significant differences between the healthy and the periodontitis stages (MWU test
684 $p < 0.01$; Figure 7a-e); however, there were no significant differences between the periodontitis stages.
685 This emphasizes how essential it is to classify the salivary microbiome compositions and distinguish
686 between the stages of periodontitis using machine learning approaches.

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

692 **3.3.3 DAT among multiple periodontitis severities and their correlation**

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

- 696 • Group 1
 - 697 1. *Treponema* spp.
 - 698 2. *Prevotella* sp. HMT 304
 - 699 3. *Prevotella* sp. HMT 526
 - 700 4. *Peptostreptococcaceae [XI][G-5] saphenum*
 - 701 5. *Treponema* sp. HMT 260
 - 702 6. *Mycoplasma faecium*
 - 703 7. *Peptostreptococcaceae [XI][G-9] brachy*
 - 704 8. *Lachnospiraceae [G-8] bacterium* HMT 500
 - 705 9. *Peptostreptococcaceae [XI][G-6] nodatum*
 - 706 10. *Fretibacterium* spp.

- 707 • Group 2
- 708 1. *Porphyromonas gingivalis*
- 709 2. *Campylobacter showae*
- 710 3. *Filifactor alocis*
- 711 4. *Treponema putidum*
- 712 5. *Tannerella forsythia*
- 713 6. *Prevotella intermedia*
- 714 7. *Porphyromonas* sp. HMT 285

- 715 • Group 3
- 716 1. *Actinomyces* spp.
- 717 2. *Corynebacterium durum*
- 718 3. *Actinomyces graevenitzii*

719 Ten DAT that were significant enriched in stage II and stage III, but deficient in healthy formed Group
 720 1 (Figure 8). Furthermore, in comparison to the healthy, the seven DAT of Group 2 were significantly
 721 enriched in each of the stages of periodontitis. On the other hand, three DAT in Group 3 were deficient in
 722 stage II and stage III, but significantly enriched in healthy. The relative proportions of the DAT further
 723 supported these findings (Figure 8b), suggesting that the DAT is primarily linked to periodontitis rather
 724 than other salivary bacteria.

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

727 3.3.4 Classification of periodontitis severities by random forest models

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

732 Based on the proportion of DAT, random forest classifier were trained to classify the periodontitis
 733 severities (Table 6). We conducted multi-label classification for the multiple periodontitis severities,
 734 namely healthy, stage I, stage II, and stage III. In this setting, we classified multiple periodontitis
 735 severities with the highest BA of 0.779 ± 0.029 (Table 4). AUC ranged between 0.81 and 0.94 (Figure
 736 10b).

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

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

743 perform a multi-label classification.

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

750 Finally, to confirm the reliability and consistency of our random forest classifier, we validated our
751 classification model using openly accessible 16S rRNA gene sequencing from Spanish participants
752 (Iniesta et al., 2023) and Portuguese participants (Relvas et al., 2021) (Figure 11). Although some
753 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 III/IV		
	Rank	Taxa	Importance	Taxa	Importance	Taxa	Importance	Taxa	Importance	Taxa	Importance	
1	<i>Porphyromonas gingivalis</i>	0.297	<i>Actinomyces spp.</i>	0.195	<i>Porphyromonas gingivalis</i>	0.360	<i>Porphyromonas gingivalis</i>	0.426	<i>Porphyromonas gingivalis</i>	0.461		
2	<i>Actinomyces spp.</i>	0.195	<i>Actinomyces graevenitzii</i>	0.054	<i>Actinomyces spp.</i>	0.125	<i>Actinomyces spp.</i>	0.244	<i>Actinomyces spp.</i>	0.257		
3	<i>Prevotella intermedia</i>	0.054	<i>Actinomyces graevenitzii</i>	0.052	<i>Porphyromonas sp. HMT 285</i>	0.095	<i>Actinomyces graevenitzii</i>	0.049	<i>Actinomyces graevenitzii</i>	0.059		
4	<i>Actinomyces graevenitzii</i>	0.052	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.050	<i>Porphyromonas sp. HMT 285</i>	0.062	<i>Corynebacterium durum</i>	0.046	<i>Corynebacterium durum</i>	0.035		
5	<i>Filifactor alocis</i>	0.050	<i>Campylobacter showae</i>	0.042	<i>Campylobacter showae</i>	0.052	<i>Filifactor alocis</i>	0.036	<i>Filifactor alocis</i>	0.032		
6	<i>Campylobacter showae</i>	0.042	<i>Porphyromonas sp. HMT 285</i>	0.040	<i>Corynebacterium durum</i>	0.050	<i>Prevotella intermedia</i>	0.033	<i>Campylobacter showae</i>	0.023		
7	<i>Porphyromonas sp. HMT 285</i>	0.040	<i>Treponema spp.</i>	0.032	<i>Treponema spp.</i>	0.038	<i>Tannerella forsythia</i>	0.025	<i>Porphyromonas sp. HMT 285</i>	0.022		
8	<i>Corynebacterium durum</i>	0.032	<i>Tannerella forsythia</i>	0.026	<i>Tannerella forsythia</i>	0.037	<i>Prevotella intermedia</i>	0.023	<i>Prevotella intermedia</i>	0.022		
9	<i>Treponema spp.</i>	0.032	<i>Prevotella intermedia</i>	0.025	<i>Prevotella intermedia</i>	0.029	<i>Treponema spp.</i>	0.021	<i>Treponema spp.</i>	0.022		
10	<i>Tannerella forsythia</i>	0.026	<i>Prevotella intermedia</i>	0.025	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.026	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.018	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.015		
11	<i>Treponema putidum</i>	0.025	<i>Freibacterium spp.</i>	0.023	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.018	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.014	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.010		
12	<i>Freibacterium spp.</i>	0.023	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.021	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.018	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.011	<i>Tannerella forsythia</i>	0.009		
13	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.021	<i>Treponema putidum</i>	0.019	<i>Treponema putidum</i>	0.014	<i>Treponema putidum</i>	0.010	<i>Freibacterium spp.</i>	0.009		
14	<i>Treponema sp. HMT 260</i>	0.019	<i>Prevotella sp. HMT 526</i>	0.018	<i>Prevotella sp. HMT 526</i>	0.011	<i>Prevotella sp. HMT 526</i>	0.009	<i>Prevotella sp. HMT 526</i>	0.006		
15	<i>Prevotella sp. HMT 526</i>	0.018	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.018	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.008	<i>Freibacterium spp.</i>	0.008	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.004		
16	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.018	<i>Prevotella sp. HMT 304</i>	0.017	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.008	<i>Treponema sp. HMT 260</i>	0.008	<i>Treponema sp. HMT 260</i>	0.004		
17	<i>Prevotella sp. HMT 304</i>	0.017	<i>Mycoplasma faecium</i>	0.014	<i>Mycoplasma faecium</i>	0.004	<i>Prevotella sp. HMT 304</i>	0.005	<i>Mycoplasma faecium</i>	0.003		
18	<i>Mycoplasma faecium</i>	0.014	<i>Prevotella sp. HMT 304</i>	0.014	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.003	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.005	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.002		
19	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.014	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.013	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.003	<i>Prevotella sp. HMT 304</i>	0.004	<i>Prevotella sp. HMT 304</i>	0.001		
20	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.013										

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

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

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

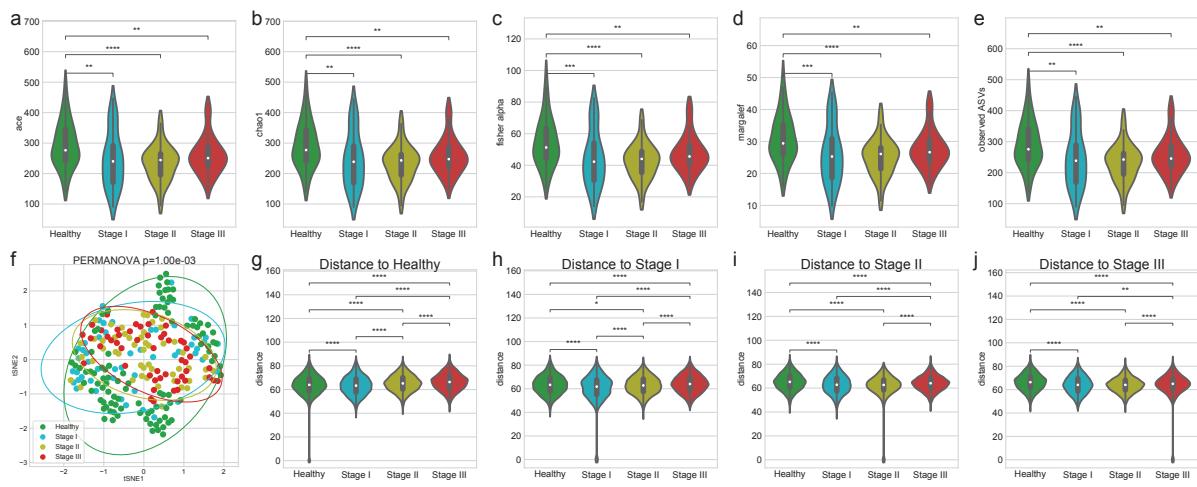


Figure 7: Diversity indices.

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

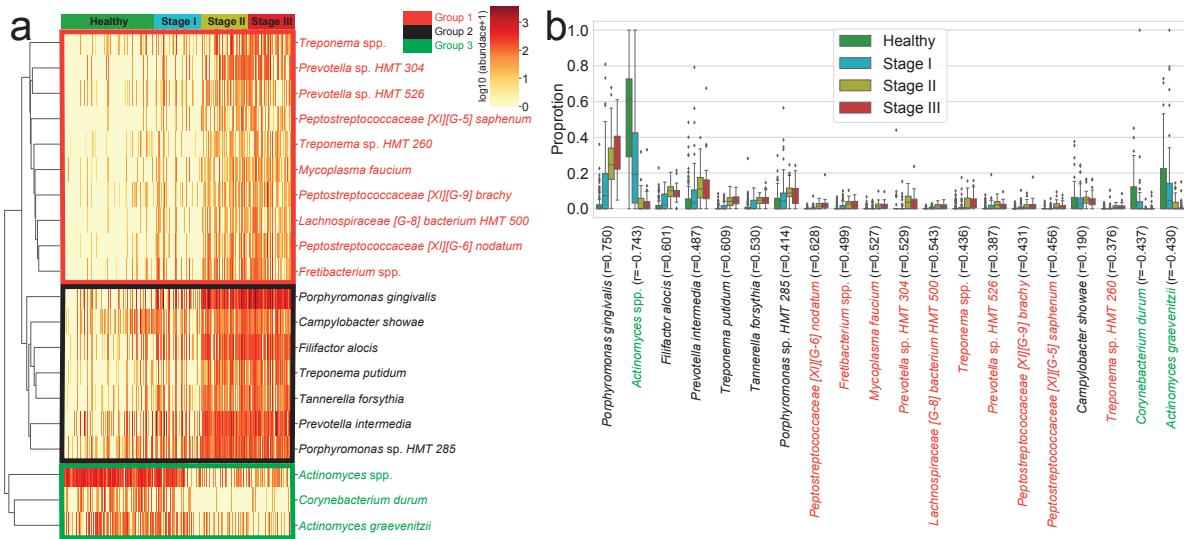


Figure 8: **Differentially abundant taxa (DAT).**

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.

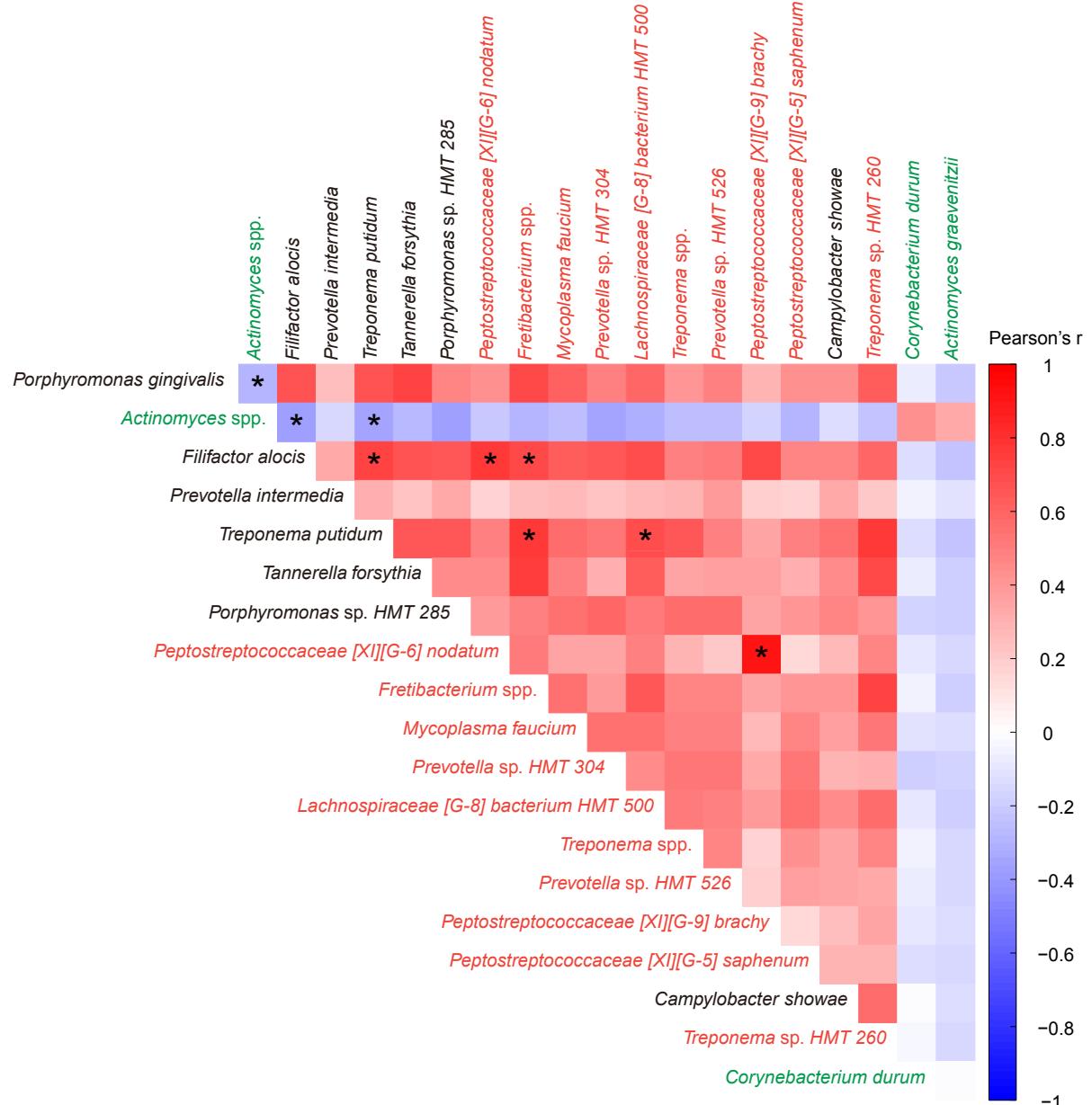


Figure 9: Correlation heatmap.

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

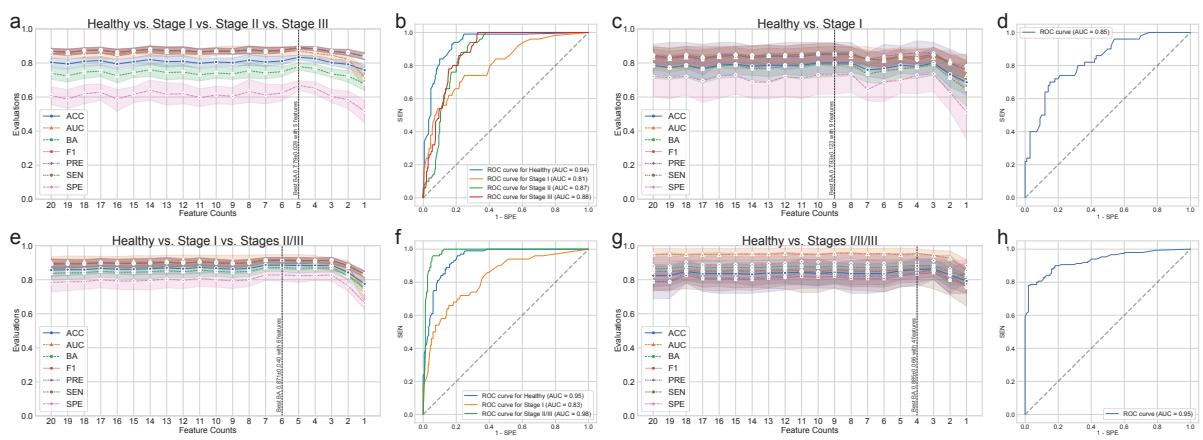


Figure 10: Random forest classification metrics.

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).

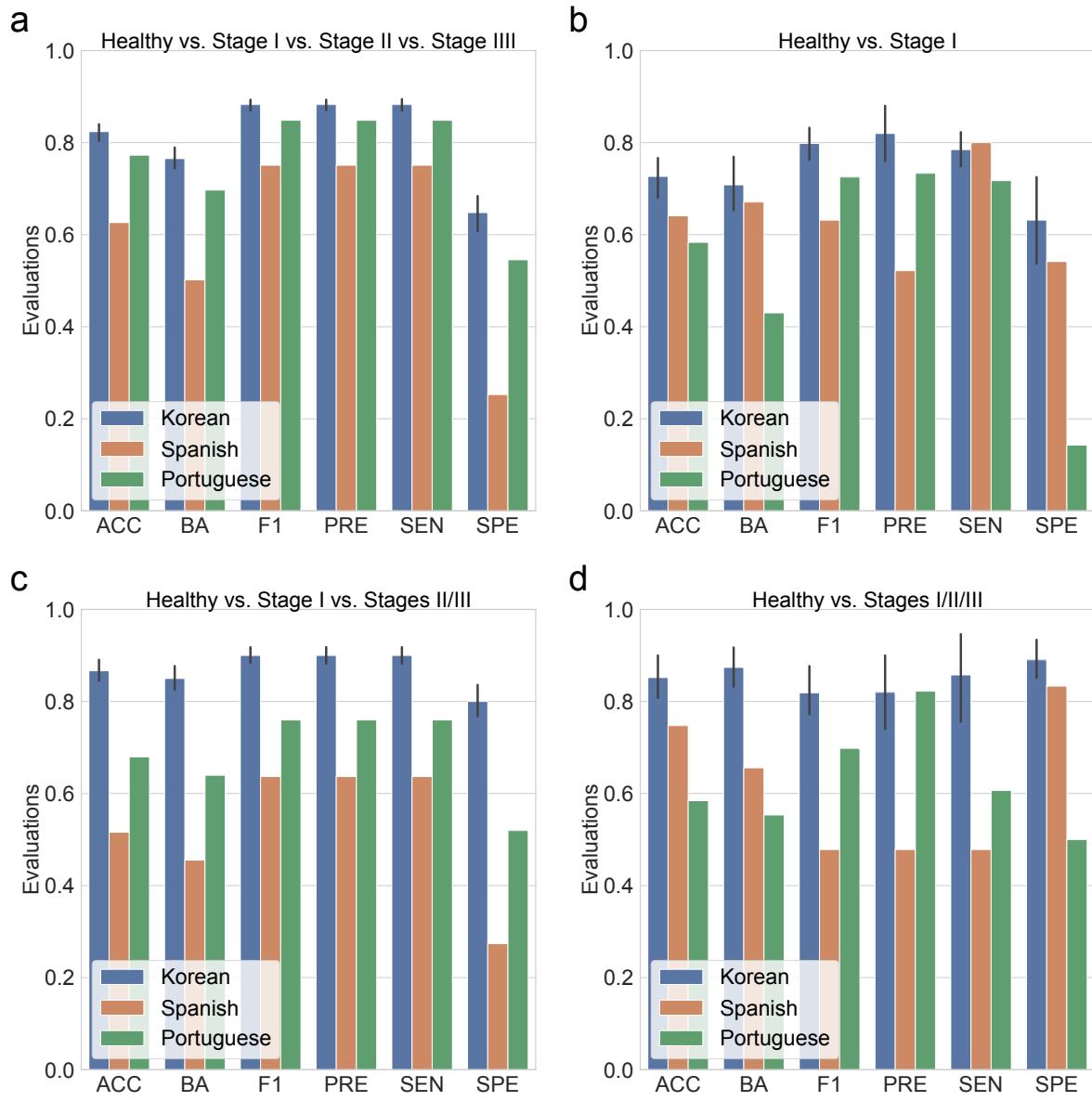


Figure 11: **Random forest classification metrics from external datasets.**

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.

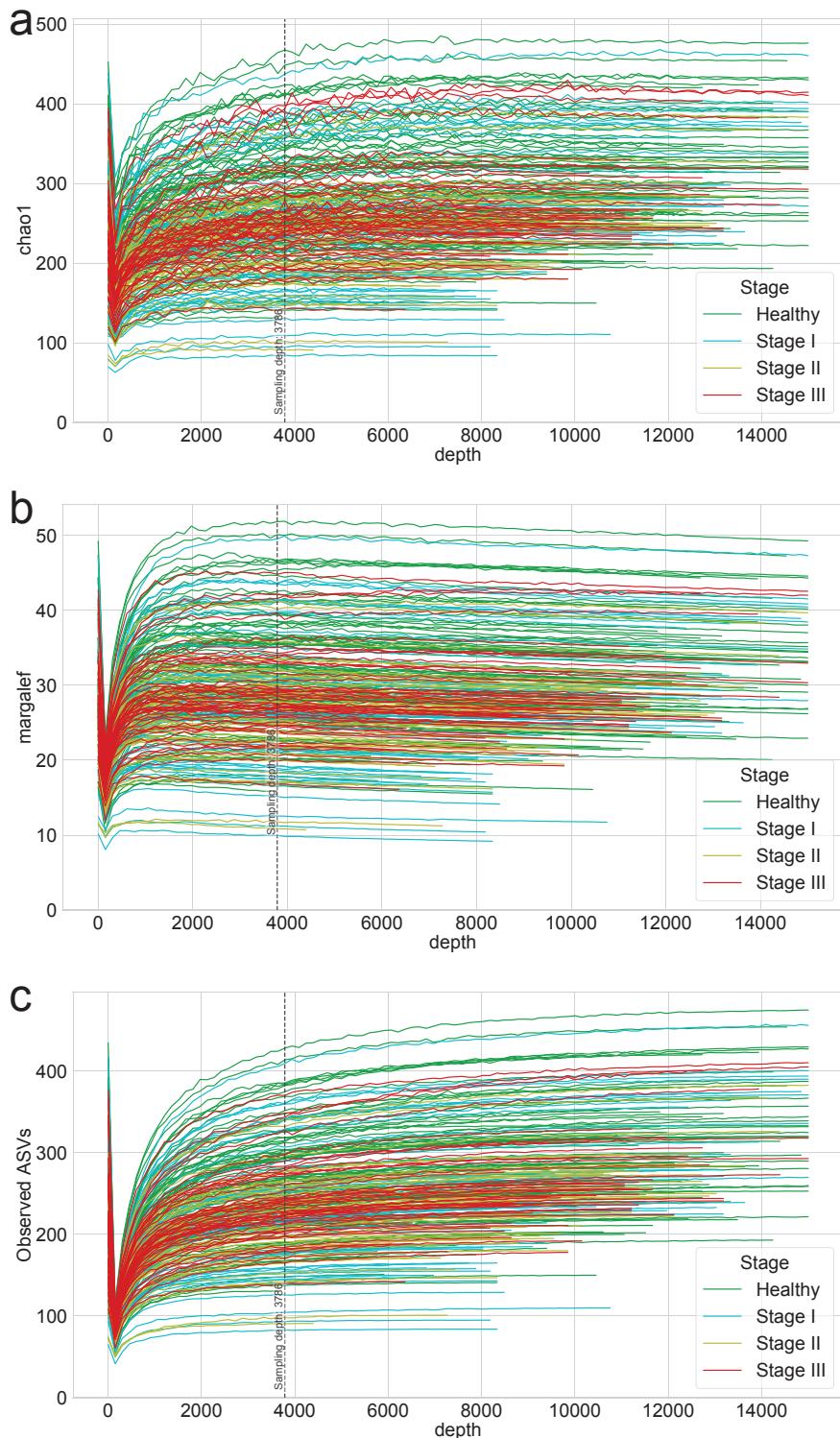


Figure 12: Rarefaction curves for alpha-diversity indices.

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

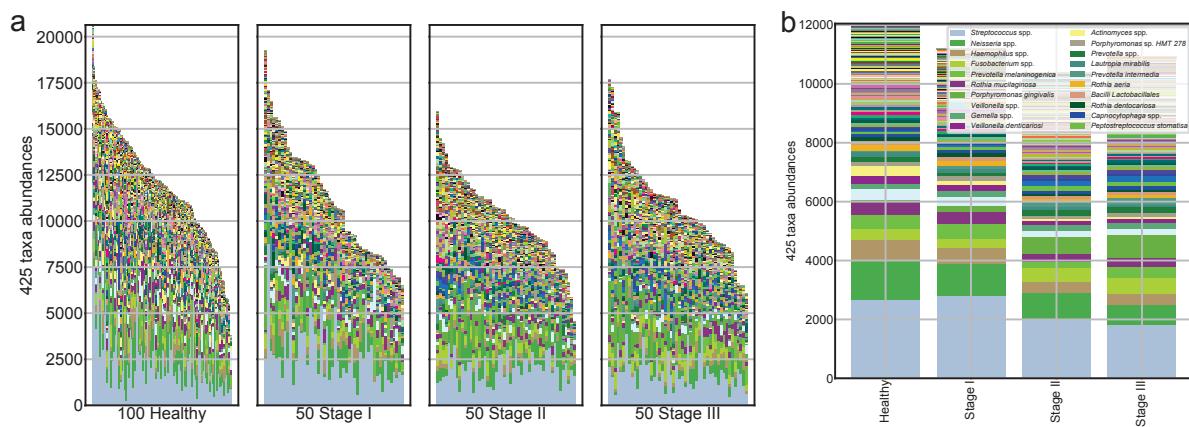


Figure 13: Salivary microbiome compositions in the different periodontal statuses.

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**).

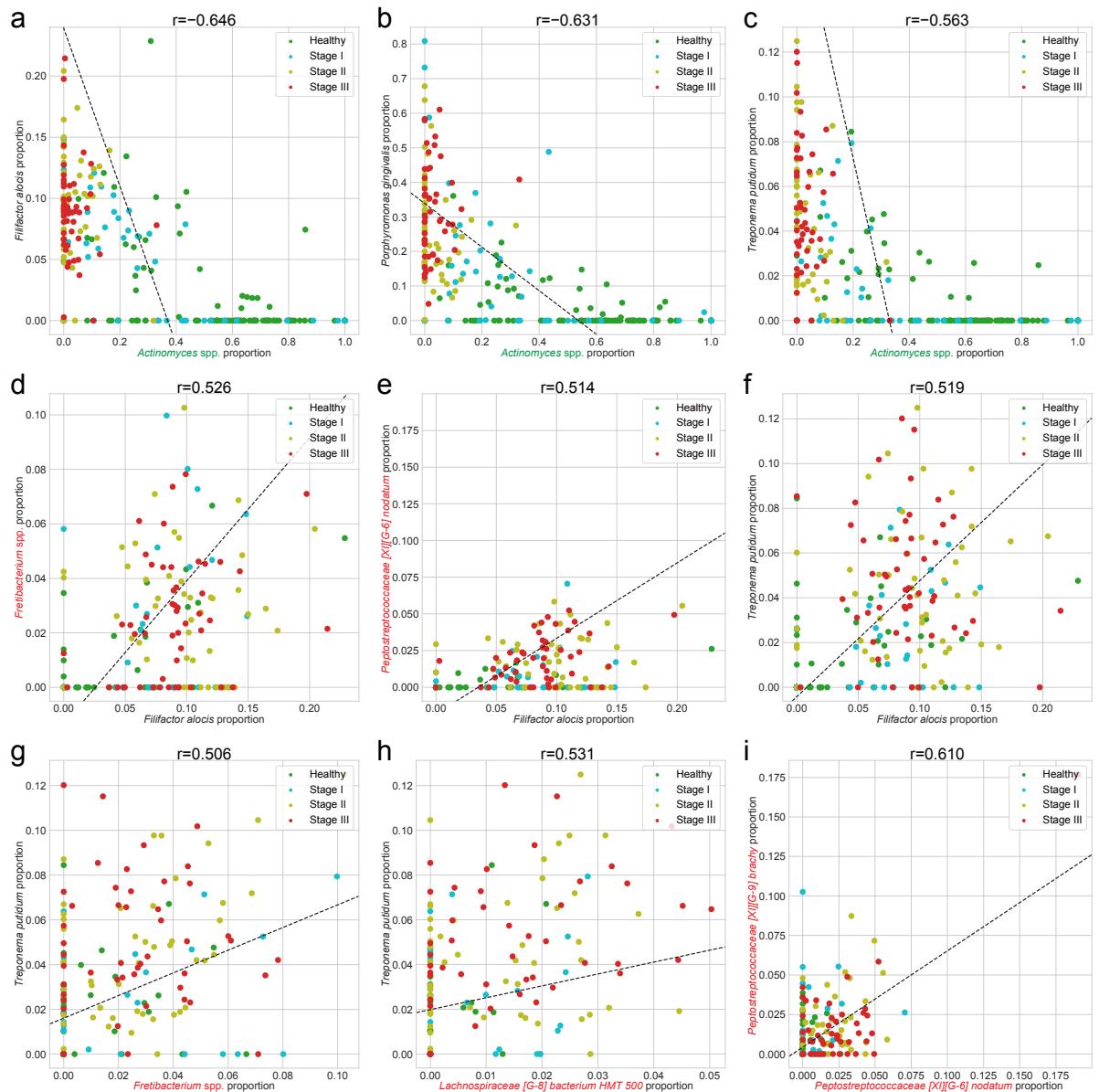


Figure 14: Correlation plots for differentially abundant taxa.

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).

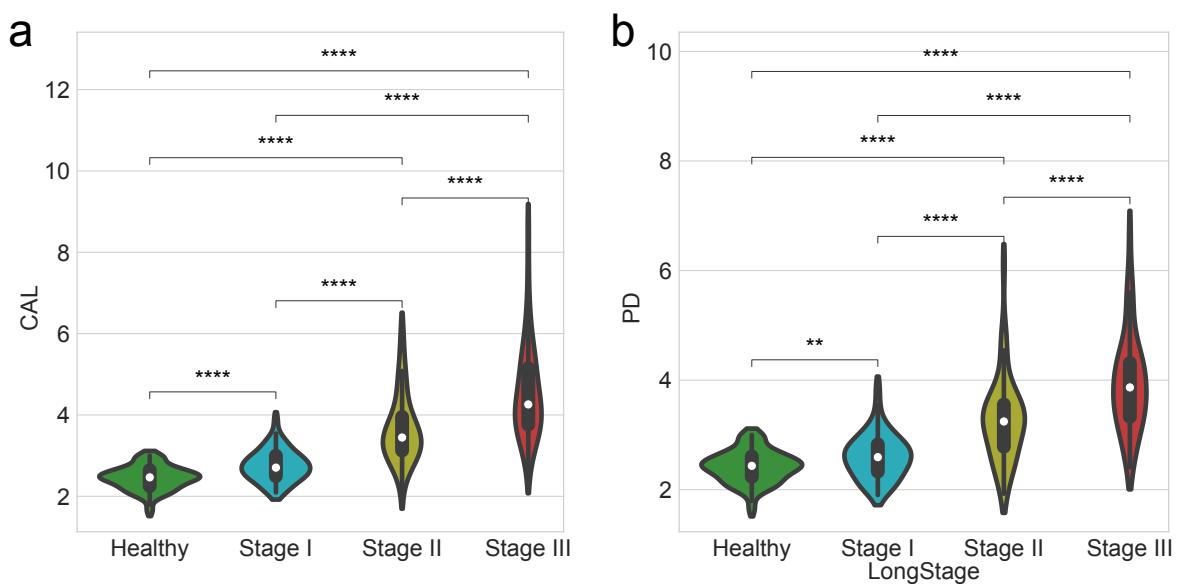


Figure 15: Clinical measurements by the periodontitis statuses.

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 $p < 0.0001$ (****).

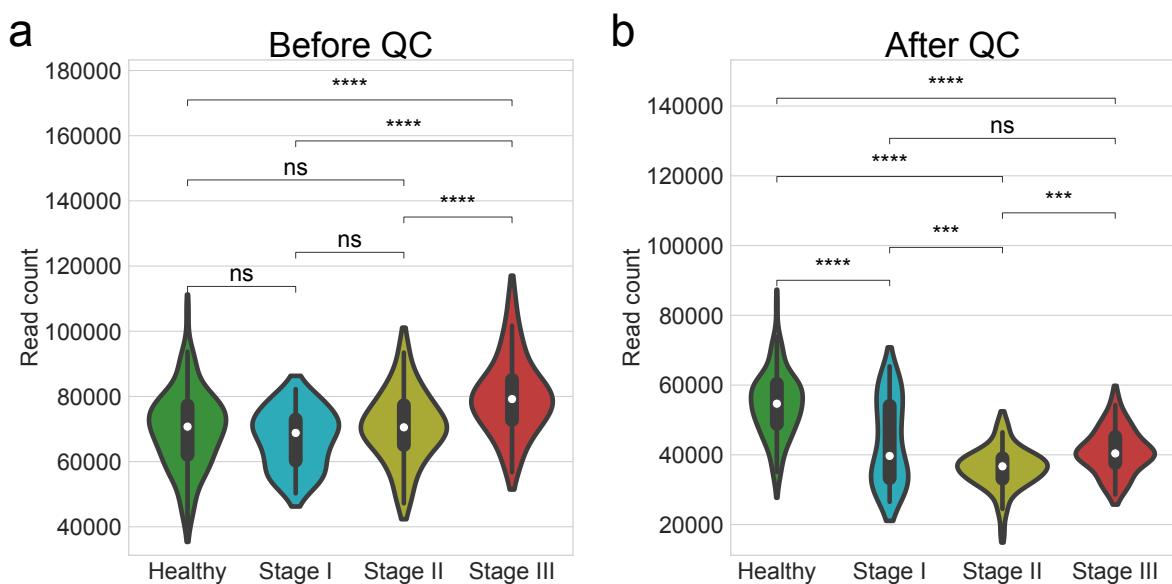


Figure 16: **Number of read counts by the periodontitis statuses.**

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 $p < 0.0001$ (****).

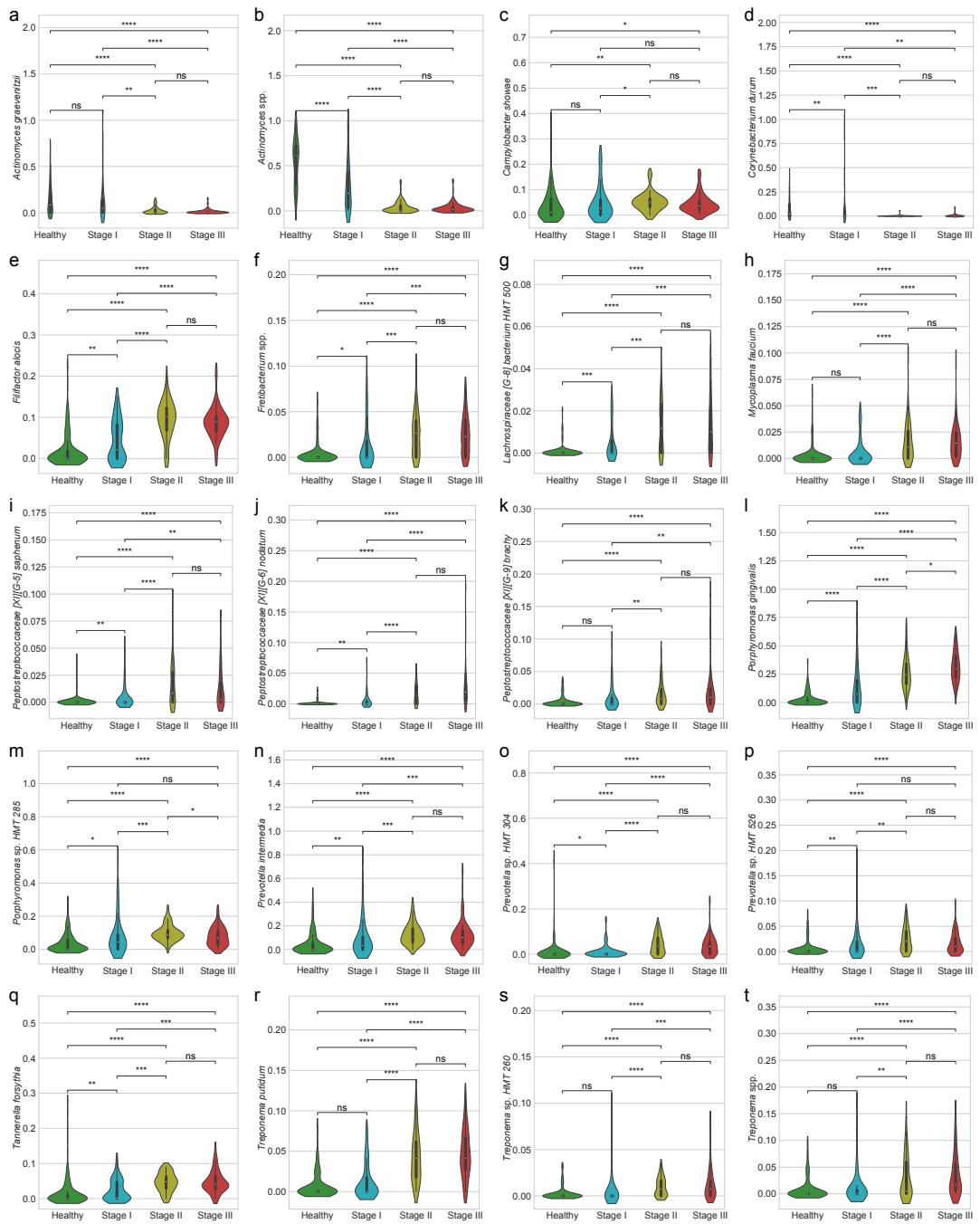


Figure 17: Proportion of DAT.

(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 $p < 0.0001$ (****).

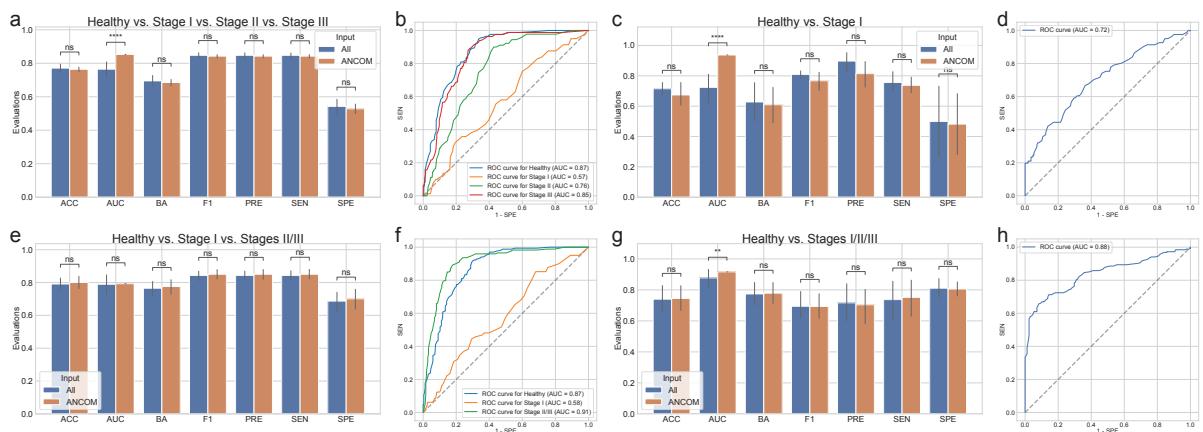


Figure 18: Random forest classification metrics with the full microbiome compositions and ANCOM-selected DAT compositions.

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 $p < 0.0001$ (****).

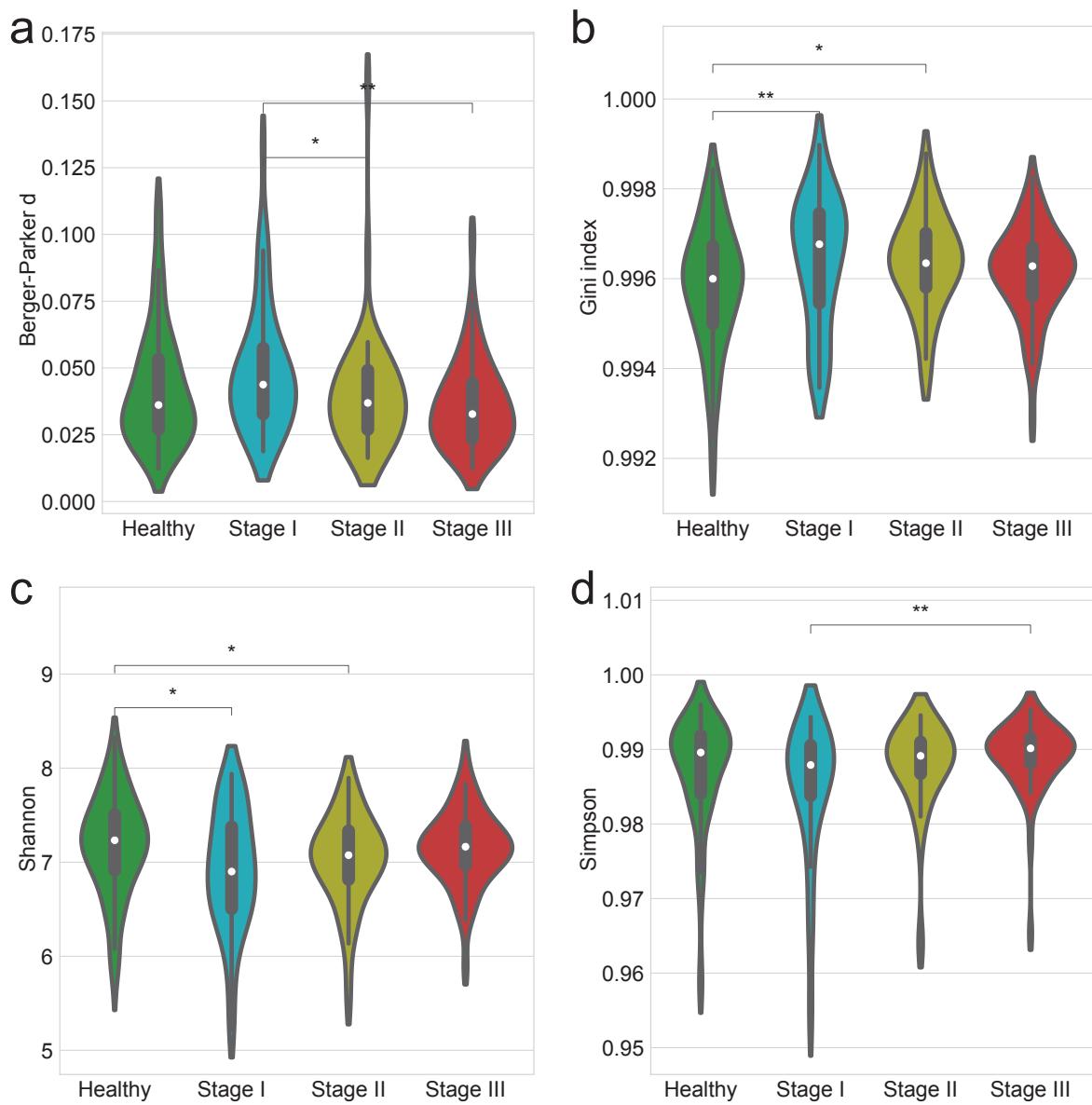


Figure 19: Alpha-diversity indices account for evenness.

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 $p < 0.01$ (**)

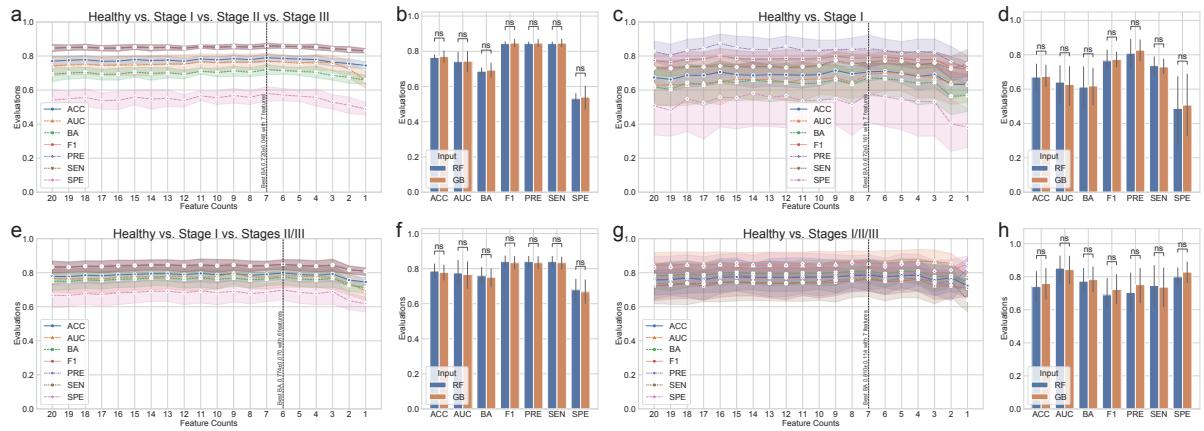


Figure 20: Gradient Boosting classification metrics.

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)

754 **3.4 Discussion**

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

763 Previous research identified the red complex as the primary pathogens of periodontitis (Listgarten,
764 1986): *Porphyromonas gingivalis*, *Tannerella forsythia*, and *Treponema denticola*. Other studies, however,
765 have shown that periodontal pathogens communicate with other bacteria in the salivary microbiome
766 networks to generate dental plaque prior to the pathogenesis and development of periodontitis (Lamont &
767 Jenkinson, 2000; Rosan & Lamont, 2000; Yoshimura, Murakami, Nishikawa, Hasegawa, & Kawaminami,
768 2009).

769 Using subgingival plaque collections, recent researches have suggested a connection between the
770 periodontitis severity and the salivary microbiome compositions (Altabtbaei et al., 2021; Iniesta et al.,
771 2023; Nemoto et al., 2021). Therefore, we have examined the salivary microbiome compositions of
772 patients with multiple severities of periodontitis and periodontally healthy controls, extending on earlier
773 studies.

774 According to our findings, the salivary microbiome compositions have 425 taxa (Figure 13). We
775 computed the alpha-diversity indices to determine the variability within each salivary microbiome
776 composition, including ace (Chao & Lee, 1992), chao1 (Chao, 1984), fisher alpha (Fisher et al., 1943),
777 margalef (Magurran, 2021), observed ASVs (DeSantis et al., 2006), Berger-Parker *d* (Berger & Parker,
778 1970), Gini index (Gini, 1912), Shannon (Weaver, 1963), and Simpson (Simpson, 1949) (Figure 7 and
779 Figure 19). Alpha-diversity indices suggested that the microbial richness of periodontally healthy controls
780 was higher than that of patients with periodontitis (Figure 7a-e and Figure 19). These results are in line with
781 findings with that patients with advanced periodontitis, namely stage II and stage III, have less diversified
782 communities than periodontally healthy controls (Jorth et al., 2014). Recognizing that the periodontitis
783 severity increases the amount of *Porphyromonas gingivalis*, the salivary microbiome compositions from
784 periodontally healthy controls conserved microbial networks dominated by *Streptococcus* spp. (Figure
785 13). *Porphyromonas gingivalis* is one of the known periodontal pathogen that could cause dysbiosys
786 in the salivary microbiomes, suggesting in the pathophysiology of periodontitis. Despite this finding,
787 earlier research found that subgingival microbiome of patients with periodontitis had a greater alpha-
788 diversity index (observed ASVs) than that of healthy controls (Iniesta et al., 2023), might due to the
789 different sampling sites between saliva and subgingival plaque. On the other hand, another research
790 has addressed significant discrepancies in alpha-diversity indices from subgingival plaque, saliva, and
791 tongue biofilms from healthy controls and periodontitis patients, resulting the highest alpha-diversity

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

801 Therefore, we calculated beta-diversity indices to analyze the differences between the study partici-
802 pants. The distances for the multiple stages of periodontitis, including stage I, stage II, and stage III, as
803 well as healthy controls (Figure 4g-j and Table 7), suggesting notable differences among the multiple
804 periodontitis severities. In other words, the composition of the salivary microbiome compositions varies
805 depending on the periodontitis stages, so that supporting the findings from a previous study (Iniesta et al.,
806 2023). Taken together that it is nearly impossible to fully restore the attachment level after it has been lost
807 due to the progression and development of periodontitis, the ability to rapidly screen for periodontitis in
808 its early phases using saliva collections would be highly beneficial for effective disease management and
809 treatment.

810 Of the total of 425 taxa in the salivary microbiome composition that have been identified (Figure 13),
811 ANCOM was applied to select 20 taxa as the DAT that indicated notable abundance variation among
812 the periodontitis severities (Figure 8 and Table 5). Three sub-groups were formed from the DAT using
813 hierarchical clustering (Figure 8a). Surprisingly, two of the red complex pathogens (Rôças, Siqueira Jr,
814 Santos, Coelho, & de Janeiro, 2001), *Porphyromonas gingivalis* and *Tannerella forsythia*, were classified
815 in Group 2 and were more prevalent in stage II and stage II periodontitis compared to healthy controls.
816 *Campylobacter showae* was additionally placed in Group 2 of the orange complex pathogens (Gambin et
817 al., 2021). Furthermore, some of the DAT in Group 2 have reported their crucial roles in pathogenesis
818 and development of periodontitis: *Filifactor alocis* (Aruni et al., 2015), *Treponema putidum* (Wyss et
819 al., 2004), *Tannerella forsythia* (Stafford, Roy, Honma, & Sharma, 2012; W. Zhu & Lee, 2016), and
820 *Prevotella intermedia* (Karched, Bhardwaj, Qudeimat, Al-Khabbaz, & Ellepolo, 2022). Taken together,
821 this indicates that DAT in Group 2 is essential to periodontitis. The portion of some Group 1 DAT,
822 including *Peptostreptococcaceae[XI][G-5] saphenum*, *Peptostreptococcaceae[XI][G-6] nodatum*, and
823 *Peptostreptococcaceae[XI][G-9] brachy*, in healthy controls and patients with periodontitis significantly
824 differed, according to earlier research (Lafaurie et al., 2022). These outcomes support our research,
825 implying that Group 1 DAT are also essential to the etiology and progression of periodontitis. However,
826 in contrast to patients with periodontitis, Group 3 DAT, namely *Corynebacterium durum* and *Actinomyces*
827 *graevenitzii*, were enriched in healthy controls, which is consistent with earlier research (Redanz et al.,
828 2021; Nibali et al., 2020).

829 In our correlation analysis (Figure 9), we have discovered strongly negative correlations (coefficient \leq
830 -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%

- 870 • BA: 84%
871 • SEN: 94%
872 • SPE: 74%
- 873 By separating patients with periodontitis from healthy controls using only four DAT, *e.g.* *Actinomyces*
874 *graevenitzii*, *Actinomyces* spp., *Corynebacterium durum*, and *Porphyromonas gingivalis*, our machine
875 learning model performed better than previously published models (Figure 10, Table 4, and Table 6):
- 876 • AUC: $95.3\% \pm 4.9\%$
877 • BA: $88.5\% \pm 6.6\%$
878 • SEN: $86.4\% \pm 15.7\%$
879 • SPE: $90.5\% \pm 7.0\%$
- 880 This result showed that by detecting Group 3 bacteria that were substantially abundant in health
881 controls than patients with periodontitis, our study increased BA by at least 5% and SPE by at least 17%.
- 882 Furthermore, we have validated our machine-learning prediction model using openly accessible 16S
883 gene rRNA sequencing data from Portuguese (Iniesta et al., 2023) and Spanish participants (Relvas et
884 al., 2021) in order to ensure the consistency of our random forest classification model (Figure 11). Our
885 classification models employed in this study were primarily developed and assessed on Korean study par-
886 ticipants, which may limit their generalizability to other ethnic groups with different salivary microbiome
887 compositions (Premaraj et al., 2020; Renson et al., 2019). Therefore, the evaluations of this periodonti-
888 tis classification models can be affected by ethnic-specific variances and differences, highlighting the
889 necessity for additional validation and adjustment across a spectrum of ethnic backgrounds.
- 890 Regarding the clinical characteristics and potential confounders influencing the analysis of salivary
891 microbiome compositions connected with periodontitis severity, this study had a number of limitations
892 that were pointed out. We did not offer clinical information, such as the percentage of teeth, the percentage
893 of bleeding on probing, nor dental furcation involvement, even though we did gather information on
894 attachment level, probing depth, plaque index, and gingival index (Renvert & Persson, 2002); this might
895 have it challenging to present thorough and in-depth data about periodontal health. Moreover, the broad age
896 range may make it tougher to evaluate the relationship between age and periodontitis statuses, providing
897 the necessity for future studies to consider into account more comprehensive clinical characteristics
898 associated with periodontitis. Additionally, potential confounders—*e.g.* body mass index (Bombin, Yan,
899 Bombin, Mosley, & Ferguson, 2022) and e-cigarette use (Suzuki, Nakano, Yoneda, Hirofushi, & Hanioka,
900 2022)—which might have affected dental health and salivary microbiome composition were disregarding
901 consideration in addition to smoking status and systemic diseases. Thus, future research incorporating
902 these components would offer a more thorough knowledge of how lifestyle factors interact and affect the
903 salivary microbiome composition and periodontal health. Throughout, resolving these limitations will
904 advance our understanding in pathogenesis and development of periodontitis, offering significant novel
905 insights on the causal connection between systemic diseases and the salivary microbiome compositions.

906 4 Metagenomic signature analysis of Korean colorectal cancer

907 4.1 Introduction

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

921 The gut microbiome plays a fundamental role in maintaining host health by helping digestion
922 (Joscelyn & Kasper, 2014; Cerqueira, Photenhauer, Pollet, Brown, & Koropatkin, 2020), regulating
923 metabolism (Dabke, Hendrick, Devkota, et al., 2019; Utzschneider, Kratz, Damman, & Hullarg, 2016;
924 Magnúsdóttir & Thiele, 2018), adjusting immune function (Kau, Ahern, Griffin, Goodman, & Gordon,
925 2011; Shi, Li, Duan, & Niu, 2017; Broom & Kogut, 2018), and even coordinating neurological processes
926 by the brain-gut axis (Martin et al., 2018; Aziz & Thompson, 1998; R. Li et al., 2024). Comprising
927 these gut microbiota, including, archaea, bacteria, fungi, and viruses, the gut microbiome contributes
928 to the synthesis of essential vitamins, and production of fatty acids, which influence intestinal integrity
929 and immune responses. Thus, well-balanced gut microbiome composition modulates systemic immune
930 function by interacting with gut-associated lymphoid tissue, shaping immune tolerance and response
931 to infections. Hence, emerging evidence suggests that dysbiosis in the gut microbiome composition are
932 associated not only a narrow range of diseases, e.g. diarrhea and enteritis (Paganini & Zimmermann,
933 2017; Gao, Yin, Xu, Li, & Yin, 2019) but also a wide range of diseases, e.g. obesity, diabetes, and cancers
934 (Barlow et al., 2015; Hartstra et al., 2015; Helmink et al., 2019; Cullin et al., 2021).

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

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

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

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

966 **4.2 Materials and methods**

967 **4.2.1 Study participants enrollment**

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

983 **4.2.2 DNA extraction procedure**

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

993 **4.2.3 Bioinformatics analysis**

994 To identify microbial signatures associated with CRC, we employed PathSeq (version 4.1.8.1) (Kostic
995 et al., 2011; Walker et al., 2018), a computational pipeline designed for metagenomic analysis of high-
996 throughput sequencing data including the whole-genome sequences. After processing these sequencing
997 data through the PathSeq pipeline, a comprehensive bioinformatics analyses were conducted to characterize
998 microbial signatures associated with CRC.

999 Prevalent taxa identification was performed by determining microbial taxa present in the majority of
1000 the study participants, filtering out low-abundance and rare taxa to ensure robust downstream analyses.

1001 To assess microbial community structure, diversity indices were calculated, including alpha-diversity
1002 to evaluate single-sample diversity and beta-diversity to compare microbial composition between the
1003 tumor tissues and their corresponding adjacent normal tissues. Following alpha-diversity indices were
1004 calculated using the scikit-bio Python package (version 0.6.3) (Rideout et al., 2018), and these alpha-
1005 diversity indices were compared using the MWU test:

- 1006 1. Berger-Parker d (Berger & Parker, 1970)
- 1007 2. Chao1 (Chao, 1984)
- 1008 3. Dominance
- 1009 4. Doubles
- 1010 5. Fisher (Fisher et al., 1943)
- 1011 6. Good's coverage (Good, 1953)
- 1012 7. Margalef (Magurran, 2021)
- 1013 8. McIntosh e (Heip, 1974)
- 1014 9. Observed ASVs (DeSantis et al., 2006)
- 1015 10. Simpson d
- 1016 11. Singles
- 1017 12. Strong (Strong, 2002)

1018 Furthermore, these beta-diversity indices were measured and compared using the PERMANOVA
1019 test (Anderson, 2014; Kelly et al., 2015). To demonstrate multi-dimensional data from the beta-diversity
1020 indices, we utilized the t-SNE algorithm (Van der Maaten & Hinton, 2008).

- 1021 1. Bray-Curtis (Sorensen, 1948)
- 1022 2. Canberra
- 1023 3. Cosine (Ochiai, 1957)
- 1024 4. Hamming (Hamming, 1950)
- 1025 5. Jaccard (Jaccard, 1908)
- 1026 6. Sokal-Sneath (Sokal & Sneath, 1963)

1027 Differentially abundant taxa (DAT) were identified using statistical method, (DESeq2 (Love et al.,
1028 2014), ANCOM (Lin & Peddada, 2020)), adjusting for sequencing depth and potential confounders to
1029 highlight taxa significantly associated with CRC. To explore functional implications, microbial pathway
1030 prediction was performed using (PICRUSt3, HUMAN3), linking microbial composition to metabolic and
1031 functional pathways relevant to carcinogenesis and progression of CRC. This multi-layered bioinformatics
1032 approach enabled a comprehensive investigation of gut microbiome alteration in CRC, facilitating the
1033 identification of potential microbial biomarkers for diagnosis and prognosis of CRC.

1034 4.2.4 Data and code availability

1035 All sequences from the 211 study participants have been published to the Korea Bioinformation Center
1036 (data ID KGD10008857): <https://kbds.re.kr/KGD10008857>. Docker image that employed through-
1037 out this study is available in the DockerHub: <https://hub.docker.com/repository/docker/>

1038 fumire/unist-crc-copm/general. Every code used in this study can be found on GitHub: <https://github.com/CompbioLabUnist/CoPM-ColonCancer>.
1039

₁₀₄₀ **4.3 Results**

₁₀₄₁ **4.3.1 Summary of clinical characteristics**

₁₀₄₂ **4.3.2 Gut microbiome compositions**

₁₀₄₃ **4.3.3 Diversity indices**

₁₀₄₄ **4.3.4 DAT selection**

₁₀₄₅ **4.3.5 ML prediction**

Table 8: Clinical characteristics of CRC study participants

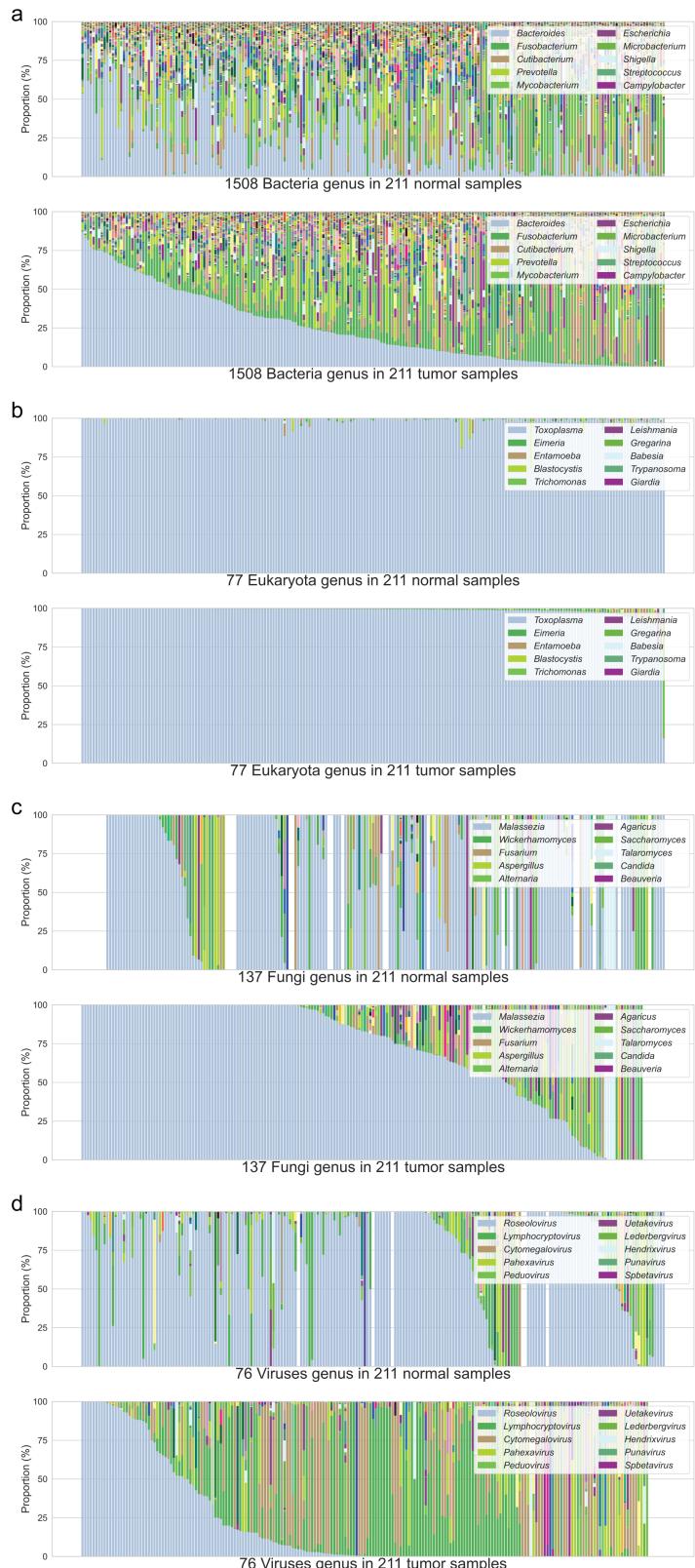


Figure 21: Gut microbiome compositions in genus level.

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

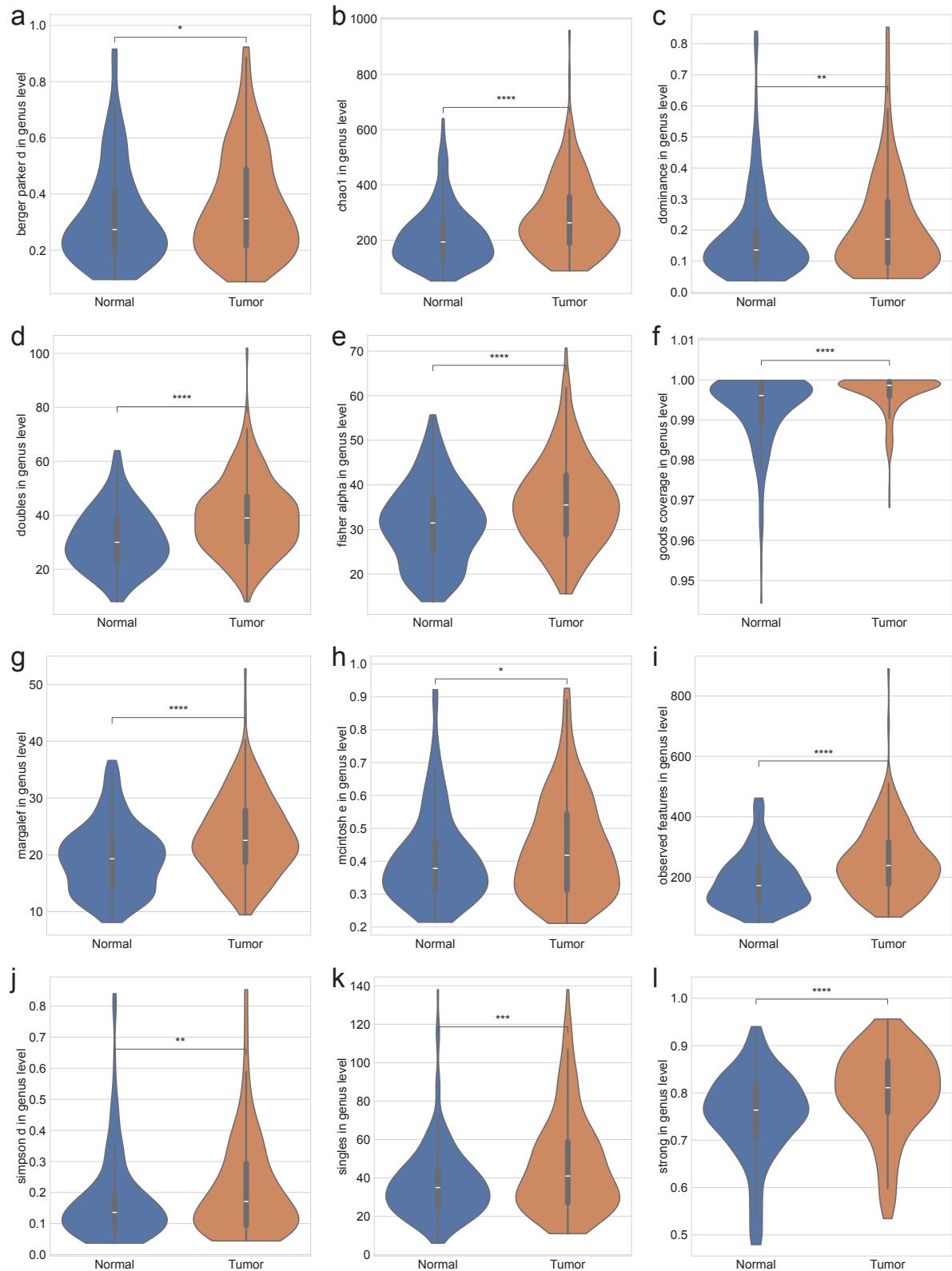


Figure 22: Alpha-diversity indices in genus level.

(a) Berger-Parker d (b) Chao1 (c) Dominance (d) Doubles (e) Fisher α (f) Good's coverage (g) Margalef (h) McIntosh (i) Observed features (j) Simpson d (k) Singles (l) Strong. MWU test: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), and $p < 0.0001$ (****)

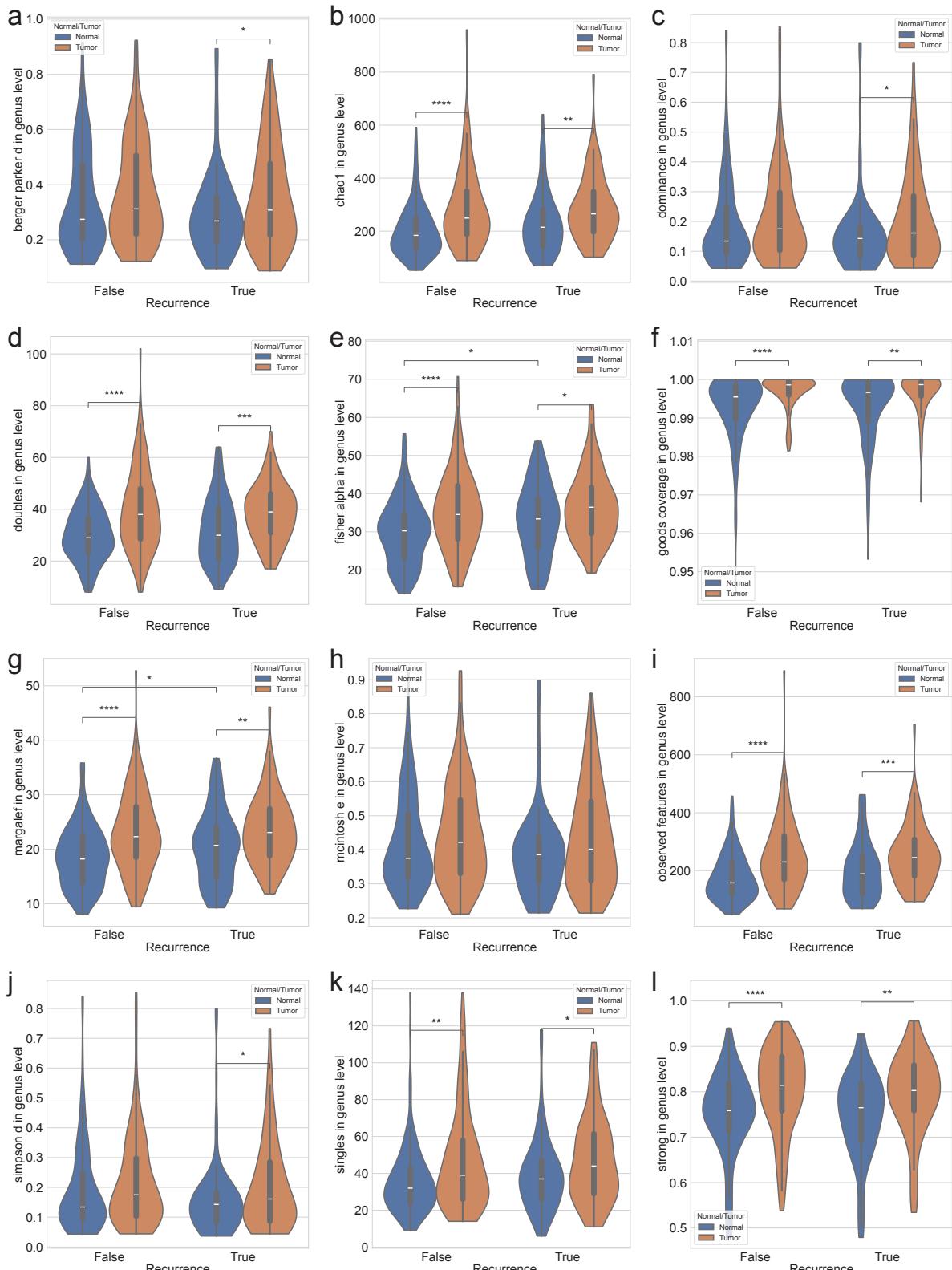


Figure 23: Alpha-diversity indices with recurrence in genus level.

(a) Berger-Parker d (b) Chao1 (c) Dominance (d) Doubles (e) Fisher α (f) Good's coverage (g) Margalef (h) McIntosh (i) Observed features (j) Simpson d (k) Singles (l) Strong. MWU test: $p < 0.05 (*)$, $p < 0.01 (**)$, $p < 0.001 (***)$, and $p < 0.0001 (****)$

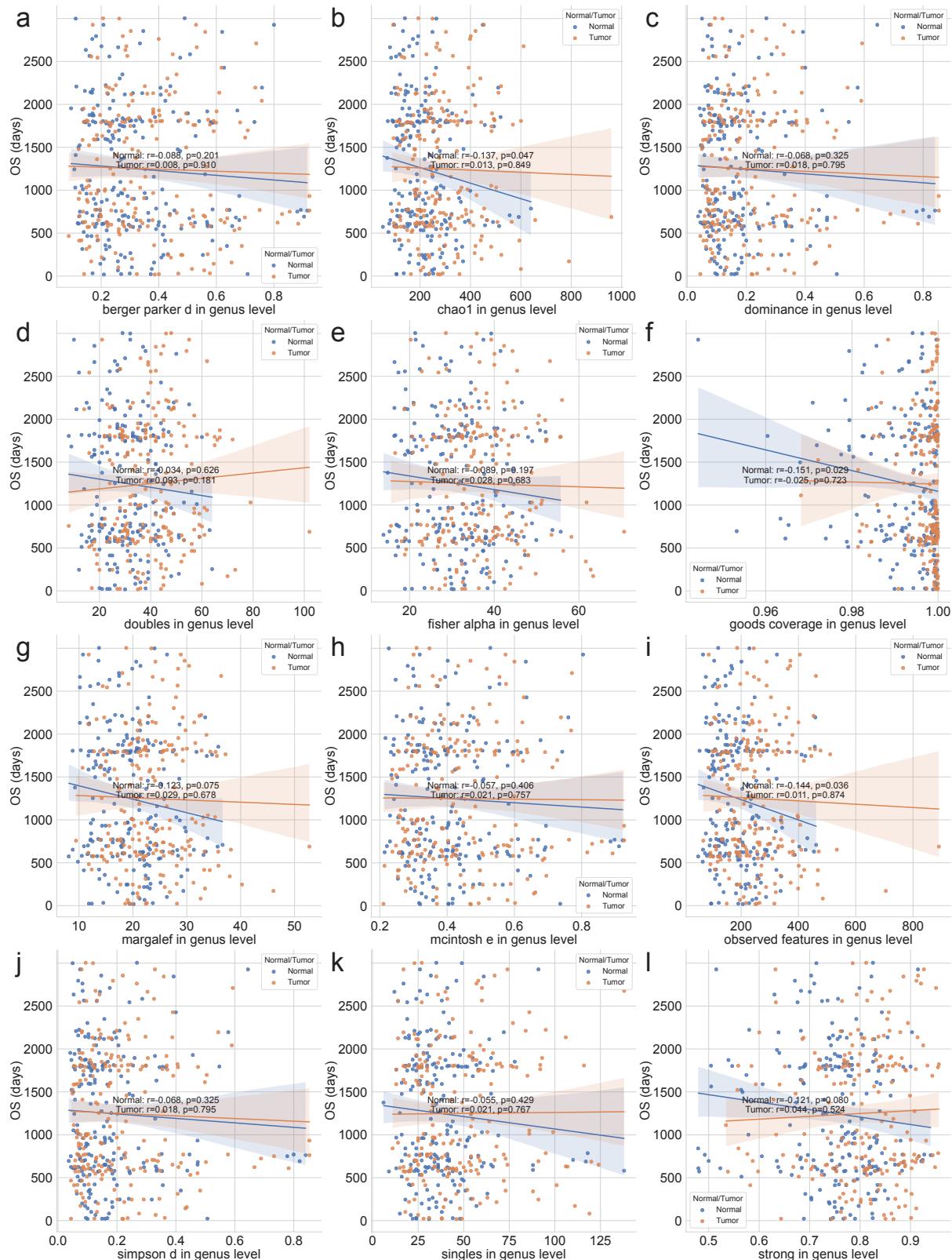


Figure 24: Alpha-diversity indices with OS in genus level.

(a) Berger-Parker d (b) Chao1 (c) Dominance (d) Doubles (e) Fisher α (f) Good's coverage (g) Margalef (h) McIntosh (i) Observed features (j) Simpson d (k) Singles (l) Strong. Statistical significance was calculated by the Spearman correlation.

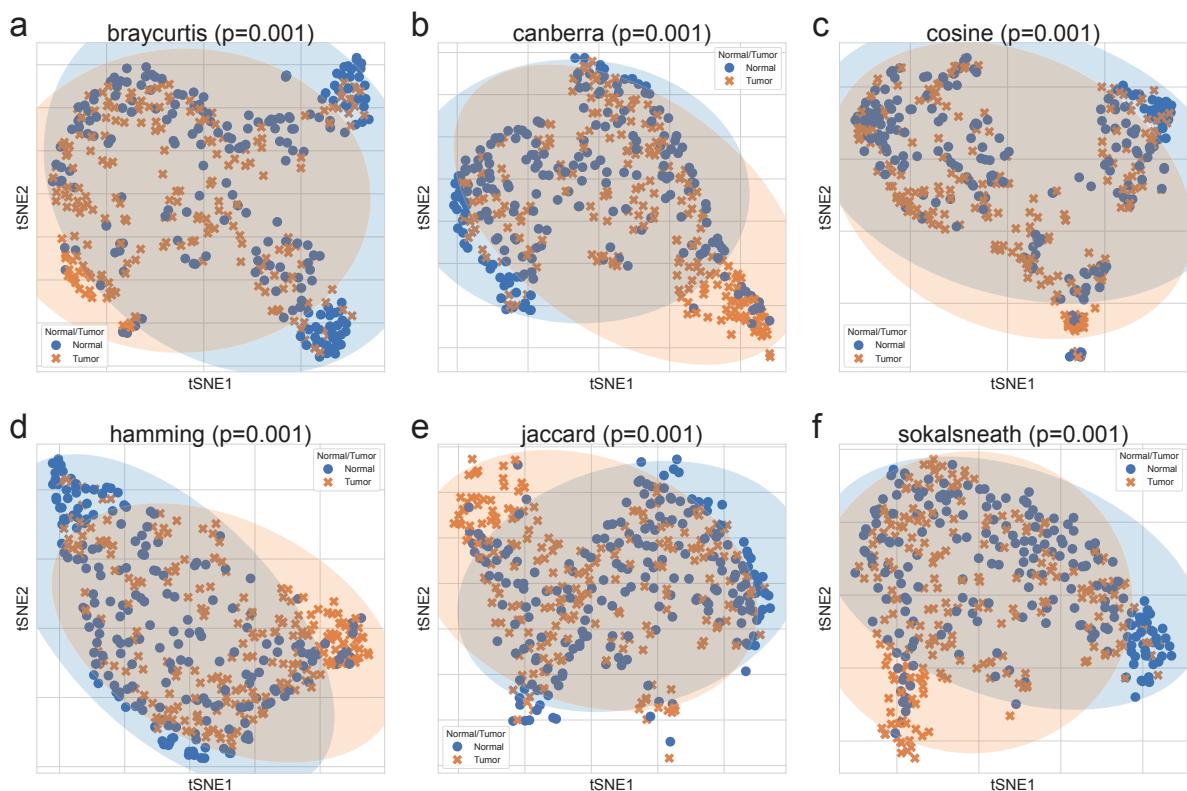


Figure 25: Beta-diversity indices in genus level.

Beta-diversity indices were visualized using a tSNE-transformed plot. The confidence ellipses are shown to display the distribution of each sub-group (Normal or Tumor). **(a)** Bray-Curtis **(b)** Canberra **(c)** Cosine **(d)** Hamming **(e)** Jaccard **(f)** Sokal-Sneath. Statistical significance were determined by PERMANOVA test.

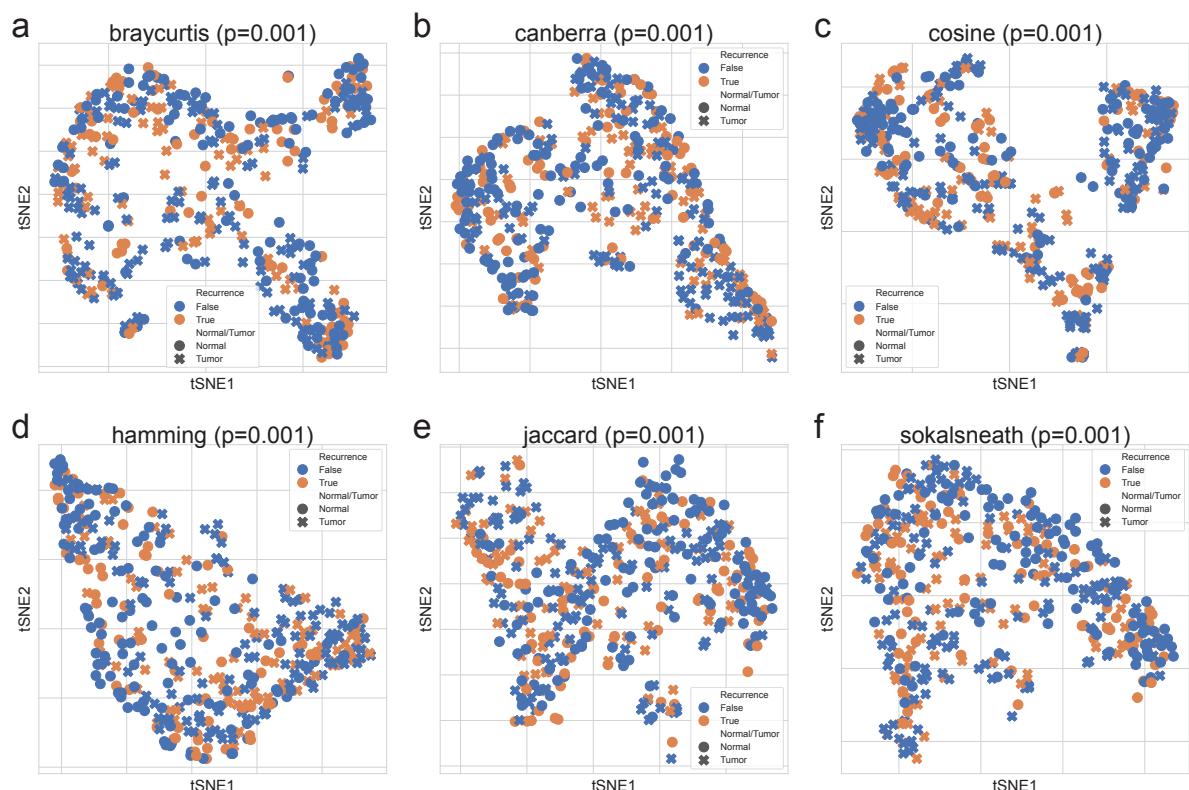


Figure 26: Beta-diversity indices with recurrence in genus level.

Beta-diversity indices were visualized using a tSNE-transformed plot. **(a)** Bray-Curtis **(b)** Canberra **(c)** Cosine **(d)** Hamming **(e)** Jaccard **(f)** Sokal-Sneath. Statistical significance were determined by PERMANOVA test.

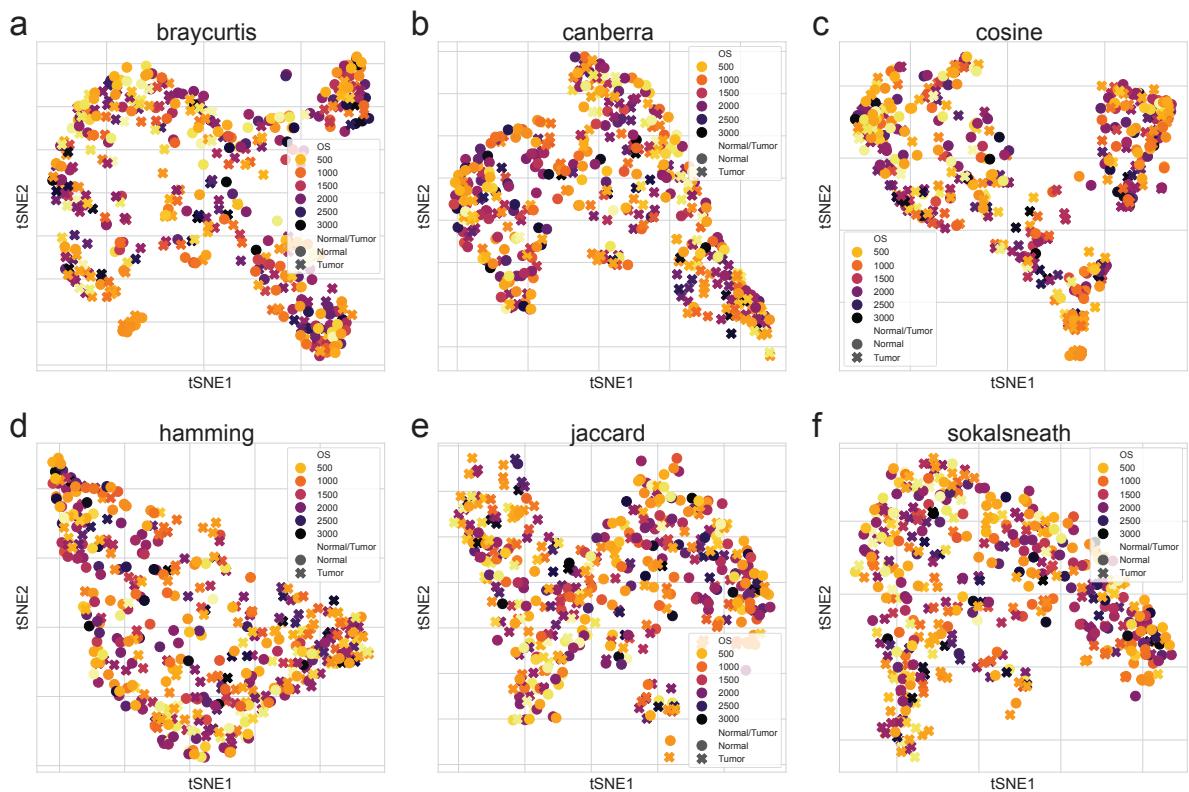


Figure 27: Beta-diversity indices with OS in genus level.

Beta-diversity indices were visualized using a tSNE-transformed plot. (a) Bray-Curtis (b) Canberra (c) Cosine (d) Hamming (e) Jaccard (f) Sokal-Sneath.

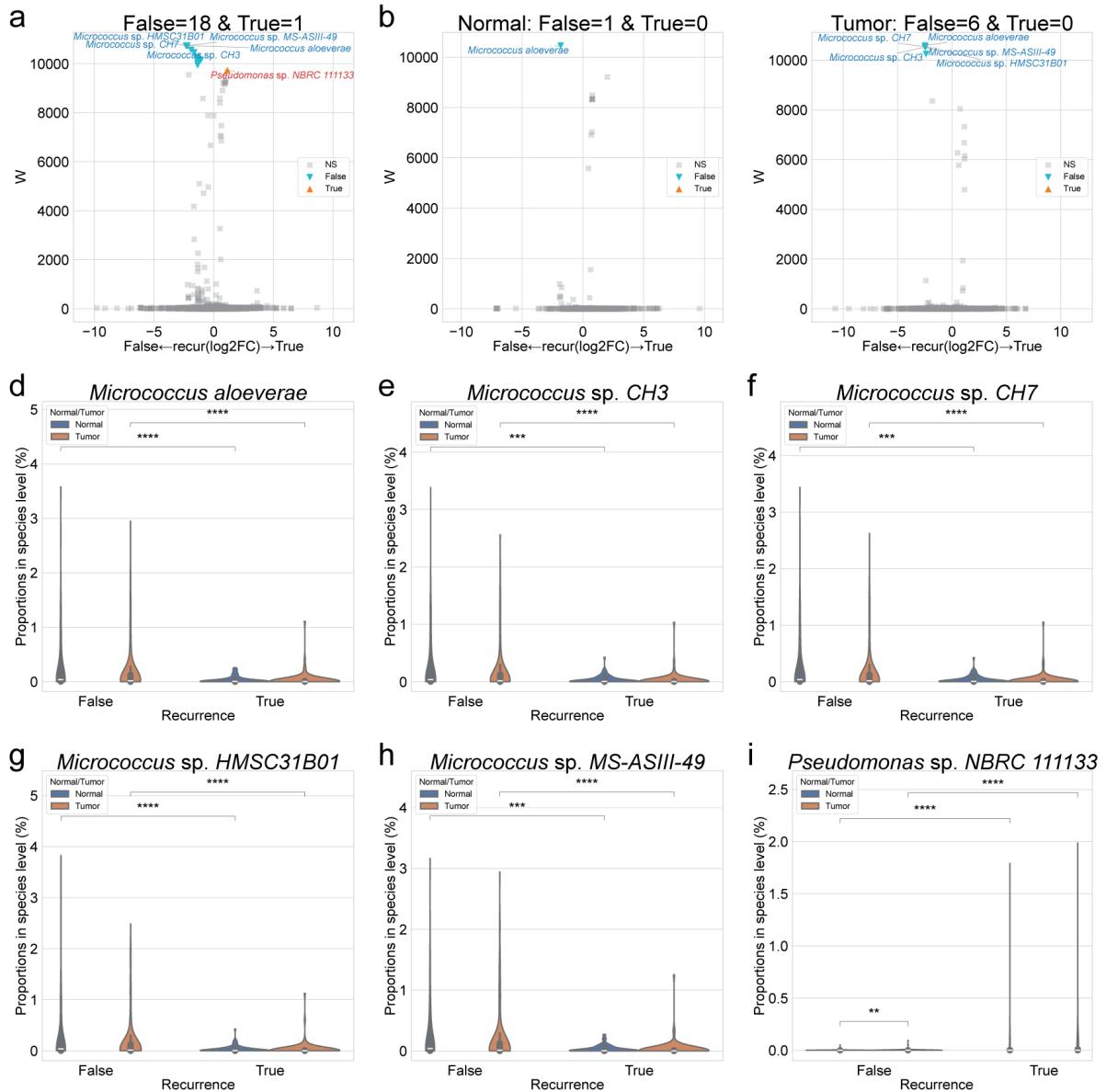


Figure 28: DAT with recurrence in genus level.

(a-c) Volcano plots with recurrence. x-axis indicates $\log_2(\text{Fold Change})$ on recurrence, and y-axis indicates ANCOM significance (W). **(a)** Total **(b)** Normal samples **(c)** Tumor samples. **(d-i)** Violin plots of each taxon proportion with recurrence. **(d)** *Micrococcus aloeverae* **(e)** *Micrococcus sp. CH3* **(f)** *Micrococcus sp. CH7* **(g)** *Micrococcus sp. HMSC31B01* **(h)** *Micrococcus sp. MS-ASIII-49* **(i)** *Pseudomonas sp. NBRC 111133*. MWU test: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***) , and $p < 0.0001$ (****)

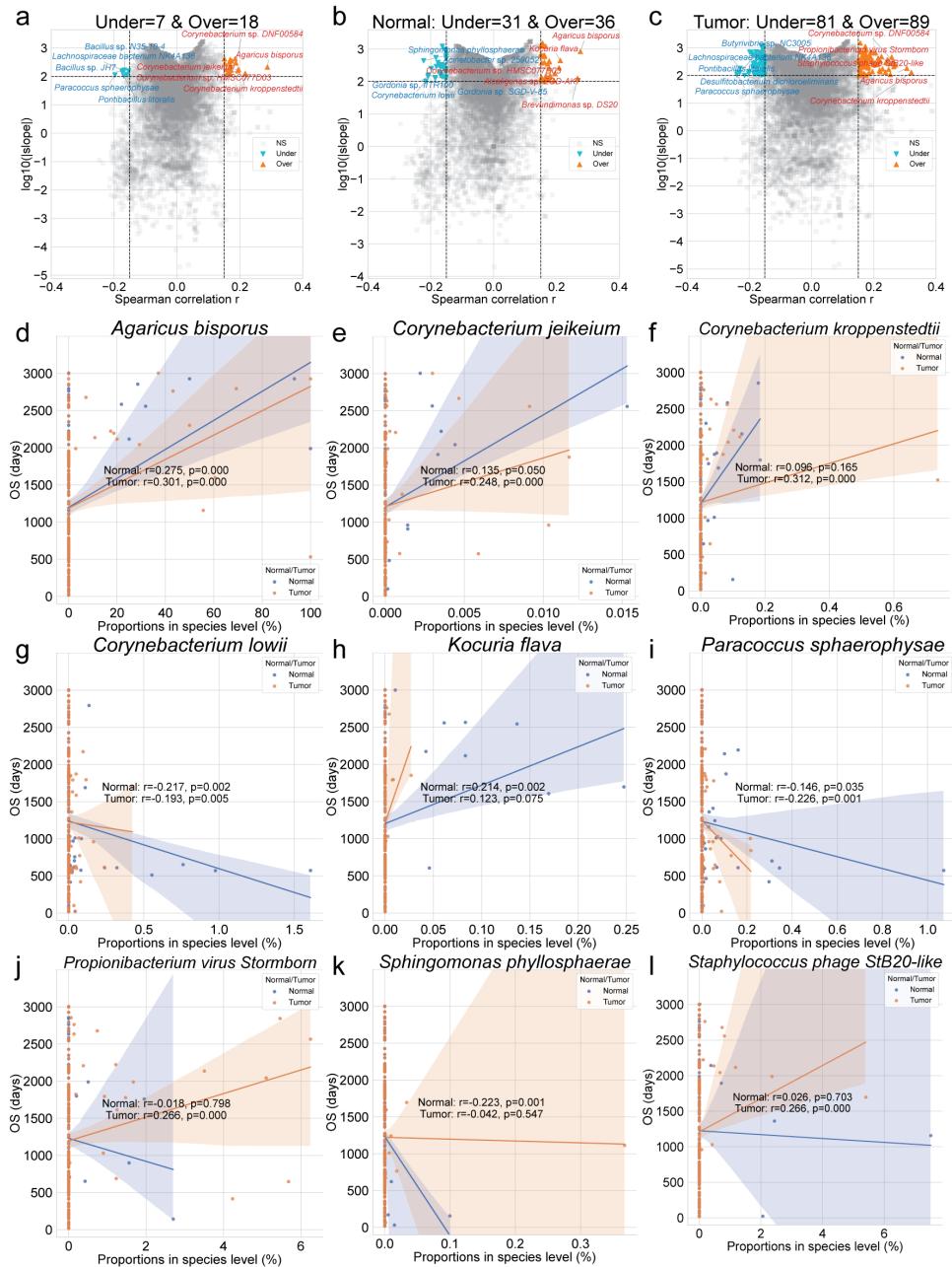


Figure 29: DAT with OS in genus level.

(a-c) Volcano plots with OS. x-axis indicates Spearman correlation coefficient (r), and y-axis indicates $\log_{10}(|\text{slope}|)$. **(a)** Total **(b)** Normal samples **(c)** Tumor samples. **(d-l)** Scatter plots of each taxon proportion with OS. **(d)** *Agaricus bisporus* **(e)** *Corynebacterium jeikeium* **(f)** *Corynebacterium kroppenstedtii* **(g)** *Corynebacterium lowii* **(h)** *Kocuria flava* **(i)** *Paracoccus sphaerophysae* **(j)** *Propionibacterium virus Stormborn* **(k)** *Sphingomonas phyllosphaerae* **(l)** *Staphylococcus phage StB20-like*. Statistical significance were calculated with Spearman correlation (r and p).

Figure 30: **ML classification.**

Figure 31: **ML regression.**

1046 **4.4 Discussion**

¹⁰⁴⁷ **5 Conclusion**

¹⁰⁴⁸ In conclusion, the research described in this doctoral dissertation was conducted to identify significant ...

¹⁰⁴⁹ In the section 2, I show that

1050 References

- 1051 Aagaard, K., Ma, J., Antony, K. M., Ganu, R., Petrosino, J., & Versalovic, J. (2014). The placenta harbors
1052 a unique microbiome. *Science translational medicine*, 6(237), 237ra65–237ra65.
- 1053 Abu-Ghazaleh, N., Chua, W. J., & Gopalan, V. (2021). Intestinal microbiota and its association with
1054 colon cancer and red/processed meat consumption. *Journal of gastroenterology and hepatology*,
1055 36(1), 75–88.
- 1056 Abusleme, L., Hoare, A., Hong, B.-Y., & Diaz, P. I. (2021). Microbial signatures of health, gingivitis,
1057 and periodontitis. *Periodontology 2000*, 86(1), 57–78.
- 1058 Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., & Pawlowsky-Glahn, V. (2000). Logratio
1059 analysis and compositional distance. *Mathematical geology*, 32, 271–275.
- 1060 Aja, E., Mangar, M., Fletcher, H., & Mishra, A. (2021). Filifactor alocis: recent insights and advances.
1061 *Journal of dental research*, 100(8), 790–797.
- 1062 Alelyani, S. (2021). Stable bagging feature selection on medical data. *Journal of Big Data*, 8(1), 11.
- 1063 Altabtbaei, K., Maney, P., Ganesan, S. M., Dabdoub, S. M., Nagaraja, H. N., & Kumar, P. S. (2021). Anna
1064 karenina and the subgingival microbiome associated with periodontitis. *Microbiome*, 9, 1–15.
- 1065 Altingöz, S. M., Kurgan, Ş., Önder, C., Serdar, M. A., Ünlütürk, U., Uyanık, M., ... Günhan, M.
1066 (2021). Salivary and serum oxidative stress biomarkers and advanced glycation end products in
1067 periodontitis patients with or without diabetes: A cross-sectional study. *Journal of periodontology*,
1068 92(9), 1274–1285.
- 1069 Alverdy, J., Hyoju, S., Weigerinck, M., & Gilbert, J. (2017). The gut microbiome and the mechanism of
1070 surgical infection. *Journal of British Surgery*, 104(2), e14–e23.
- 1071 An, S., & Park, S. (2022). Association of physical activity and sedentary behavior with the risk of
1072 colorectal cancer. *Journal of Korean Medical Science*, 37(19).
- 1073 Anderson, M. J. (2014). Permutational multivariate analysis of variance (permanova). *Wiley statsref:
1074 statistics reference online*, 1–15.
- 1075 Aruni, A. W., Mishra, A., Dou, Y., Chioma, O., Hamilton, B. N., & Fletcher, H. M. (2015). Filifactor
1076 alocis—a new emerging periodontal pathogen. *Microbes and infection*, 17(7), 517–530.
- 1077 Aziz, Q., & Thompson, D. G. (1998). Brain-gut axis in health and disease. *Gastroenterology*, 114(3),
1078 559–578.
- 1079 Bai, X., Wei, H., Liu, W., Coker, O. O., Gou, H., Liu, C., ... others (2022). Cigarette smoke promotes
1080 colorectal cancer through modulation of gut microbiota and related metabolites. *Gut*, 71(12),

- 1081 2439–2450.
- 1082 Baldelli, V., Scaldaferri, F., Putignani, L., & Del Chierico, F. (2021). The role of enterobacteriaceae in
1083 gut microbiota dysbiosis in inflammatory bowel diseases. *Microorganisms*, 9(4), 697.
- 1084 Bardou, M., Rouland, A., Martel, M., Loffroy, R., Barkun, A. N., & Chapelle, N. (2022). Obesity and
1085 colorectal cancer. *Alimentary Pharmacology & Therapeutics*, 56(3), 407–418.
- 1086 Barlow, G. M., Yu, A., & Mathur, R. (2015). Role of the gut microbiome in obesity and diabetes mellitus.
1087 *Nutrition in clinical practice*, 30(6), 787–797.
- 1088 Basavaprabhu, H., Sonu, K., & Prabha, R. (2020). Mechanistic insights into the action of probiotics
1089 against bacterial vaginosis and its mediated preterm birth: An overview. *Microbial pathogenesis*,
1090 141, 104029.
- 1091 Belstrøm, D., Constancias, F., Drautz-Moses, D. I., Schuster, S. C., Veleba, M., Mahé, F., & Givskov, M.
1092 (2021). Periodontitis associates with species-specific gene expression of the oral microbiota. *npj
1093 Biofilms and Microbiomes*, 7(1), 76.
- 1094 Berger, W. H., & Parker, F. L. (1970). Diversity of planktonic foraminifera in deep-sea sediments.
1095 *Science*, 168(3937), 1345–1347.
- 1096 Berghella, V. (2012). Universal cervical length screening for prediction and prevention of preterm birth.
1097 *Obstetrical & gynecological survey*, 67(10), 653–657.
- 1098 Blencowe, H., Cousens, S., Oestergaard, M. Z., Chou, D., Moller, A.-B., Narwal, R., ... others (2012).
1099 National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends
1100 since 1990 for selected countries: a systematic analysis and implications. *The lancet*, 379(9832),
1101 2162–2172.
- 1102 Boland, C. R., & Goel, A. (2010). Microsatellite instability in colorectal cancer. *Gastroenterology*,
1103 138(6), 2073–2087.
- 1104 Boleij, A., Hechenbleikner, E. M., Goodwin, A. C., Badani, R., Stein, E. M., Lazarev, M. G., ... others
1105 (2015). The bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer
1106 patients. *Clinical Infectious Diseases*, 60(2), 208–215.
- 1107 Bolstad, A., Jensen, H. B., & Bakken, V. (1996). Taxonomy, biology, and periodontal aspects of
1108 fusobacterium nucleatum. *Clinical microbiology reviews*, 9(1), 55–71.
- 1109 Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... others
1110 (2019). Reproducible, interactive, scalable and extensible microbiome data science using qiime 2.
1111 *Nature biotechnology*, 37(8), 852–857.
- 1112 Bombin, A., Yan, S., Bombin, S., Mosley, J. D., & Ferguson, J. F. (2022). Obesity influences composition
1113 of salivary and fecal microbiota and impacts the interactions between bacterial taxa. *Physiological
1114 reports*, 10(7), e15254.
- 1115 Bonnet, M., Buc, E., Sauvanet, P., Darcha, C., Dubois, D., Pereira, B., ... Darfeuille-Michaud, A. (2014).
1116 Colonization of the human gut by e. coli and colorectal cancer risk. *Clinical Cancer Research*,
1117 20(4), 859–867.
- 1118 Breiman, L. (2001). Random forests. *Machine learning*, 45, 5–32.
- 1119 Brennan, C. A., & Garrett, W. S. (2019). Fusobacterium nucleatum—symbiont, opportunist and

- 1120 oncobacterium. *Nature Reviews Microbiology*, 17(3), 156–166.
- 1121 Broom, L. J., & Kogut, M. H. (2018). The role of the gut microbiome in shaping the immune system of
1122 chickens. *Veterinary immunology and immunopathology*, 204, 44–51.
- 1123 Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: improving accuracy of classifier
1124 ensembles by using random feature subsets. *Pattern recognition*, 36(6), 1291–1302.
- 1125 Bullman, S., Pedamallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., ... others (2017). Analysis
1126 of fusobacterium persistence and antibiotic response in colorectal cancer. *Science*, 358(6369),
1127 1443–1448.
- 1128 Burt, R. W., Leppert, M. F., Slattery, M. L., Samowitz, W. S., Spirio, L. N., Kerber, R. A., ... others
1129 (2004). Genetic testing and phenotype in a large kindred with attenuated familial adenomatous
1130 polyposis. *Gastroenterology*, 127(2), 444–451.
- 1131 Cai, Y., Li, Y., Xiong, Y., Geng, X., Kang, Y., & Yang, Y. (2024). Diabetic foot exacerbates gut
1132 mycobiome dysbiosis in adult patients with type 2 diabetes mellitus: revealing diagnostic markers.
1133 *Nutrition & Diabetes*, 14(1), 71.
- 1134 Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016).
1135 Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7),
1136 581–583.
- 1137 Canakci, V., & Canakci, C. F. (2007). Pain levels in patients during periodontal probing and mechanical
1138 non-surgical therapy. *Clinical oral investigations*, 11, 377–383.
- 1139 Cappellato, M., Baruzzo, G., & Di Camillo, B. (2022). Investigating differential abundance methods in
1140 microbiome data: A benchmark study. *PLoS computational biology*, 18(9), e1010467.
- 1141 Castaner, O., Goday, A., Park, Y.-M., Lee, S.-H., Magkos, F., Shiow, S.-A. T. E., & Schröder, H. (2018).
1142 The gut microbiome profile in obesity: a systematic review. *International journal of endocrinology*,
1143 2018(1), 4095789.
- 1144 Center, M. M., Jemal, A., Smith, R. A., & Ward, E. (2009). Worldwide variations in colorectal cancer.
1145 *CA: a cancer journal for clinicians*, 59(6), 366–378.
- 1146 Centor, R. M. (1991). Signal detectability: the use of roc curves and their analyses. *Medical decision
1147 making*, 11(2), 102–106.
- 1148 Cerqueira, F. M., Photenhauer, A. L., Pollet, R. M., Brown, H. A., & Koropatkin, N. M. (2020). Starch
1149 digestion by gut bacteria: crowdsourcing for carbs. *Trends in Microbiology*, 28(2), 95–108.
- 1150 Champagne, C., McNairn, H., Daneshfar, B., & Shang, J. (2014). A bootstrap method for assessing
1151 classification accuracy and confidence for agricultural land use mapping in canada. *International
1152 Journal of Applied Earth Observation and Geoinformation*, 29, 44–52.
- 1153 Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian
1154 Journal of statistics*, 265–270.
- 1155 Chao, A., & Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the
1156 American statistical Association*, 87(417), 210–217.
- 1157 Chapple, I. L., Mealey, B. L., Van Dyke, T. E., Bartold, P. M., Dommisch, H., Eickholz, P., ... others
1158 (2018). Periodontal health and gingival diseases and conditions on an intact and a reduced

- 1159 periodontium: Consensus report of workgroup 1 of the 2017 world workshop on the classification
1160 of periodontal and peri-implant diseases and conditions. *Journal of periodontology*, 89, S74–S84.
- 1161 Chen, T., Marsh, P., & Al-Hebshi, N. (2022). Smdi: an index for measuring subgingival microbial
1162 dysbiosis. *Journal of dental research*, 101(3), 331–338.
- 1163 Chen, T., Yu, W.-H., Izard, J., Baranova, O. V., Lakshmanan, A., & Dewhirst, F. E. (2010). The human
1164 oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and
1165 genomic information. *Database*, 2010.
- 1166 Chen, X., D’Souza, R., & Hong, S.-T. (2013). The role of gut microbiota in the gut-brain axis: current
1167 challenges and perspectives. *Protein & cell*, 4, 403–414.
- 1168 Chen, X., Jansen, L., Guo, F., Hoffmeister, M., Chang-Claude, J., & Brenner, H. (2021). Smoking,
1169 genetic predisposition, and colorectal cancer risk. *Clinical and translational gastroenterology*,
1170 12(3), e00317.
- 1171 Chen, X., Li, H., Guo, F., Hoffmeister, M., & Brenner, H. (2022). Alcohol consumption, polygenic risk
1172 score, and early-and late-onset colorectal cancer risk. *EClinicalMedicine*, 49.
- 1173 Chew, R. J. J., Tan, K. S., Chen, T., Al-Hebshi, N. N., & Goh, C. E. (2024). Quantifying periodontitis-
1174 associated oral dysbiosis in tongue and saliva microbiomes—an integrated data analysis. *Journal
1175 of Periodontology*.
- 1176 Čižmárová, B., Tomečková, V., Hubková, B., Hurajtová, A., Ohlasová, J., & Birková, A. (2022). Salivary
1177 redox homeostasis in human health and disease. *International Journal of Molecular Sciences*,
1178 23(17), 10076.
- 1179 Cullin, N., Antunes, C. A., Straussman, R., Stein-Thoeringer, C. K., & Elinav, E. (2021). Microbiome
1180 and cancer. *Cancer Cell*, 39(10), 1317–1341.
- 1181 Dabke, K., Hendrick, G., Devkota, S., et al. (2019). The gut microbiome and metabolic syndrome. *The
1182 Journal of clinical investigation*, 129(10), 4050–4057.
- 1183 DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... Andersen, G. L.
1184 (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with
1185 arb. *Applied and environmental microbiology*, 72(7), 5069–5072.
- 1186 Doyle, R., Alber, D., Jones, H., Harris, K., Fitzgerald, F., Peebles, D., & Klein, N. (2014). Term and
1187 preterm labour are associated with distinct microbial community structures in placental membranes
1188 which are independent of mode of delivery. *Placenta*, 35(12), 1099–1101.
- 1189 Fahmy, C. A., Gamal-Eldeen, A. M., El-Hussieny, E. A., Raafat, B. M., Mehanna, N. S., Talaat, R. M., &
1190 Shaaban, M. T. (2019). Bifidobacterium longum suppresses murine colorectal cancer through the
1191 modulation of oncomirs and tumor suppressor mirnas. *Nutrition and cancer*, 71(4), 688–700.
- 1192 Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological conservation*, 61(1),
1193 1–10.
- 1194 Fettweis, J. M., Serrano, M. G., Brooks, J. P., Edwards, D. J., Girerd, P. H., Parikh, H. I., ... others
1195 (2019). The vaginal microbiome and preterm birth. *Nature medicine*, 25(6), 1012–1021.
- 1196 Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and
1197 the number of individuals in a random sample of an animal population. *The Journal of Animal*

- 1198 *Ecology*, 42–58.
- 1199 Flanagan, L., Schmid, J., Ebert, M., Soucek, P., Kunicka, T., Liska, V., ... others (2014). Fusobacterium
1200 nucleatum associates with stages of colorectal neoplasia development, colorectal cancer and disease
1201 outcome. *European journal of clinical microbiology & infectious diseases*, 33, 1381–1390.
- 1202 Fortenberry, J. D. (2013). The uses of race and ethnicity in human microbiome research. *Trends in
1203 microbiology*, 21(4), 165–166.
- 1204 Francescone, R., Hou, V., & Grivennikov, S. I. (2014). Microbiome, inflammation, and cancer. *The
1205 Cancer Journal*, 20(3), 181–189.
- 1206 Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4),
1207 367–378.
- 1208 Gamin, D. J., Vitali, F. C., De Carli, J. P., Mazzon, R. R., Gomes, B. P., Duque, T. M., & Trentin, M. S.
1209 (2021). Prevalence of red and orange microbial complexes in endodontic-periodontal lesions: a
1210 systematic review and meta-analysis. *Clinical Oral Investigations*, 1–14.
- 1211 Gao, J., Yin, J., Xu, K., Li, T., & Yin, Y. (2019). What is the impact of diet on nutritional diarrhea
1212 associated with gut microbiota in weaning piglets: a system review. *BioMed research international*,
1213 2019(1), 6916189.
- 1214 Ghanavati, R., Akbari, A., Mohammadi, F., Asadollahi, P., Javadi, A., Talebi, M., & Rohani, M. (2020).
1215 Lactobacillus species inhibitory effect on colorectal cancer progression through modulating the
1216 wnt/β-catenin signaling pathway. *Molecular and Cellular Biochemistry*, 470, 1–13.
- 1217 Ghorbani, E., Avan, A., Ryzhikov, M., Ferns, G., Khazaei, M., & Soleimanpour, S. (2022). Role of
1218 lactobacillus strains in the management of colorectal cancer: An overview of recent advances.
1219 *Nutrition*, 103, 111828.
- 1220 Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current
1221 understanding of the human microbiome. *Nature medicine*, 24(4), 392–400.
- 1222 Gini, C. (1912). Variabilità e mutabilità (variability and mutability). *Tipografia di Paolo Cuppini,
1223 Bologna, Italy*, 156.
- 1224 Goldenberg, R. L., Culhane, J. F., Iams, J. D., & Romero, R. (2008). Epidemiology and causes of preterm
1225 birth. *The lancet*, 371(9606), 75–84.
- 1226 Gonçalves, L., Subtil, A., Oliveira, M. R., & de Zea Bermudez, P. (2014). Roc curve estimation: An
1227 overview. *REVSTAT-Statistical journal*, 12(1), 1–20.
- 1228 Good, I. J. (1953). The population frequencies of species and the estimation of population parameters.
1229 *Biometrika*, 40(3-4), 237–264.
- 1230 Goodyear, M. D., Krleza-Jeric, K., & Lemmens, T. (2007). *The declaration of helsinki* (Vol. 335) (No.
1231 7621). British Medical Journal Publishing Group.
- 1232 Haffajee, A., Teles, R., & Socransky, S. (2006). Association of eubacterium nodatum and treponema
1233 denticola with human periodontitis lesions. *Oral microbiology and immunology*, 21(5), 269–282.
- 1234 Hajishengallis, G. (2015). Periodontitis: from microbial immune subversion to systemic inflammation.
1235 *Nature reviews immunology*, 15(1), 30–44.
- 1236 Hamjane, N., Mechita, M. B., Nourouti, N. G., & Barakat, A. (2024). Gut microbiota dysbiosis-associated

- 1237 obesity and its involvement in cardiovascular diseases and type 2 diabetes. a systematic review.
1238 *Microvascular Research*, 151, 104601.
- 1239 Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*,
1240 29(2), 147–160.
- 1241 Hampel, H., Frankel, W. L., Martin, E., Arnold, M., Khanduja, K., Kuebler, P., ... others (2008).
1242 Feasibility of screening for lynch syndrome among patients with colorectal cancer. *Journal of*
1243 *Clinical Oncology*, 26(35), 5783–5788.
- 1244 Han, Y. W. (2015). Fusobacterium nucleatum: a commensal-turned pathogen. *Current opinion in*
1245 *microbiology*, 23, 141–147.
- 1246 Han, Y. W., & Wang, X. (2013). Mobile microbiome: oral bacteria in extra-oral infections and
1247 inflammation. *Journal of dental research*, 92(6), 485–491.
- 1248 Hand, D. J. (2012). Assessing the performance of classification methods. *International Statistical Review*,
1249 80(3), 400–414.
- 1250 Hartstra, A. V., Bouter, K. E., Bäckhed, F., & Nieuwdorp, M. (2015). Insights into the role of the
1251 microbiome in obesity and type 2 diabetes. *Diabetes care*, 38(1), 159–165.
- 1252 Hashemi Goradel, N., Heidarzadeh, S., Jahangiri, S., Farhood, B., Mortezaee, K., Khanlarkhani, N., &
1253 Negahdari, B. (2019). Fusobacterium nucleatum and colorectal cancer: A mechanistic overview.
1254 *Journal of Cellular Physiology*, 234(3), 2337–2344.
- 1255 Heip, C. (1974). A new index measuring evenness. *Journal of the Marine Biological Association of the*
1256 *United Kingdom*, 54(3), 555–557.
- 1257 Helmink, B. A., Khan, M. W., Hermann, A., Gopalakrishnan, V., & Wargo, J. A. (2019). The microbiome,
1258 cancer, and cancer therapy. *Nature medicine*, 25(3), 377–388.
- 1259 Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2),
1260 427–432.
- 1261 Hiranmayi, K. V., Sirisha, K., Rao, M. R., & Sudhakar, P. (2017). Novel pathogens in periodontal
1262 microbiology. *Journal of Pharmacy and Bioallied Sciences*, 9(3), 155–163.
- 1263 Honda, K., & Littman, D. R. (2012). The microbiome in infectious disease and inflammation. *Annual*
1264 *review of immunology*, 30(1), 759–795.
- 1265 Honest, H., Forbes, C., Durée, K., Norman, G., Duffy, S., Tsourapas, A., ... others (2009). Screening to
1266 prevent spontaneous preterm birth: systematic reviews of accuracy and effectiveness literature with
1267 economic modelling. *Health Technol Assess*, 13(43), 1–627.
- 1268 Hong, Y. M., Lee, J., Cho, D. H., Jeon, J. H., Kang, J., Kim, M.-G., ... J. K. (2023). Predicting preterm
1269 birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1), 21105.
- 1270 Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations.
1271 *International journal of data mining & knowledge management process*, 5(2), 1.
- 1272 Huang, R.-Y., Lin, C.-D., Lee, M.-S., Yeh, C.-L., Shen, E.-C., Chiang, C.-Y., ... Fu, E. (2007). Mandibular
1273 disto-lingual root: a consideration in periodontal therapy. *Journal of periodontology*, 78(8), 1485–
1274 1490.
- 1275 Iams, J. D., & Berghella, V. (2010). Care for women with prior preterm birth. *American journal of*

- 1276 *obstetrics and gynecology*, 203(2), 89–100.
- 1277 Ide, M., & Papapanou, P. N. (2013). Epidemiology of association between maternal periodontal
1278 disease and adverse pregnancy outcomes—systematic review. *Journal of clinical periodontology*,
1279 40, S181–S194.
- 1280 Iniesta, M., Chamorro, C., Ambrosio, N., Marín, M. J., Sanz, M., & Herrera, D. (2023). Subgingival
1281 microbiome in periodontal health, gingivitis and different stages of periodontitis. *Journal of*
1282 *Clinical Periodontology*, 50(7), 905–920.
- 1283 Inra, J. A., Steyerberg, E. W., Grover, S., McFarland, A., Syngal, S., & Kastrinos, F. (2015). Racial
1284 variation in frequency and phenotypes of apc and mutyh mutations in 6,169 individuals undergoing
1285 genetic testing. *Genetics in Medicine*, 17(10), 815–821.
- 1286 Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44,
1287 223–270.
- 1288 Janda, J. M., & Abbott, S. L. (2007). 16s rrna gene sequencing for bacterial identification in the diagnostic
1289 laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9), 2761–2764.
- 1290 Jiang, W., & Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach
1291 for estimating the prediction error in microarray classification. *Statistics in medicine*, 26(29),
1292 5320–5334.
- 1293 John, G. K., & Mullin, G. E. (2016). The gut microbiome and obesity. *Current oncology reports*, 18,
1294 1–7.
- 1295 Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., . . . others (2019).
1296 Evaluation of 16s rrna gene sequencing for species and strain-level microbiome analysis. *Nature*
1297 *communications*, 10(1), 5029.
- 1298 Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N., & Whiteley, M. (2014). Metatranscriptomics
1299 of the human oral microbiome during health and disease. *MBio*, 5(2), 10–1128.
- 1300 Joscelyn, J., & Kasper, L. H. (2014). Digesting the emerging role for the gut microbiome in central
1301 nervous system demyelination. *Multiple Sclerosis Journal*, 20(12), 1553–1559.
- 1302 Kang, Y., Kang, X., Yang, H., Liu, H., Yang, X., Liu, Q., . . . others (2022). Lactobacillus acidophilus ame-
1303 liorates obesity in mice through modulation of gut microbiota dysbiosis and intestinal permeability.
1304 *Pharmacological research*, 175, 106020.
- 1305 Karched, M., Bhardwaj, R. G., Qudeimat, M., Al-Khabbaz, A., & Ellepola, A. (2022). Proteomic analysis
1306 of the periodontal pathogen prevotella intermedia secretomes in biofilm and planktonic lifestyles.
1307 *Scientific Reports*, 12(1), 5636.
- 1308 Katz, J., Chegini, N., Shiverick, K., & Lamont, R. (2009). Localization of p. gingivalis in preterm delivery
1309 placenta. *Journal of dental research*, 88(6), 575–578.
- 1310 Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., & Gordon, J. I. (2011). Human nutrition, the
1311 gut microbiome and the immune system. *Nature*, 474(7351), 327–336.
- 1312 Kelly, B. J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J. D., Collman, R. G., . . . Li, H. (2015).
1313 Power and sample-size estimation for microbiome studies using pairwise distances and permanova.
1314 *Bioinformatics*, 31(15), 2461–2468.

- 1315 Kennedy, J., Alexander, P., Taillie, L. S., & Jaacks, L. M. (2024). Estimated effects of reductions in
1316 processed meat consumption and unprocessed red meat consumption on occurrences of type 2
1317 diabetes, cardiovascular disease, colorectal cancer, and mortality in the usa: a microsimulation
1318 study. *The Lancet Planetary Health*, 8(7), e441–e451.
- 1319 Kim, B.-R., Shin, J., Guevarra, R. B., Lee, J. H., Kim, D. W., Seol, K.-H., ... Isaacson, R. E. (2017).
1320 Deciphering diversity indices for a better understanding of microbial communities. *Journal of
1321 Microbiology and Biotechnology*, 27(12), 2089–2093.
- 1322 Kim, C. H. (2018). Immune regulation by microbiome metabolites. *Immunology*, 154(2), 220–229.
- 1323 Kim, E.-H., Kim, S., Kim, H.-J., Jeong, H.-o., Lee, J., Jang, J., ... others (2020). Prediction of chronic
1324 periodontitis severity using machine learning models based on salivary bacterial copy number.
1325 *Frontiers in Cellular and Infection Microbiology*, 10, 571515.
- 1326 Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and
1327 bootstrap. *Computational statistics & data analysis*, 53(11), 3735–3745.
- 1328 Kinane, D. F., Stathopoulou, P. G., & Papapanou, P. N. (2017). Periodontal diseases. *Nature reviews
1329 Disease primers*, 3(1), 1–14.
- 1330 Kindinger, L. M., Bennett, P. R., Lee, Y. S., Marchesi, J. R., Smith, A., Caciato, S., ... MacIntyre,
1331 D. A. (2017). The interaction between vaginal microbiota, cervical length, and vaginal progesterone
1332 treatment for preterm birth risk. *Microbiome*, 5, 1–14.
- 1333 Kogut, M. H., Lee, A., & Santin, E. (2020). Microbiome and pathogen interaction with the immune
1334 system. *Poultry science*, 99(4), 1906–1913.
- 1335 Kostic, A. D., Ojesina, A. I., Pedamallu, C. S., Jung, J., Verhaak, R. G., Getz, G., & Meyerson, M. (2011).
1336 Pathseq: software to identify or discover microbes by deep sequencing of human tissue. *Nature
1337 biotechnology*, 29(5), 393–396.
- 1338 Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification
1339 and combining techniques. *Artificial Intelligence Review*, 26, 159–190.
- 1340 Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., ... Watanabe, T.
1341 (2015). Colorectal cancer. *Nature reviews. Disease primers*, 1, 15065.
- 1342 Lafaurie, G. I., Neuta, Y., Ríos, R., Pacheco-Montealegre, M., Pianeta, R., Castillo, D. M., ... oth-
1343 ers (2022). Differences in the subgingival microbiome according to stage of periodontitis: A
1344 comparison of two geographic regions. *PloS one*, 17(8), e0273523.
- 1345 Lamont, R. J., & Jenkinson, H. F. (2000). Subgingival colonization by porphyromonas gingivalis. *Oral
1346 Microbiology and Immunology: Mini-review*, 15(6), 341–349.
- 1347 Lamont, R. J., Koo, H., & Hajishengallis, G. (2018). The oral microbiota: dynamic communities and
1348 host interactions. *Nature reviews microbiology*, 16(12), 745–759.
- 1349 Leitich, H., & Kaider, A. (2003). Fetal fibronectin—how useful is it in the prediction of preterm birth?
1350 *BJOG: An International Journal of Obstetrics & Gynaecology*, 110, 66–70.
- 1351 Le Leu, R. K., Hu, Y., Brown, I. L., Woodman, R. J., & Young, G. P. (2010). Synbiotic intervention of
1352 bifidobacterium lactis and resistant starch protects against colorectal cancer development in rats.
1353 *Carcinogenesis*, 31(2), 246–251.

- 1354 León, R., Silva, N., Ovalle, A., Chaparro, A., Ahumada, A., Gajardo, M., ... Gamonal, J. (2007).
1355 Detection of porphyromonas gingivalis in the amniotic fluid in pregnant women with a diagnosis
1356 of threatened premature labor. *Journal of periodontology*, 78(7), 1249–1255.
- 1357 Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform.
1358 *bioinformatics*, 25(14), 1754–1760.
- 1359 Li, N., Lu, B., Luo, C., Cai, J., Lu, M., Zhang, Y., ... Dai, M. (2021). Incidence, mortality, survival,
1360 risk factor and screening of colorectal cancer: A comparison among china, europe, and northern
1361 america. *Cancer letters*, 522, 255–268.
- 1362 Li, R., Miao, Z., Liu, Y., Chen, X., Wang, H., Su, J., & Chen, J. (2024). The brain–gut–bone axis in
1363 neurodegenerative diseases: insights, challenges, and future prospects. *Advanced Science*, 11(38),
1364 2307971.
- 1365 Li, X., Yu, D., Wang, Y., Yuan, H., Ning, X., Rui, B., ... Li, M. (2021). The intestinal dysbiosis of
1366 mothers with gestational diabetes mellitus (gdm) and its impact on the gut microbiota of their
1367 newborns. *Canadian Journal of Infectious Diseases and Medical Microbiology*, 2021(1), 3044534.
- 1368 Li, Y., Qian, F., Cheng, X., Wang, D., Wang, Y., Pan, Y., ... Tian, Y. (2023). Dysbiosis of oral microbiota
1369 and metabolite profiles associated with type 2 diabetes mellitus. *Microbiology spectrum*, 11(1),
1370 e03796–22.
- 1371 Lim, J. W., Park, T., Tong, Y. W., & Yu, Z. (2020). The microbiome driving anaerobic digestion and
1372 microbial analysis. In *Advances in bioenergy* (Vol. 5, pp. 1–61). Elsevier.
- 1373 Lin, H., Eggesbø, M., & Peddada, S. D. (2022). Linear and nonlinear correlation estimators unveil
1374 undescribed taxa interactions in microbiome data. *Nature communications*, 13(1), 4946.
- 1375 Lin, H., & Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nature
1376 communications*, 11(1), 3514.
- 1377 Lin, H., & Peddada, S. D. (2024). Multigroup analysis of compositions of microbiomes with covariate
1378 adjustments and repeated measures. *Nature Methods*, 21(1), 83–91.
- 1379 Listgarten, M. A. (1986). Pathogenesis of periodontitis. *Journal of clinical periodontology*, 13(5),
1380 418–425.
- 1381 Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome
1382 medicine*, 8, 1–11.
- 1383 López-Aladid, R., Fernández-Barat, L., Alcaraz-Serrano, V., Bueno-Freire, L., Vázquez, N., Pastor-
1384 Ibáñez, R., ... Torres, A. (2023). Determining the most accurate 16s rrna hypervariable region for
1385 taxonomic identification from respiratory samples. *Scientific reports*, 13(1), 3974.
- 1386 Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for
1387 rna-seq data with deseq2. *Genome biology*, 15, 1–21.
- 1388 Magnúsdóttir, S., & Thiele, I. (2018). Modeling metabolism of the human gut microbiome. *Current
1389 opinion in biotechnology*, 51, 90–96.
- 1390 Magurran, A. E. (2021). Measuring biological diversity. *Current Biology*, 31(19), R1174–R1177.
- 1391 Mandic, M., Safizadeh, F., Niedermaier, T., Hoffmeister, M., & Brenner, H. (2023). Association of
1392 overweight, obesity, and recent weight loss with colorectal cancer risk. *JAMA network Open*, 6(4),

- 1393 e239556–e239556.
- 1394 Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically
1395 larger than the other. *The annals of mathematical statistics*, 50–60.
- 1396 Manolis, A. A., Manolis, T. A., Melita, H., & Manolis, A. S. (2022). Gut microbiota and cardiovascular
1397 disease: symbiosis versus dysbiosis. *Current Medicinal Chemistry*, 29(23), 4050–4077.
- 1398 Martin, C. R., Osadchiy, V., Kalani, A., & Mayer, E. A. (2018). The brain-gut-microbiome axis. *Cellular
1399 and molecular gastroenterology and hepatology*, 6(2), 133–148.
- 1400 Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine
1401 learning. *Journal of Applied Science and Technology Trends*, 1(2), 140–147.
- 1402 Mayer, E. A., Tillisch, K., Gupta, A., et al. (2015). Gut/brain axis and the microbiota. *The Journal of
1403 clinical investigation*, 125(3), 926–938.
- 1404 Melguizo-Rodríguez, L., Costela-Ruiz, V. J., Manzano-Moreno, F. J., Ruiz, C., & Illescas-Montes, R.
1405 (2020). Salivary biomarkers and their application in the diagnosis and monitoring of the most
1406 common oral pathologies. *International journal of molecular sciences*, 21(14), 5173.
- 1407 Merrill, L. C., & Mangano, K. M. (2023). Racial and ethnic differences in studies of the gut microbiome
1408 and osteoporosis. *Current Osteoporosis Reports*, 21(5), 578–591.
- 1409 Miller, C. S., Ding, X., Dawson III, D. R., & Ebersole, J. L. (2021). Salivary biomarkers for discriminating
1410 periodontitis in the presence of diabetes. *Journal of clinical periodontology*, 48(2), 216–225.
- 1411 Morita, T., Yamazaki, Y., Mita, A., Takada, K., Seto, M., Nishinoue, N., ... Maeno, M. (2010). A cohort
1412 study on the association between periodontal disease and the development of metabolic syndrome.
1413 *Journal of periodontology*, 81(4), 512–519.
- 1414 Na, H. S., Kim, S. Y., Han, H., Kim, H.-J., Lee, J.-Y., Lee, J.-H., & Chung, J. (2020). Identification of
1415 potential oral microbial biomarkers for the diagnosis of periodontitis. *Journal of clinical medicine*,
1416 9(5), 1549.
- 1417 Nemoto, T., Shiba, T., Komatsu, K., Watanabe, T., Shimogishi, M., Shibasaki, M., ... others (2021).
1418 Discrimination of bacterial community structures among healthy, gingivitis, and periodontitis
1419 statuses through integrated metatranscriptomic and network analyses. *Msystems*, 6(6), e00886–21.
- 1420 Nesbitt, M. J., Reynolds, M. A., Shiau, H., Choe, K., Simonsick, E. M., & Ferrucci, L. (2010). Association
1421 of periodontitis and metabolic syndrome in the baltimore longitudinal study of aging. *Aging clinical
1422 and experimental research*, 22, 238–242.
- 1423 Network, C. G. A., et al. (2012). Comprehensive molecular characterization of human colon and rectal
1424 cancer. *Nature*, 487(7407), 330.
- 1425 Nibali, L., Sousa, V., Davrandi, M., Spratt, D., Alyahya, Q., Dopico, J., & Donos, N. (2020). Differences
1426 in the periodontal microbiome of successfully treated and persistent aggressive periodontitis.
1427 *Journal of Clinical Periodontology*, 47(8), 980–990.
- 1428 Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Ž., & Tomović, M. (2017). Evaluation of classification
1429 models in machine learning. *Theory and Applications of Mathematics & Computer Science*, 7(1),
1430 39.
- 1431 Obuchowski, N. A., & Bullen, J. A. (2018). Receiver operating characteristic (roc) curves: review of

- 1432 methods with applications in diagnostic medicine. *Physics in Medicine & Biology*, 63(7), 07TR01.
- 1433 Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in japan and its neighbouring
1434 regions. *Bulletin of Japanese Society of Scientific Fisheries*, 22, 526–530.
- 1435 Offenbacher, S., Katz, V., Fertik, G., Collins, J., Boyd, D., Maynor, G., ... Beck, J. (1996). Periodontal
1436 infection as a possible risk factor for preterm low birth weight. *Journal of periodontology*, 67,
1437 1103–1113.
- 1438 Ojesina, A. I., Pedamallu, C. S., Kostic, A., Jung, J., Auclair, D., Lohr, J., ... Meyerson, M. (2013). High
1439 throughput sequencing-based pathogen discovery in multiple myeloma. *Blood*, 122(21), 5322.
- 1440 Omundiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019). Machine learning classification techniques
1441 for breast cancer diagnosis. In *Iop conference series: materials science and engineering* (Vol. 495,
1442 p. 012033).
- 1443 O'Sullivan, D. E., Sutherland, R. L., Town, S., Chow, K., Fan, J., Forbes, N., ... Brenner, D. R. (2022).
1444 Risk factors for early-onset colorectal cancer: a systematic review and meta-analysis. *Clinical
1445 gastroenterology and hepatology*, 20(6), 1229–1240.
- 1446 Paganini, D., & Zimmermann, M. B. (2017). The effects of iron fortification and supplementation on the
1447 gut microbiome and diarrhea in infants and children: a review. *The American journal of clinical
1448 nutrition*, 106, 1688S–1693S.
- 1449 Pan, A. Y. (2021). Statistical analysis of microbiome data: the challenge of sparsity. *Current Opinion in
1450 Endocrine and Metabolic Research*, 19, 35–40.
- 1451 Papapanou, P. N., Sanz, M., Buduneli, N., Dietrich, T., Feres, M., Fine, D. H., ... others (2018).
1452 Periodontitis: Consensus report of workgroup 2 of the 2017 world workshop on the classification of
1453 periodontal and peri-implant diseases and conditions. *Journal of periodontology*, 89, S173–S182.
- 1454 Parizadeh, M., & Arrieta, M.-C. (2023). The global human gut microbiome: genes, lifestyles, and diet.
1455 *Trends in Molecular Medicine*.
- 1456 Park, J., Park, S. H., Lee, D., Lee, J. E., Lee, D., Na, K. J., ... Im, H.-J. (2024). Detecting cancer microbiota
1457 using unmapped rna reads on spatial transcriptomics. *Cancer Research*, 84(6_Supplement), 4881–
1458 4881.
- 1459 Payne, M. S., Newnham, J. P., Doherty, D. A., Furfaro, L. L., Pendal, N. L., Loh, D. E., & Keelan, J. A.
1460 (2021). A specific bacterial dna signature in the vagina of australian women in midpregnancy
1461 predicts high risk of spontaneous preterm birth (the predict1000 study). *American journal of
1462 obstetrics and gynecology*, 224(2), 206–e1.
- 1463 Peirce, J. M., & Alviña, K. (2019). The role of inflammation and the gut microbiome in depression and
1464 anxiety. *Journal of neuroscience research*, 97(10), 1223–1241.
- 1465 Peltomaki, P. (2003). Role of dna mismatch repair defects in the pathogenesis of human cancer. *Journal
1466 of clinical oncology*, 21(6), 1174–1179.
- 1467 Pezzino, S., Sofia, M., Greco, L. P., Litrico, G., Filippello, G., Sarvà, I., ... Latteri, S. (2023). Microbiome
1468 dysbiosis: a pathological mechanism at the intersection of obesity and glaucoma. *International
1469 Journal of Molecular Sciences*, 24(2), 1166.
- 1470 Premaraj, T. S., Vella, R., Chung, J., Lin, Q., Hunter, P., Underwood, K., ... Zhou, Y. (2020). Ethnic

- 1471 variation of oral microbiota in children. *Scientific reports*, 10(1), 14788.
- 1472 Raut, J. R., Schöttker, B., Holleczek, B., Guo, F., Bhardwaj, M., Miah, K., ... Brenner, H. (2021).
1473 A microrna panel compared to environmental and polygenic scores for colorectal cancer risk
1474 prediction. *Nature Communications*, 12(1), 4811.
- 1475 Rebersek, M. (2021). Gut microbiome and its role in colorectal cancer. *BMC cancer*, 21(1), 1325.
- 1476 Redanz, U., Redanz, S., Treerat, P., Prakasam, S., Lin, L.-J., Merritt, J., & Kreth, J. (2021). Differential
1477 response of oral mucosal and gingival cells to corynebacterium durum, streptococcus sanguinis, and
1478 porphyromonas gingivalis multispecies biofilms. *Frontiers in cellular and infection microbiology*,
1479 11, 686479.
- 1480 Relvas, M., Regueira-Iglesias, A., Balsa-Castro, C., Salazar, F., Pacheco, J., Cabral, C., ... Tomás, I.
1481 (2021). Relationship between dental and periodontal health status and the salivary microbiome:
1482 bacterial diversity, co-occurrence networks and predictive models. *Scientific reports*, 11(1), 929.
- 1483 Renson, A., Jones, H. E., Beghini, F., Segata, N., Zolnik, C. P., Usyk, M., ... others (2019). Sociodemo-
1484 graphic variation in the oral microbiome. *Annals of epidemiology*, 35, 73–80.
- 1485 Renvert, S., & Persson, G. (2002). A systematic review on the use of residual probing depth, bleeding on
1486 probing and furcation status following initial periodontal therapy to predict further attachment and
1487 tooth loss. *Journal of clinical periodontology*, 29, 82–89.
- 1488 Rideout, J. R., Caporaso, G., Bolyen, E., McDonald, D., Baeza, Y. V., Alastuey, J. C., ... Sharma, K.
1489 (2018, December). *biocore/scikit-bio: scikit-bio 0.5.5: More compositional methods added*. Zenodo.
1490 Retrieved from <https://doi.org/10.5281/zenodo.2254379> doi: 10.5281/zenodo.2254379
- 1491 Rôças, I. N., Siqueira Jr, J. F., Santos, K. R., Coelho, A. M., & de Janeiro, R. (2001). “red complex” (bacteroides forsythus, porphyromonas gingivalis, and treponema denticola) in endodontic
1492 infections: a molecular approach. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology,*
1493 *and Endodontology*, 91(4), 468–471.
- 1494 Romero, R., Dey, S. K., & Fisher, S. J. (2014). Preterm labor: one syndrome, many causes. *Science*,
1495 345(6198), 760–765.
- 1496 Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., ... others (2014). The
1497 composition and stability of the vaginal microbiota of normal pregnant women is different from
1498 that of non-pregnant women. *Microbiome*, 2, 1–19.
- 1499 Rosan, B., & Lamont, R. J. (2000). Dental plaque formation. *Microbes and infection*, 2(13), 1599–1607.
- 1500 Schwabe, R. F., & Jobin, C. (2013). The microbiome and cancer. *Nature Reviews Cancer*, 13(11),
1501 800–812.
- 1502 Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011).
1503 Metagenomic biomarker discovery and explanation. *Genome biology*, 12, 1–18.
- 1504 Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A
1505 survey and review. In *Emerging technology in modelling and graphics: Proceedings of iem graph*
1506 2018 (pp. 99–111).
- 1507 Sepich-Poore, G. D., Zitvogel, L., Straussman, R., Hasty, J., Wargo, J. A., & Knight, R. (2021). The
1508 microbiome and human cancer. *Science*, 371(6536), eabc4552.

- 1510 Sharma, S., & Tripathi, P. (2019). Gut microbiome and type 2 diabetes: where we are and where to go?
1511 *The Journal of nutritional biochemistry*, 63, 101–108.
- 1512 Shi, N., Li, N., Duan, X., & Niu, H. (2017). Interaction between the gut microbiome and mucosal
1513 immune system. *Military Medical Research*, 4, 1–7.
- 1514 Simpson, E. (1949). Measurement of diversity. *Nature*, 163.
- 1515 Sokal, R. R., & Sneath, P. H. (1963). Principles of numerical taxonomy.
- 1516 Song, M., Chan, A. T., & Sun, J. (2020). Influence of the gut microbiome, diet, and environment on risk
1517 of colorectal cancer. *Gastroenterology*, 158(2), 322–340.
- 1518 Söreide, K., Janssen, E., Söiland, H., Körner, H., & Baak, J. (2006). Microsatellite instability in colorectal
1519 cancer. *Journal of British Surgery*, 93(4), 395–406.
- 1520 Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on
1521 similarity of species content and its application to analyses of the vegetation on danish commons.
1522 *Biologiske skrifter*, 5, 1–34.
- 1523 Sotiriadis, A., Papatheodorou, S., Kavvadias, A., & Makrydimas, G. (2010). Transvaginal cervical
1524 length measurement for prediction of preterm birth in women with threatened preterm labor: a
1525 meta-analysis. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International
1526 Society of Ultrasound in Obstetrics and Gynecology*, 35(1), 54–64.
- 1527 Spss, I., et al. (2011). Ibm spss statistics for windows, version 20.0. *New York: IBM Corp*, 440, 394.
- 1528 Stafford, G., Roy, S., Honma, K., & Sharma, A. (2012). Sialic acid, periodontal pathogens and tannerella
1529 forsythia: stick around and enjoy the feast! *Molecular Oral Microbiology*, 27(1), 11–22.
- 1530 Stout, M. J., Conlon, B., Landeau, M., Lee, I., Bower, C., Zhao, Q., ... Mysorekar, I. U. (2013).
1531 Identification of intracellular bacteria in the basal plate of the human placenta in term and preterm
1532 gestations. *American journal of obstetrics and gynecology*, 208(3), 226–e1.
- 1533 Strong, W. (2002). Assessing species abundance unevenness within and between plant communities.
1534 *Community Ecology*, 3(2), 237–246.
- 1535 Sultan, S., El-Mowafy, M., Elgaml, A., Ahmed, T. A., Hassan, H., & Mottawea, W. (2021). Metabolic
1536 influences of gut microbiota dysbiosis on inflammatory bowel disease. *Frontiers in physiology*, 12,
1537 715506.
- 1538 Suzuki, N., Nakano, Y., Yoneda, M., Hirofumi, T., & Hanioka, T. (2022). The effects of cigarette
1539 smoking on the salivary and tongue microbiome. *Clinical and Experimental Dental Research*, 8(1),
1540 449–456.
- 1541 Swidsinski, A., Khilkin, M., Kerjaschki, D., Schreiber, S., Ortner, M., Weber, J., & Lochs, H. (1998).
1542 Association between intraepithelial escherichia coli and colorectal cancer. *Gastroenterology*,
1543 115(2), 281–286.
- 1544 Swift, D., Cresswell, K., Johnson, R., Stilianoudakis, S., & Wei, X. (2023). A review of normalization
1545 and differential abundance methods for microbiome counts data. *Wiley Interdisciplinary Reviews:
1546 Computational Statistics*, 15(1), e1586.
- 1547 Tanner, A. C., Kent Jr, R., Kanasi, E., Lu, S. C., Paster, B. J., Sonis, S. T., ... Van Dyke, T. E. (2007).
1548 Clinical characteristics and microbiota of progressing slight chronic periodontitis in adults. *Journal*

- 1549 *of clinical periodontology*, 34(11), 917–930.
- 1550 Tanner, A. C., Paster, B. J., Lu, S. C., Kanasi, E., Kent Jr, R., Van Dyke, T., & Sonis, S. T. (2006).
1551 Subgingival and tongue microbiota during early periodontitis. *Journal of dental research*, 85(4),
1552 318–323.
- 1553 Tejeda, M., Farrell, J., Zhu, C., Haines, J. L., Wang, L.-S., Schellenberg, G. D., . . . others (2021). Multiple
1554 viruses detected in human dna are associated with alzheimer disease risk. *Alzheimer's & Dementia*,
1555 17, e054585.
- 1556 Teles, F., Wang, Y., Hajishengallis, G., Hasturk, H., & Marchesan, J. T. (2021). Impact of systemic
1557 factors in shaping the periodontal microbiome. *Periodontology 2000*, 85(1), 126–160.
- 1558 Thaiss, C. A., Zmora, N., Levy, M., & Elinav, E. (2016). The microbiome and innate immunity. *Nature*,
1559 535(7610), 65–74.
- 1560 Tian, R., Liu, H., Feng, S., Wang, H., Wang, Y., Wang, Y., . . . Zhang, S. (2021). Gut microbiota dysbiosis
1561 in stable coronary artery disease combined with type 2 diabetes mellitus influences cardiovascular
1562 prognosis. *Nutrition, Metabolism and Cardiovascular Diseases*, 31(5), 1454–1466.
- 1563 Tilg, H., Kaser, A., et al. (2011). Gut microbiome, obesity, and metabolic dysfunction. *The Journal of
1564 clinical investigation*, 121(6), 2126–2132.
- 1565 Tonetti, M. S., Greenwell, H., & Kornman, K. S. (2018). Staging and grading of periodontitis: Framework
1566 and proposal of a new classification and case definition. *Journal of periodontology*, 89, S159–S172.
- 1567 Tringe, S. G., & Hugenholtz, P. (2008). A renaissance for the pioneering 16s rRNA gene. *Current opinion
1568 in microbiology*, 11(5), 442–446.
- 1569 Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., . . . others (2017). A
1570 guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological
1571 Reviews*, 92(2), 698–715.
- 1572 Ulger Toprak, N., Yagci, A., Gulluoglu, B., Akin, M., Demirkalem, P., Celenk, T., & Soyletir, G. (2006).
1573 A possible role of bacteroides fragilis enterotoxin in the aetiology of colorectal cancer. *Clinical
1574 microbiology and infection*, 12(8), 782–786.
- 1575 Ursell, L. K., Metcalf, J. L., Parfrey, L. W., & Knight, R. (2012). Defining the human microbiome.
1576 *Nutrition reviews*, 70(suppl_1), S38–S44.
- 1577 Utzschneider, K. M., Kratz, M., Damman, C. J., & Hullarg, M. (2016). Mechanisms linking the gut
1578 microbiome and glucose metabolism. *The Journal of Clinical Endocrinology & Metabolism*,
1579 101(4), 1445–1454.
- 1580 Vander Haar, E. L., So, J., Gyamfi-Bannerman, C., & Han, Y. W. (2018). Fusobacterium nucleatum and
1581 adverse pregnancy outcomes: epidemiological and mechanistic evidence. *Anaerobe*, 50, 55–59.
- 1582 Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning
1583 research*, 9(11).
- 1584 Vasen, H. F., Mecklin, J.-P., Khan, P. M., & Lynch, H. T. (1991). The international collaborative group
1585 on hereditary non-polyposis colorectal cancer (icg-hnppcc). *Diseases of the Colon & Rectum*, 34(5),
1586 424–425.
- 1587 Vilar, E., & Gruber, S. B. (2010). Microsatellite instability in colorectal cancer—the stable evidence.

- 1588 *Nature reviews Clinical oncology*, 7(3), 153–162.
- 1589 Walker, M. A., Pedamallu, C. S., Ojesina, A. I., Bullman, S., Sharpe, T., Whelan, C. W., & Meyerson, M.
1590 (2018). Gatk pathseq: a customizable computational tool for the discovery and identification of
1591 microbial sequences in libraries from eukaryotic hosts. *Bioinformatics*, 34(24), 4287–4289.
- 1592 Weaver, W. (1963). *The mathematical theory of communication*. University of Illinois Press.
- 1593 Whiteside, S. A., Razvi, H., Dave, S., Reid, G., & Burton, J. P. (2015). The microbiome of the urinary
1594 tract—a role beyond infection. *Nature Reviews Urology*, 12(2), 81–90.
- 1595 Witkin, S. (2019). Vaginal microbiome studies in pregnancy must also analyse host factors. *BJOG: An
1596 International Journal of Obstetrics & Gynaecology*, 126(3), 359–359.
- 1597 Wong, T.-T., & Yeh, P.-Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE
1598 Transactions on Knowledge and Data Engineering*, 32(8), 1586–1594.
- 1599 Wyss, C., Moter, A., Choi, B.-K., Dewhirst, F., Xue, Y., Schüpbach, P., ... Guggenheim, B. (2004).
1600 Treponema putidum sp. nov., a medium-sized proteolytic spirochaete isolated from lesions of
1601 human periodontitis and acute necrotizing ulcerative gingivitis. *International journal of systematic
1602 and evolutionary microbiology*, 54(4), 1117–1122.
- 1603 Xia, Y. (2023). Statistical normalization methods in microbiome data with application to microbiome
1604 cancer research. *Gut Microbes*, 15(2), 2244139.
- 1605 Yaman, E., & Subasi, A. (2019). Comparison of bagging and boosting ensemble machine learning methods
1606 for automated emg signal classification. *BioMed research international*, 2019(1), 9152506.
- 1607 Yang, I., Claussen, H., Arthur, R. A., Hertzberg, V. S., Geurs, N., Corwin, E. J., & Dunlop, A. L. (2022).
1608 Subgingival microbiome in pregnancy and a potential relationship to early term birth. *Frontiers in
1609 cellular and infection microbiology*, 12, 873683.
- 1610 Yildiz, B., Bilbao, J. I., & Sproul, A. B. (2017). A review and analysis of regression and machine learning
1611 models on commercial building electricity load forecasting. *Renewable and Sustainable Energy
1612 Reviews*, 73, 1104–1122.
- 1613 Yoshimura, F., Murakami, Y., Nishikawa, K., Hasegawa, Y., & Kawaminami, S. (2009). Surface
1614 components of *porphyromonas gingivalis*. *Journal of periodontal research*, 44(1), 1–12.
- 1615 Zhang, C.-Z., Cheng, X.-Q., Li, J.-Y., Zhang, P., Yi, P., Xu, X., & Zhou, X.-D. (2016). Saliva in the
1616 diagnosis of diseases. *International journal of oral science*, 8(3), 133–137.
- 1617 Zhou, X., Wang, L., Xiao, J., Sun, J., Yu, L., Zhang, H., ... others (2022). Alcohol consumption,
1618 dna methylation and colorectal cancer risk: Results from pooled cohort studies and mendelian
1619 randomization analysis. *International journal of cancer*, 151(1), 83–94.
- 1620 Zhu, W., & Lee, S.-W. (2016). Surface interactions between two of the main periodontal pathogens:
1621 *Porphyromonas gingivalis* and *tannerella forsythia*. *Journal of periodontal & implant science*,
1622 46(1), 2–9.
- 1623 Zhu, X., Han, Y., Du, J., Liu, R., Jin, K., & Yi, W. (2017). Microbiota-gut-brain axis and the central
1624 nervous system. *Oncotarget*, 8(32), 53829.
- 1625 Zhuang, Y., Wang, H., Jiang, D., Li, Y., Feng, L., Tian, C., ... others (2021). Multi gene mutation
1626 signatures in colorectal cancer patients: predict for the diagnosis, pathological classification, staging

¹⁶²⁷

and prognosis. *BMC cancer*, 21, 1–16.

Acknowledgments

1629 I would like to disclose my earnest appreciation for my advisor, Professor **Semin Lee**, who provided
 1630 solicitous supervision and cherished opportunities throughout the course of my research. His advice and
 1631 consultation encouraged me to become as a researcher and to receive all humility and gentleness. I am also
 1632 grateful to all of my committee members, Professor **Taejoon Kwon**, Professor **Eunhee Kim**, Professor
 1633 **Kyemyung Park**, and Professor **Min Hyuk Lim**, for their meaningful mentions and suggestions.

1634 I extend my deepest gratitude to my Lord, **the Flying Spaghetti Monster**, His Noodly Appendage
 1635 has guided me through the twist and turns of this academic journey. His presence, ever comforting and
 1636 mysterious, has been a source of strength and humor during both highs and lows. In moments of doubt, I
 1637 found solace in the belief that you were there, gently reminding me to keep faith in the process. His Holy
 1638 Noodle has nourished my mind, and for that, I am truly overwhelmed. May His Holy Noodle continue to
 1639 guide me in all my future endeavors. *R’Amen.*

1640 I would like to extend my heartfelt gratitude to Professor **You Mi Hong** for her invaluable guidance
 1641 and insightful advice on PTB study. Her expertise in maternal and fetal health, along with her deep under-
 1642 standing of statistical and clinical interpretations, greatly contributed to refining the analytical framework
 1643 of this study. Her constructive feedback and thoughtful discussions provided critical perspectives that
 1644 enhanced the robustness and relevance of the research findings. I sincerely appreciate her generosity
 1645 in sharing her knowledge and effort, as well as her encouragement throughout my Ph.D. journey. Her
 1646 support has been instrumental in strengthening this work, and I am truly grateful for her contributions.

1647 I also would like to express my sincere gratitude for Professor **Jun Hyeok Lim** for his invaluable
 1648 guidance and insightful advice on lung cancer study. His expertise in cancer genomics and data interpreta-
 1649 tion provided essential perspectives that greatly enriched the analytical approach of my Ph.D. journey. His
 1650 constructive feedback and thoughtful discussion helped refine methodologies and enhance the scientific
 1651 rigor of the research. I deeply appreciate his willingness to share his knowledge and expertise, which has
 1652 been instrumental in shaping key aspects of this work. His support and encouragement have been truly
 1653 inspiring, and I am grateful for the opportunity to have benefited from his mentorship.

1654 I would like to extend my heartfelt gratitude to my colleagues of the **Computational Biology Lab @**
 1655 **UNIST**, whose collaboration, friendship, brotherhood, and support have been an invaluable part of my
 1656 journey. Your willingness to share insights, engage in thoughtful discussions, and offer encouragement
 1657 during the challenging moments of research has significantly shaped my academic experience. The
 1658 camaraderie in Computational Biology Lab made even the most demanding days more enjoyable, and I
 1659 am deeply grateful for the collaborative environment we created together. I appreciate you for standing
 1660 by my side throughout this Ph.D. journey.

1661 I would like to express my heartfelt gratitude to **my family**, whose unwavering support has been the
 1662 foundation of everything I have achieved. Your love, encouragement, and belief in me have sustained me
 1663 through every challenge, and I could not have come this far without you. From your words of wisdom to
 1664 your patience and understanding, each of you has played a vital role in helping me navigate this journey.
 1665 The strength and comfort I have drawn from our family bond have been my greatest source of resilience.

1666 Your presence, both near and far, has filled my life with warmth and motivation. I am deeply grateful for
1667 your unconditional love and for always being there when I needed you the most. Thank you for being my
1668 constant source of strength and inspiration.

1669 I am incredibly pleased to my friends, especially my GSHS alumni (이망특), for their unwavering
1670 support and encouragement throughout this journey. The bonds we formed back in our school days have
1671 only grown stronger over the years, and I am fortunate to have had such loyal and understanding friends
1672 by my side. Your constant words of motivation, and even moments of levity during stressful times have
1673 helped keep me grounded. Whether it was a late-night conversations, a shared laugh, or a simple message
1674 of reassurance, you all have played a vital role in keeping me focused and motivated. I am relieved for the
1675 ways you celebrated each small achievement with me and how you patiently listened to my worries. The
1676 memories of our shared past provided me with comfort and a sense of stability when the road ahead felt
1677 uncertain. I could not have reached this point without the love and friendship that you all have generously
1678 given. Each of your, in your unique way, has contributed to this dissertation, even if indirectly, and for
1679 that, I am forever beholden. I look forward to continuing our friendship as we all grow in our individual
1680 paths, knowing that the support we share is something truly special.

1681 I would like to express my deepest recognition to **my girlfriend (expected)** for her unwavering
1682 support, patience, and companionship throughout my Ph.D. journey. Her presence has been a constant
1683 source of comfort and motivation, helping me navigate the challenges of research and writing with
1684 renewed energy. Through moments of frustration and accomplishment alike, her encouragement has
1685 reminded me of the importance of balance and perseverance. Her kindness, understanding, and belief
1686 in me have been invaluable, making even the most difficult days feel lighter. I am truly grateful for her
1687 support and for sharing this journey with me, and I look forward to all the moments we will continue to
1688 experience together.

1689 I would like to express my sincere gratitude to the amazing members of my animal protection groups,
1690 DRDR (두루두루) and UNIMALS (유니멀스), whose dedication and compassion have been a constant
1691 source of motivation. Your unwavering commitment to improving the lives of animals has inspired me
1692 throughout this journey. I am also thankful for the beautiful cats we have cared for, whose presence
1693 brought both joy and purpose to our allegiance. Their playful spirits and gentle companionship served as
1694 daily reminders of why we continue to fight for animal rights. The bond we share, both with each other
1695 and with the animals we protect, has enriched my life in countless ways. I appreciate you all again for
1696 your support, dedication, and for being part of this meaningful cause.

1697 I would like to express my deepest gratitude to **everyone** I have had the honor of meeting throughout
1698 this journey. Your kindness, encouragement, and support have carried me through both the challenging
1699 and rewarding moments of my life. Whether through a kind word, thoughtful advice, or simply being
1700 there when I needed it most, your presence has made all the difference. I am incredibly fortunate to have
1701 received such generosity and warmth from those around me, and I do not take it for granted. Every act
1702 of kindness, no matter how big or small, has been a source of strength and motivation for me. To all
1703 my friends, colleagues, mentors, and beloved ones, thank you for your unwavering support. I am truly
1704 grateful for each of you, and your kindness has left an indelible mark on my journey.

1705 My Lord, *the Flying Spaghetti Monster*,
1706 give us grace to accept with serenity the things that cannot be changed,
1707 courage to change the things that should be changed,
1708 and the wisdom to distinguish the one from the other.

1709
1710 Glory be to *the Meatball*, to *the Sauce*, and to *the Holy Noodle*.
1711 As it was in the beginning, is now, and ever shall be.

1712 *R'Amen.*



May your progress be evident to all

