

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

Microbiota in Human Diseases

A thesis/dissertation submitted to
Ulsan National Institute of Science and Technology
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Jaewoong Lee

04.16.2025 of submission

Approved by

Advisor

Semin Lee

Microbiota in Human Diseases

Jaewoong Lee

This certifies that the thesis/dissertation of Jaewoong Lee is approved.

04.16.2025 of submission

Signature

Advisor: Semin Lee

Signature

Taejoon Kwon

Signature

Eunhee Kim

Signature

Kyemyung Park

Signature

Min Hyuk Lim

Abstract

14 The human microbiome plays a critical role in diseases, influencing immune response, metabolism,
15 and disease progression. Recent advances in microbiome sequencing techniques have highlighted its
16 potential as a diagnostic, prognostic, and therapeutic strategies in various diseases, including preterm
17 birth (Section 2), periodontitis (Section 3), and colorectal cancer (Section 4). Dysbiosis, characterized
18 by alterations in microbiome composition, has been linked to pathogenesis, disease progression, and
19 treatment outcome, emphasizing the need for comprehensive metagenomic analyses. By investigating
20 microbiome profiling, researchers can uncover microbial biomarkers and host-microbiome interactions
21 that contribute to underlying mechanisms of disease. Thus, understanding these complex relationships
22 not only enhances early detection and risk stratification but also paves the way for microbiome-based
23 therapeutic interventions and personalized medicine strategies. Ultimately, as microbiome research
24 continues to evolve, its integration with genomics, metabolomics, and immunology suggests promise for
25 transforming disease management and improving treatment outcomes.

26 Section 2 investigated the association between the prenatal salivary microbiome and preterm birth
27 using 16S rRNA gene sequencing and developed a random forest-based prediction model for risk of
28 preterm birth. A total of 59 pregnant women were included as the study participants, with 30 in the
29 preterm birth group and 29 in the full-term birth group. Salivary microbiome samples were collected via
30 mouthwash within 25 hours before delivery, and 16S rRNA gene sequencing was performed to analyze
31 microbial taxonomic composition. Differentially abundant taxa were identified by DESeq2, revealing the
32 25 significant taxa, including three PTB-enriched taxa and 22 FTB-enriched taxa, suggesting distinct
33 microbial differences between the two groups. A random forest classifier was applied to predict PTB
34 risk based on salivary microbiome composition, achieving the high balanced accuracy (0.765 ± 0.071)
35 using the nine most important taxa. These findings indicate that salivary microbiome profiling may
36 serve as a novel predictive tool for PTB risk assessment, complementing existing clinical predictors.
37 However, further multi-center and large-scale studies are necessary to validate these findings before
38 clinical application.

39 Section 3 characterized salivary microbiome compositions to classify periodontal health and different
40 stages of periodontitis using 16S rRNA gene sequencing. A total of 250 study participants were included,
41 comprising 100 periodontally healthy controls and 150 periodontitis patients equally classified into stage
42 I, stage II, and stage III. Microbial diversity indices were calculated, and ANCOM was used to identify 20
43 differentially abundant taxa among the multiple periodontitis stages. A random forest machine learning
44 model was developed to classify periodontitis stages based on the proportions of differentially abundant
45 taxa, achieving an area-under-curve of 0.870 ± 0.079 . Among the identified differentially abundant taxa,
46 *Porphyromonas gingivalis* and *Actinomyces* spp. were the most important features in distinguishing
47 periodontitis stages. Random forest classifier also effectively distinguished healthy individuals from stage
48 I periodontitis with an area-under-curve of 0.852 ± 0.103 and detected periodontitis patients from healthy

49 controls with an area-under-curve of 0.953pm0.049. External validation with Spanish and Portuguese
50 datasets showed a slight performance decrease, likely due to ethnic variations in salivary microbiome
51 composition, emphasizing the need for population-specific models. These findings suggest that salivary
52 microbiome composition profiling may serve as a non-invasive diagnostic technique for periodontitis,
53 aiding in early detection and personalized dental care.

54 Section 4 conducted a comprehensive metagenomic analysis of colorectal cancer using PathSeq,
55 focusing on key clinical outcomes, including recurrence history and overall survival duration. Significant
56 differences in alpha-diversity and beta-diversity indices were observed between tumor and its adjacent
57 normal tissues, with further stratification revealing distinct microbial diversity patterns associated with
58 recurrence status and survival outcomes. Differentially abundant taxa were identified, highlighting
59 microbial signatures may influence CRC progression and prognosis. To evaluate the predictive potential
60 of these selected differentially abundant taxa, we developed a random forest-based machine learning
61 model for CRC recurrence risk and survival duration. While the classification model for recurrence
62 prediction achieved moderate accuracy (about 60%), and the regression model of OS duration showed
63 moderated errors (about 700 days), these results suggest that gut microbiome composition alone may not
64 be sufficient for personalized clinical predictions. These findings emphasize the need for multi-omics
65 integration, combining host genomic alterations, *e.g.* somatic and germline mutations with gut microbiome
66 compositions, to improve CRC risk stratification and personalized medicine applications. This study
67 highlights the potential role of gut microbiome for biomarkers in CRC diagnosis and prognosis while
68 underscoring the complexity of host-microbiome interaction in CRC progression.

69 Together, these studies demonstrate the clinical relevance of microbiome profiling in three distinct
70 yet interconnected diseases by analyzing microbial diversity, identifying differentially abundant taxa, and
71 leveraging machine learning for predictive modeling. While each condition exhibited unique microbial
72 signatures, the findings collectively underscore the broader impact of dysbiosis on pathogenesis and
73 disease progression. These results suggest that microbial biomarkers could serve as valuable tools for
74 early detection, risk assessment, and personalized medicine strategies across multiple disease contexts.
75 However, the predictive performance of machine learning models highlights the requirement for multi-
76 omics integration, incorporating host genomic data to improve the accuracy of disease prediction and
77 personalized therapeutic interventions. Moving forward, further large-scale and multi-cohort validation
78 studies will be essential to refine microbiome-based biomarkers and ensure their clinical applicability in
79 therapeutic guidance. By deepening our understanding of host-microbiome interactions, this dissertation
80 contributes to the growing field of microbiome-driven personalized medicine, paving a novel approaches
81 in disease prevention and management.

82
83 **This doctoral dissertation is an addition based on the following papers that the author has already**
84 **published:**

- 85 • Hong, Y. M., **Lee, Jaewoong**, Cho, D. H., Jeon, J. H., Kang, J., Kim, M. G., ... & Kim, J. K. (2023).
86 Predicting preterm birth using machine learning techniques in oral microbiome. *Scientific Reports*,
87 13(1), 21105.

Contents

89	1	Introduction	1
90	2	Predicting preterm birth using random forest classifier in salivary microbiome	8
91	2.1	Introduction	8
92	2.2	Materials and methods	10
93	2.2.1	Study design and study participants	10
94	2.2.2	Clinical data collection and grouping	10
95	2.2.3	Salivary microbiome sample collection	10
96	2.2.4	16s rRNA gene sequencing	10
97	2.2.5	Bioinformatics analysis	11
98	2.2.6	Data and code availability	11
99	2.3	Results	12
100	2.3.1	Overview of clinical information	12
101	2.3.2	Comparison of salivary microbiomes composition	12
102	2.3.3	Random forest classification to predict PTB risk	12
103	2.4	Discussion	20
104	3	Random forest prediction model for periodontitis statuses based on the salivary microbiomes	22
105	3.1	Introduction	22
106	3.2	Materials and methods	24
107	3.2.1	Study participants enrollment	24
108	3.2.2	Periodontal clinical parameter diagnosis	24
109	3.2.3	Saliva sampling and DNA extraction procedure	26
110	3.2.4	Bioinformatics analysis	26
111	3.2.5	Data and code availability	27
112	3.3	Results	29

113	3.3.1	Summary of clinical information and sequencing data	29
114	3.3.2	Diversity indices reveal differences among the periodontitis severities .	29
115	3.3.3	DAT among multiple periodontitis severities and their correlation . .	29
116	3.3.4	Classification of periodontitis severities by random forest models . .	30
117	3.4	Discussion	51
118	4	Metagenomic signature analysis of Korean colorectal cancer	55
119	4.1	Introduction	55
120	4.2	Materials and methods	57
121	4.2.1	Study participants enrollment	57
122	4.2.2	DNA extraction procedure	57
123	4.2.3	Bioinformatics analysis	57
124	4.2.4	Data and code availability	59
125	4.3	Results	60
126	4.3.1	Summary of clinical characteristics	60
127	4.3.2	Gut microbiome compositions	60
128	4.3.3	Diversity indices	61
129	4.3.4	DAT selection	62
130	4.3.5	Random forest prediction	64
131	4.4	Discussion	82
132	5	Conclusion	88
133	References	90
134	Acknowledgments	109

135

List of Figures

136	1	DAT volcano plot for PTB prediction	14
137	2	Salivary microbiome compositions over DAT for PTB prediction	15
138	3	Random forest-based PTB prediction model	16
139	4	Diversity indices	17
140	5	PROM-related DAT between FTB and PTB	18
141	6	Validation of random forest-based PTB prediction model	19
142	7	Diversity indices for periodontitis	37
143	8	DAT for periodontitis	38
144	9	Correlation heatmap between periodontitis DAT	39
145	10	Random forest classification metrics for periodontitis prediction	40
146	11	Random forest classification metrics from external datasets	41
147	12	Rarefaction curves for alpha-diversity indices	42
148	13	Salivary microbiome compositions in the different periodontal statuses	43
149	14	Correlation plots for periodontitis DAT	44
150	15	Clinical measurements by the periodontitis statuses	45
151	16	Number of read counts by the periodontitis statuses	46
152	17	Proportions of periodontitis DAT	47

153	18	Random forest classification metrics with the full microbiome compositions and ANCOM-selected DAT compositions	48
154			
155	19	Alpha-diversity indices account for evenness	49
156	20	Gradient Boosting classification metrics for periodontitis prediction	50
157	21	Gut microbiome compositions in genus level	72
158	22	Alpha-diversity indices in genus level	73
159	23	Alpha-diversity indices with recurrence in genus level	74
160	24	Alpha-diversity indices with OS in genus level	75
161	25	Beta-diversity indices in genus level	76
162	26	Beta-diversity indices with recurrence in genus level	77
163	27	Beta-diversity indices with recurrence in genus level	78
164	28	DAT with recurrence in species level	79
165	29	DAT with OS in species level	80
166	30	Random forest classification and regression	81

List of Tables

168	1	Confusion matrix	6
169	2	Standard clinical information of PTB study participants	13
170	3	Clinical characteristics of the study participants	32
171	4	Feature combinations and their evaluations	33
172	5	List of DAT among the periodontally healthy and periodontitis stages	34
173	6	Feature the importance of taxa in the classification of different periodontal statuses.	35
174	7	Beta-diversity pairwise comparisons on the periodontitis statuses	36
175	8	Clinical characteristics of CRC study participants	66
176	9	DAT list for CRC recurrence	67
177	10	DAT list for CRC OS	68
178	11	Random forest classification and their evaluations	70
179	12	Random forest regression and their evaluations	71

List of Abbreviations

- 181 **ACC** Accuracy
182 **ACE** Abundance-based coverage estimator
183 **ASV** Amplicon sequence variant
184 **AUC** Area-under-curve
185 **BA** Balanced accuracy
186 **BMI** Body mass index
187 **C-section** Cesarean section
188 **DAT** Differentially abundant taxa
189 **F1** F1 score
190 **Faith PD** Faith's phylogenetic diversity
191 **FC** Fold change
192 **FN** False negative
193 **FP** False positive
194 **FTB** Full-term birth
195 **GA** Gestational age
196 **MAE** Mean absolute error
197 **MSI** Microsatellite instability
198 **MSI-H** MSI-High
199 **MSI-L** MSI-Low
200 **MSS** Microsatellite stable
201 **MWU test** Mann-Whitney U-test
202 **OS** Overall survival
203 **PRE** Precision
204 **PROM** Prelabor rupture of membrane
205 **PTB** Preterm birth

- 206 **qPCR** quantitative-PCR
- 207 **RMSE** Root mean squared error
- 208 **ROC curve** Receiver-operating characteristics curve
- 209 **rRNA** Ribosomal RNA
- 210 **SD** Standard deviation
- 211 **SEN** Sensitivity
- 212 **SPE** Specificity
- 213 **t-SNE** t-distributed stochastic neighbor embedding
- 214 **TN** True negative
- 215 **TP** True positive

216 1 Introduction

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

234 The microbiome participates a crucial role in overall health, influencing not only digestion and immune
235 function but also systemic and neurological processes through the brain-gut axis (Martin, Osadchiy,
236 Kalani, & Mayer, 2018; Aziz & Thompson, 1998; R. Li et al., 2024). The gut microbiota interact with
237 the host through metabolic byproducts, immune signaling, and the production of neurotransmitters, *e.g.*
238 serotonin and dopamine, which are essential for brain function and cognition. Disruptions in microbial
239 composition, known as dysbiosis, have been linked to various diseases, including inflammatory bowel
240 disease (Sultan et al., 2021; Baldelli, Scaldaferrri, Putignani, & Del Chierico, 2021), obesity (Kang et al.,
241 2022; Hamjane, Mechita, Nourouti, & Barakat, 2024; Pezzino et al., 2023), diabetes (Cai et al., 2024;
242 X. Li et al., 2021; Y. Li et al., 2023), and cardiovascular diseases (Manolis, Manolis, Melita, & Manolis,
243 2022; Tian et al., 2021). Furthermore, the brain-gut axis, a bidirectional communication system between
244 the gut microbiome composition and the central nervous system, has been implicated in mental disorders,
245 *e.g.* anxiety disorder, depressive disorder, and neurodegenerative diseases. Emerging evidence suggested
246 that alterations in the host microbiome can influence mood, cognitive function, and even behavior through
247 immune modulation, vagus nerve signaling, and microbial metabolites. These findings highlight the
248 microbiome as a critical factor in maintaining host health and suggest that targeted interventions, namely
249 probiotics, antibiotics, dietary modification, and microbiome-based therapies, may hold promise for
250 improving both physical and mental comfort. Hence, understanding the microbial effects could lead to
251 novel therapeutic strategies for a wide range of health conditions.

252 16S ribosomal RNA (rRNA) gene sequencing is one of the most extensively applied methods for
253 characterizing microbial communities by targeting the conserved 16S rRNA gene, which contains both

254 highly conserved and variable regions in bacteria (Tringe & Hugenholtz, 2008; Janda & Abbott, 2007).
255 The conserved regions enable universal primer binding, while the variable regions provide the specificity
256 needed to differentiate microbial taxa. Among these regions, the V3-V4 region is frequently selected for
257 sequencing due to its balance between phylogenetic resolution and sequencing efficiency (Johnson et al.,
258 2019; López-Aladid et al., 2023). Therefore, the V3-V4 region offers sufficient variability to classify a
259 wide range of bacteria taxa while maintaining compatibility with widely used sequencing platforms.

260 The Anna Karenina principle, originally derived from literature of Leo Tolstoy, has been applied
261 to microbiome research to describe the manner that microbial communities in patients with diseases
262 tend to be more variable and unstable compared to those in healthy individuals (Ma, 2020; W. Li &
263 Yang, 2025). This Anna Karenina principle suggests that while healthy microbiomes exhibit relatively
264 stable and uniform compositions, while disease-associated microbiomes become highly dysregulated due
265 to various environmental, genetic, and pathological influences. Dysbiosis-driven mechanisms, such as
266 inflammation, genotoxic metabolic production, and immune modulation, can contribute pathogenesis
267 and progression of diseases, including periodontitis. In the context of cancer, this Anna Karenina
268 principle suggests that gut microbiome dysbiosis does not follow a single uniform pattern in patients
269 with CRC but rather presents as diverse and individualized disruption in microbial composition. This
270 instability may play a role in field cancerization, where microbial alteration extend beyond the tumor
271 site to adjacent normal-appearing tissues (Curtius, Wright, & Graham, 2018; Rubio, Lang-Schwarz,
272 & Vieth, 2022), potentially priming the tumor microenvironment for malignancy. Therefore, the high
273 inter-individual variability in microbiome alteration across these disease supports the Anna Karenina
274 principle, highlighting the complexity of dysbiosis-driven diseases and the necessity for personalized
275 microbiome-based diagnostic and interventions. Investigating the shared and disease-specific microbial
276 disruptions across these conditions may offer novel insights into microbiome-driven pathogenesis and
277 therapeutic strategies.

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

292 Diversity indices are essential techniques for evaluating the complexity and variety of microbial

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

Differentially abundant taxa (DAT) detection is a key analytical approach in microbiome study to identify microbial taxa that significantly differ in abundance between distinct study participant groups. This DAT detection method is particularly valuable for understanding how microbial communities vary across different conditions, such as disease states, environmental factors, and/or experimental treatments. Various statistical and computational techniques, *e.g.* LEfSe (Segata et al., 2011), DESeq2 (Love, Huber, & Anders, 2014), ANCOM (Lin & Peddada, 2020), and ANCOM-BC (Lin, Eggesbø, & Peddada, 2022; Lin & Peddada, 2024), are commonly used to assess differential abundance while accounting for compositional and sparsity-related challenges in microbiome composition data (Swift, Cresswell, Johnson, Stilianoudakis, & Wei, 2023; Cappellato, Baruzzo, & Di Camillo, 2022). Thus, identifying DAT can 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; Geurts, Ernst, & Wehenkel, 2006). 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,

332 which lowers the possibility of overfitting in comparison to a single decision tree. Furthermore, random
333 forest classifier offers several advantages, including its robustness to outliers and its ability to calculate
334 the feature importance.

335 Furthermore, k -fold cross-validation is a widely applied resampling technique that enhances the
336 reliability and robustness of machine learning models by iteratively evaluating their performance across
337 multiple data partitions (Wong & Yeh, 2019; Ghojogh & Crowley, 2019). Instead of relying on a single
338 train-test split, k -fold cross-validation divides the dataset into equally sized k folds, where the machine
339 learning model is trained on $k - 1$ folds and tested on the remaining fold in an iterative manner. This
340 process is repeated k times, with each fold serving as the test set once, and the final performance is
341 averaged across all iterations to provide a more generalizable estimate of model metrics. By reducing the
342 risk of overfitting and minimizing variance in performance evaluation, k -fold cross-validation ensures
343 that the machine learning model is not overly dependent on a specific train-test split. By applying k -fold
344 cross-validation, researchers can ensure that their machine learning models are both robust and reliable,
345 leading to more accurate and reproducible results (Fushiki, 2011).

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

357 The receiver-operating characteristics (ROC) curve is a graphical representation used to evaluate
358 the performance of a classification model by plotting the sensitivity against (1-specificity) at multiple
359 threshold setting (Gonçalves, Subtil, Oliveira, & de Zea Bermudez, 2014; Obuchowski & Bullen, 2018;
360 Centor, 1991). The ROC curve illustrates the trade-off between detecting true positives while minimizing
361 false positives, suggesting determining the optimal decision threshold for classification. A key metric
362 derived from the ROC curve is the area-under-curve (AUC), which quantifies overall ability of the
363 classification model to discriminate between positive and negative predictions. An AUC value of 0.5
364 indicates a model performing no better than random chance, while value closer to 1.0 suggests high
365 predictive accuracy. Thus, by analyzing the AUC value of the ROC curve, researchers can compare
366 different models and select the better classification model that offers the best balance between sensitivity
367 and specificity for a given application.

368 Regression is a powerful predictive machine learning approach used to analyze complex relationships
369 between variables and make continuous value predictions (Maulud & Abdulazeez, 2020; Yıldız, Bilbao, &
370 Sproul, 2017). Beside classification, which assigns discrete labels, regression models estimate numerical

371 outcomes based on input features, making them particularly useful in biological research and clinical
372 applications for predicting disease risk, patient outcomes, and biomarker selection. By leveraging high-
373 throughput biological techniques and clinical information, regression model enables the discovery of
374 hidden patterns and the development of precision medicine strategies. As computational methods advance,
375 integrating regression models with metagenomic data can improve predictive accuracy and facilitate
376 data-driven therapeutic guide in healthcare.

377 Evaluating the performance of machine learning regression models requires assessing their prediction
378 errors using appropriate metrics. Mean absolute error (MAE; Equation 7) and root mean squared error
379 (RMSE; Equation 8) are commonly used measures for quantifying the accuracy of regression models. By
380 optimizing regression models based on MAE and RMSE, researchers can improve prediction accuracy
381 and enhance the reliability of machine learning regression models.

382 This dissertation present a comprehensive, multi-disease human microbiome analysis, bridging the
383 association between preterm birth (PTB) (Section 2), periodontitis (Section 3), and colorectal cancer
384 (CRC) (Section 4) through a unified metagenomic approach. While previous studies have examined the
385 role and characteristics of human microbiome in these diseases individually, this dissertation uniquely
386 integrates human microbiome-driven insights across these diseases to identify shared and disease-specific
387 microbial signatures. By applying high-throughput metagenomic sequencing, microbial diversity analysis,
388 and advanced bioinformatics techniques, this dissertation aims to uncover novel microbiome-based
389 biomarkers and mechanistic insights into how microbial communities influence these conditions. These
390 findings contribute to a broader understanding of microbiome-mediated disease interactions and pave the
391 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)

392

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

393

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

394

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

395

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

396

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

397

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

398

$$MAE = \sum_{i=1}^n |Prediction_i - Real_i| / n \quad (7)$$

$$RMSE = \sqrt{\sum_{i=1}^n (Prediction_i - Real_i)^2 / n} \quad (8)$$

399 **2 Predicting preterm birth using random forest classifier in salivary mi-**
400 **crobiome**

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

402 Hong, Y. M., **Lee, Jaewoong**, Cho, D. H., Jeon, J. H., Kang, J., Kim, M. G., ... & Kim, J. K. (2023).
403 Predicting preterm birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1),
404 21105.

405 **2.1 Introduction**

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

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

423 Preterm prelabor rupture of membranes (PROM) brought on by gestational inflammation and infection
424 contribute to about 70% of PTB cases (Romero, Dey, & Fisher, 2014). Nevertheless, as antibiotics and
425 anti-inflammatory therapeutic strategies were ineffective to decrease PTB occurrence rates, the pathology
426 of PTB has not been entirely elucidated by inflammatory and infectious pathways (Romero, Hassan, et al.,
427 2014). Recent researches on maternal microbiomes were beginning to examine unidentified connections
428 of PTB as a consequence of developmental processes in molecular biological technology (Fettweis et al.,
429 2019).

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

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

438 Multiple unfavorable birth outcomes, including PROM and PTB, have been linked to periodontitis
439 as an independence risk factor, according to numerous epidemiological researches (Offenbacher et al.,
440 1996). It is expected that the oral microbiome will be able to explain additional hematogenous pathways
441 in light of these precedents; however, the oral microbiome composition of fetuses is limited understood.

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

445 **2.2 Materials and methods**

446 **2.2.1 Study design and study participants**

447 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
448 Declaration of Helsinki (Goodyear, Krleza-Jeric, & Lemmens, 2007). The Institutional Review Board
449 authorized this study (IRB file No. 2019-01-024). Participants who were admitted for elective cesarean
450 sections (C-sections) or induction births, as well as those who had written informed consent obtained
451 with premature labor or PROM, were eligible.
452

453 **2.2.2 Clinical data collection and grouping**

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

- 456 • maternal age at delivery
- 457 • diabetes mellitus
- 458 • hypertension
- 459 • overweight and obesity
- 460 • C-section
- 461 • history PROM or PTB
- 462 • gestational week on delivery
- 463 • birth weight
- 464 • sex

465 **2.2.3 Salivary microbiome sample collection**

466 Salivary microbiome samples were collected 24 hours before to delivery using mouthwash. The standard
467 methods of sterilizing were performed. Medical experts oversaw each stage of the sample collecting
468 procedure. Participants received instruction not to eat, drink, or brush their teeth for 30 minutes before
469 sampling salivary microbiome. Saliva samples were gathered by washing the mouth for 30 seconds with
470 12 mL of a mouthwash solution (E-zен Gargle, JN Pharm, Pyeongtaek, Gyeonggi, Korea). The samples
471 were tagged with the anonymous ID for each participant and kept in low temperature (4 °C) until they
472 underwent further processing. Genomic DNA was extracted using an ExgeneTM Clinic SV kit (GeneAll
473 Biotechnology, Seoul, Korea) following with the manufacturer instructions and store at -20 °C.

474 **2.2.4 16s rRNA gene sequencing**

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

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

481 **2.2.5 Bioinformatics analysis**

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

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

493 Taxonomic assignment were implemented with HOMD (version 15.22) (T. Chen et al., 2010).
494 Afterward, DESeq2 was implemented to identify differentially abundant taxa (DAT) that could dis-
495 tinguish between salivary microbiome from PTB and FTB participants (Love et al., 2014). Taxa with
496 $|\log_2 \text{FoldChange}| > 1$ and $p < 0.05$ were considered as statistically significant.

497 The taxa for predicting PTB using salivary microbiome data were determined using a random forest
498 classifier (Breiman, 2001). Through stratified *k*-fold cross-validation (*k* = 5) that preserves the existence
499 rate of PTB and FTB participants, consistency and trustworthy classification were ensured (Wong & Yeh,
500 2019).

501 **2.2.6 Data and code availability**

502 All sequences from the 59 study participants have been published to the Sequence Read Archives
503 (project ID PRJNA985119): <https://dataview.ncbi.nlm.nih.gov/object/PRJNA985119>. Docker
504 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.

507 **2.3 Results**

508 **2.3.1 Overview of clinical information**

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

515 **2.3.2 Comparison of salivary microbiomes composition**

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

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

525 A significant negative correlation was found using Pearson correlation analysis between GW and
526 differences between PTB-enriched DAT and FTB-enriched DAT (Pearson correlation $r = -0.542$ and
527 $p = 7.8e-6$; Figure 5).

528 **2.3.3 Random forest classification to predict PTB risk**

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

Table 2: Standard clinical information of PTB 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

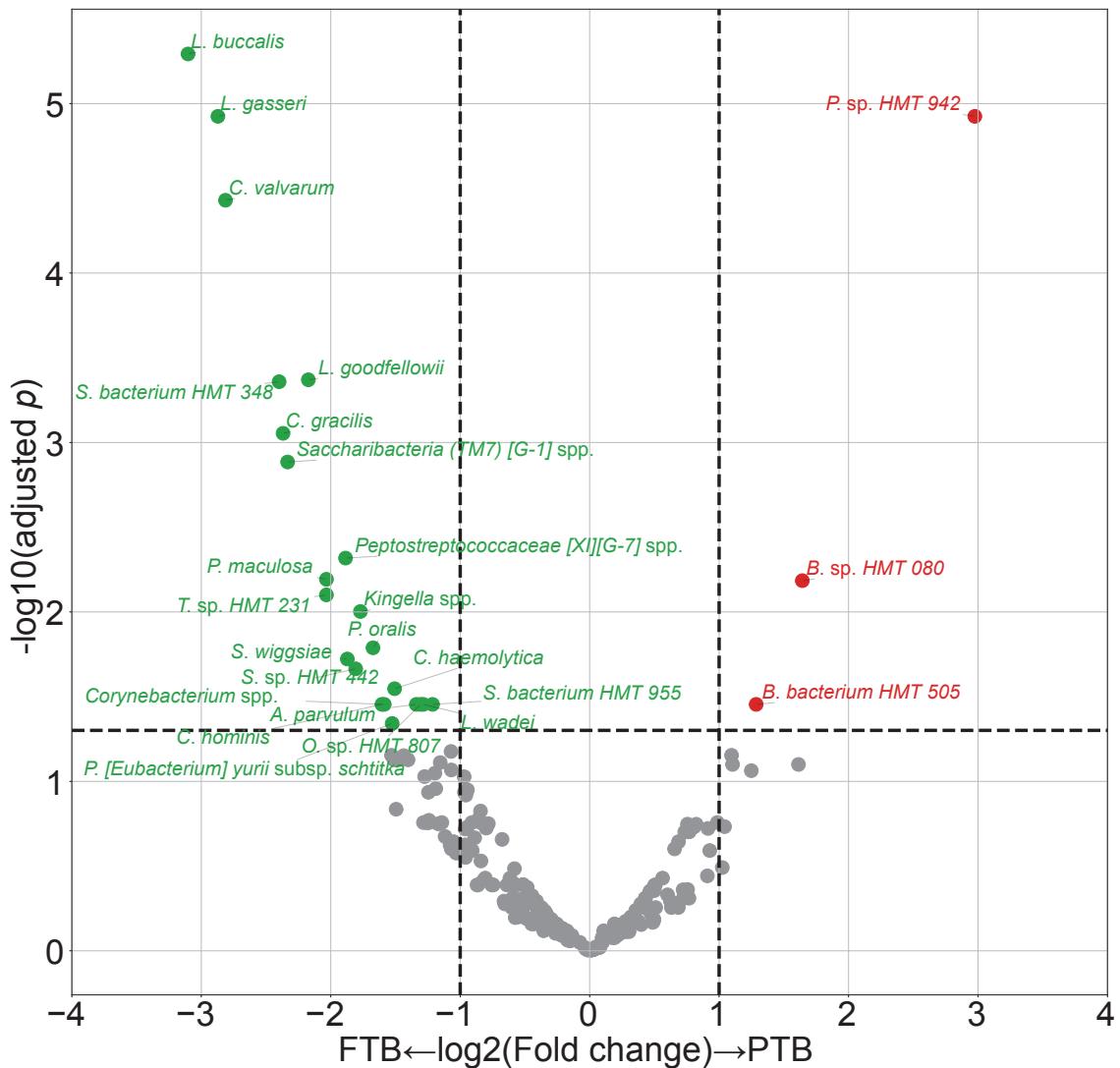


Figure 1: DAT volcano plot for PTB prediction.

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

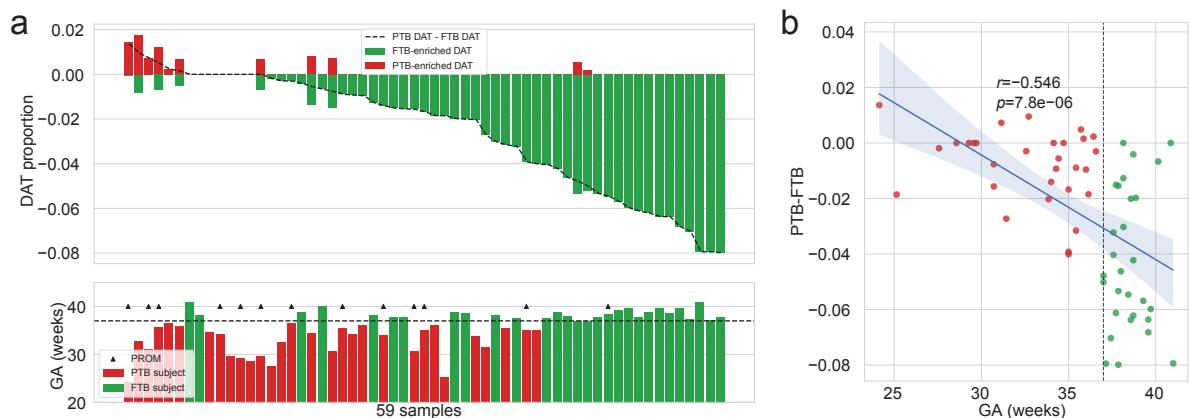


Figure 2: **Salivary microbiome compositions over DAT for PTB prediction.**

(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.

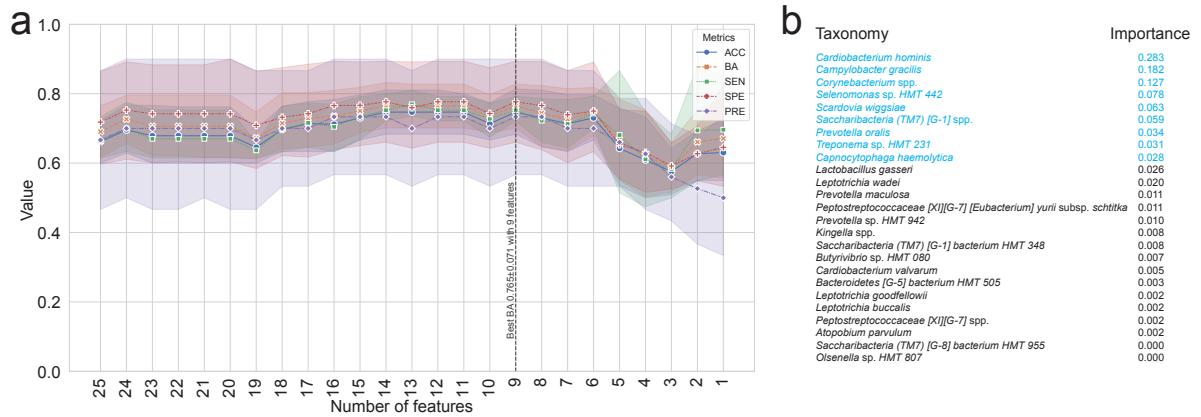


Figure 3: **Random forest-based PTB prediction model.**

(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.

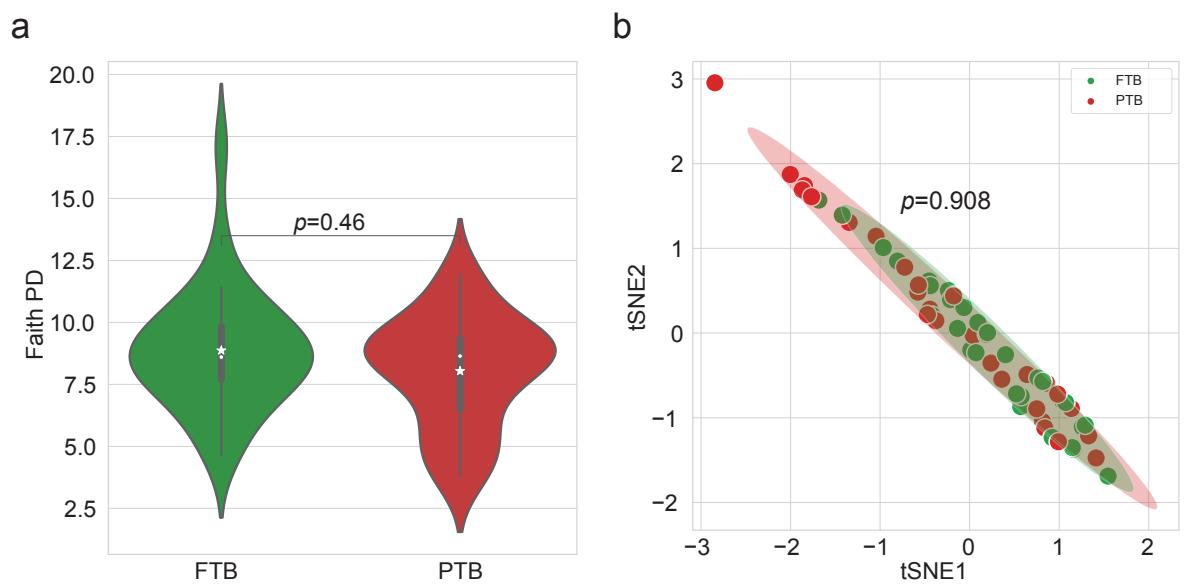


Figure 4: **Diversity indices.**

(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 $p = 0.908$)

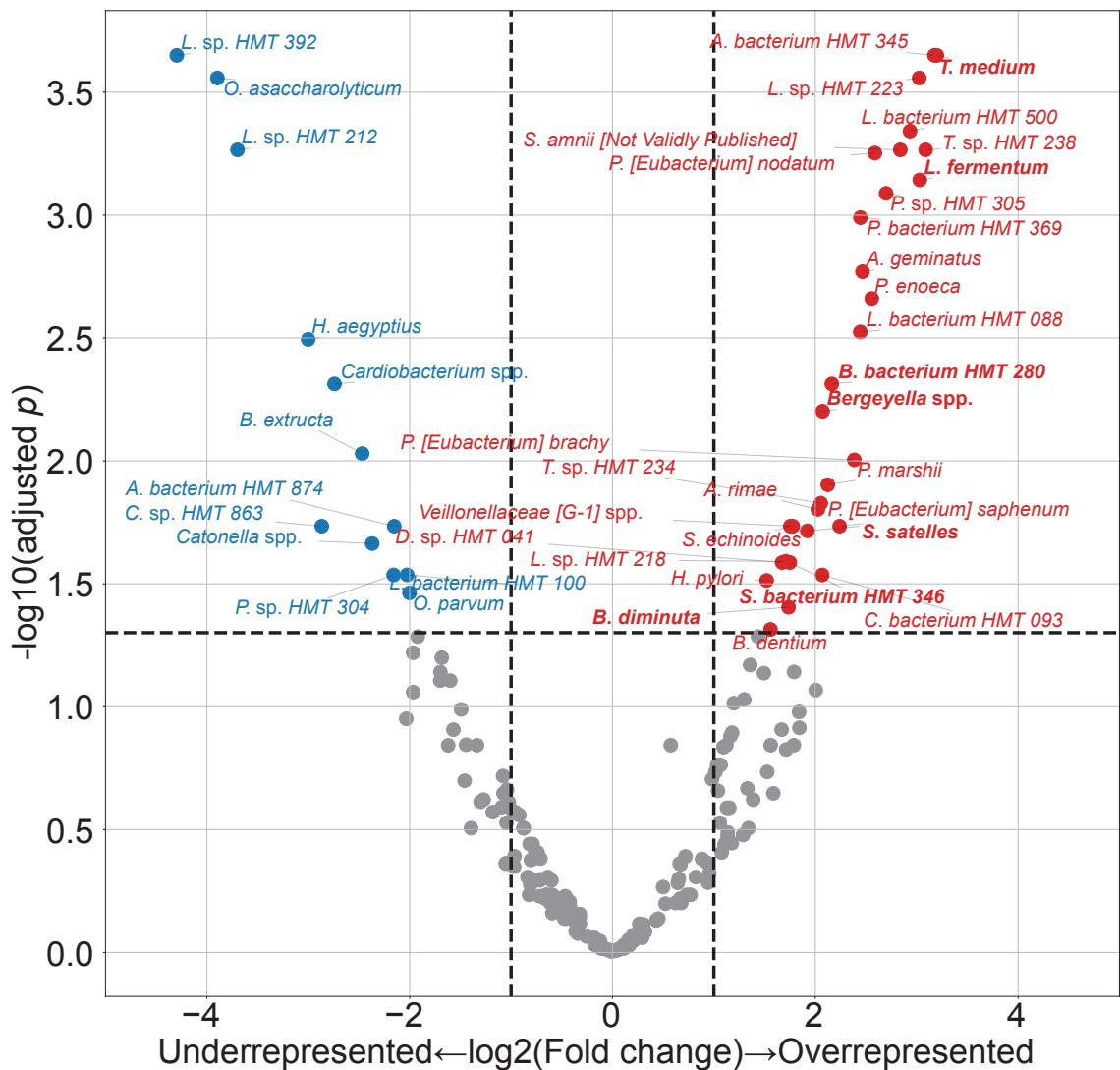


Figure 5: PROM-related DAT between FTB and PTB.

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.

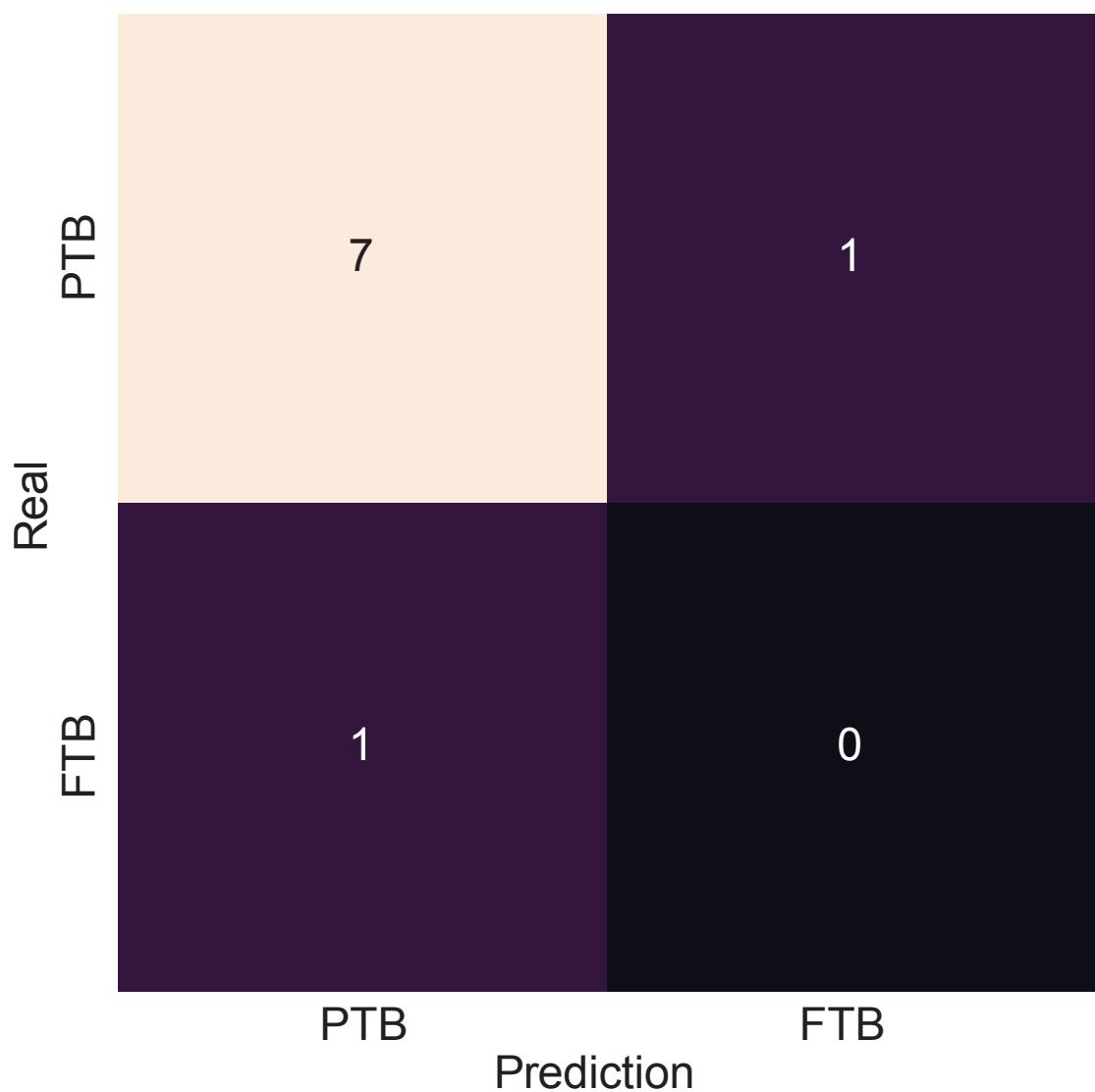


Figure 6: Validation of random forest-based PTB prediction model.

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)

536 **2.4 Discussion**

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

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

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

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

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

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

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

588 **3 Random forest prediction model for periodontitis statuses based on the**
589 **salivary microbiomes**

590 **3.1 Introduction**

591 Saliva microbial dysbiosis brought on by the accumulation of plaque results in periodontitis, a chronic
592 inflammatory disease of the tissue that surrounds the tooth (Kinane, Stathopoulou, & Papapanou, 2017).
593 Loss of periodontal attachment is a consequence of periodontitis, which may lead to irreversible bone loss
594 and, eventually, permanent tooth loss if left untreated. A new classification criterion of periodontal diseases
595 was created in 2018, about 20 years after the 1999 statements of the previous one (Papapanou et al.,
596 2018). Even with this evolution, radiographic and clinical markers of periodontitis progression remain the
597 primary methods for diagnosing periodontitis (Papapanou et al., 2018). Such tools, nevertheless, frequently
598 demonstrate the prior damage from periodontitis rather than its present condition. Certain individuals have
599 a higher risk of periodontitis, a higher chance of developing severe generalized periodontitis, and a worse
600 response to common salivary bacteria control techniques utilized to prevent and treat periodontitis. As a
601 result, the 2017 framework for diagnosing periodontitis additionally allows for the potential development
602 of biomarkers to enhance diagnosis and treatment of periodontitis (Tonetti, Greenwell, & Kornman, 2018).
603 Instead of only depending on the progression of periodontitis, a new etiological indication based on the
604 current state must be introduced in order to enable appropriate intervention through early detection of
605 periodontitis. Thus, the current clinical diagnostic techniques that rely on periodontal probing can be
606 uncomfortable for patients with periodontitis (Canakci & Canakci, 2007).

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

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

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

639 Recently, we employed multiplex quantitative-PCR (qPCR) and machine learning-based classification
640 model to predict the severity of periodontitis based on the amount of nine pathogens of periodontitis from
641 saliva collections (E.-H. Kim et al., 2020). On the other hand, the fact that we focused merely at nine
642 pathogens for periodontitis and neglected the variety bacterial species associated to the various severities
643 of periodontitis constrained the breadth of our investigation. By developing a machine learning model
644 that could classify multiple severities of periodontitis based on the salivary microbiome composition,
645 this study aims to fill these knowledge gaps and produce more accurate and therapeutically useful
646 guidance to evaluate progression of periodontitis. Hence, in order to examine the salivary microbiome
647 composition of both healthy controls and patients with periodontitis in multiple stages, we applied
648 16S rRNA gene sequencing. Furthermore, employing the 2018 classification criteria, we sought to find
649 biomarkers (bacterial species) for the precise prediction of periodontitis severities (Papapanou et al.,
650 2018; Chapple et al., 2018).

651 **3.2 Materials and methods**

652 **3.2.1 Study participants enrollment**

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

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

670 **3.2.2 Periodontal clinical parameter diagnosis**

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

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

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

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

688 0. Normal gingiva: without inflammation nor discoloration.

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

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

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

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

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

700 • Healthy:

701 1. Bleeding sites < 10%

702 2. Probing depth: \leq 3 mm

703 • Stage I:

704 1. No tooth loss because of periodontitis.

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

706 3. Radiographic bone loss: < 15%

707 • Stage II:

708 1. No tooth loss because of periodontitis.

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

710 3. Radiographic bone loss: 15-33%

711 • Stage III:

712 1. Teeth loss because of periodontitis: \leq 3 teeth

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

714 3. Radiographic bone loss: > 33%

715 **3.2.3 Saliva sampling and DNA extraction procedure**

716 All study participants received instructions to avoid eating, drinking, brushing, and using mouthwash for
717 at least an hour prior to the saliva sample collection process. These collections were conducted between
718 09:00 and 11:00. Mouth rinse was collected by rinsing the mouth for 30 seconds with 12 mL of a solution
719 (E-zen Gargle, JN Pharm, Korea). All saliva samples were tagged with anonymous ID and stored at -4 °C.

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

- 724 • Forward: 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNNGCWGCAG-3'
725 • Reverse: 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'

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

- 728 1. Heat activation for 30 seconds at 95 °C.
729 2. 25 cycles for 30 seconds at 95 °C.
730 3. 30 seconds at 55 °C.
731 4. 30 seconds at 72 °C.

732 NexteraXT Indexed Primer was applied to amplification 10 µL of the purified initial PCR products for
733 the final library creation. The second PCR used the same conditions as the first PCR conditions but with
734 10 cycles. 16S rRNA gene sequencing was performed via 2×300 bp paired-end sequencing at Macrogen
735 Inc. (Macrogen, Seoul, Korea) using Illumina MiSeq platform (Illumina, San Diego, CA, USA).

736 **3.2.4 Bioinformatics analysis**

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

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

- 748 • Shannon (Weaver, 1963)
749 • Simpson (Simpson, 1949)

750 Aitchison index for a beta-diversity index was calculated using QIIME2 (version 2020.8) (Aitchison,
751 Barceló-Vidal, Martín-Fernández, & Pawlowsky-Glahn, 2000; Bolyen et al., 2019). We employed the
752 t-SNE algorithm to illustrate multi-dimensional data from the beta-diversity index computation (Van der
753 Maaten & Hinton, 2008). The beta-diversity index was compared using the PERMANOVA test (Anderson,
754 2014; Kelly et al., 2015) and MWU test.

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

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

767 As previously stated (E.-H. Kim et al., 2020), we used stratified k -fold cross-validation ($k = 10$)
768 by severity of periodontitis to achieve consistent and trustworthy classification results (Wong & Yeh,
769 2019). Additionally, we utilized various features with confusion matrices and their derivations to evaluate
770 the classification outcomes in order to identify which features optimize classification evaluations and
771 decrease sequencing efforts. Using the DAT discovered by ANCOM, we iteratively removed the least
772 significant taxa from the input features (taxa) of the random forest (Breiman, 2001) and gradient boosting
773 (Friedman, 2002) classification models using the backward elimination method. Random forest classifier
774 builds multiple decision trees independently using bootstrapped samples and aggregates their predictions,
775 enhancing stability and reducing overfitting problems. In contrast, Gradient boosting constructs trees
776 sequentially, where each new tree improves the errors of the previous ones using gradient descent, leading
777 to higher classification evaluations.

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

782 **3.2.5 Data and code availability**

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

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

788 **3.3 Results**

789 **3.3.1 Summary of clinical information and sequencing data**

790 Among clinical information of the study participants, clinical attachment level, probing depth, plaque
791 index, and gingival index, were significantly increased with periodontitis severity (Kruskal-Wallis test
792 $p < 0.001$), while sex were observed no significant difference (Table 2). Notably, clinical attachment level
793 and probing depth have significant differences among the periodontitis severities (MWU test $p < 0.01$;
794 Figure 15). Additionally, 71461.00 ± 11792.30 and 45909.78 ± 11404.65 reads per sample were obtained
795 before and after filtering low-quality reads and trimming extra-long tails, respectively (Figure 16). In 250
796 study subjects, we have found a total of 425 bacterial taxa (Figure 13).

797 **3.3.2 Diversity indices reveal differences among the periodontitis severities**

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

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

808 **3.3.3 DAT among multiple periodontitis severities and their correlation**

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

- 812 • Group 1
 - 813 1. *Treponema* spp.
 - 814 2. *Prevotella* sp. HMT 304
 - 815 3. *Prevotella* sp. HMT 526
 - 816 4. *Peptostreptococcaceae [XI][G-5]* saphenum
 - 817 5. *Treponema* sp. HMT 260
 - 818 6. *Mycoplasma faecium*
 - 819 7. *Peptostreptococcaceae [XI][G-9]* brachy
 - 820 8. *Lachnospiraceae [G-8]* bacterium HMT 500
 - 821 9. *Peptostreptococcaceae [XI][G-6]* nodatum
 - 822 10. *Fretibacterium* spp.

- 823 • Group 2
- 824 1. *Porphyromonas gingivalis*
- 825 2. *Campylobacter showae*
- 826 3. *Filifactor alocis*
- 827 4. *Treponema putidum*
- 828 5. *Tannerella forsythia*
- 829 6. *Prevotella intermedia*
- 830 7. *Porphyromonas* sp. HMT 285

- 831 • Group 3
- 832 1. *Actinomyces* spp.
- 833 2. *Corynebacterium durum*
- 834 3. *Actinomyces graevenitzii*

835 Ten DAT that were significant enriched in stage II and stage III, but deficient in healthy formed Group
 836 1 (Figure 8). Furthermore, in comparison to the healthy, the seven DAT of Group 2 were significantly
 837 enriched in each of the stages of periodontitis. On the other hand, three DAT in Group 3 were deficient in
 838 stage II and stage III, but significantly enriched in healthy. The relative proportions of the DAT further
 839 supported these findings (Figure 8b), suggesting that the DAT is primarily linked to periodontitis rather
 840 than other salivary bacteria.

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

843 3.3.4 Classification of periodontitis severities by random forest models

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

848 Based on the proportion of DAT, random forest classifier were trained to classify the periodontitis
 849 severities (Table 6). We conducted multi-label classification for the multiple periodontitis severities,
 850 namely healthy, stage I, stage II, and stage III. In this setting, we classified multiple periodontitis
 851 severities with the highest BA of 0.779 ± 0.029 (Table 4). AUC ranged between 0.81 and 0.94 (Figure
 852 10b).

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

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

859 perform a multi-label classification.

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

866 Finally, to confirm the reliability and consistency of our random forest classifier, we validated our
867 classification model using openly accessible 16S rRNA gene sequencing from Spanish participants
868 (Iniesta et al., 2023) and Portuguese participants (Relvas et al., 2021) (Figure 11). Although some
869 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	0.798±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±0.025	0.914±0.025	0.828±0.051	
Healthy vs. Stages I/II/III	<i>P.gingivalis</i>	0.792±0.114	0.856±0.105	0.819±0.088	0.776±0.089	0.840±0.092	0.756±0.175	0.883±0.054
	<i>P.gingivalis+Act.</i>	0.828±0.121	0.926±0.074	0.847±0.116	0.797±0.123	0.800±0.126	0.830±0.191	0.864±0.074
Top 4 taxa		0.860±0.078	0.953±0.049	0.885±0.066	0.832±0.079	0.840±0.128	0.864±0.157	0.905±0.070

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

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

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

Condition	Healthy vs. Stage I vs. Stage II vs. Stage III			Healthy vs. Stage I			Healthy vs. Stage I vs. Stage II/III			Healthy vs. Stage I/II/III		
	Rank	Taxa	Importance	Taxa	Importance	Taxa	Importance	Taxa	Importance	Taxa	Importance	
1	<i>Porphyromonas gingivalis</i>	0.297	<i>Actinomyces spp.</i>	0.195	<i>Porphyromonas gingivalis</i>	0.360	<i>Porphyromonas gingivalis</i>	0.426	<i>Porphyromonas gingivalis</i>	0.461		
2	<i>Actinomyces spp.</i>	0.195	<i>Actinomyces graevenitzii</i>	0.054	<i>Actinomyces spp.</i>	0.125	<i>Actinomyces spp.</i>	0.244	<i>Actinomyces spp.</i>	0.257		
3	<i>Prevotella intermedia</i>	0.054	<i>Actinomyces graevenitzii</i>	0.052	<i>Porphyromonas sp. HMT 285</i>	0.055	<i>Actinomyces graevenitzii</i>	0.049	<i>Actinomyces graevenitzii</i>	0.059		
4	<i>Actinomyces graevenitzii</i>	0.052	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.050	<i>Porphyromonas sp. HMT 285</i>	0.062	<i>Corynebacterium durum</i>	0.046	<i>Corynebacterium durum</i>	0.035		
5	<i>Filifactor alocis</i>	0.050	<i>Campylobacter showae</i>	0.042	<i>Campylobacter showae</i>	0.052	<i>Filifactor alocis</i>	0.036	<i>Filifactor alocis</i>	0.032		
6	<i>Campylobacter showae</i>	0.042	<i>Porphyromonas sp. HMT 285</i>	0.040	<i>Corynebacterium durum</i>	0.052	<i>Prevotella intermedia</i>	0.033	<i>Campylobacter showae</i>	0.023		
7	<i>Porphyromonas sp. HMT 285</i>	0.040	<i>Treponema spp.</i>	0.032	<i>Treponema spp.</i>	0.038	<i>Tannerella forsythia</i>	0.025	<i>Porphyromonas sp. HMT 285</i>	0.022		
8	<i>Corynebacterium durum</i>	0.032	<i>Tannerella forsythia</i>	0.026	<i>Tannerella forsythia</i>	0.037	<i>Prevotella intermedia</i>	0.023	<i>Prevotella intermedia</i>	0.022		
9	<i>Treponema spp.</i>	0.032	<i>Prevotella intermedia</i>	0.025	<i>Prevotella intermedia</i>	0.029	<i>Treponema spp.</i>	0.021	<i>Treponema spp.</i>	0.022		
10	<i>Tannerella forsythia</i>	0.026	<i>Prevotella intermedia</i>	0.025	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.026	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.018	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.015		
11	<i>Treponema putidum</i>	0.025	<i>Freibacterium spp.</i>	0.023	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.018	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.014	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.010		
12	<i>Freibacterium spp.</i>	0.023	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.021	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.018	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.011	<i>Tannerella forsythia</i>	0.009		
13	<i>Peptostreptococcaceae (XII)(G-9) brachy</i>	0.021	<i>Treponema putidum</i>	0.019	<i>Treponema putidum</i>	0.014	<i>Treponema putidum</i>	0.010	<i>Freibacterium spp.</i>	0.009		
14	<i>Treponema sp. HMT 260</i>	0.019	<i>Prevotella sp. HMT 526</i>	0.018	<i>Prevotella sp. HMT 526</i>	0.011	<i>Prevotella sp. HMT 526</i>	0.009	<i>Prevotella sp. HMT 526</i>	0.006		
15	<i>Prevotella sp. HMT 526</i>	0.018	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.018	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.008	<i>Freibacterium spp.</i>	0.008	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.004		
16	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.018	<i>Prevotella sp. HMT 304</i>	0.017	<i>Peptostreptococcaceae (XII)(G-6) nodatum</i>	0.008	<i>Treponema sp. HMT 260</i>	0.008	<i>Treponema sp. HMT 260</i>	0.004		
17	<i>Prevotella sp. HMT 304</i>	0.017	<i>Mycoplasma faecium</i>	0.014	<i>Mycoplasma faecium</i>	0.004	<i>Prevotella sp. HMT 304</i>	0.005	<i>Mycoplasma faecium</i>	0.003		
18	<i>Mycoplasma faecium</i>	0.014	<i>Prevotella sp. HMT 304</i>	0.014	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.003	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.005	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.002		
19	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.014	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.013	<i>Peptostreptococcaceae (XII)(G-5) saphenum</i>	0.003	<i>Prevotella sp. HMT 304</i>	0.004	<i>Prevotella sp. HMT 304</i>	0.001		
20	<i>Lachnospiraceae (G-8) bacterium HMT 500</i>	0.013										

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

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

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

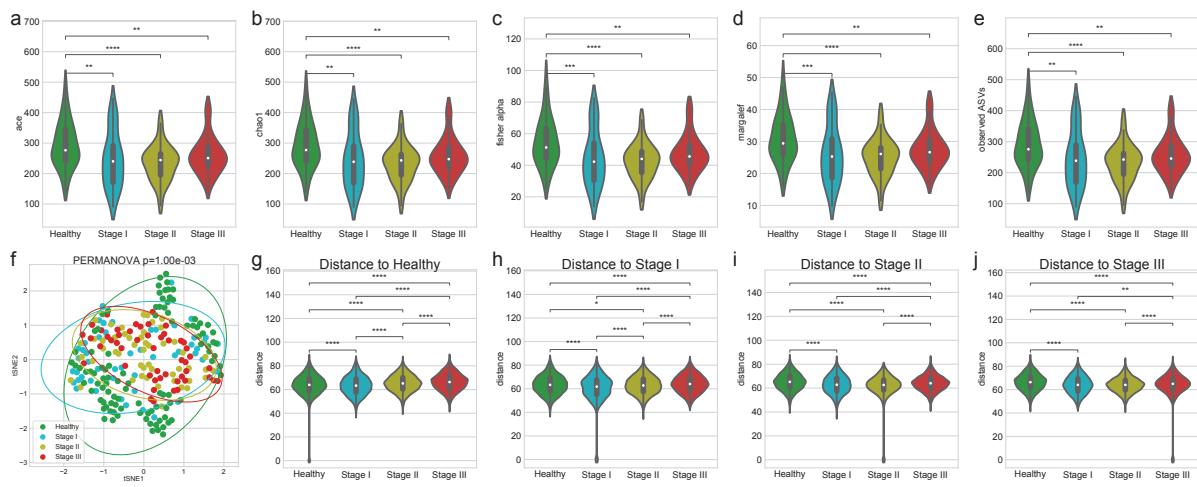


Figure 7: Diversity indices for periodontitis.

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: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***) ≤ 0.0001 (****).

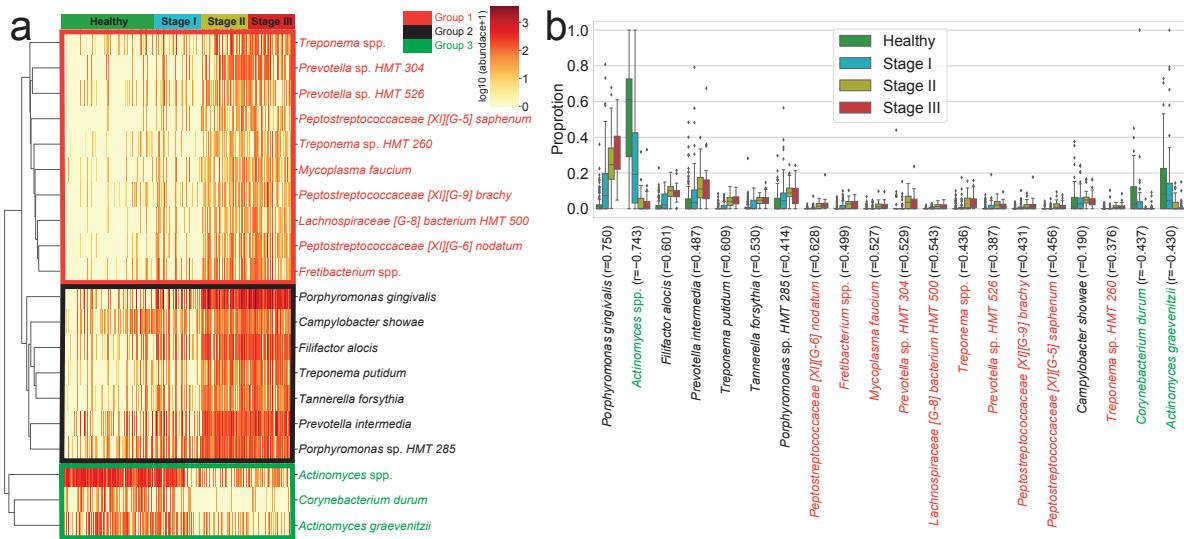


Figure 8: DAT for periodontitis.

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.

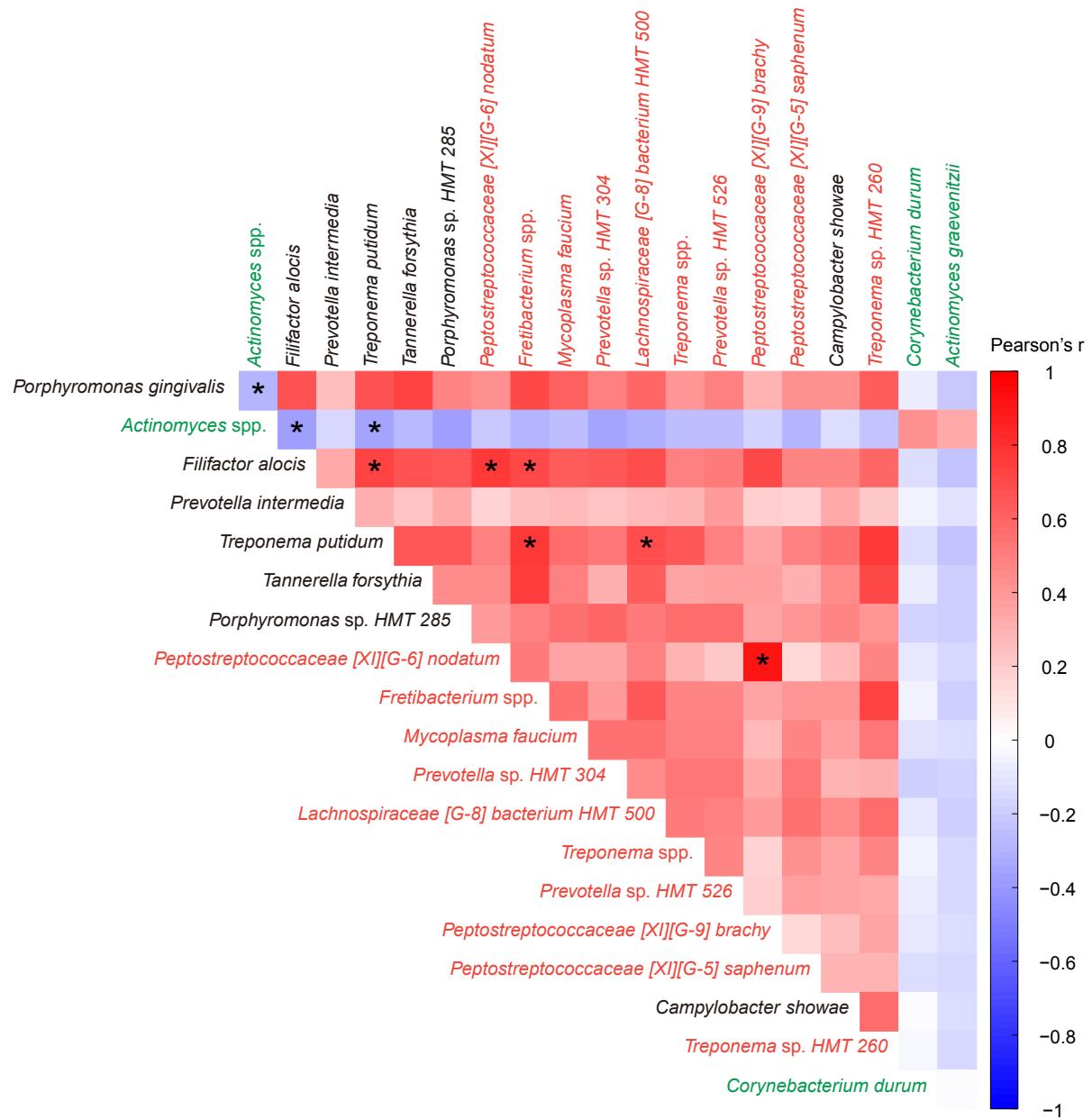


Figure 9: Correlation heatmap between periodontitis DAT.

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

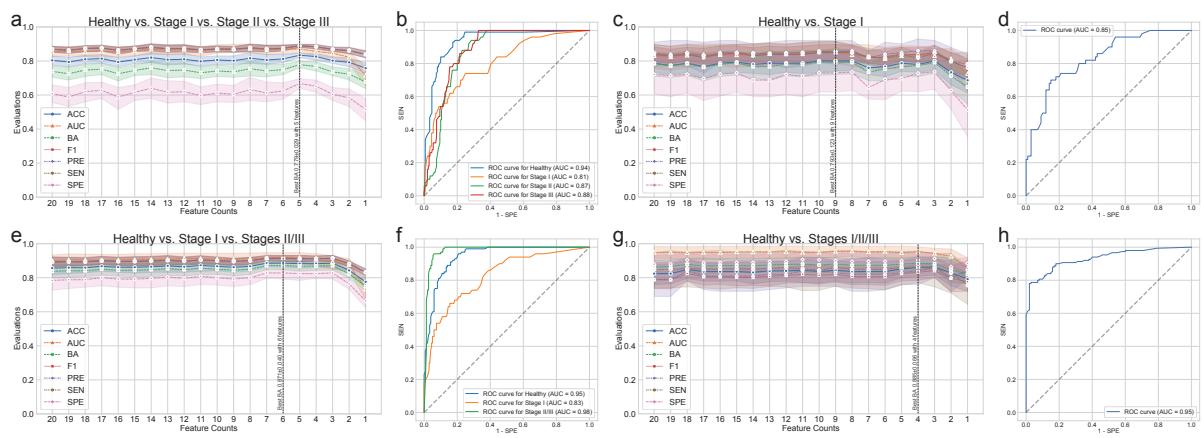


Figure 10: Random forest classification metrics for periodontitis prediction.

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).

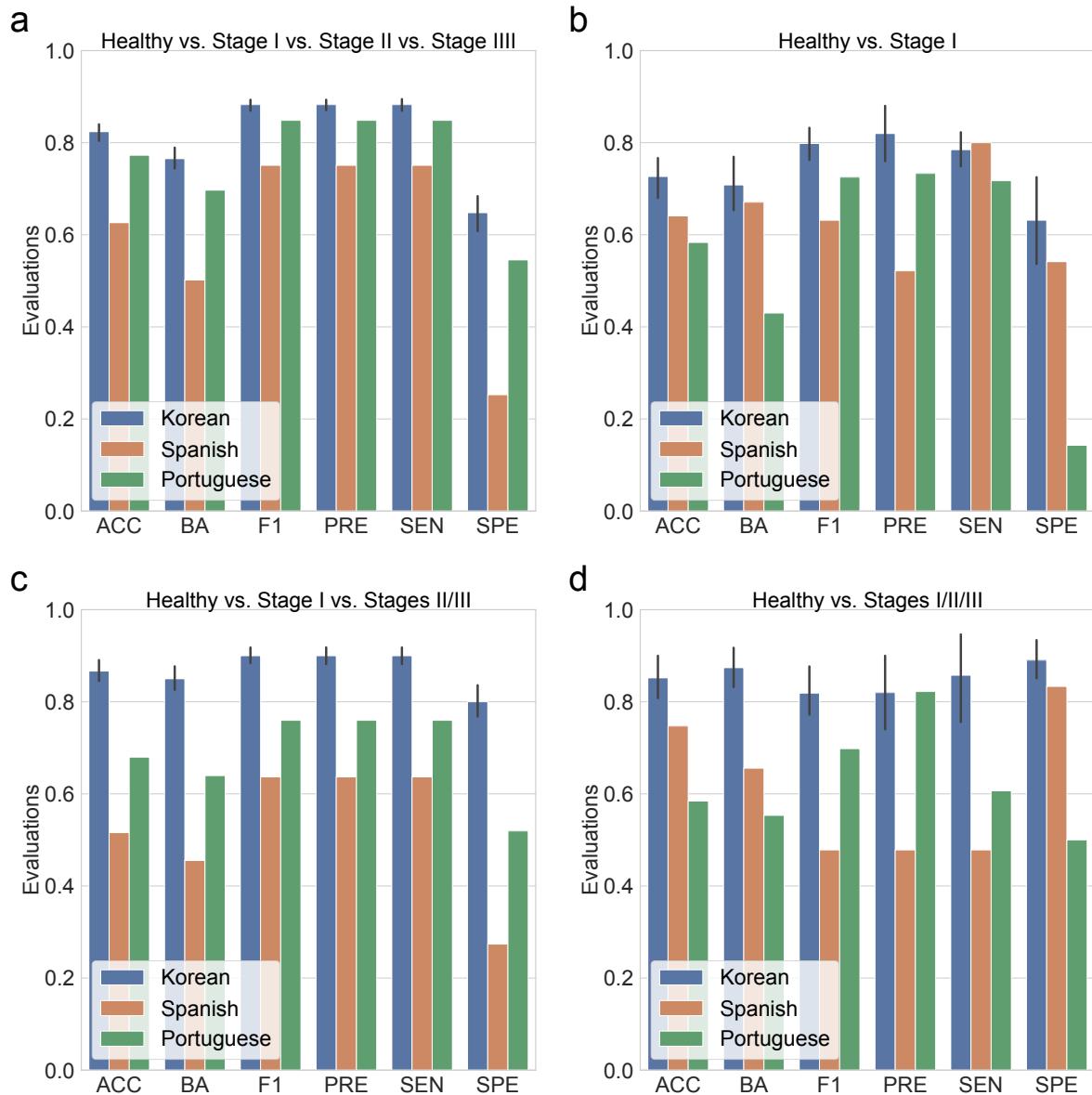


Figure 11: **Random forest classification metrics from external datasets.**

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.

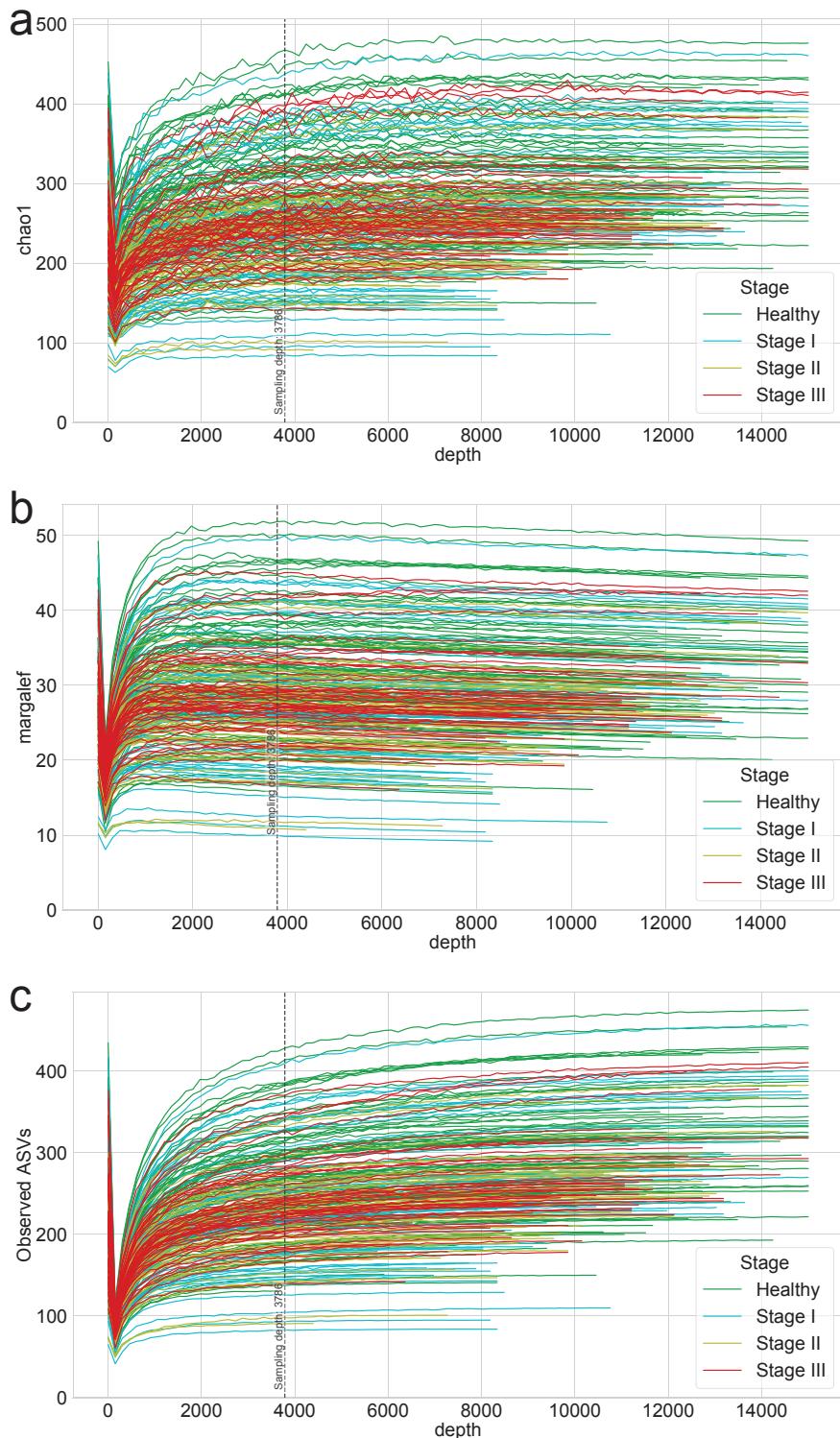


Figure 12: Rarefaction curves for alpha-diversity indices.

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

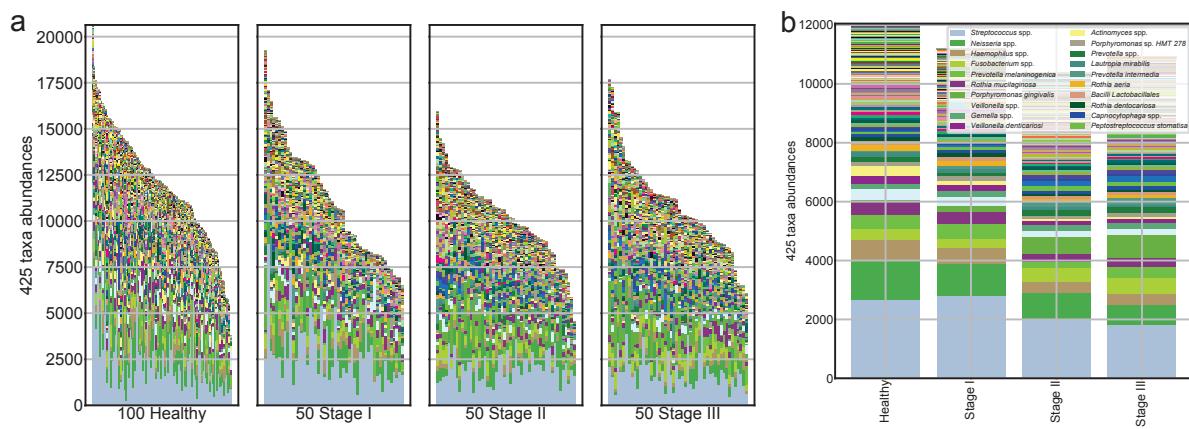


Figure 13: Salivary microbiome compositions in the different periodontal statuses.

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).

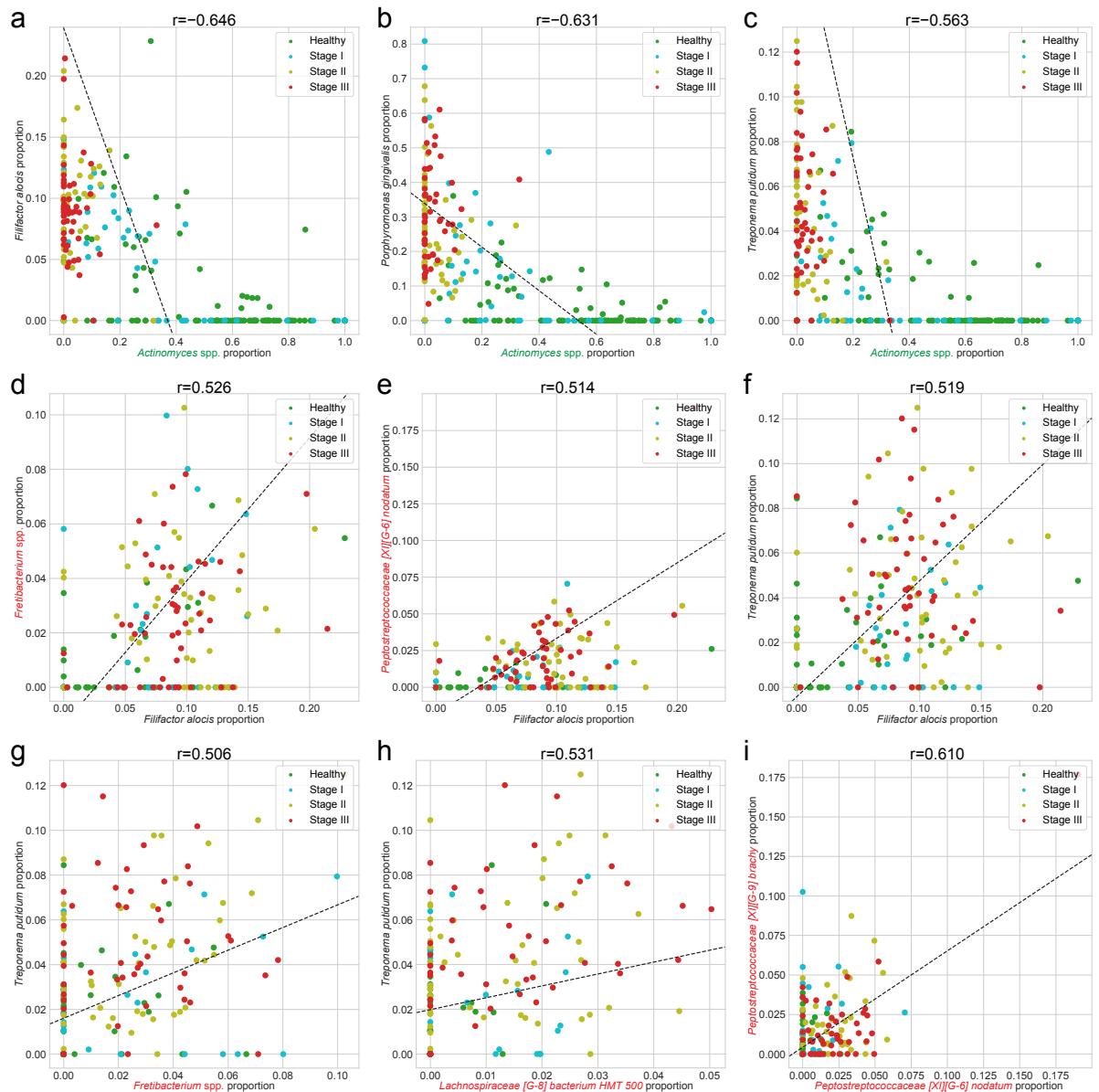


Figure 14: Correlation plots for periodontitis DAT.

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).

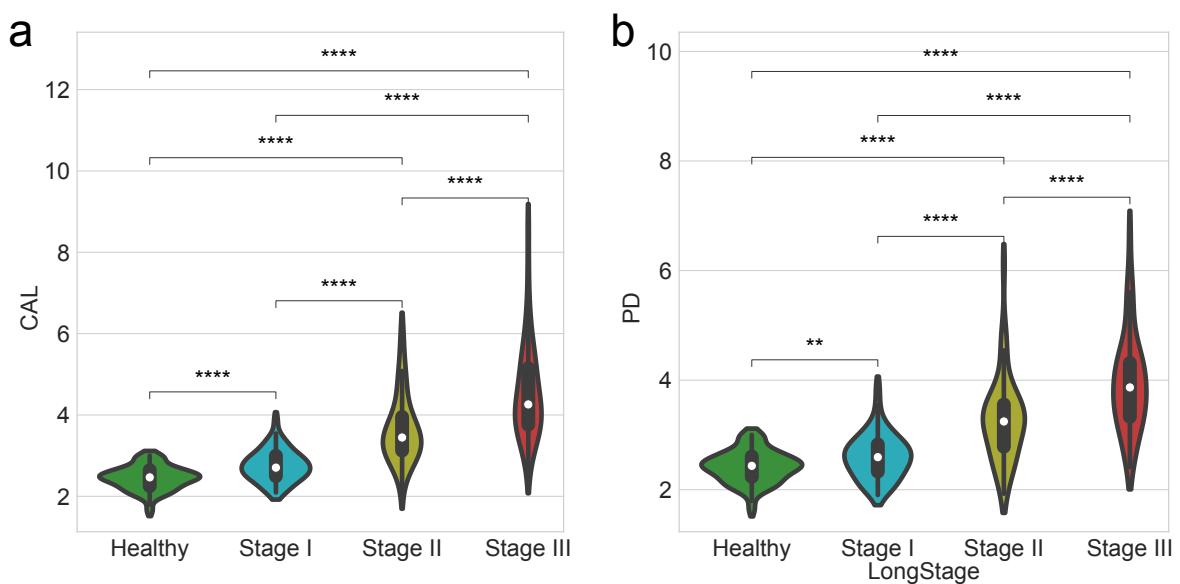


Figure 15: Clinical measurements by the periodontitis statuses.

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

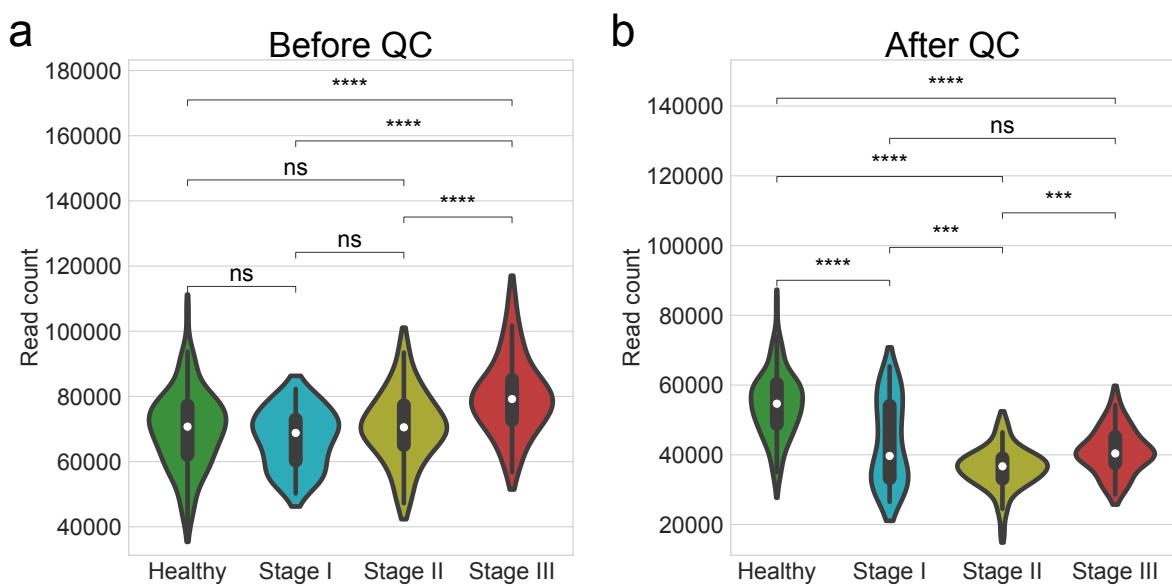


Figure 16: **Number of read counts by the periodontitis statuses.**

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

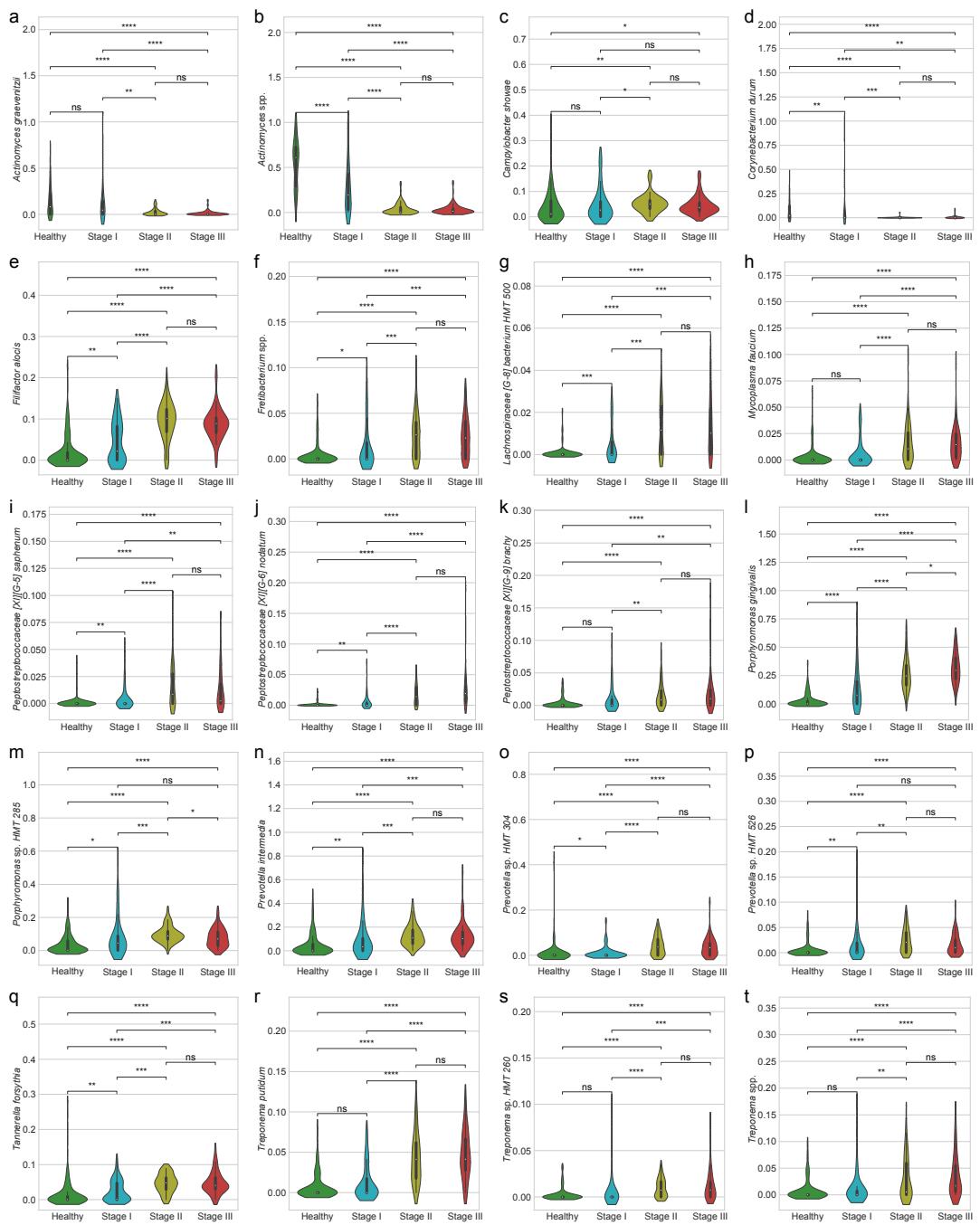


Figure 17: Proportions of periodontitis DAT.

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

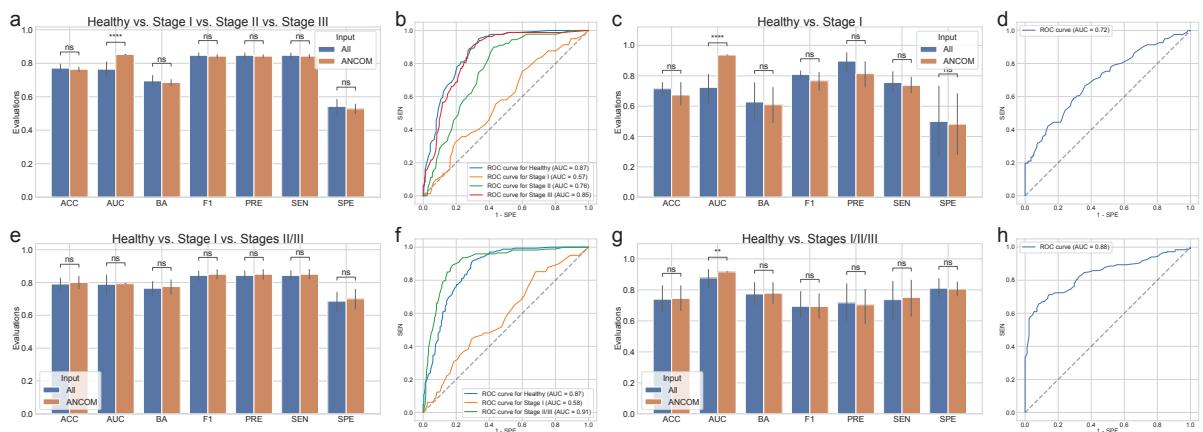


Figure 18: Random forest classification metrics with the full microbiome compositions and ANCOM-selected DAT compositions.

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

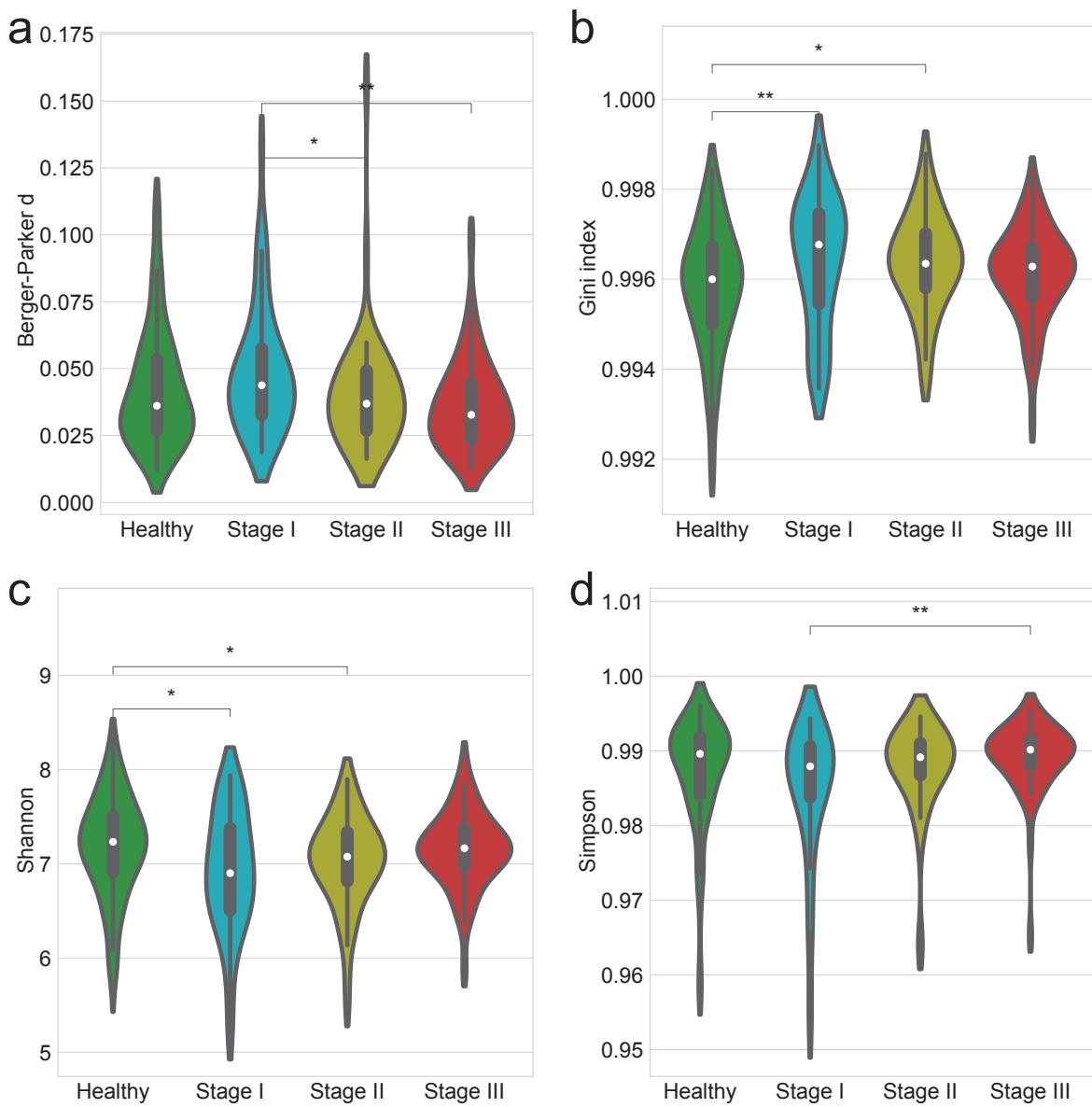


Figure 19: Alpha-diversity indices account for evenness.

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 $p < 0.01$ (**)

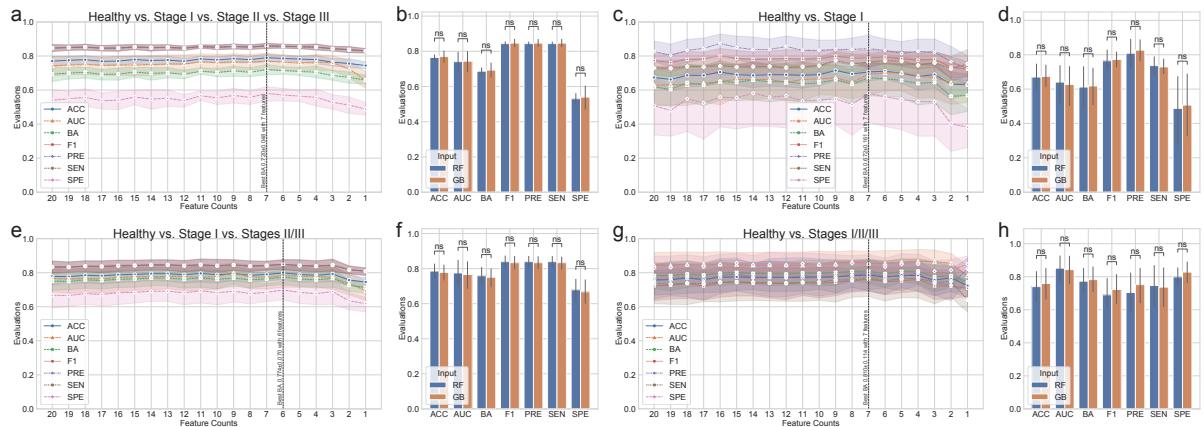


Figure 20: Gradient Boosting classification metrics for periodontitis prediction.

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)

870 **3.4 Discussion**

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

879 Previous research identified the red complex as the primary pathogens of periodontitis (Listgarten,
880 1986): *Porphyromonas gingivalis*, *Tannerella forsythia*, and *Treponema denticola*. Other studies, however,
881 have shown that periodontal pathogens communicate with other bacteria in the salivary microbiome
882 networks to generate dental plaque prior to the pathogenesis and development of periodontitis (Lamont &
883 Jenkinson, 2000; Rosan & Lamont, 2000; Yoshimura, Murakami, Nishikawa, Hasegawa, & Kawaminami,
884 2009).

885 Using subgingival plaque collections, recent researches have suggested a connection between the
886 periodontitis severity and the salivary microbiome compositions (Altabtbaei et al., 2021; Iniesta et al.,
887 2023; Nemoto et al., 2021). Therefore, we have examined the salivary microbiome compositions of
888 patients with multiple severities of periodontitis and periodontally healthy controls, extending on earlier
889 studies.

890 According to our findings, the salivary microbiome compositions have 425 taxa (Figure 13). We
891 computed the alpha-diversity indices to determine the variability within each salivary microbiome
892 composition, including ace (Chao & Lee, 1992), chao1 (Chao, 1984), fisher alpha (Fisher et al., 1943),
893 margalef (Magurran, 2021), observed ASVs (DeSantis et al., 2006), Berger-Parker *d* (Berger & Parker,
894 1970), Gini (Gini, 1912), Shannon (Weaver, 1963), and Simpson (Simpson, 1949) (Figure 7 and Figure
895 19). Alpha-diversity indices suggested that the microbial richness of periodontally healthy controls was
896 higher than that of patients with periodontitis (Figure 7a-e and Figure 19). These results are in line with
897 findings with that patients with advanced periodontitis, namely stage II and stage III, have less diversified
898 communities than periodontally healthy controls (Jorth et al., 2014). Recognizing that the periodontitis
899 severity increases the amount of *Porphyromonas gingivalis*, the salivary microbiome compositions from
900 periodontally healthy controls conserved microbial networks dominated by *Streptococcus* spp. (Figure
901 13). *Porphyromonas gingivalis* is one of the known periodontal pathogen that could cause dysbiosys
902 in the salivary microbiomes, suggesting in the pathophysiology of periodontitis. Despite this finding,
903 earlier research found that subgingival microbiome of patients with periodontitis had a greater alpha-
904 diversity index (observed ASVs) than that of healthy controls (Iniesta et al., 2023), might due to the
905 different sampling sites between saliva and subgingival plaque. On the other hand, another research
906 has addressed significant discrepancies in alpha-diversity indices from subgingival plaque, saliva, and
907 tongue biofilms from healthy controls and periodontitis patients, resulting the highest alpha-diversity

908 index in saliva collections (Belstrøm et al., 2021). Moreover, early-stage periodontitis, namely stage I,
909 did not determine statistically significant differences in alpha-diversity indices compared to advanced
910 periodontitis, including stage II and stage III (Figure 7a-e). Accordingly, saliva collection of stage I
911 periodontitis may exhibit heterogeneity, indicating a midpoint condition between a healthy state and
912 advanced periodontitis (stage II and stage III). Likewise, gingivitis is often associated with low abundances
913 of the majority of periodontal pathogens, including *Porphyromonas gingivalis*, *Tannerella forsythia*, and
914 *Treponema denticola* (Abusleme et al., 2021). Compared to healthy controls, patients with stage I
915 periodontitis have higher detection rates of *Porphyromonas gingivalis* and *Tannerella forsythia* (Tanner et
916 al., 2006, 2007).

917 Therefore, we calculated beta-diversity indices to analyze the differences between the study partici-
918 pants. The distances for the multiple stages of periodontitis, including stage I, stage II, and stage III, as
919 well as healthy controls (Figure 4g-j and Table 7), suggesting notable differences among the multiple
920 periodontitis severities. In other words, the composition of the salivary microbiome compositions varies
921 depending on the periodontitis stages, so that supporting the findings from a previous study (Iniesta et al.,
922 2023). Taken together that it is nearly impossible to fully restore the attachment level after it has been lost
923 due to the progression and development of periodontitis, the ability to rapidly screen for periodontitis in
924 its early phases using saliva collections would be highly beneficial for effective disease management and
925 treatment.

926 Of the total of 425 taxa in the salivary microbiome composition that have been identified (Figure 13),
927 ANCOM was applied to select 20 taxa as the DAT that indicated notable abundance variation among
928 the periodontitis severities (Figure 8 and Table 5). Three sub-groups were formed from the DAT using
929 hierarchical clustering (Figure 8a). Surprisingly, two of the red complex pathogens (Rôças, Siqueira Jr,
930 Santos, Coelho, & de Janeiro, 2001), *Porphyromonas gingivalis* and *Tannerella forsythia*, were classified
931 in Group 2 and were more prevalent in stage II and stage III periodontitis compared to healthy controls.
932 *Campylobacter showae* was additionally placed in Group 2 of the orange complex pathogens (Gambin et
933 al., 2021). Furthermore, some of the DAT in Group 2 have reported their crucial roles in pathogenesis
934 and development of periodontitis: *Filifactor alocis* (Aruni et al., 2015), *Treponema putidum* (Wyss et
935 al., 2004), *Tannerella forsythia* (Stafford, Roy, Honma, & Sharma, 2012; W. Zhu & Lee, 2016), and
936 *Prevotella intermedia* (Karched, Bhardwaj, Qudeimat, Al-Khabbaz, & Ellepolo, 2022). Taken together,
937 this indicates that DAT in Group 2 is essential to periodontitis. The portion of some Group 1 DAT,
938 including *Peptostreptococcaceae[XI][G-5] saphenum*, *Peptostreptococcaceae[XI][G-6] nodatum*, and
939 *Peptostreptococcaceae[XI][G-9] brachy*, in healthy controls and patients with periodontitis significantly
940 differed, according to earlier research (Lafaurie et al., 2022). These outcomes support our research,
941 implying that Group 1 DAT are also essential to the etiology and progression of periodontitis. However,
942 in contrast to patients with periodontitis, Group 3 DAT, namely *Corynebacterium durum* and *Actinomyces*
943 *graevenitzii*, were enriched in healthy controls, which is consistent with earlier research (Redanz et al.,
944 2021; Nibali et al., 2020).

945 In our correlation analysis (Figure 9), we have discovered strongly negative correlations (coefficient \leq
946 -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 (Table 6). 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%

- 986 • BA: 84%
- 987 • SEN: 94%
- 988 • SPE: 74%

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

- 992 • AUC: $95.3\% \pm 4.9\%$
- 993 • BA: $88.5\% \pm 6.6\%$
- 994 • SEN: $86.4\% \pm 15.7\%$
- 995 • SPE: $90.5\% \pm 7.0\%$

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

998 Furthermore, we have validated our machine-learning prediction model using openly accessible 16S
999 rRNA gene sequencing data from Portuguese (Iniesta et al., 2023) and Spanish participants (Relvas et
1000 al., 2021) in order to ensure the consistency of our random forest classification model (Figure 11). Our
1001 classification models employed in this study were primarily developed and assessed on Korean study par-
1002 ticipants, which may limit their generalizability to other ethnic groups with different salivary microbiome
1003 compositions (Premaraj et al., 2020; Renson et al., 2019). Therefore, the evaluations of this periodonti-
1004 tis classification models can be affected by ethnic-specific variances and differences, highlighting the
1005 necessity for additional validation and adjustment across a spectrum of ethnic backgrounds.

1006 Regarding the clinical characteristics and potential confounders influencing the analysis of salivary
1007 microbiome compositions connected with periodontitis severity, this study had a number of limitations
1008 that were pointed out. We did not offer clinical information, such as the percentage of teeth, the percentage
1009 of bleeding on probing, nor dental furcation involvement, even though we did gather information on
1010 attachment level, probing depth, plaque index, and gingival index (Renvert & Persson, 2002); this might
1011 have it challenging to present thorough and in-depth data about periodontal health. Moreover, the broad age
1012 range may make it tougher to evaluate the relationship between age and periodontitis statuses, providing
1013 the necessity for future studies to consider into account more comprehensive clinical characteristics
1014 associated with periodontitis. Additionally, potential confounders—*e.g.* body mass index (Bombin, Yan,
1015 Bombin, Mosley, & Ferguson, 2022) and e-cigarette use (Suzuki, Nakano, Yoneda, Hirofumi, & Hanioka,
1016 2022)—which might have affected dental health and salivary microbiome composition were disregarding
1017 consideration in addition to smoking status and systemic diseases. Thus, future research incorporating
1018 these components would offer a more thorough knowledge of how lifestyle factors interact and affect the
1019 salivary microbiome composition and periodontal health. Throughout, resolving these limitations will
1020 advance our understanding in pathogenesis and development of periodontitis, offering significant novel
1021 insights on the causal connection between systemic diseases and the salivary microbiome compositions.

1022 **4 Metagenomic signature analysis of Korean colorectal cancer**

1023 **4.1 Introduction**

1024 Colorectal cancer (CRC) is one of the most prevalent and life-threatening malignancies worldwide
1025 (Kuipers et al., 2015; Center, Jemal, Smith, & Ward, 2009; N. Li et al., 2021), with its incidence
1026 influenced by a combination of genetic (Zhuang et al., 2021; Peltomaki, 2003), environmental (O'Sullivan
1027 et al., 2022; Raut et al., 2021), and lifestyle factors (X. Chen et al., 2021; Bai et al., 2022; Zhou et
1028 al., 2022; X. Chen, Li, Guo, Hoffmeister, & Brenner, 2022). Established risk factors include a often
1029 diet in red and processed meats (Kennedy, Alexander, Taillie, & Jaacks, 2024; Abu-Ghazaleh, Chua, &
1030 Gopalan, 2021), obesity (Mandic, Safizadeh, Niedermaier, Hoffmeister, & Brenner, 2023; Bardou et
1031 al., 2022), cigarette smoking (X. Chen et al., 2021; Bai et al., 2022), alcohol consumption (Zhou et al.,
1032 2022; X. Chen et al., 2022), and a sedentary lifestyle (S. An & Park, 2022), all of which contribute to
1033 chronic inflammation, mutagenesis, and metabolic regulation. Additionally, underlying conditions, e.g.
1034 Lynch syndrome (Vasen, Mecklin, Khan, & Lynch, 1991; Hampel et al., 2008) and familial adenomatous
1035 polyposis (Inra et al., 2015; Burt et al., 2004), significantly increase risk of CRC due to persistent mucosal
1036 inflammation and somatic mutations that promote tumorigenesis.

1037 The gut microbiome plays a fundamental role in maintaining host health by helping digestion
1038 (Joscelyn & Kasper, 2014; Cerqueira, Photenhauer, Pollet, Brown, & Koropatkin, 2020), regulating
1039 metabolism (Dabke, Hendrick, Devkota, et al., 2019; Utzschneider, Kratz, Damman, & Hullarg, 2016;
1040 Magnúsdóttir & Thiele, 2018), adjusting immune function (Kau, Ahern, Griffin, Goodman, & Gordon,
1041 2011; Shi, Li, Duan, & Niu, 2017; Broom & Kogut, 2018), and even coordinating neurological processes
1042 by the brain-gut axis (Martin et al., 2018; Aziz & Thompson, 1998; R. Li et al., 2024). Comprising
1043 these gut microbiota, including, archaea, bacteria, fungi, and viruses, the gut microbiome contributes
1044 to the synthesis of essential vitamins, and production of fatty acids, which influence intestinal integrity
1045 and immune responses. Thus, well-balanced gut microbiome composition modulates systemic immune
1046 function by interacting with gut-associated lymphoid tissue, shaping immune tolerance and response
1047 to infections. Hence, emerging evidence suggests that dysbiosis in the gut microbiome composition are
1048 associated not only a narrow range of diseases, e.g. diarrhea and enteritis (Paganini & Zimmermann,
1049 2017; J. Gao, Yin, Xu, Li, & Yin, 2019) but also a wide range of diseases, e.g. obesity, diabetes, and
1050 cancers (Barlow et al., 2015; Hartstra et al., 2015; Helmink et al., 2019; Cullin et al., 2021).

1051 Recent studies have highlighted the crucial role of the gut microbiome in tumorigenesis and progres-
1052 sion of CRC (Song, Chan, & Sun, 2020; Rebersek, 2021), with dysbiosis emerging as a potential risk
1053 factor. Dysbiosis in gut microbiome compositions can promote tumorigenesis of many cancers, including
1054 CRC, through several signaling cascades, including inflammation, mutagenesis, and altered metabolism
1055 in host. Certain bacteria species, such as *Fusobacterium* genus (Hashemi Goradel et al., 2019; Bullman et
1056 al., 2017; Flanagan et al., 2014), *Bacteroides* genus (Ulger Toprak et al., 2006; Boleij et al., 2015), and
1057 *Escherichia coli* (Swidsinski et al., 1998; Bonnet et al., 2014), have been associated with development
1058 and progression of CRC by producing pro-inflammatory signals, generating toxins including mutagens,

1059 and disrupting the intestinal barriers including mucous surface. In contrast, beneficial bacteria, such as
1060 *Lactobacillus* genus (Ghorbani et al., 2022; Ghanavati et al., 2020) and *Bifidobacterium* genus (Le Leu,
1061 Hu, Brown, Woodman, & Young, 2010; Fahmy et al., 2019), are regarded to apply protective roles by
1062 maintaining homeostasis of gut microbiome compositions and regulating immune responses including
1063 inflammation.

1064 Furthermore, identifying metagenome biomarkers in Korean CRC patients is essential, as the gut
1065 microbiome compositions significantly vary by ethnicity due to genetic, dietary, and environmental
1066 factor (Fortenberry, 2013; Merrill & Mangano, 2023; Parizadeh & Arrieta, 2023). Additionally, ethnicity-
1067 specific microbiome composition signatures may affect the reliability of previously established biomarkers
1068 derived from predominantly Western CRC cohorts (Network et al., 2012), necessitating population-
1069 specific investigations. By identifying metagenomic biomarkers tailored to Korean CRC patients, we
1070 can improve early detection rate of early-stage CRC, develop more accurate risk of CRC, and explore
1071 microbiome-targeted therapies that consider host-microbiome interactions within the Korean population.

1072 Accordingly, this study aims to identify microbiome-based biomarkers specific to CRC within
1073 the Korean population, addressing the critical demand for ethnicity-specific microbiome research. By
1074 leveraging metagenomic sequencing and advanced computational biology analysis, this study seeks to
1075 uncover novel microbial signatures associated with Korean CRC patients. As part of the larger "Multi-
1076 genomic analysis for biomarker development in colon cancer" project (NTIS No. 1711055951), this study
1077 investigates microbial signatures within next-generation sequencing data to enhance precision medicine
1078 approaches for CRC and to develop robust microbiome-based biomarkers for early detection, prognosis,
1079 and therapeutic stratification, complementing genomic and epigenomic markers. Hence, this research
1080 represents a crucial step toward personalized cancer diagnostic and therapeutic strategies tailored to the
1081 Korean population.

1082 **4.2 Materials and methods**

1083 **4.2.1 Study participants enrollment**

1084 To achieve metagenomic observations of CRC, a total of 211 Korean CRC patients were enrolled (Table
1085 8). The tissue samples were collected from both the tumor lesion and its corresponding adjacent normal
1086 lesion to enable comparative metagenomic analyses. Tumor tissue samples were obtained from confirmed
1087 CRC lesions, ensuring adequate representation of CRC-associated microbial alterations. Adjacent normal
1088 tissues were collected from non-cancerous regions away from the tumor margin to serve as a control
1089 for baseline molecular and microbial composition. Moreover, clinical information was collected for all
1090 study participants included in this study to investigate potential associations between gut microbiome
1091 compositions and clinical outcomes. Key clinical characteristics recorded included overall survival (OS)
1092 and recurrence. These clinical parameters were integrated with metagenomic data to explore potential
1093 microbiome-based biomarkers for CRC prognosis and progression. Ethical approval was obtained for
1094 clinical data collection, and all patient information was anonymized to ensure confidentiality in accordance
1095 with institutional guidelines.

1096 **4.2.2 DNA extraction procedure**

1097 Tissue samples were immediately processed under sterile conditions to prevent contamination and
1098 preserved in low temperature (-80°C) storage for downstream DNA extraction and whole-genome
1099 sequencing. Furthermore, produced sequencing data were provided by the "Multi-genomic analysis
1100 for biomarker development in colon cancer" project (NTIS No. 1711055951) in mapped BAM format,
1101 aligned to the hg38 human reference genome. The preprocessing pipeline utilized by the main project
1102 included high-throughput whole-genome sequencing using standardized alignment algorithm, BWA
1103 (H. Li & Durbin, 2009). In addition to the mapped human sequences, our whole-genome sequencing
1104 data retained unmapped sequences, which contain potential microbial reads that were not aligned to the
1105 human reference genome.

1106 **4.2.3 Bioinformatics analysis**

1107 To identify microbial signatures associated with CRC, we employed PathSeq (version 4.1.8.1) (Kostic
1108 et al., 2011; Walker et al., 2018), a computational pipeline designed for metagenomic analysis of high-
1109 throughput sequencing data including the whole-genome sequences. After processing these sequencing
1110 data through the PathSeq pipeline, a comprehensive bioinformatics analyses were conducted to characterize
1111 microbial signatures associated with CRC.

1112 Prevalent taxa identification was performed by determining microbial taxa present in the majority of
1113 the study participants, filtering out low-abundance and rare taxa to ensure robust downstream analyses.

1114 To assess microbial community structure, diversity indices were calculated, including alpha-diversity
1115 to evaluate single-sample diversity and beta-diversity to compare microbial composition between the
1116 tumor tissues and their corresponding adjacent normal tissues. Following alpha-diversity indices were

1117 calculated using the scikit-bio Python package (version 0.6.3) (Rideout et al., 2018), and these alpha-
1118 diversity indices were compared using the MWU test:

- 1119 1. Berger-Parker d (Berger & Parker, 1970)
- 1120 2. Chao1 (Chao, 1984)
- 1121 3. Dominance
- 1122 4. Doubles
- 1123 5. Fisher (Fisher et al., 1943)
- 1124 6. Good's coverage (Good, 1953)
- 1125 7. Margalef (Magurran, 2021)
- 1126 8. McIntosh e (Heip, 1974)
- 1127 9. Observed ASVs (DeSantis et al., 2006)
- 1128 10. Simpson d
- 1129 11. Singles
- 1130 12. Strong (Strong, 2002)

1131 Furthermore, these beta-diversity indices were measured and compared using the PERMANOVA
1132 test (Anderson, 2014; Kelly et al., 2015). To demonstrate multi-dimensional data from the beta-diversity
1133 indices, we utilized the t-SNE algorithm (Van der Maaten & Hinton, 2008).

- 1134 1. Bray-Curtis (Sorensen, 1948)
- 1135 2. Canberra
- 1136 3. Cosine (Ochiai, 1957)
- 1137 4. Hamming (Hamming, 1950)
- 1138 5. Jaccard (Jaccard, 1908)
- 1139 6. Sokal-Sneath (Sokal & Sneath, 1963)

1140 Differentially abundant taxa (DAT) were identified using statistical method, ANCOM (Lin & Peddada,
1141 2020), adjusting for sequencing depth and potential confounders to highlight taxa significantly associated
1142 with categorical clinical information in CRC, such as recurrence. Furthermore, to point attention to
1143 taxa that are substantially linked to continuous clinical measurement in CRC, including OS, DAT were
1144 found using the Spearman correlation and slope from linear regression (Equation 9). Note that both the
1145 Spearman correlation and the slope from linear regression were utilized to provide a more comprehensive
1146 assessment of the relationship between DAT proportions and OS. While the correlation coefficient
1147 measures the strength and direction of a linear relationship between these variables, it does not convey
1148 information about the magnitude of change in independent variable relative to dependent variable. The
1149 slope of the linear regression model, on the other hand, quantifies this change by indicating how much
1150 the dependent variable is expected to increase or decrease per unit change in the independent variable. By
1151 incorporating both the correlation coefficient and the slope from the linear regression, we ensured that
1152 the analysis captured not only whether two variables were associated but also the extent to which one
1153 variable influenced the other. This dual approach enhances the interpretability of results, particularly in
1154 biological and clinical studies where both statistical association and biological effect size are crucial for
1155 meaningful suggestions.

$$\text{slope} = \frac{\Delta \text{OS}}{\Delta \text{DAT proportion}} \quad (9)$$

1156 To assess the predictive potential of microbial signatures in CRC prognosis, we employed a random
1157 forest machine learning model using DAT proportions as input features. Random forest classification was
1158 utilized to predict CRC recurrence, where the classification model was trained to distinguish between
1159 CRC patients with or without recurrence based on the gut microbiome compositions. Additionally,
1160 random forest regression was applied to predict OS by estimating survival time as a continuous clinical
1161 outcome based on microbiome features. This approach allowed for the identification of microbial taxa
1162 that contribute significantly to CRC prognosis, offering insights into potential gut microbiome-based
1163 biomarkers for cancer progression. By integrating these random forest machine learning models, we
1164 aimed to improve CRC risk stratification and precision medicine strategies.

1165 This multi-layered bioinformatics approach enabled a comprehensive investigation of gut microbiome
1166 alteration in CRC, facilitating the identification of potential microbial biomarkers for diagnosis and
1167 prognosis of CRC.

1168 **4.2.4 Data and code availability**

1169 All sequences from the 211 study participants have been published to the Korea Bioinformation Center
1170 (data ID KGD10008857): <https://kbds.re.kr/KGD10008857>. Docker image that employed through-
1171 out this study is available in the DockerHub: <https://hub.docker.com/repository/docker/fumire/unist-crc-copm/general>. Every code used in this study can be found on GitHub: <https://github.com/CompbioLabUnist/CoPM-ColonCancer>.

1174 **4.3 Results**

1175 **4.3.1 Summary of clinical characteristics**

1176 Microsatellite instability (MSI) is one of the key molecular features and risk factors in CRC, resulting
1177 from defects in the DNA mismatch repair system (Boland & Goel, 2010). MSI leads to the accumulation
1178 of mutations in short repetitive DNA sequences (microsatellites), contributing to genomic instability and
1179 tumor development (Søreide, Janssen, Söiland, Körner, & Baak, 2006; Vilar & Gruber, 2010). Therefore,
1180 we compared clinical measurements with MSI status, including microsatellite stable (MSS), MSI-low
1181 (MSI-L), and MSI-high (MSI-H). There were no significant differences in the clinical measurements, *e.g.*
1182 recurrence, sex, OS, and age in diagnosis, in the total of 211 study participants (Table 8).

1183 **4.3.2 Gut microbiome compositions**

1184 In the total of 211 CRC study participants, these ten kingdoms were found in the gut microbiome
1185 composition:

- 1186 1. Archaea kingdom: 31 genera
- 1187 2. Bacteria kingdom: 1508 genera
- 1188 3. Bamfordvirae kingdom: 1 genus
- 1189 4. Eukaryota kingdom: 77 genera
- 1190 5. Fungi kingdom: 137 genera
- 1191 6. Loebvirae kingdom: 2 genera
- 1192 7. Orthornavirae kingdom: 1 genus
- 1193 8. Parnavirae kingdom: 3 genera
- 1194 9. Shotokuvirae kingdom: 6 genera
- 1195 10. Viruses kingdom: 76 genera

1196 Among these kingdoms, the proportions of four major kingdoms, which have at least 50 genera, in
1197 the gut microbiome composition were displayed (Figure 21): Bacteria kingdom, Eukaryota kingdom,
1198 Fungi kingdom, and Viruses kingdom. In the Bacteria kingdom (Figure 21a), *Bacteroides* genus is the
1199 most prevalent genus in the tumor tissue samples, followed by *Fusobacterium* and *Cutibacterium* genera.
1200 *Toxoplasma* and *Malassezia* genera were the dominant genus, which have over 90% of proportions, in
1201 the Eukaryota kingdom (Figure 21b) and the Fungi kingdom (Figure 21c), respectively. On the other
1202 hand, *Roseolovirus* genus is the most popular genus of the Viruses kingdom in the normal tissue samples
1203 (Figure 21d); contrarily, *Lymphocryptovirus* and *Cytomegalovirus* genera had been dominant genera in
1204 the tumor tissue samples. Taken together, these results suggest that the Anna Karenina principle (Ma,
1205 2020; W. Li & Yang, 2025), *i.e.* in human microbiome-associated diseases, every disease-associated
1206 microbiome, including dysbiosis, is unique and patient-specific, whereas all healthy microbiomes are
1207 similar, also applies to CRC.

1208 **4.3.3 Diversity indices**

1209 In alpha-diversity analysis, which measures within-sample microbial community, revealed a significant
1210 increase in tumor tissue samples compared to adjacent normal tissue samples (Figure 22). Alpha-diversity
1211 indices, including Chao1, Fisher α , and observed features, were consistently higher in CRC tumor tissues
1212 (MWU test $p < 0.05$), indicating a more heterogeneous microbial community, *e.g.* the Anna Karenina
1213 principle, potentially influenced by tumor-associated dysbiosis.

1214 To assess the microbial impact on CRC recurrence, alpha-diversity indices compared between normal
1215 and tumor tissue samples in accordance with recurrence information (Figure 23). In the recurrence
1216 patients, most alpha-diversity indices (11 out of 12), except McIntosh index, exhibited increasing in
1217 tumor tissue samples than normal tissue samples (MWU test $p < 0.05$; Figure 23); In the non-recurrence
1218 patients, on the other hand, some alpha-diversity indices (8 out of 12) amplified in tumor tissue samples
1219 than normal tissue samples (MWU test $p < 0.05$; Figure 23). What is interesting about the alpha-diversity
1220 analysis in this figure is that a few indices, namely Fisher α (Figure 28e) and Margalef (Figure 23g),
1221 presented augmentation in normal tissue sample of the recurrence patients than that of the non-recurrence
1222 patients (MWU test $p < 0.05$). Overall, these alpha-diversity results demonstrate that tumor tissue samples
1223 have more diverse microbiome composition than normal tissue samples. Furthermore, although only
1224 two indices significantly increased, the recurrence patients have diversified microbiome compositions
1225 than the non-recurrence patients in normal sample tissue, not in tumor sample tissues, indicating field
1226 cancerization by the gut microbiome leads to unfavorable prognosis such as recurrence (Curtius et al.,
1227 2018; Rubio et al., 2022).

1228 To determine the microbial impact on OS of CRC patients, the Spearman correlation compared
1229 between alpha-diversity indices and OS duration (Figure 24). No significant Spearman correlation was
1230 found between every alpha-diversity indices and OS (Spearman correlation $p \geq 0.1$; Figure 24). However,
1231 a few alpha-diversity indices, *e.g.* Chao1 (Figure 24b), Good's coverage (Figure 24f), and observed
1232 features (Figure 24i), showed negative correlations with OS (Spearman correlation $p < 0.05$). Together
1233 these correlation results provide important insights into heterogeneous microbiome leads to shorter OS,
1234 suggesting the Anna Karenina principle and the field cancerization.

1235 In beta-diversity analysis, which calculates inter-sample microbial community, explain significant
1236 disparity between tumor tissue samples and normal tissue samples (Figure 25). Every six beta-diversity
1237 indices presented discrepancy between normal tissue samples and tumor tissue samples (PERMANOVA
1238 test $p < 0.001$), implying that tumor tissue samples have distinct microbiome compositions from normal
1239 tissue samples.

1240 Beta-diversity indices were evaluated between normal and tumor tissue samples along with recurrence
1241 history in order to evaluate the microbial influence on CRC recurrence (Figure 26). All six beta-diversity
1242 indices examined significant difference in microbial community structure between the recurrence patients
1243 and the non-recurrence patients (PERMANOVA test $p < 0.001$; Figure 26), indicating that tumor-
1244 associated gut microbiome composition varies resulting on recurrence status. tSNE-transformed plots
1245 further illustrated clear clustering patterns (Figure 26), suggesting again that the recurrence patients

1246 harbor dissimilar microbial communities compared to the non-recurrence patients. These observed
1247 differences in beta-diversity represent that microbial shifts, including dysbiosis, may be associated with
1248 CRC progression and recurrence risk, possibly due to specific taxa contributing to a tumor-promoting
1249 microenvironment.

1250 Moreover, beta-diversity analysis suggested a potential associated with OS duration in CRC patients.
1251 In all six beta-diversity indices, tSNE-transformed plots showed clear clustering patterns along OS
1252 duration (Figure 27), implying that possible microbiome composition shifts related to survival outcomes
1253 in CRC. However, since OS is a continuous variable, statistical significance testing could not be directly
1254 performed for these clustering patterns. Despite this limitation, the observed microbial community
1255 variations suggest that alterations in the gut microbiome composition may be associated to CRC prognosis
1256 and survival duration.

1257 Together, diversity indices analyses revealed significant microbial community alterations between
1258 normal and tumor tissue samples, as well as between the recurrence and non-recurrence CRC patients.
1259 Alpha-diversity indices significantly increased in tumor tissue samples than normal tissue samples (MWU
1260 test $p < 0.05$; Figure 22). This increase was more pronounced in the recurrence patients (11 of 12
1261 indices) compared to non-recurrence patients (8 of 12 indices) (Figure 23), indicating a potential link
1262 between microbial diversity and CRC recurrence. Additionally, negative correlation between OS and
1263 alpha-diversity indices were observed in normal samples (Spearman correlation $p < 0.05$; Figure 24),
1264 suggesting that lower microbial diversity may be associated with longer survival in CRC. On the other
1265 hand, beta-diversity indices analysis, showed significant separation between tumor and tumor tissue
1266 samples across all six beta-diversity indices (PERMANOVA test $p < 0.001$; Figure 25). Furthermore,
1267 the recurrence and non-recurrence patients displayed significantly discrete microbial compositions
1268 (PERMONOVA test $p < 0.001$; Figure 26), implying that microbial community shifts may reflect CRC
1269 progression and recurrence risk. These findings highlight the importance of microbiome diversity and
1270 gut microbiome composition in CRC prognosis and warrant further investigation into their potential as
1271 predictive biomarkers.

1272 4.3.4 DAT selection

1273 The selection of differentially abundant taxa (DAT) aimed to identify microbial taxa that exhibit significant
1274 differences in relative abundance between clinical information, such as recurrence history or OS in CRC
1275 patients. Identifying and selection these microbial discrepancies is crucial for understanding the role of
1276 the gut microbiome composition in CRC progression, prognosis, and potential therapeutic interventions.

1277 We identified 19 DAT associated with recurrence history across the total samples by ANCOM (Figure
1278 28a), including 18 non-recurrence-enriched DAT and a recurrence-related DAT. When stratified by sample
1279 type, one DAT was enriched in normal samples of the non-recurrence patients (Figure 28b), whereas six
1280 DAT exhibited significant differential abundance in tumor samples (Figure 28c). These findings suggest
1281 that microbial composition variations in the tumor microenvironment are more pronounced in relation to
1282 recurrence status (Table 9), potentially indicating a microbial signature linked to CRC progression. These

1283 identified DAT may contribute to tumor-associated dysbiosis, influencing the likelihood of CRC recurrence
1284 through mechanisms such as inflammation, metabolic modulation, or immune system interaction.

1285 The non-recurrence-enriched DAT have decreased proportions both in normal and tumor samples of
1286 the recurrence patients than those in the non-recurrence patients (MWU test $p < 0.001$; Figure 28d-h).
1287 What is interesting about these non-recurrence-enriched DAT is that they belong to the *Micrococcus* genus.
1288 Among them, *Micrococcus aloeverae* was consistently identified in all three settings—total (Figure 28a),
1289 normal (Figure 28b), and tumor samples (Figure 28c)—indicating its stable presence regardless of tissue
1290 type. Variation in relative proportions of *Micrococcus aloeverae* (Figure 28d) suggests potential ecological
1291 adaptability within tumor microenvironment of CRC. The remaining *Micrococcus*-related DAT showed
1292 less variation between the recurrence and non-recurrence patients, reinforcing their limited associations
1293 with CRC recurrence. Moreover, only one taxon, *Pseudomonas* sp. *NBRC 111133*, was identified as
1294 recurrence-enriched DAT (Figure 28a). This suggests a potential association between *Pseudomonas* sp.
1295 *NBRC 111133* and CRC recurrence, indicating that its presence may contribute to a tumor-supportive
1296 microbial environment. *Pseudomonas* sp. *NBRC 111133* had higher relative proportions both in normal
1297 and tumor tissue samples of the recurrence patients than those of the non-recurrence patients (Figure
1298 28i). Likewise, *Pseudomonas* sp. *NBRC 111133* were prevalent in tumor tissue samples than normal
1299 tissue samples of the non-recurrence patients (MWU test $p < 0.01$; Figure 28i); however, no significant
1300 difference between normal and tumor tissue samples of the recurrence patients.

1301 These findings imply that while certain species belong to *Micrococcus* genus may be prevalent in CRC
1302 tissues, their roles in cancer progression and recurrence risk remain uncertain. Species of *Pseudomonas*
1303 genus are known for their metabolic involvement in biofilm formation, antibiotic resistance, and immune
1304 modulation, which could play a role in CRC progression.

1305 Furthermore, correlation analysis between DAT abundance and OS duration identified a total of 16
1306 over-represented DAT in the total samples (Figure 29a). When analyzed separately, 11 DAT, which consist
1307 of four under-represented and seven over-represented DAT showed significant correlations with OS in
1308 normal samples (Figure 29b), while four under-represented and 45 over-represented DAT were identified
1309 in tumor samples (Figure 29c), indicating that microbial composition shifts in tumor tissues may have a
1310 stronger association with survival outcomes. The higher number of survival-associated DAT in tumor
1311 tissue suggests that the tumor microbiome plays a more dynamic role in progression and prognosis
1312 of CRC. These findings highlight the potential of gut microbial composition as a prognostic indicator
1313 in CRC, warranting further investigation into the functional roles of these DAT in influencing clinical
1314 outcomes.

1315 Among a total of 57 OS-correlated DAT (Table 10) with Spearman correlation and the slope (Equation
1316 9). *Agaricus bisporus* (Figure 29d) and *Corynebacterium* sp. *KPLI824* (Figure 29h) are identified as
1317 over-represented DAT both in normal samples and tumor samples (Spearman correlation $p < 0.05$),
1318 whereas *Corynebacterium lowii* (Figure 29g) and *Paracoccus sphaerophysae* (Figure 29i) are selected
1319 as under-represented DAT both in normal samples and tumor samples (Spearman correlation $p < 0.05$).
1320 On the other hand, *Clostridiales bacterium* (Figure 29e) is classified as under-represented DAT only in
1321 normal samples (Spearman correlation $p < 0.01$), while *Corynebacterium kroppenstedtii* (Figure 29f) is

1322 described as over-represented DAT only in tumor samples (Spearman correlation $p < 0.001$).

1323 These findings highlight the potential influence of microbial dysbiosis on cancer progression and
1324 prognosis. The presence of these OS-correlated DAT in tumor and/or adjacent normal tissues suggests
1325 that microbial alterations may contribute to field cancerization, a phenomenon where histopathologically
1326 benign tissues surrounding the tumor undergo molecular, inflammatory, and microbial shifts, creating
1327 a microenvironment conducive to tumor development and progression. Therefore, these discoveries
1328 reinforce the importance of investigating the gut microbiome as a prognostic biomarker and suggest that
1329 targeting microbial dysbiosis could offer new therapeutic strategies for improving clinical outcomes of
1330 CRC.

1331 **4.3.5 Random forest prediction**

1332 We employed the random forest-based machine learning prediction to assess the predictive power of DAT
1333 from gut microbiome composition for CRC prognosis. To achieve this aim, we utilized random forest
1334 classification to predict recurrence status, training the model to differentiate between recurrence and
1335 non-recurrence patients based on microbial abundance patterns. Additionally, we applied random forest
1336 regression to predict OS, aiming to identify microbial taxa associated with survival duration. By leveraging
1337 random forest models, this study aimed to establish a microbiome-based predictive machine learning
1338 models for CRC recurrence risk assessment and survival prognosis, contributing to the development of
1339 prediction medicine strategies based on gut microbial signatures.

1340 To evaluate the predictive power of gut microbiome composition in CRC recurrence, we implemented
1341 a random forest classification model using two different input sets (Figure 30a-f): the entire gut mi-
1342 crobiome composition and DAT. Comparing these models allowed us to assess whether focusing on
1343 DAT-selected microbial features enhances classification performance. While the DAT-based classification
1344 models showed slightly improved classification metrics (MWU test $p \geq 0.05$), including ACC, AUC, and
1345 BA, over the entire microbiome-based model in the total sample (Figure 30a and Figure 29d), normal sam-
1346 ples (Figure 30b and Figure 30e), and tumor samples (Figure 30c and Figure 30f), overall classification
1347 metrics remained around 60%, suggesting moderated predictive capability. This relatively low metrics
1348 highlight the complexity of CRC recurrence, indicating that while dysbiosis may contribute to CRC
1349 progression, it is likely interwinded with host genetic factors such as germline and somatic mutations.
1350 Thus, the interplay between microbial shifts and tumor genomic alterations warrants further investigation,
1351 as integrating microbiome and genomic sequencing data may improve therapeutic strategies.

1352 To assess the predictive capability of the gut microbiome composition in OS of CRC patients, we
1353 implemented a random forest regression model, comparing two different input sets (Figure 30g-i): the
1354 entire gut microbiome composition and DAT. This comparison also aimed to determine whether focusing
1355 on key microbial features (DAT) enhances predictive accuracy. While DAT-based model showed a slight
1356 improvement over the entire microbiome-based model in normal samples (Figure 30h) and tumor samples
1357 (Figure 30i), the regression error remained high (about 700 days), indicating substantial variability in
1358 survival outcomes that cannot be fully explained by gut microbiome composition alone. This result

1359 suggest that while gut microbial dysbiosis may influence CRC progression, survival duration (OS) is
1360 likely also driven by host genetic factors, highlighting the requirement for multi-omics integration, where
1361 combining microbiome and genomic sequencing data may provide a more accurate and comprehensive
1362 predictive model for CRC patients survival.

Table 8: Clinical characteristics of CRC study participants.

Statistical significance were assessed using the χ -squared test for categorical values and the Kruskal-Wallis test for continuous values. OS: overall survival.

	Overall	MSS	MSI-L	MSI-H	p-value
n	211	181	7	18	
Recurrence, n (%)	False	132 (62.6%)	112 (61.9%)	4 (57.1%)	0.657
	True	79 (37.4%)	69 (38.1%)	3 (42.9%)	
Sex, n (%)	Male	137 (64.9%)	119 (65.7%)	6 (85.7%)	0.357
	Female	74 (35.1%)	62 (34.3%)	1 (14.3%)	
OS, mean±SD	1248.5±770.3	1268.1±769.5	1416.6±496.3	1097.7±903.2	0.580
Age, mean±SD	61.2±13.1	61.7±12.4	60.1±15.6	60.2±19.4	0.867

Table 9: DAT list for CRC recurrence.

Significance threshold is $|\log_2 \text{FC}| > 1.0| \wedge W > 9600$. Non-significant values remain blank. DAT are sorted in alphabetical order. FC: fold change

Taxonomy name	Entire-log ₂ FC	Entire-W	Normal-log ₂ FC	Normal-W	Tumor-log ₂ FC	Tumor-W
<i>Cutibacterium acnes</i>	-1.878	10570				
<i>Cutibacterium avidum</i>	-1.383	10266				
<i>Cutibacterium granulosum</i>	-1.476	10271				
<i>Micrococcus aloeverae</i>	-2.280	10740	-1.821	10462	-2.481	10591
<i>Micrococcus luteus</i>	-2.216	10744				
<i>Micrococcus</i> sp. <i>CH3</i>	-2.323	10740			-2.493	10527
<i>Micrococcus</i> sp. <i>CH7</i>	-2.321	10740			-2.493	10542
<i>Micrococcus</i> sp. <i>HMSC31B01</i>	-2.282	10739			-2.458	10519
<i>Micrococcus</i> sp. <i>MS-ASIII-49</i>	-2.284	10740			-2.470	10527
<i>Pseudomonas</i> sp. <i>NBRC 111133</i>	1.139	9732				
<i>Pseudonocardia</i> sp. <i>P2</i>	-2.200	10736			-2.394	10253
<i>Staphylococcus</i> sp. <i>HMSC034A07</i>	-1.341	10050				
<i>Staphylococcus</i> sp. <i>HMSC063F03</i>	-1.322	10001				
<i>Staphylococcus</i> sp. <i>HMSC064E11</i>	-1.064	10163				
<i>Staphylococcus</i> sp. <i>HMSC067B04</i>	-1.343	9952				
<i>Staphylococcus</i> sp. <i>HMSC068G12</i>	-1.344	10173				
<i>Staphylococcus</i> sp. <i>HMSC072H01</i>	-1.298	10197				
<i>Staphylococcus</i> sp. <i>HMSC077C03</i>	-1.331	10115				
<i>Treponema endosymbiont of Eucomonympha</i> sp.	-1.629	10472				

Table 10: DAT list for CRC OS.

Significance threshold is $\log_{10}|\text{slope}| > 2.0 \wedge |r| > 0.2$. Non-significant values remain blank. DAT are sorted in alphabetical order.

Taxonomy name	Entire-slope	Entire-r	Normal-slope	Normal-r	Tumor-slope	Tumor-r
<i>Acinetobacter venetianus</i>					3.087	0.203
<i>Actinotalea ferrariae</i>					2.574	0.200
<i>Agaricus bisporus</i>	2.329	0.287	2.925	0.276	2.258	0.306
<i>Bifidobacterium boum</i>					2.096	-0.216
<i>Brevundimonas</i> sp. <i>DS20</i>			2.180	0.279		
<i>Clostridiales bacterium</i>			2.631	-0.203		
<i>Corynebacterium kroppenstedtii</i>	2.117	0.220			2.117	0.302
<i>Corynebacterium lipophiloflavum</i>			2.137	0.227		
<i>Corynebacterium lowii</i>			2.006	-0.216		
<i>Corynebacterium</i> sp. <i>KPL1818</i>	2.101	0.209	2.487	0.220	2.044	0.215
<i>Corynebacterium</i> sp. <i>KPL1824</i>	2.057	0.207	2.511	0.212	2.003	0.226
<i>Corynebacterium</i> sp. <i>KPL1986</i>					2.205	0.202
<i>Corynebacterium</i> sp. <i>KPL1996</i>					2.205	0.202
<i>Corynebacterium</i> sp. <i>KPL1998</i>					2.205	0.202
<i>Corynebacterium</i> sp. <i>KPL2004</i>					2.205	0.202
<i>Kocuria flava</i>			2.729	0.214		
<i>Kytococcus sedentarius</i>					2.267	0.206
<i>Lachnospiraceae bacterium AD3010</i>			2.609	-0.203		
<i>Lachnospiraceae bacterium NK4A136</i>					2.538	-0.220
<i>Methylorum extorquens</i>					2.068	0.295
<i>Microbacterium barkeri</i>			2.071	0.389		
<i>Paracoccus sphaerophysae</i>					2.012	-0.209
<i>Pontibacillus litoralis</i>					2.580	-0.209
<i>Porphyromonas macacae</i>			2.476	-0.200		
<i>Pseudomonas balearica</i>					2.117	0.203
<i>Pseudomonas monteilii</i>					2.183	0.228
<i>Rodentibacter myodis</i>					2.444	0.245
<i>Roseovarius tolerans</i>					2.295	0.221
<i>Staphylococcus epidermidis</i>					2.243	0.214
<i>Staphylococcus</i> sp. <i>HMSC034A07</i>					2.183	0.209
<i>Staphylococcus</i> sp. <i>HMSC034D07</i>	2.278	0.206			2.252	0.253
<i>Staphylococcus</i> sp. <i>HMSC034G11</i>	2.362	0.208			2.357	0.261
<i>Staphylococcus</i> sp. <i>HMSC036A09</i>					2.308	0.239
<i>Staphylococcus</i> sp. <i>HMSC055A10</i>					2.168	0.222
<i>Staphylococcus</i> sp. <i>HMSC055B03</i>	2.134	0.202			2.134	0.266
<i>Staphylococcus</i> sp. <i>HMSC058E12</i>					2.106	0.216
<i>Staphylococcus</i> sp. <i>HMSC061C10</i>					2.882	0.207
<i>Staphylococcus</i> sp. <i>HMSC062B11</i>	2.391	0.203			2.377	0.253
<i>Staphylococcus</i> sp. <i>HMSC062D04</i>	2.278	0.202			2.274	0.259
<i>Staphylococcus</i> sp. <i>HMSC063F03</i>	2.376	0.201			2.367	0.251
<i>Staphylococcus</i> sp. <i>HMSC063F05</i>	2.387	0.210			2.381	0.266
<i>Staphylococcus</i> sp. <i>HMSC064E11</i>					2.276	0.218
<i>Staphylococcus</i> sp. <i>HMSC065D11</i>					2.329	0.245

Table 10 continued from previous page

Taxonomy name	Entire-slope	Entire-r	Normal-slope	Normal-r	Tumor-slope	Tumor-r
<i>Staphylococcus</i> sp. <i>HMSC066G04</i>					2.181	0.218
<i>Staphylococcus</i> sp. <i>HMSC067B04</i>	2.332	0.205			2.329	0.260
<i>Staphylococcus</i> sp. <i>HMSC068G12</i>					2.294	0.226
<i>Staphylococcus</i> sp. <i>HMSC070A07</i>	2.360	0.216			2.362	0.287
<i>Staphylococcus</i> sp. <i>HMSC073C02</i>	2.352	0.205			2.334	0.246
<i>Staphylococcus</i> sp. <i>HMSC073E10</i>					2.366	0.255
<i>Staphylococcus</i> sp. <i>HMSC074D07</i>	2.330	0.218			2.308	0.270
<i>Staphylococcus</i> sp. <i>HMSC076H12</i>					2.200	0.219
<i>Staphylococcus</i> sp. <i>HMSC077C03</i>					2.258	0.207
<i>Staphylococcus</i> sp. <i>HMSC077D09</i>					2.245	0.230
<i>Staphylococcus</i> sp. <i>HMSC077G12</i>	2.335	0.200			2.345	0.276
<i>Staphylococcus</i> sp. <i>HMSC077H01</i>					2.214	0.241
<i>Streptomyces cinnamoneus</i>					2.787	0.208
<i>Thauera terpenica</i>					2.975	0.226

Table 11: Random forest classification and their evaluations.

	Dataset	ACC	AUC	BA	F1	PRE	SEN	SPE
Entire	Total	0.544±0.139	0.667±0.141	0.561±0.141	0.544±0.139	0.559±0.152	0.562±0.192	0.559±0.152
	Normal	0.464±0.214	0.571±0.182	0.484±0.210	0.464±0.214	0.515±0.200	0.454±0.255	0.515±0.200
	Tumor	0.481±0.176	0.615±0.087	0.497±0.181	0.481±0.176	0.464±0.189	0.530±0.212	0.464±0.189
DAT	Total	0.582±0.112	0.656±0.109	0.592±0.120	0.582±0.112	0.558±0.114	0.626±0.167	0.558±0.114
	Normal	0.530±0.117	0.567±0.102	0.553±0.123	0.530±0.117	0.501±0.117	0.604±0.194	0.501±0.117
	Tumor	0.478±0.122	0.570±0.164	0.504±0.143	0.478±0.122	0.527±0.240	0.480±0.119	0.527±0.240

Table 12: **Random forest regression and their evaluations.**

Dataset		MAE	RMSE
Entire	Total	704.909 ± 249.010	894.943 ± 246.192
	Normal	803.487 ± 145.365	979.334 ± 158.813
	Tumor	811.505 ± 204.788	1005.182 ± 197.351
DAT	Total	823.700 ± 141.448	994.698 ± 157.983
	Normal	663.414 ± 147.203	825.461 ± 151.120
	Tumor	729.302 ± 179.940	884.863 ± 181.154

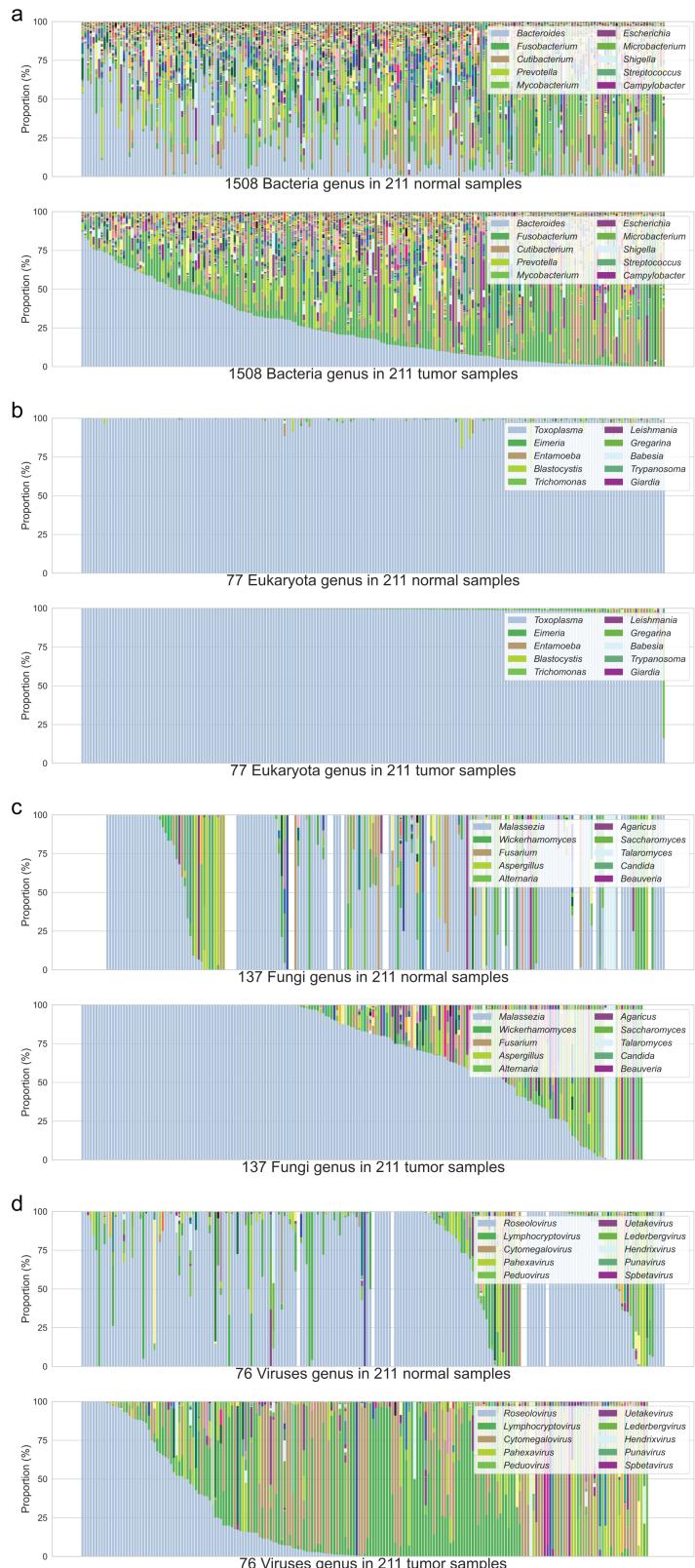


Figure 21: Gut microbiome compositions in genus level.

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

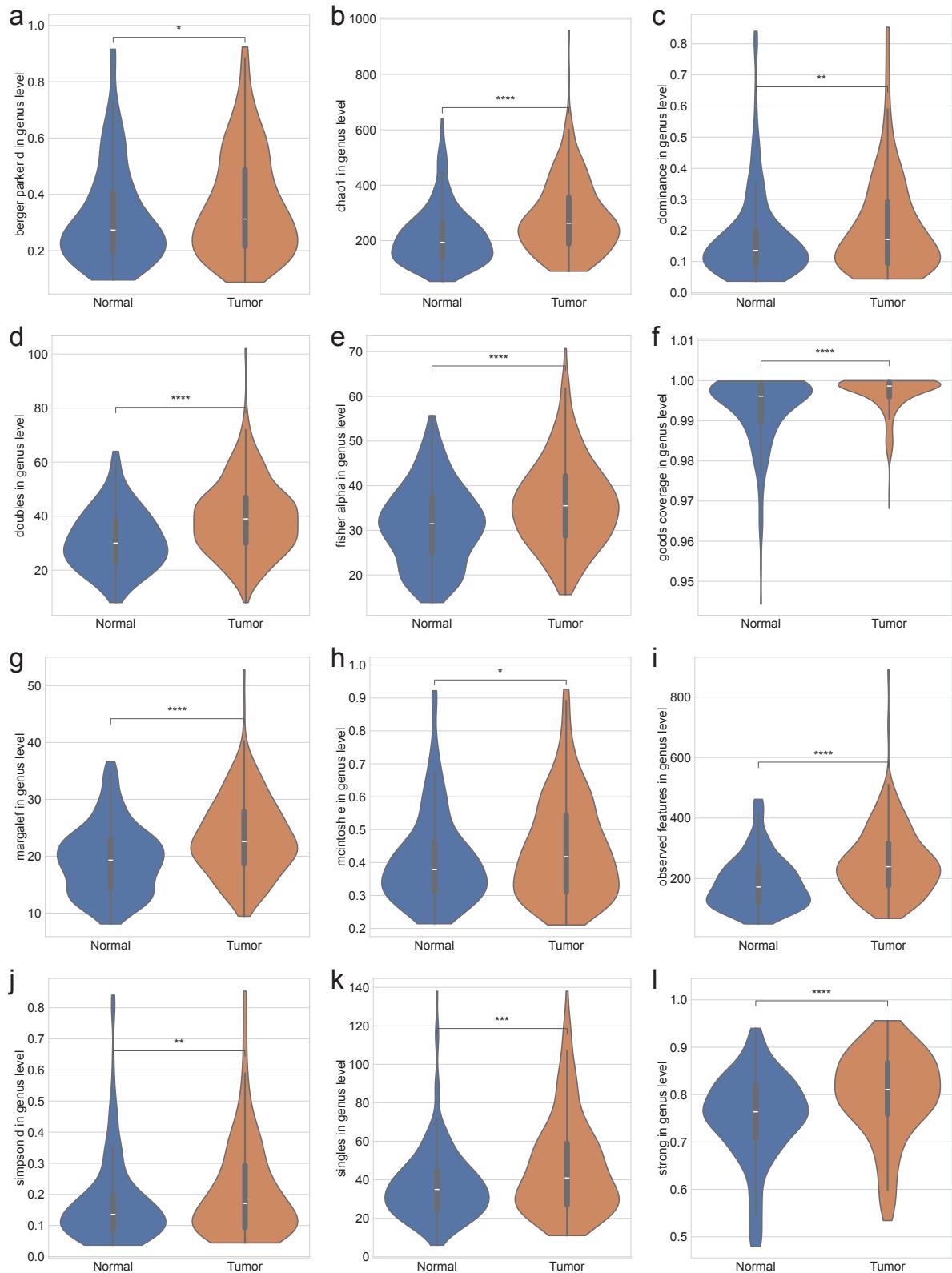


Figure 22: Alpha-diversity indices in genus level.

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

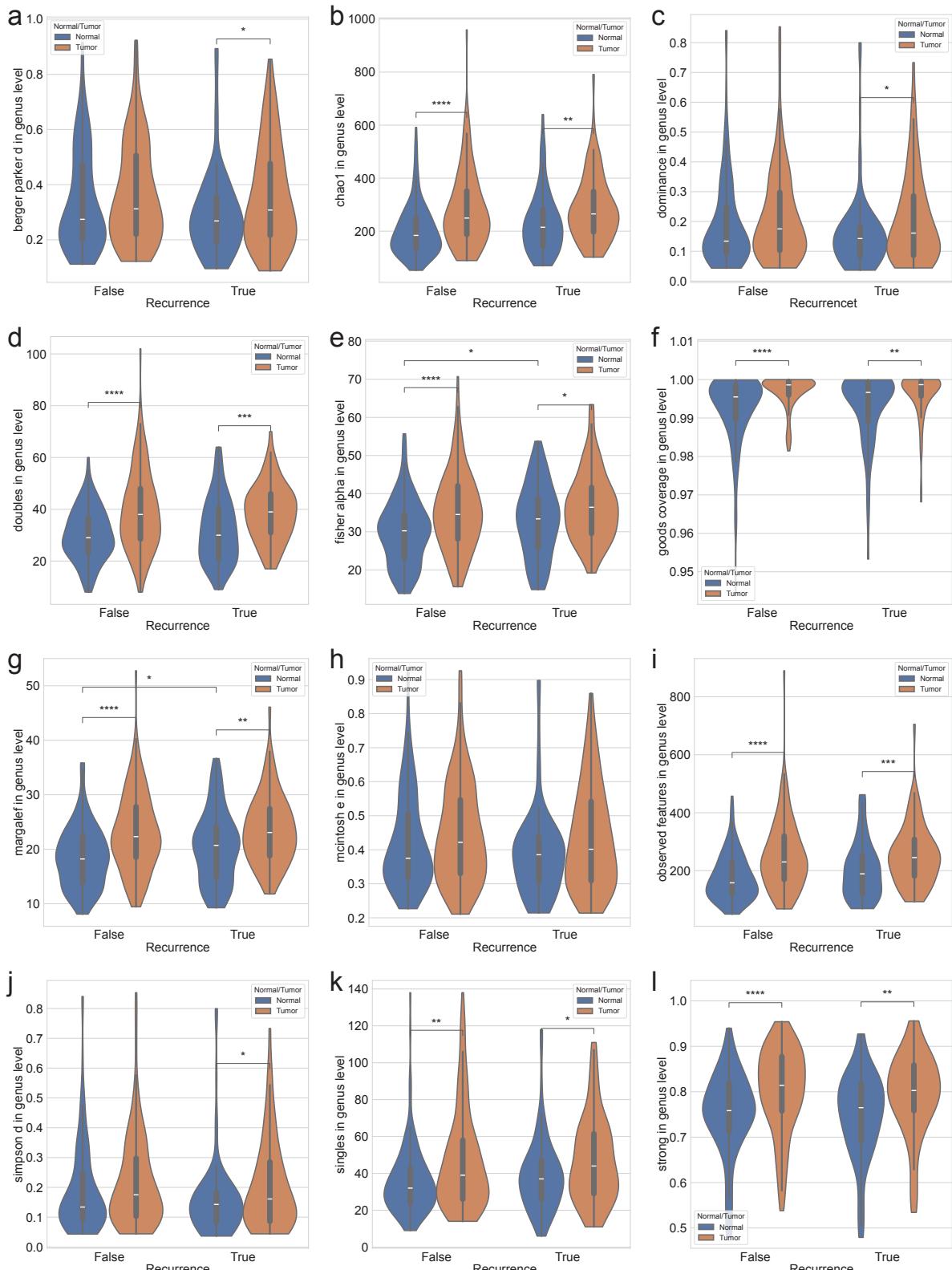


Figure 23: Alpha-diversity indices with recurrence in genus level.

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

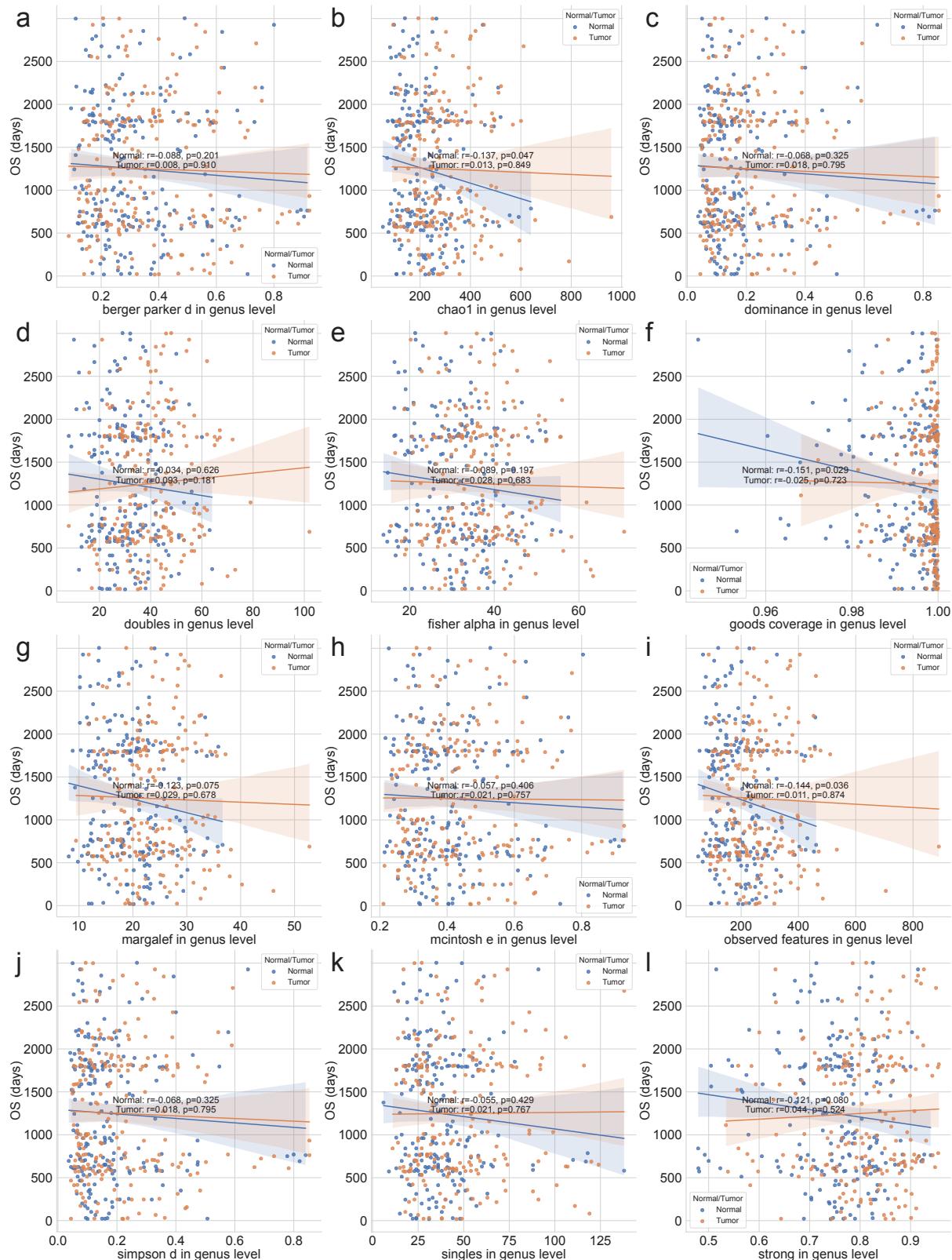


Figure 24: Alpha-diversity indices with OS in genus level.

(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.

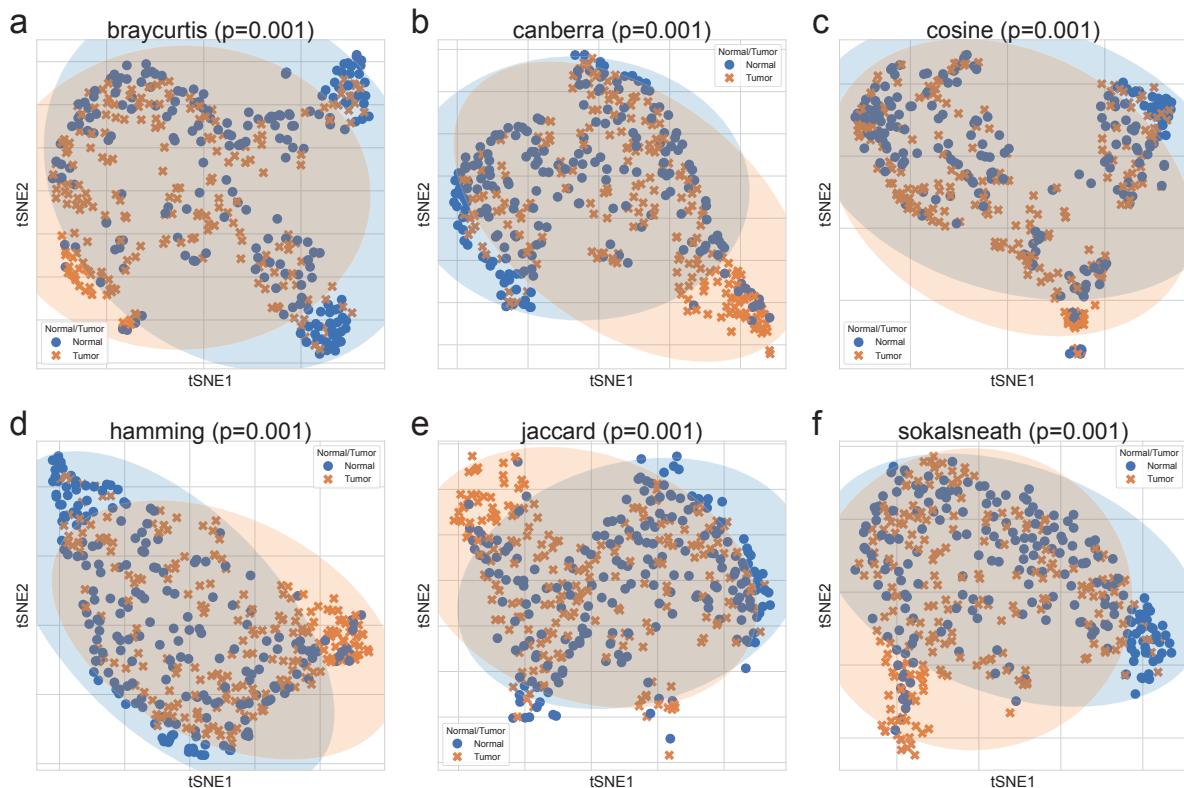


Figure 25: Beta-diversity indices in genus level.

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.

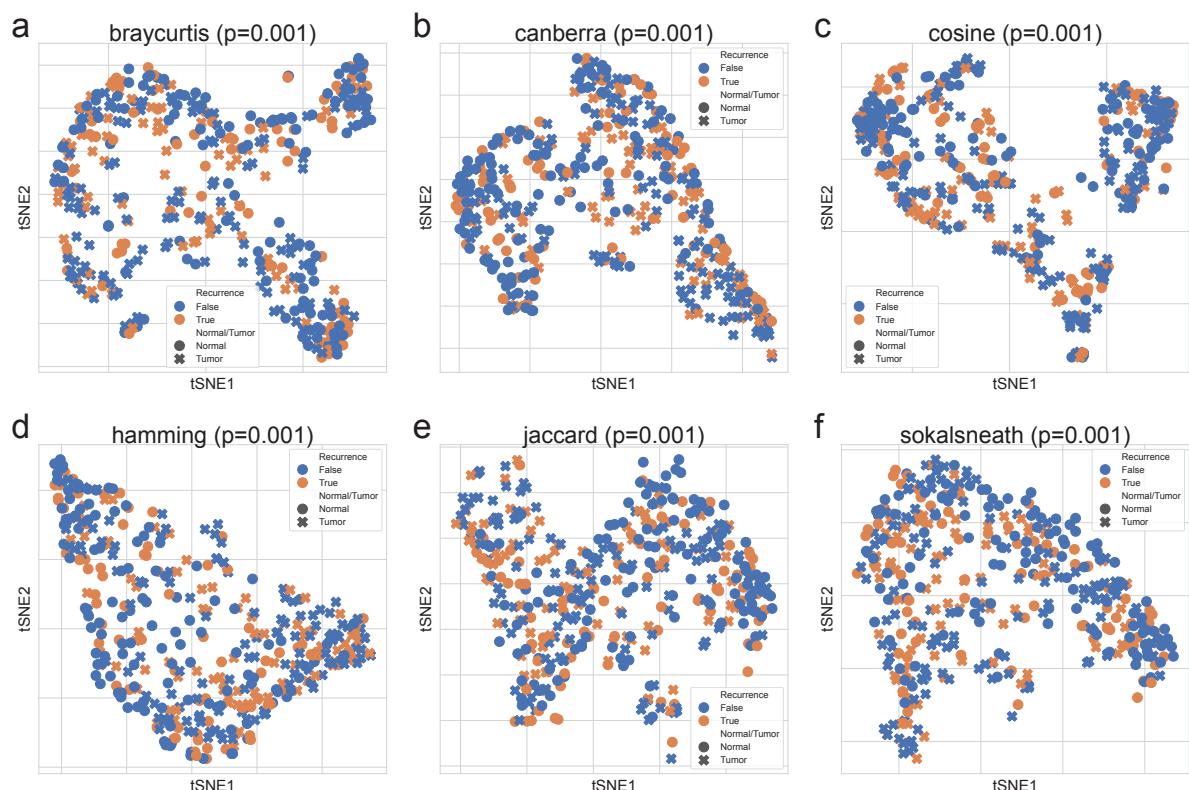


Figure 26: Beta-diversity indices with recurrence in genus level.

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.

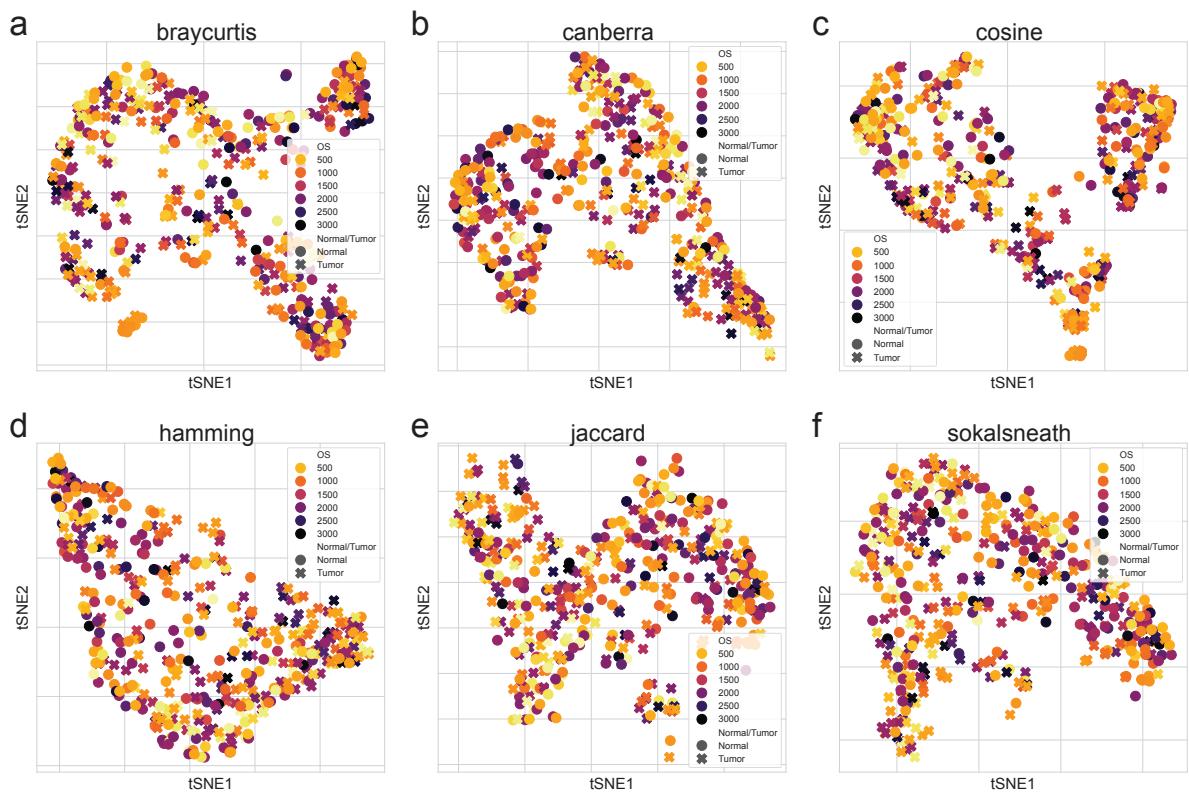


Figure 27: Beta-diversity indices with OS in genus level.

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

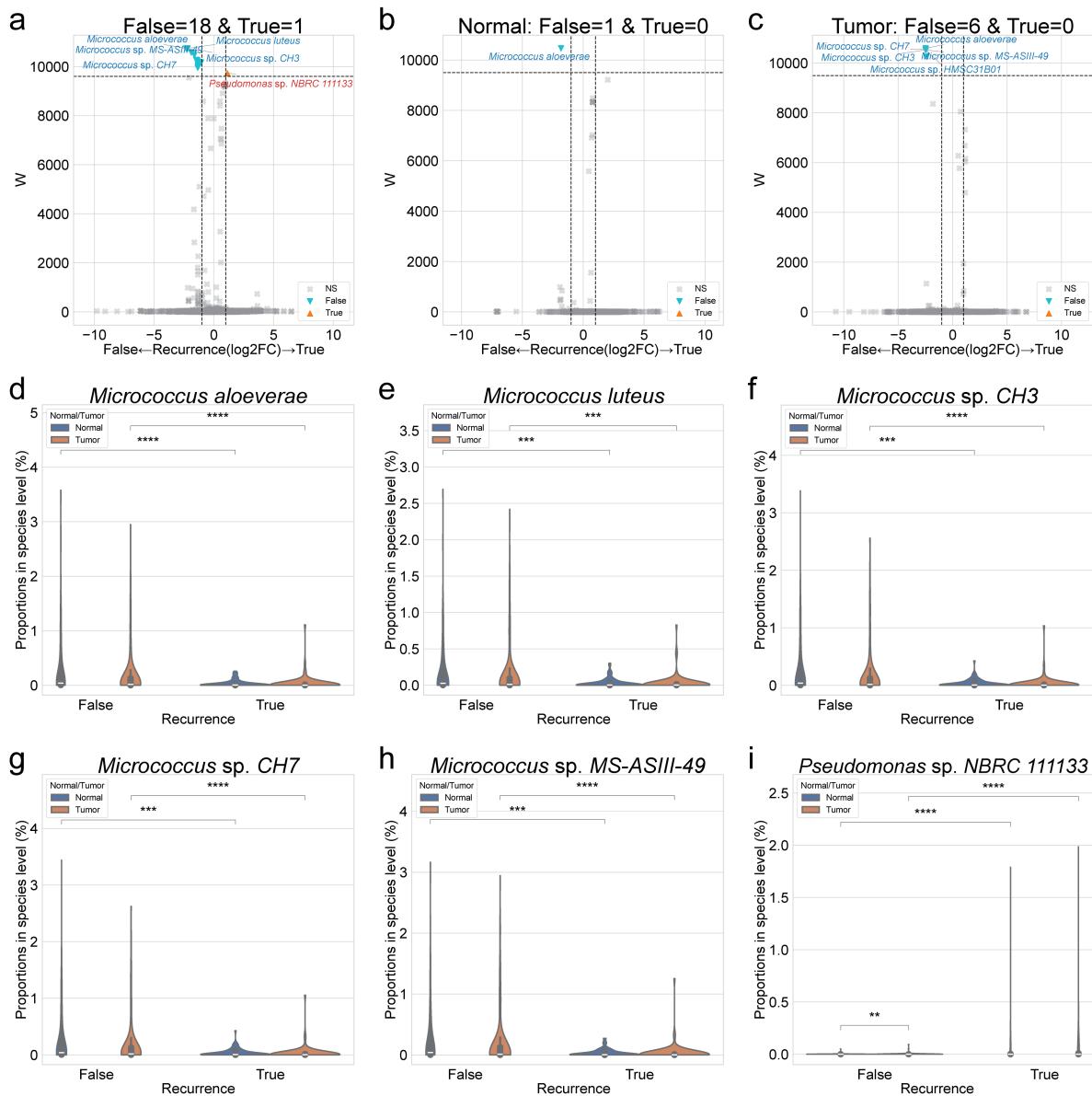


Figure 28: DAT with recurrence in species level.

(a-c) Volcano plots with recurrence. x-axis indicates \log_2 (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 luteus* **(f)** *Micrococcus* sp. CH3 **(g)** *Micrococcus* sp. CH7 **(h)** *Micrococcus* sp. MS-ASIII-49 **(i)** *Pseudomonas* sp. NBRC 111133. MWU test: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***)¹, and $p < 0.0001$ (****)

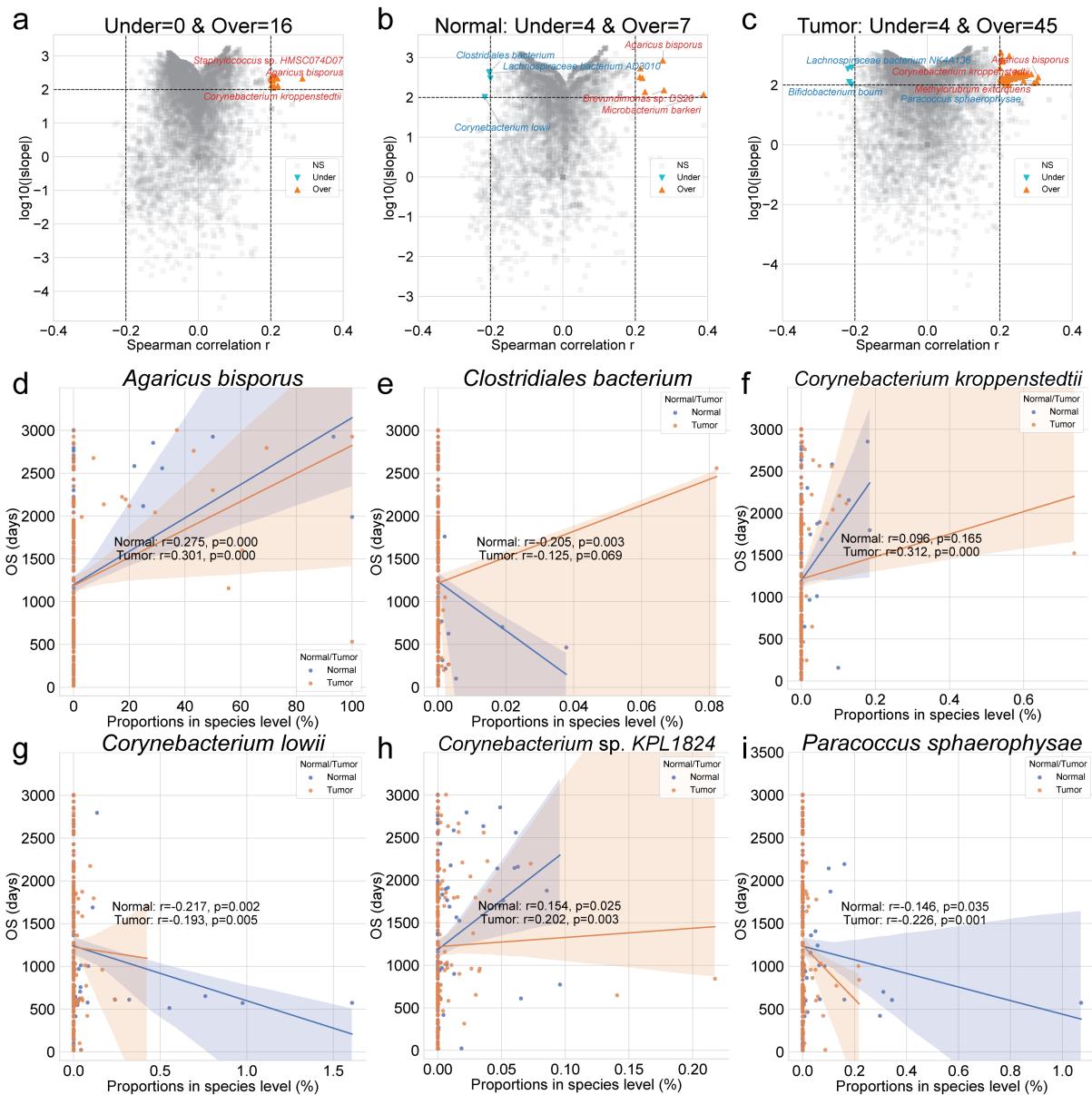


Figure 29: DAT with OS in species level.

(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-li)** Scatter plots of each taxon proportion with OS. **(d)** *Agaricus bisporus* **(e)** *Clostridiales bacterium* **(f)** *Corynebacterium kroppenstedtii* **(g)** *Corynebacterium lowii* **(h)** *Corynebacterium sp. KPL1824* **(i)** *Paracoccus sphaerophysae*. Statistical significance were calculated with Spearman correlation (r and p).

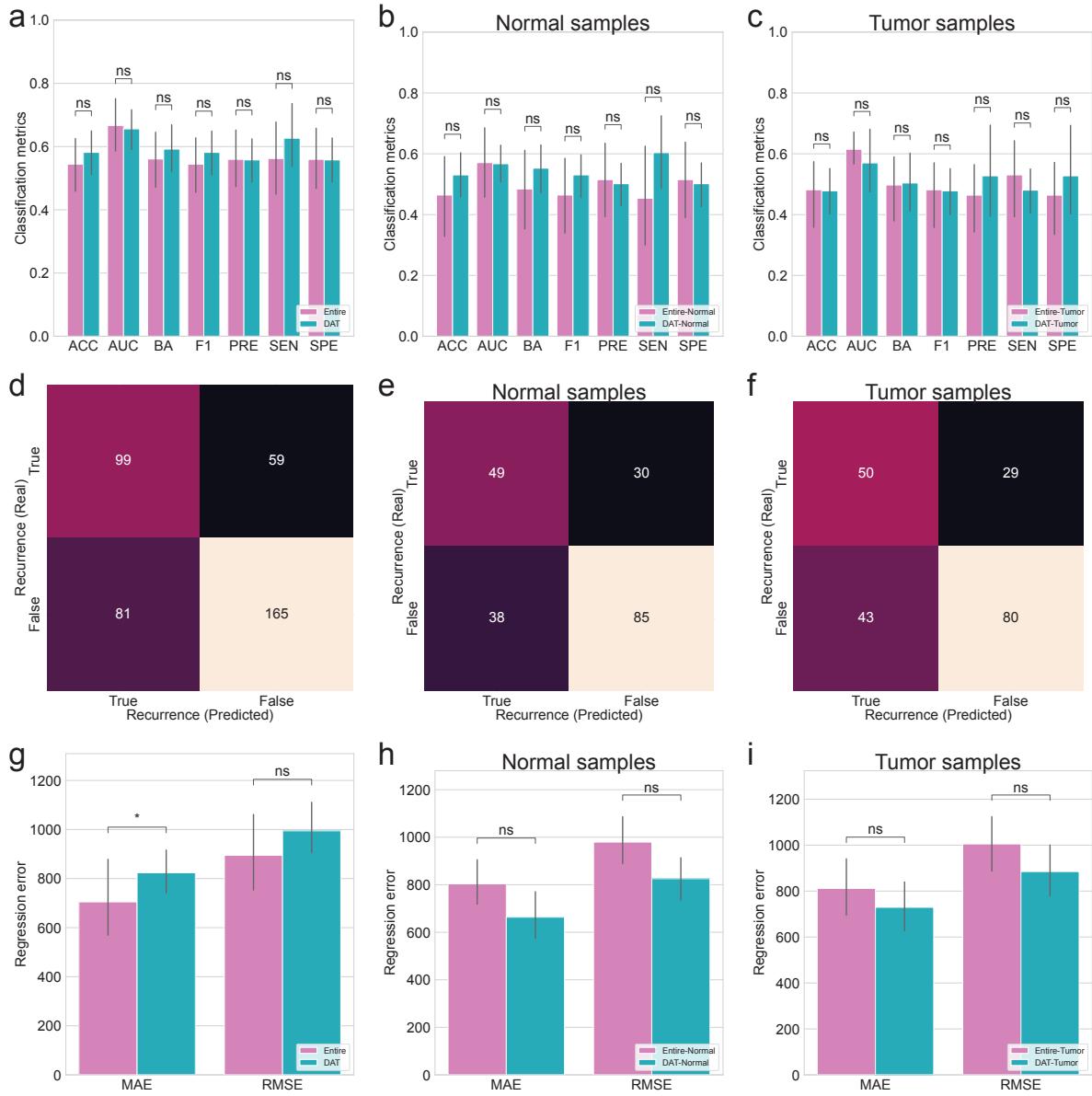


Figure 30: **Random forest classification and regression.**

(a-c) Random forest classification metrics for recurrence. **(a)** Total **(b)** Normal samples **(c)** Tumor samples. **(d-f)** Random forest classification confusion matrices for recurrence. **(d)** Total **(e)** Normal samples **(f)** Tumor samples. **(g-i)** Random forest regression errors for OS. **(g)** Total **(h)** Normal samples **(i)** Tumor samples. MWU test: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***) $p < 0.0001$ (****)

1363 **4.4 Discussion**

1364 This study provides a comprehensive metagenomic signature analysis of Korean CRC patients by
1365 examining prevalent microbial taxa, diversity indices, DAT selection and random forest-based predictions
1366 for recurrence and survival outcomes. Our analysis revealed distinct prevalent microbial communities in
1367 CRC patients (Figure 21), with significant difference between tumor tissues and adjacent matched normal
1368 tissues. Alpha-diversity indices analyses showed an overall shift in microbial diversity within tumor
1369 samples (Figure 22, Figure 23, and Figure 24), while beta-diversity analyses indicated significant changes
1370 in microbial composition associated with recurrence history and survival duration (Figure 25, Figure
1371 26, and Figure 27). Through DAT selection by ANCOM and Spearman correlation, we identified key
1372 microbial taxa link to recurrence history (Table 9 and Figure 28) and OS duration (Table 10 and Figure
1373 29), highlighting potential microbial biomarkers for CRC prognosis. To evaluate the predictive capacity of
1374 these microbial features, we implemented random forest-based machine learning models, where random
1375 forest classification demonstrated moderate accuracy (about 60%) for CRC recurrence prediction (Table
1376 11 and Figure 30) and random forest regression showed slightly high error (about 700 days) for OS
1377 prediction (Table 12 and Figure 30), suggesting that gut microbiome alterations alone are insufficient
1378 for precise prognosis and may interact with host genetic factors such as germline and somatic mutations.
1379 These findings underscore the potential of microbial biomarkers in CRC risk stratification, emphasizing
1380 the need for multi-omics integration to improve predictive models and personalized medicine strategies
1381 in CRC.

1382 In the Bacteria kingdom (Figure 21a), *Bacteroides* genus is the most frequent genus in tumor tis-
1383 sues, then came *Fusobacterium* and *Cutibacterium* genera. These results also accord with previous
1384 studies, which showed that *Bacteroides fragilis* (Scott, Whittle, Jeraldo, & Chia, 2022; Purcell, Permain,
1385 & Keenan, 2022), *Fusobacterium nucleatum* (Wang & Fang, 2023; Zepeda-Rivera et al., 2024), and
1386 *Cutibacterium acnes* (Benej et al., 2024) have significant roles in tumorigenesis and development of CRC.
1387 Further, not only those bacterium genera individually, the association between *Bacteroides* genus and
1388 *Fusobacterium* genus is reported (Viljoen, Dakshinamurthy, Goldberg, & Blackburn, 2015; Joo et al.,
1389 2024; Duy et al., 2024; Conde-Pérez et al., 2024), suggesting possible contribution to CRC pathogenesis
1390 through mechanisms such as biofilm formation, immune evasion, and/or metabolic interactions with
1391 other dysbiotic taxa. Given that *Fusobacterium* genus has been shown to co-aggregate with *Bacteroides*
1392 genus, it is plausible that *Cutibacterium* might interact with these genera to influence inflammation,
1393 epithelial barrier integrity, and tumor progression. Thus, further studies integrating functional metage-
1394 nomics, metabolomics, and host-microbiome interactions are warranted to elucidate the precise role of
1395 *Cutibacterium* genus and its relationship with CRC-associated microbial networks.

1396 Analysis of eukaryotic and fungal microbial compositions revealed that the *Toxoplasma* genus was
1397 prevalent in both normal and tumor samples (Figure 21b), while the *Malassezia* genus was more prevalent
1398 in tumor samples (Figure 21c). The consistent presence of *Toxoplasma* genus across both sample types
1399 suggests that this intracellular pathogen may be a stable component of the gut microbiome, although its
1400 role in CRC pathogenesis remains unclear (Yu et al., 2020; Zavareh et al., 2021). In contrast, the increase

1401 prevalence of *Malassezia* genus in tumor tissue aligns with emerging evidence that certain fungal genus
1402 may contribute to CRC-promoting inflammation and metabolic alterations (R. Gao et al., 2017; Yuan et
1403 al., 2025), suggesting a potential role in CRC development and progression. These findings highlight the
1404 need for further investigation into the functional impact of eukaryotic and fungal microbiota in CRC for
1405 shaping the tumor microenvironment.

1406 In normal tissue samples, *Roseolovirus* genus was the most prevalent viral taxon (Figure 21d),
1407 indicating its stable presence in the gut virome of healthy colonic tissues. However, in tumor tissue
1408 samples, *Lymphocryptovirus* and *Cytomegalovirus* genera were more prevalent viral taxa, suggesting
1409 an alteration in viral community structure associated with CRC progression. This viral compositional
1410 shift aligns with the Anna Karenina principle (Ma, 2020; W. Li & Yang, 2025), implying that microbial
1411 communities in diseased states exhibit greater instability and variability compare to their adjacent normal
1412 tissues. The emergence of *Lymphocryptovirus* (Mjelle, Castro, & Aass, 2025; De Flora & Bonanni,
1413 2011) and *Cytomegalovirus* (Harkins et al., 2002; Taher et al., 2014; Bender et al., 2009) genera in
1414 tumor samples raises the possibility that oncogenic viruses may contribute to CRC carcinogenesis by
1415 promoting chronic inflammation, immune modulation, and/or direct viral-host interactions affecting
1416 cellular transformation. Moreover, the detection of tumor-associated viral alterations in adjacent normal
1417 tissues supports the concept of field cancerization (Curtius et al., 2018; Rubio et al., 2022), where viral
1418 dysbiosis may extend beyond the tumor itself, creating a pro-tumorigenic microenvironment even before
1419 malignant transformation occurs. These findings underscore the potential impact of viral communities in
1420 CRC and highlight the requirement for further research into their functional roles in carcinogenesis of
1421 CRC.

1422 Alpha-diversity indices revealed a significant increase in microbial diversity in tumor samples com-
1423 pared to its adjacent normal tissues (MWU test $p < 0.05$; Figure 22), suggesting CRC is associated with
1424 a more heterogeneous gut microbiome (Liu et al., 2021). The increase in alpha-diversity indices within
1425 tumor tissues may support the Anna Karenina principle and/or the concept of field cancerization, where
1426 microbial alterations extend beyond the tumor site and contribute to a pre-malignant microenvironment.
1427 The enrichment of distinct bacterial, eukaryotic, fungal, and viral taxa within tumor samples suggest that
1428 microbial dysbiosis in CRC is not limited to a single pathogenic genus or species but rather involves
1429 complex community-level changes.

1430 Furthermore, alpha-diversity indices in relation to recurrence history revealed distinct microbial
1431 diversity patterns between normal and tumor tissue samples (Figure 23). In recurrence patients, tumor
1432 samples exhibited a greater increase in alpha-diversity indices compared to their adjacent normal tissues
1433 (11/12 indices, 92% indices; Figure 23), suggesting that a more heterogeneous microbial community may
1434 be linked to tumor aggressiveness and recurrence potential (Huo et al., 2022; Vigneswaran & Shogan,
1435 2020). This trend aligns with a highly diverse but dysregulated microbiome in tumor samples may
1436 contribute to immune evasion, chronic inflammation, and tumor-promoting metabolic changes. In non-
1437 recurrence patients, although tumor samples still exhibited increased alpha-diversity indices compared to
1438 normal tissues, the difference was less pronounced (8/12 indices, 67% indices; Figure 23), suggesting that
1439 a relatively more stable microbiome in tumor tissues may be associated with favorable survival outcomes

1440 (Avuthu & Guda, 2022). These findings reinforce the concept that tumor microbiome diversity changes are
1441 inconsistent across CRC patients, supporting the Anna Karenina principle. Additionally, the differences in
1442 alpha-diversity indices of normal tissues between recurrence and non-recurrence patients further suggest
1443 (Figure 23e and Figure 23g) that specific microbial communities may influence post-treatment disease
1444 progression.

1445 Moreover, alpha-diversity indices and OS duration in CRC patients revealed distinct patterns between
1446 normal and tumor tissues (Figure 24), suggesting that microbial diversity in non-cancerous lesions may
1447 play a role in cancer prognosis (Galeano Niño et al., 2022). While no significant correlation was found
1448 between tumor-associated microbiome and OS duration, three of the 12 alpha-diversity indices exhibited
1449 negative correlations with OS in normal tissues (Figure 29b, Figure 29f, and Figure 29i), indicating
1450 that lower microbial heterogeneity in normal lesions was associated with longer survival. This finding
1451 suggests that a more heterogeneous microbial community in normal colon tissues may contribute to a
1452 microenvironment that fosters tumor progression, aligning with the field cancerization. Therefore, the
1453 negative correlations observed only in normal tissues suggests that pre-onset dysbiosis in non-cancerous
1454 regions could influence prognosis of CRC, potentially serving as an early indicator of cancer progression
1455 risk.

1456 Beta-diversity indices revealed significant differences in gut microbiome compositions between tumor
1457 and normal tissues (Figure 25), aligning with the alpha-diversity indices and further confirming the
1458 presence of dysbiosis in gut microbiome of CRC. The distinct clustering of tumor and normal samples in
1459 beta-diversity indices (PERMANOVA $p < 0.001$) suggests that CRC is associated with a major alteration
1460 in microbial structure. This transformation may be driven by the expansion of tumor-associated taxa and
1461 the shrinkage of protective taxa, resulting in a tumor-supportive microenvironment. This clear separation
1462 in beta-diversity indices between tumor and normal tissues supports again the field cancerization, where
1463 microbial alterations extend beyond tumor lesions and affect surrounding non-cancerous lesions.

1464 Furthermore, beta-diversity indices demonstrated significant microbial composition shifts between
1465 normal and tumor tissues in accordance with recurrence status (Figure 26), suggesting that dysbiosis in the
1466 gut microbiome may play a role in CRC progression and post-treatment recurrence. By the beta-diversity
1467 indices, the observed recurrence-associated microbial shifts highlight the potential of beta-diversity index
1468 as predictive markers for recurrence risk of CRC, warranting further studies to explore their functional
1469 significance and potential integration into microbiome-based prognostic models.

1470 Moreover, beta-diversity indices suggested a potential association between the gut microbiome com-
1471 position and OS in CRC patients (Figure 27), as distinct clustering were observed in relation to survival
1472 duration. However, due to the continuous nature of survival duration, direct statistical comparison using
1473 PERMANOVA test could be not performed, limiting the ability to formally quantify these differences.
1474 Despite this limitation, the observed separation of microbial communities along OS suggests that the
1475 gut microbiome composition may play a major role in CRC prognosis, potentially influencing immune
1476 response, tumor progression, and treatment outcomes. This lack of statistical validation highlights the
1477 need for alternative approaches to better assess the relationship between microbiome structure and sur-
1478 vival outcome. Further investigation is required to determine whether specific microbial taxa drive these

1479 compositional shifts and whether gut microbiome profiles could serve as prognostic biomarkers for CRC
1480 survival outcomes.

1481 To identify DAT associated with recurrence risk in CRC, we applied ANCOM to compare the gut
1482 microbiome compositions between recurrence and non-recurrence patients (Table 9 and Figure 28). By
1483 applying ANCOM separately to total samples (Figure 28a), normal samples (Figure 28b), and tumor
1484 samples (Figure 28c), we identified both global and tissue-specific microbial shifts linked to CRC
1485 recurrence. Among these 19 recurrence-related DAT (Table 9), several DAT belonging to the *Micrococcus*
1486 and *Staphylococcus* genera were significantly more abundant in non-recurrence patients. *Micrococcus*
1487 genus has been reported with anti-bacterial, anti-fungal, and anti-inflammatory activities (Tizabi &
1488 Hill, 2023), and another study has found that the production of carotenoid pigments from *Micrococcus*
1489 *luteus* (Figure 28e) exhibited promising antibiotics agents (Hegazy, Abu-Hussien, Elsenosy, El-Sayed,
1490 & Abo El-Naga, 2024). In this CRC study participants, *Cutibacterium acnes* was selected one of the
1491 non-recurrence-enriched DAT (Table 9). This finding is consistent with previous studies which have
1492 suggested that *Cutibacterium acnes* inhibits the activities of pathogens, such as *Staphylococcus aureus*,
1493 and suppresses tumor growth (Benej et al., 2024; Ding, Lian, Tam, & Oh, 2024). On the other hands,
1494 in this CRC study participants, many *Staphylococcus* species have chosen as non-recurrence-enriched
1495 DAT (Table 9); however, this outcome is contrary to previous studies which have described that cancer-
1496 promoting activity of *Staphylococcus aureus* (Z. Li, Zhuang, Wang, Wang, & Dong, 2021; Cuervo et
1497 al., 2010), suggesting opposite behaviors between *Staphylococcus aureus* and other *Staphylococcus*
1498 spp. Last but not least, *Pseudomonas* sp. NBRC 111133 has been found as the only recurrence-enriched
1499 DAT (Figure 28i). This also accords with earlier studies, which showed that *Pseudomonas aeruginosa*
1500 infections in cancer patients (Ohmagari et al., 2005; Paprocka et al., 2022).

1501 To determine the DAT correlated to survival duration in CRC, we applied Spearman correlation to
1502 measure effects of the gut microbiome composition with OS (Table 10 and Figure 29). By implementing
1503 Spearman correlation to total samples (Figure 29a), normal samples (Figure 29b), and tumor samples
1504 (Figure 29c), we found that CRC survival is associated with both tissue type-specific and global microbial
1505 alterations. Among these 57 OS-related DAT (Table 29), several DAT from the *Corynebacterium* and
1506 *Staphylococcus* genera have significant correlations with survival duration of CRC. *Agaricus bisporus*
1507 has positive correlation with OS both in normal and tumor samples (Figure 29d). In accordance with
1508 this finding, previous studies have demonstrated that a polysaccharide produced from *Agaricus bisporus*
1509 exhibited anti-cancerous activity in colon cancer (Dong, Wang, Tang, Liu, & Gao, 2024; El-Deeb et
1510 al., 2022; N. Zhang, Liu, Tang, Yang, & Wang, 2023). Furthermore, most of *Corynebacterium* genus,
1511 including *Corynebacterium kroppenstedtii* (Figure 29f) and *Corynebacterium* sp. KPL1824 (Figure
1512 29h), have positive correlations with OS; however, *Corynebacterium lowii* (Figure 29g) has negative
1513 correlation with OS both in normal and tumor samples. Comparison of the findings with those of other
1514 studies confirms a breast cancer risk factor of *Corynebacterium afermentans* (J. An, Kwon, Oh, & Kim,
1515 2025), an increasing of *Corynebacterium appendicis* in CRC (Hasan et al., 2022), an inhibition role of
1516 *Corynebacterium matruchotii* of cancer growth in oral squamous cell carcinoma (Shen et al., 2022), and
1517 a promoting cancer cell apoptosis of *Corynebacterium durum* (S. Kim et al., 2024), warranting future

1518 investigations to selecting pro-tumorigenic and anti-tumorigenic species of *Corynebacterium* genus. *Clostridiales*
1519 *bacterium* has negative correlation with OS in normal samples (Figure 29e). However, this result does not
1520 support previous researches which have demonstrated that anti-cancer activities with immune modulation
1521 of *Clostridiales* genus (Montalban-Arques et al., 2021; Minton, 2003), suggesting different roles from
1522 *Clostridiales* species for immune response against cancer. Many species from *Staphylococcus* genus
1523 have positive correlations with OS (Table 10). Although, these results differ from some published
1524 studies which indicated cancer prevention and treatment via reduction of *Staphylococcus epidermidis*
1525 (Bernardo et al., 2023; Kepp, Zitzvogel, & Kroemer, 2023), they are consistent with other published
1526 researches which suggested that other species of *Staphylococcus* genus exhibited anti-cancer activities
1527 (Hassan, Mustafa, Rahim, & Isa, 2016; M. Zhang et al., 2022). Moreover, *Lachnospiraceae bacterium*
1528 *AD3010* and *Porphyromonas macacae* have negative correlations with OS in normal samples, while
1529 *Lachnospiraceae bacterium NK4A136* and *Paracoccus sphaerophysae* have negative correlations with OS
1530 in tumor samples (Table 10). Previous studies have addressed that high abundance of *Lachnospiraceae*
1531 genus in the gut microbiome showed anti-tumor roles in the CRC (Hexun et al., 2023; X. Zhang et al.,
1532 2023), indicating that more comprehensive investigation of species from *Lachnospiraceae* genus might be
1533 required. *Porphyromonas gingivalis*, a well-known periodontitis pathogens from *Porphyromonas* genus,
1534 was also reported promoting cancer resistance and development on CRC, lung cancer, and oesophageal
1535 cancer (León et al., 2007; Katz et al., 2009; S. Gao et al., 2021), providing a warrant to elucidate cancer-
1536 related roles of not only *Porphyromonas gingivalis* but also other *Porphyromonas* genus. *Paracoccus*
1537 *sphaerophysae* displayed negative correlation with OS in tumor samples and insignificantly negative
1538 correlation with OS (Spearman $|r| \leq 0.2$) in normal samples (Figure 29i), it is consistent with the literature
1539 which have shown *Paracoccus* genus is more prevalent in nasopharyngeal carcinoma group than healthy
1540 individuals (Lu et al., 2024).

1541 To assess the predictive potential of recurrence-related DAT in CRC recurrence risk, we implemented a
1542 random forest classification model (Table 11 and Figure 30). The classification model achieved moderated
1543 classification performance (about 60%), indicating that while gut microbial provides some predictive
1544 value, it is likely insufficient as a standalone biomarker for recurrence risk of CRC. This limited predictive
1545 accuracy may be attributed to the complex and dynamic nature of gut microbiome network, where
1546 epigenetic modifications and immune modulation collectively influence development and progression of
1547 CRC. Additionally, host-microbiome interactions, including metabolic pathways, may further contribute
1548 to recurrence of cancer, warranting a more integrative multi-omics approach. Therefore, future studies
1549 incorporating genomic sequencing data, e.g. somatic mutations and host immune signatures, could provide
1550 a more comprehensive understanding of how microbial dysbiosis interacts with tumor biology.

1551 To evaluate the predictive potential of OS-related DAT in survival duration of CRC, we employed
1552 a random forest regression model (Table 12 and Figure 30). The regression model exhibited moderate
1553 regression error (about 700 days), suggesting that while gut microbiome composition provides some
1554 predictive values for cancer patient survival, it is likely influenced by additional host-specific and
1555 environmental factors. The complex interplay between the gut microbiome and CRC progression involves
1556 sophisticated microbial networks, metabolic interactions, and immune response, making it difficult

1557 to capture survival outcomes solely based on microbiome features. Furthermore, host-microbiome
1558 interactions, including MSI, tumor mutational burden, and epigenetic modifications, likely play a crucial
1559 role in determine favorable or unfavorable survival. These findings highlight the need for multi-omics
1560 integration, combining genomic sequencing data and metagenomic functional analysis, to gain deeper
1561 insights into how microbial dysbiosis interacts with tumor biology and clinical outcomes. Future studies
1562 incorporating machine learning models with multi-layered biological data may improve the accuracy of
1563 survival prediction and contribute to personalized medicine appraoches for CRC therapeutics.

1564 **5 Conclusion**

1565 This dissertation underscores the critical character of microbiome research in understanding disease
1566 mechanisms, predicting health outcomes, and advancing personalized medicine. By investigating PTB,
1567 periodontitis, and CRC, this dissertation demonstrated how microbial diversity alters, DAT, and ma-
1568 chine learning-based modeling contribute to disease classification and prognosis. While each condition
1569 exhibited unique microbiome alterations, the findings collectively support the Anna Karenina princi-
1570 ple, which suggests that microbial communities in patients with disease become more variable and
1571 dysregulated compared to their relatively stable and uniform counterparts in healthy individuals. The
1572 Anna Karenina principle was evident in all three diseases examine, where dysbiosis not only disrupted
1573 microbial homeostasis but also contributed to disease progression and development. The ability to identify
1574 disease-specific microbial signature reinforces the importance of microbiome profiling as a non-invasive
1575 therapeutic guidance.

1576 In the PTB study (Section 2), salivary microbiome profiling revealed distinct microbial shifts between
1577 PTB and FTB, with a random forest-based model achieving high accuracy in assessment PTB risk.
1578 Similarly, the periodontitis study (Section 3) identified salivary microbial markers that classified between
1579 healthy individuals and multiple stages of periodontitis, suggesting the potential for salivary microbiome-
1580 based diagnostics in management and treatment of periodontitis. The CRC study (Section 4) revealed
1581 significant alpha-diversity and beta-diversity indices differences between tumor and adjacent normal
1582 tissues, with distinct microbial compositions associated with recurrence status and survival duration.
1583 However, while random forest models for predicting recurrence risk and survival duration provided
1584 moderate accuracy, the findings suggest that gut microbiome composition alone may not be sufficient for
1585 precise clinical instruction.

1586 The Anna Karenina principle was particularly evident in the CRC study (Section 4), where gut
1587 microbial alterations were highly individualized among tumor sample, with recurrence status and survival
1588 duration related to divergent microbial community structures. This aligns with the concept of field
1589 cancerization, where dysbiosis extends beyond the tumor lesion, affecting adjacent non-cancerous lesions
1590 and potentially contributing to tumorigenesis and cancer development. These findings reinforce the
1591 complexity of host-microbiome interactions, where microbial imbalances may not only reflect disease
1592 status but actively participate in disease etiology through inflammation, metabolic alterations, and immune
1593 modulation. The variability in microbial community alterations across patients highlights the need for
1594 multi-omics integration, combining host genomic data to enhance personalized treatment and management
1595 strategies.

1596 Despite the promising insights gained from microbiome analyses, this dissertation acknowledges
1597 several limitations. The moderated predictive performance of machine learning models suggests that
1598 microbial features along may not fully capture disease mechanisms. Future research should integrate
1599 multi-omics datasets, including host genomic mutations, metabolic profiles, and immune signatures,
1600 to improve biomarker discovery and disease prediction models. Additionally, population-specific mi-
1601 crobiome differences must be considered, as external validation in the periodontitis study (Section 3)

1602 showed variations in salivary microbiome composition between different ethnic groups. Large-scale
1603 and multi-center studies are essential to validate microbiome-based biomarkers and ensure their clinical
1604 applicability across diverse populations.

1605 Overall, this dissertation contributes to the growing field of microbiome-driven personalized medicine,
1606 demonstrating the potential of microbiome profiling, diversity analysis, identification DAT, and machine
1607 learning-based modeling in assessing disease risk and progression. By furthering our understanding of
1608 host-microbiome interactions, these findings pave a novel microbiome-targeted therapeutic strategies, ad-
1609 vancing personalized disease prevention and treatment. Moving forward, integrating microbiome research
1610 with genomics, metabolomics, and immunology holds the potential to transform disease management and
1611 personalized medicine, ultimately improving treatment outcomes across a broad spectrum of diseases.

¹⁶¹² References

- ¹⁶¹³ Aagaard, K., Ma, J., Antony, K. M., Ganu, R., Petrosino, J., & Versalovic, J. (2014). The placenta harbors
¹⁶¹⁴ a unique microbiome. *Science translational medicine*, 6(237), 237ra65–237ra65.
- ¹⁶¹⁵ Abu-Ghazaleh, N., Chua, W. J., & Gopalan, V. (2021). Intestinal microbiota and its association with
¹⁶¹⁶ colon cancer and red/processed meat consumption. *Journal of gastroenterology and hepatology*,
¹⁶¹⁷ 36(1), 75–88.
- ¹⁶¹⁸ Abusleme, L., Hoare, A., Hong, B.-Y., & Diaz, P. I. (2021). Microbial signatures of health, gingivitis,
¹⁶¹⁹ and periodontitis. *Periodontology 2000*, 86(1), 57–78.
- ¹⁶²⁰ Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., & Pawlowsky-Glahn, V. (2000). Logratio
¹⁶²¹ analysis and compositional distance. *Mathematical geology*, 32, 271–275.
- ¹⁶²² Aja, E., Mangar, M., Fletcher, H., & Mishra, A. (2021). Filifactor alocis: recent insights and advances.
¹⁶²³ *Journal of dental research*, 100(8), 790–797.
- ¹⁶²⁴ Alelyani, S. (2021). Stable bagging feature selection on medical data. *Journal of Big Data*, 8(1), 11.
- ¹⁶²⁵ Altabtbaei, K., Maney, P., Ganesan, S. M., Dabdoub, S. M., Nagaraja, H. N., & Kumar, P. S. (2021). Anna
¹⁶²⁶ karenina and the subgingival microbiome associated with periodontitis. *Microbiome*, 9, 1–15.
- ¹⁶²⁷ Altingöz, S. M., Kurgan, Ş., Önder, C., Serdar, M. A., Ünlütürk, U., Uyanık, M., ... Günhan, M.
¹⁶²⁸ (2021). Salivary and serum oxidative stress biomarkers and advanced glycation end products in
¹⁶²⁹ periodontitis patients with or without diabetes: A cross-sectional study. *Journal of periodontology*,
¹⁶³⁰ 92(9), 1274–1285.
- ¹⁶³¹ Alverdy, J., Hyoju, S., Weigerinck, M., & Gilbert, J. (2017). The gut microbiome and the mechanism of
¹⁶³² surgical infection. *Journal of British Surgery*, 104(2), e14–e23.
- ¹⁶³³ An, J., Kwon, H., Oh, S.-Y., & Kim, Y. J. (2025). Association between breast cancer risk factors and
¹⁶³⁴ blood microbiome in patients with breast cancer. *Scientific Reports*, 15(1), 6115.
- ¹⁶³⁵ An, S., & Park, S. (2022). Association of physical activity and sedentary behavior with the risk of
¹⁶³⁶ colorectal cancer. *Journal of Korean Medical Science*, 37(19).
- ¹⁶³⁷ Anderson, M. J. (2014). Permutational multivariate analysis of variance (permanova). *Wiley statsref:
1638 statistics reference online*, 1–15.
- ¹⁶³⁹ Aruni, A. W., Mishra, A., Dou, Y., Chioma, O., Hamilton, B. N., & Fletcher, H. M. (2015). Filifactor
¹⁶⁴⁰ alocis—a new emerging periodontal pathogen. *Microbes and infection*, 17(7), 517–530.
- ¹⁶⁴¹ Avuthu, N., & Guda, C. (2022). Meta-analysis of altered gut microbiota reveals microbial and metabolic
¹⁶⁴² biomarkers for colorectal cancer. *Microbiology Spectrum*, 10(4), e00013–22.

- 1643 Aziz, Q., & Thompson, D. G. (1998). Brain-gut axis in health and disease. *Gastroenterology*, 114(3),
1644 559–578.
- 1645 Bai, X., Wei, H., Liu, W., Coker, O. O., Gou, H., Liu, C., . . . others (2022). Cigarette smoke promotes
1646 colorectal cancer through modulation of gut microbiota and related metabolites. *Gut*, 71(12),
1647 2439–2450.
- 1648 Baldelli, V., Scaldaferrri, F., Putignani, L., & Del Chierico, F. (2021). The role of enterobacteriaceae in
1649 gut microbiota dysbiosis in inflammatory bowel diseases. *Microorganisms*, 9(4), 697.
- 1650 Bardou, M., Rouland, A., Martel, M., Loffroy, R., Barkun, A. N., & Chapelle, N. (2022). Obesity and
1651 colorectal cancer. *Alimentary Pharmacology & Therapeutics*, 56(3), 407–418.
- 1652 Barlow, G. M., Yu, A., & Mathur, R. (2015). Role of the gut microbiome in obesity and diabetes mellitus.
1653 *Nutrition in clinical practice*, 30(6), 787–797.
- 1654 Basavaprabhu, H., Sonu, K., & Prabha, R. (2020). Mechanistic insights into the action of probiotics
1655 against bacterial vaginosis and its mediated preterm birth: An overview. *Microbial pathogenesis*,
1656 141, 104029.
- 1657 Belstrøm, D., Constancias, F., Drautz-Moses, D. I., Schuster, S. C., Veleba, M., Mahé, F., & Givkov, M.
1658 (2021). Periodontitis associates with species-specific gene expression of the oral microbiota. *npj
1659 Biofilms and Microbiomes*, 7(1), 76.
- 1660 Bender, C., Zipeto, D., Bidoia, C., Costantini, S., Zamò, A., Menestrina, F., & Bertazzoni, U. (2009).
1661 Analysis of colorectal cancers for human cytomegalovirus presence. *Infectious agents and cancer*,
1662 4, 1–6.
- 1663 Benej, M., Hoyd, R., Kreamer, M., Wheeler, C. E., Grencewicz, D. J., Choueiry, F., . . . others (2024). The
1664 tumor microbiome reacts to hypoxia and can influence response to radiation treatment in colorectal
1665 cancer. *Cancer research communications*, 4(7), 1690–1701.
- 1666 Berger, W. H., & Parker, F. L. (1970). Diversity of planktonic foraminifera in deep-sea sediments.
1667 *Science*, 168(3937), 1345–1347.
- 1668 Berghella, V. (2012). Universal cervical length screening for prediction and prevention of preterm birth.
1669 *Obstetrical & gynecological survey*, 67(10), 653–657.
- 1670 Bernardo, G., Le Noci, V., Ottaviano, E., De Cecco, L., Camisaschi, C., Guglielmetti, S., . . . others (2023).
1671 Reduction of staphylococcus epidermidis in the mammary tumor microbiota induces antitumor
1672 immunity and decreases breast cancer aggressiveness. *Cancer Letters*, 555, 216041.
- 1673 Blencowe, H., Cousens, S., Oestergaard, M. Z., Chou, D., Moller, A.-B., Narwal, R., . . . others (2012).
1674 National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends
1675 since 1990 for selected countries: a systematic analysis and implications. *The lancet*, 379(9832),
1676 2162–2172.
- 1677 Boland, C. R., & Goel, A. (2010). Microsatellite instability in colorectal cancer. *Gastroenterology*,
1678 138(6), 2073–2087.
- 1679 Boleij, A., Hechenbleikner, E. M., Goodwin, A. C., Badani, R., Stein, E. M., Lazarev, M. G., . . . others
1680 (2015). The bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer
1681 patients. *Clinical Infectious Diseases*, 60(2), 208–215.

- 1682 Bolstad, A., Jensen, H. B., & Bakken, V. (1996). Taxonomy, biology, and periodontal aspects of
1683 *fusobacterium nucleatum*. *Clinical microbiology reviews*, 9(1), 55–71.
- 1684 Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... others
1685 (2019). Reproducible, interactive, scalable and extensible microbiome data science using qiime 2.
1686 *Nature biotechnology*, 37(8), 852–857.
- 1687 Bombin, A., Yan, S., Bombin, S., Mosley, J. D., & Ferguson, J. F. (2022). Obesity influences composition
1688 of salivary and fecal microbiota and impacts the interactions between bacterial taxa. *Physiological
1689 reports*, 10(7), e15254.
- 1690 Bonnet, M., Buc, E., Sauvanet, P., Darcha, C., Dubois, D., Pereira, B., ... Darfeuille-Michaud, A. (2014).
1691 Colonization of the human gut by *e. coli* and colorectal cancer risk. *Clinical Cancer Research*,
1692 20(4), 859–867.
- 1693 Breiman, L. (2001). Random forests. *Machine learning*, 45, 5–32.
- 1694 Brennan, C. A., & Garrett, W. S. (2019). *Fusobacterium nucleatum*—symbiont, opportunist and
1695 *oncobacterium*. *Nature Reviews Microbiology*, 17(3), 156–166.
- 1696 Broom, L. J., & Kogut, M. H. (2018). The role of the gut microbiome in shaping the immune system of
1697 chickens. *Veterinary immunology and immunopathology*, 204, 44–51.
- 1698 Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: improving accuracy of classifier
1699 ensembles by using random feature subsets. *Pattern recognition*, 36(6), 1291–1302.
- 1700 Bullman, S., Pedamallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., ... others (2017). Analysis
1701 of *fusobacterium* persistence and antibiotic response in colorectal cancer. *Science*, 358(6369),
1702 1443–1448.
- 1703 Burt, R. W., Leppert, M. F., Slattery, M. L., Samowitz, W. S., Spirio, L. N., Kerber, R. A., ... others
1704 (2004). Genetic testing and phenotype in a large kindred with attenuated familial adenomatous
1705 polyposis. *Gastroenterology*, 127(2), 444–451.
- 1706 Cai, Y., Li, Y., Xiong, Y., Geng, X., Kang, Y., & Yang, Y. (2024). Diabetic foot exacerbates gut
1707 mycobiome dysbiosis in adult patients with type 2 diabetes mellitus: revealing diagnostic markers.
1708 *Nutrition & Diabetes*, 14(1), 71.
- 1709 Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016).
1710 Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7),
1711 581–583.
- 1712 Canakci, V., & Canakci, C. F. (2007). Pain levels in patients during periodontal probing and mechanical
1713 non-surgical therapy. *Clinical oral investigations*, 11, 377–383.
- 1714 Cappellato, M., Baruzzo, G., & Di Camillo, B. (2022). Investigating differential abundance methods in
1715 microbiome data: A benchmark study. *PLoS computational biology*, 18(9), e1010467.
- 1716 Castaner, O., Goday, A., Park, Y.-M., Lee, S.-H., Magkos, F., Shiow, S.-A. T. E., & Schröder, H. (2018).
1717 The gut microbiome profile in obesity: a systematic review. *International journal of endocrinology*,
1718 2018(1), 4095789.
- 1719 Center, M. M., Jemal, A., Smith, R. A., & Ward, E. (2009). Worldwide variations in colorectal cancer.
1720 *CA: a cancer journal for clinicians*, 59(6), 366–378.

- 1721 Centor, R. M. (1991). Signal detectability: the use of roc curves and their analyses. *Medical decision
making*, 11(2), 102–106.
- 1722
- 1723 Cerqueira, F. M., Photenhauer, A. L., Pollet, R. M., Brown, H. A., & Koropatkin, N. M. (2020). Starch
1724 digestion by gut bacteria: crowdsourcing for carbs. *Trends in Microbiology*, 28(2), 95–108.
- 1725 Champagne, C., McNairn, H., Daneshfar, B., & Shang, J. (2014). A bootstrap method for assessing
1726 classification accuracy and confidence for agricultural land use mapping in canada. *International
1727 Journal of Applied Earth Observation and Geoinformation*, 29, 44–52.
- 1728 Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian
1729 Journal of statistics*, 265–270.
- 1730 Chao, A., & Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the
1731 American statistical Association*, 87(417), 210–217.
- 1732 Chapple, I. L., Mealey, B. L., Van Dyke, T. E., Bartold, P. M., Dommisch, H., Eickholz, P., ... others
1733 (2018). Periodontal health and gingival diseases and conditions on an intact and a reduced
1734 periodontium: Consensus report of workgroup 1 of the 2017 world workshop on the classification
1735 of periodontal and peri-implant diseases and conditions. *Journal of periodontology*, 89, S74–S84.
- 1736 Chen, T., Marsh, P., & Al-Hebshi, N. (2022). Smdi: an index for measuring subgingival microbial
1737 dysbiosis. *Journal of dental research*, 101(3), 331–338.
- 1738 Chen, T., Yu, W.-H., Izard, J., Baranova, O. V., Lakshmanan, A., & Dewhirst, F. E. (2010). The human
1739 oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and
1740 genomic information. *Database*, 2010.
- 1741 Chen, X., D’Souza, R., & Hong, S.-T. (2013). The role of gut microbiota in the gut-brain axis: current
1742 challenges and perspectives. *Protein & cell*, 4, 403–414.
- 1743 Chen, X., Jansen, L., Guo, F., Hoffmeister, M., Chang-Claude, J., & Brenner, H. (2021). Smoking,
1744 genetic predisposition, and colorectal cancer risk. *Clinical and translational gastroenterology*,
1745 12(3), e00317.
- 1746 Chen, X., Li, H., Guo, F., Hoffmeister, M., & Brenner, H. (2022). Alcohol consumption, polygenic risk
1747 score, and early-and late-onset colorectal cancer risk. *EClinicalMedicine*, 49.
- 1748 Chew, R. J. J., Tan, K. S., Chen, T., Al-Hebshi, N. N., & Goh, C. E. (2024). Quantifying periodontitis-
1749 associated oral dysbiosis in tongue and saliva microbiomes—an integrated data analysis. *Journal
1750 of Periodontology*.
- 1751 Čižmárová, B., Tomečková, V., Hubková, B., Hurajtová, A., Ohlasová, J., & Birková, A. (2022). Salivary
1752 redox homeostasis in human health and disease. *International Journal of Molecular Sciences*,
1753 23(17), 10076.
- 1754 Conde-Pérez, K., Aja-Macaya, P., Buetas, E., Trigo-Tasende, N., Nasser-Ali, M., Rumbo-Feal, S., ...
1755 others (2024). The multispecies microbial cluster of fusobacterium, parvimonas, bacteroides and
1756 faecalibacterium as a precision biomarker for colorectal cancer diagnosis. *Molecular Oncology*,
1757 18(5), 1093–1122.
- 1758 Cuervo, S. I., Cortés, J. A., Sánchez, R., Rodríguez, J. Y., Silva, E., Tibavizco, D., & Arroyo, P. (2010).
1759 Risk factors for mortality caused by staphylococcus aureus bacteremia in cancer patients. *Enfer-*

- 1760 *medades infecciosas y microbiologia clinica*, 28(6), 349–354.
- 1761 Cullin, N., Antunes, C. A., Straussman, R., Stein-Thoeringer, C. K., & Elinav, E. (2021). Microbiome
1762 and cancer. *Cancer Cell*, 39(10), 1317–1341.
- 1763 Curtius, K., Wright, N. A., & Graham, T. A. (2018). An evolutionary perspective on field cancerization.
1764 *Nature Reviews Cancer*, 18(1), 19–32.
- 1765 Dabke, K., Hendrick, G., Devkota, S., et al. (2019). The gut microbiome and metabolic syndrome. *The
1766 Journal of clinical investigation*, 129(10), 4050–4057.
- 1767 De Flora, S., & Bonanni, P. (2011). The prevention of infection-associated cancers. *Carcinogenesis*,
1768 32(6), 787–795.
- 1769 DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... Andersen, G. L.
1770 (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with
1771 arb. *Applied and environmental microbiology*, 72(7), 5069–5072.
- 1772 Ding, R., Lian, S. B., Tam, Y. C., & Oh, C. C. (2024). The cutaneous microbiome in skin cancer—a
1773 systematic review. *JDDG: Journal der Deutschen Dermatologischen Gesellschaft*, 22(2), 177–184.
- 1774 Dong, K., Wang, J., Tang, F., Liu, Y., & Gao, L. (2024). A polysaccharide with a triple helix structure
1775 from agaricus bisporus: Characterization and anti-colon cancer activity. *International Journal of
1776 Biological Macromolecules*, 281, 136521.
- 1777 Doyle, R., Alber, D., Jones, H., Harris, K., Fitzgerald, F., Peebles, D., & Klein, N. (2014). Term and
1778 preterm labour are associated with distinct microbial community structures in placental membranes
1779 which are independent of mode of delivery. *Placenta*, 35(12), 1099–1101.
- 1780 Duy, T. N., Le Huy, H., Thanh, Q. Đ., Thi, H. N., Minh, H. N. T., Dang, M. N., ... Tat, T. N. (2024).
1781 Association between bacteroides fragilis and fusobacterium nucleatum infection and colorectal
1782 cancer in vietnamese patients. *Anaerobe*, 88, 102880.
- 1783 El-Deeb, N. M., Ibrahim, O. M., Mohamed, M. A., Farag, M. M., Farrag, A. A., & El-Aassar, M.
1784 (2022). Alginate/κ-carrageenan oral microcapsules loaded with agaricus bisporus polysaccharides
1785 mh751906 for natural killer cells mediated colon cancer immunotherapy. *International Journal of
1786 Biological Macromolecules*, 205, 385–395.
- 1787 Fahmy, C. A., Gamal-Eldeen, A. M., El-Hussieny, E. A., Raafat, B. M., Mehanna, N. S., Talaat, R. M., &
1788 Shaaban, M. T. (2019). Bifidobacterium longum suppresses murine colorectal cancer through the
1789 modulation of oncomirs and tumor suppressor mirnas. *Nutrition and cancer*, 71(4), 688–700.
- 1790 Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological conservation*, 61(1),
1791 1–10.
- 1792 Fettweis, J. M., Serrano, M. G., Brooks, J. P., Edwards, D. J., Girerd, P. H., Parikh, H. I., ... others
1793 (2019). The vaginal microbiome and preterm birth. *Nature medicine*, 25(6), 1012–1021.
- 1794 Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and
1795 the number of individuals in a random sample of an animal population. *The Journal of Animal
1796 Ecology*, 42–58.
- 1797 Flanagan, L., Schmid, J., Ebert, M., Soucek, P., Kunicka, T., Liska, V., ... others (2014). Fusobacterium
1798 nucleatum associates with stages of colorectal neoplasia development, colorectal cancer and disease

- 1799 outcome. *European journal of clinical microbiology & infectious diseases*, 33, 1381–1390.
- 1800 Fortenberry, J. D. (2013). The uses of race and ethnicity in human microbiome research. *Trends in*
1801 *microbiology*, 21(4), 165–166.
- 1802 Francescone, R., Hou, V., & Grivennikov, S. I. (2014). Microbiome, inflammation, and cancer. *The*
1803 *Cancer Journal*, 20(3), 181–189.
- 1804 Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4),
1805 367–378.
- 1806 Fushiki, T. (2011). Estimation of prediction error by using k-fold cross-validation. *Statistics and*
1807 *Computing*, 21, 137–146.
- 1808 Galeano Niño, J. L., Wu, H., LaCourse, K. D., Kempchinsky, A. G., Baryiames, A., Barber, B., ... others
1809 (2022). Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer. *Nature*,
1810 611(7937), 810–817.
- 1811 Gambin, D. J., Vitali, F. C., De Carli, J. P., Mazzon, R. R., Gomes, B. P., Duque, T. M., & Trentin, M. S.
1812 (2021). Prevalence of red and orange microbial complexes in endodontic-periodontal lesions: a
1813 systematic review and meta-analysis. *Clinical Oral Investigations*, 1–14.
- 1814 Gao, J., Yin, J., Xu, K., Li, T., & Yin, Y. (2019). What is the impact of diet on nutritional diarrhea
1815 associated with gut microbiota in weaning piglets: a system review. *BioMed research international*,
1816 2019(1), 6916189.
- 1817 Gao, R., Kong, C., Li, H., Huang, L., Qu, X., Qin, N., & Qin, H. (2017). Dysbiosis signature of mycobiota
1818 in colon polyp and colorectal cancer. *European Journal of Clinical Microbiology & Infectious*
1819 *Diseases*, 36, 2457–2468.
- 1820 Gao, S., Liu, Y., Duan, X., Liu, K., Mohammed, M., Gu, Z., ... others (2021). Porphyromonas gingivalis
1821 infection exacerbates oesophageal cancer and promotes resistance to neoadjuvant chemotherapy.
1822 *British Journal of Cancer*, 125(3), 433–444.
- 1823 Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63, 3–42.
- 1824 Ghanavati, R., Akbari, A., Mohammadi, F., Asadollahi, P., Javadi, A., Talebi, M., & Rohani, M. (2020).
1825 Lactobacillus species inhibitory effect on colorectal cancer progression through modulating the
1826 wnt/β-catenin signaling pathway. *Molecular and Cellular Biochemistry*, 470, 1–13.
- 1827 Ghojogh, B., & Crowley, M. (2019). The theory behind overfitting, cross validation, regularization,
1828 bagging, and boosting: tutorial. *arXiv preprint arXiv:1905.12787*.
- 1829 Ghorbani, E., Avan, A., Ryzhikov, M., Ferns, G., Khazaei, M., & Soleimanpour, S. (2022). Role of
1830 lactobacillus strains in the management of colorectal cancer: An overview of recent advances.
1831 *Nutrition*, 103, 111828.
- 1832 Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current
1833 understanding of the human microbiome. *Nature medicine*, 24(4), 392–400.
- 1834 Gini, C. (1912). Variabilità e mutabilità (variability and mutability). *Tipografia di Paolo Cuppini,*
1835 *Bologna, Italy*, 156.
- 1836 Goldenberg, R. L., Culhane, J. F., Iams, J. D., & Romero, R. (2008). Epidemiology and causes of preterm
1837 birth. *The lancet*, 371(9606), 75–84.

- 1838 Gonçalves, L., Subtil, A., Oliveira, M. R., & de Zea Bermudez, P. (2014). Roc curve estimation: An
1839 overview. *REVSTAT-Statistical journal*, 12(1), 1–20.
- 1840 Good, I. J. (1953). The population frequencies of species and the estimation of population parameters.
1841 *Biometrika*, 40(3-4), 237–264.
- 1842 Goodyear, M. D., Krleza-Jeric, K., & Lemmens, T. (2007). *The declaration of helsinki* (Vol. 335) (No.
1843 7621). British Medical Journal Publishing Group.
- 1844 Haffajee, A., Teles, R., & Socransky, S. (2006). Association of eubacterium nodatum and treponema
1845 denticola with human periodontitis lesions. *Oral microbiology and immunology*, 21(5), 269–282.
- 1846 Hajishengallis, G. (2015). Periodontitis: from microbial immune subversion to systemic inflammation.
1847 *Nature reviews immunology*, 15(1), 30–44.
- 1848 Hamjane, N., Mechita, M. B., Nourouti, N. G., & Barakat, A. (2024). Gut microbiota dysbiosis-associated
1849 obesity and its involvement in cardiovascular diseases and type 2 diabetes. a systematic review.
1850 *Microvascular Research*, 151, 104601.
- 1851 Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*,
1852 29(2), 147–160.
- 1853 Hampel, H., Frankel, W. L., Martin, E., Arnold, M., Khanduja, K., Kuebler, P., ... others (2008).
1854 Feasibility of screening for lynch syndrome among patients with colorectal cancer. *Journal of
1855 Clinical Oncology*, 26(35), 5783–5788.
- 1856 Han, Y. W. (2015). Fusobacterium nucleatum: a commensal-turned pathogen. *Current opinion in
1857 microbiology*, 23, 141–147.
- 1858 Han, Y. W., & Wang, X. (2013). Mobile microbiome: oral bacteria in extra-oral infections and
1859 inflammation. *Journal of dental research*, 92(6), 485–491.
- 1860 Hand, D. J. (2012). Assessing the performance of classification methods. *International Statistical Review*,
1861 80(3), 400–414.
- 1862 Harkins, L., Volk, A. L., Samanta, M., Mikolaenko, I., Britt, W. J., Bland, K. I., & Cobbs, C. S. (2002).
1863 Specific localisation of human cytomegalovirus nucleic acids and proteins in human colorectal
1864 cancer. *The Lancet*, 360(9345), 1557–1563.
- 1865 Hartstra, A. V., Bouter, K. E., Bäckhed, F., & Nieuwdorp, M. (2015). Insights into the role of the
1866 microbiome in obesity and type 2 diabetes. *Diabetes care*, 38(1), 159–165.
- 1867 Hasan, R., Bose, S., Roy, R., Paul, D., Rawat, S., Nilwe, P., ... Choudhury, S. (2022). Tumor tissue-
1868 specific bacterial biomarker panel for colorectal cancer: Bacteroides massiliensis, alistipes species,
1869 alistipes onderdonkii, bifidobacterium pseudocatenulatum, corynebacterium appendicis. *Archives
1870 of microbiology*, 204(6), 348.
- 1871 Hashemi Goradel, N., Heidarzadeh, S., Jahangiri, S., Farhood, B., Mortezaee, K., Khanlarkhani, N., &
1872 Negahdari, B. (2019). Fusobacterium nucleatum and colorectal cancer: A mechanistic overview.
1873 *Journal of Cellular Physiology*, 234(3), 2337–2344.
- 1874 Hassan, Z., Mustafa, S., Rahim, R. A., & Isa, N. M. (2016). Anti-breast cancer effects of live, heat-killed
1875 and cytoplasmic fractions of enterococcus faecalis and staphylococcus hominis isolated from
1876 human breast milk. *In Vitro Cellular & Developmental Biology-Animal*, 52, 337–348.

- 1877 Hegazy, A. A., Abu-Hussien, S. H., Elsenosy, N. K., El-Sayed, S. M., & Abo El-Naga, M. Y. (2024).
1878 Optimization, characterization and biosafety of carotenoids produced from whey using *micrococcus*
1879 *luteus*. *BMC biotechnology*, 24(1), 74.
- 1880 Heip, C. (1974). A new index measuring evenness. *Journal of the Marine Biological Association of the*
1881 *United Kingdom*, 54(3), 555–557.
- 1882 Helmink, B. A., Khan, M. W., Hermann, A., Gopalakrishnan, V., & Wargo, J. A. (2019). The microbiome,
1883 cancer, and cancer therapy. *Nature medicine*, 25(3), 377–388.
- 1884 Hexun, Z., Miyake, T., Maekawa, T., Mori, H., Yasukawa, D., Ohno, M., ... Tani, M. (2023). High
1885 abundance of lachnospiraceae in the human gut microbiome is related to high immunoscores in
1886 advanced colorectal cancer. *Cancer Immunology, Immunotherapy*, 72(2), 315–326.
- 1887 Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2),
1888 427–432.
- 1889 Hiranmayi, K. V., Sirisha, K., Rao, M. R., & Sudhakar, P. (2017). Novel pathogens in periodontal
1890 microbiology. *Journal of Pharmacy and Bioallied Sciences*, 9(3), 155–163.
- 1891 Honda, K., & Littman, D. R. (2012). The microbiome in infectious disease and inflammation. *Annual*
1892 *review of immunology*, 30(1), 759–795.
- 1893 Honest, H., Forbes, C., Durée, K., Norman, G., Duffy, S., Tsourapas, A., ... others (2009). Screening to
1894 prevent spontaneous preterm birth: systematic reviews of accuracy and effectiveness literature with
1895 economic modelling. *Health Technol Assess*, 13(43), 1–627.
- 1896 Hong, Y. M., Lee, J., Cho, D. H., Jeon, J. H., Kang, J., Kim, M.-G., ... J. K. (2023). Predicting preterm
1897 birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1), 21105.
- 1898 Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations.
1899 *International journal of data mining & knowledge management process*, 5(2), 1.
- 1900 Huang, R.-Y., Lin, C.-D., Lee, M.-S., Yeh, C.-L., Shen, E.-C., Chiang, C.-Y., ... Fu, E. (2007). Mandibular
1901 disto-lingual root: a consideration in periodontal therapy. *Journal of periodontology*, 78(8), 1485–
1902 1490.
- 1903 Huo, R.-X., Wang, Y.-J., Hou, S.-B., Wang, W., Zhang, C.-Z., & Wan, X.-H. (2022). Gut mucosal
1904 microbiota profiles linked to colorectal cancer recurrence. *World journal of gastroenterology*,
1905 28(18), 1946.
- 1906 Iams, J. D., & Berghella, V. (2010). Care for women with prior preterm birth. *American journal of*
1907 *obstetrics and gynecology*, 203(2), 89–100.
- 1908 Ide, M., & Papapanou, P. N. (2013). Epidemiology of association between maternal periodontal
1909 disease and adverse pregnancy outcomes—systematic review. *Journal of clinical periodontology*,
1910 40, S181–S194.
- 1911 Iniesta, M., Chamorro, C., Ambrosio, N., Marín, M. J., Sanz, M., & Herrera, D. (2023). Subgingival
1912 microbiome in periodontal health, gingivitis and different stages of periodontitis. *Journal of*
1913 *Clinical Periodontology*, 50(7), 905–920.
- 1914 Inra, J. A., Steyerberg, E. W., Grover, S., McFarland, A., Syngal, S., & Kastrinos, F. (2015). Racial
1915 variation in frequency and phenotypes of apc and mutyh mutations in 6,169 individuals undergoing

- 1916 genetic testing. *Genetics in Medicine*, 17(10), 815–821.
- 1917 Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44, 1918 223–270.
- 1919 Janda, J. M., & Abbott, S. L. (2007). 16s rrna gene sequencing for bacterial identification in the diagnostic 1920 laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9), 2761–2764.
- 1921 Jiang, W., & Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach 1922 for estimating the prediction error in microarray classification. *Statistics in medicine*, 26(29), 1923 5320–5334.
- 1924 John, G. K., & Mullin, G. E. (2016). The gut microbiome and obesity. *Current oncology reports*, 18, 1925 1–7.
- 1926 Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., . . . others (2019). 1927 Evaluation of 16s rrna gene sequencing for species and strain-level microbiome analysis. *Nature 1928 communications*, 10(1), 5029.
- 1929 Joo, J. E., Chu, Y. L., Georgeson, P., Walker, R., Mahmood, K., Clendenning, M., . . . others (2024). 1930 Intratumoral presence of the genotoxic gut bacteria pks+ e. coli, enterotoxigenic bacteroides fragilis, 1931 and fusobacterium nucleatum and their association with clinicopathological and molecular features 1932 of colorectal cancer. *British Journal of Cancer*, 130(5), 728–740.
- 1933 Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N., & Whiteley, M. (2014). Metatranscriptomics 1934 of the human oral microbiome during health and disease. *MBio*, 5(2), 10–1128.
- 1935 Joscelyn, J., & Kasper, L. H. (2014). Digesting the emerging role for the gut microbiome in central 1936 nervous system demyelination. *Multiple Sclerosis Journal*, 20(12), 1553–1559.
- 1937 Kang, Y., Kang, X., Yang, H., Liu, H., Yang, X., Liu, Q., . . . others (2022). Lactobacillus acidophilus ame- 1938 liorates obesity in mice through modulation of gut microbiota dysbiosis and intestinal permeability. 1939 *Pharmacological research*, 175, 106020.
- 1940 Karched, M., Bhardwaj, R. G., Qudeimat, M., Al-Khabbaz, A., & Ellepola, A. (2022). Proteomic analysis 1941 of the periodontal pathogen prevotella intermedia secretomes in biofilm and planktonic lifestyles. 1942 *Scientific Reports*, 12(1), 5636.
- 1943 Katz, J., Chegini, N., Shiverick, K., & Lamont, R. (2009). Localization of p. gingivalis in preterm delivery 1944 placenta. *Journal of dental research*, 88(6), 575–578.
- 1945 Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., & Gordon, J. I. (2011). Human nutrition, the 1946 gut microbiome and the immune system. *Nature*, 474(7351), 327–336.
- 1947 Kelly, B. J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J. D., Collman, R. G., . . . Li, H. (2015). 1948 Power and sample-size estimation for microbiome studies using pairwise distances and permanova. 1949 *Bioinformatics*, 31(15), 2461–2468.
- 1950 Kennedy, J., Alexander, P., Taillie, L. S., & Jaacks, L. M. (2024). Estimated effects of reductions in 1951 processed meat consumption and unprocessed red meat consumption on occurrences of type 2 1952 diabetes, cardiovascular disease, colorectal cancer, and mortality in the usa: a microsimulation 1953 study. *The Lancet Planetary Health*, 8(7), e441–e451.
- 1954 Kepp, O., Zitvogel, L., & Kroemer, G. (2023). *Prevention and treatment of cancers by tumor antigen-*

- 1955 *expressing staphylococcus epidermidis* (Vol. 12) (No. 1). Taylor & Francis.
- 1956 Kim, B.-R., Shin, J., Guevarra, R. B., Lee, J. H., Kim, D. W., Seol, K.-H., ... Isaacson, R. E. (2017).
1957 Deciphering diversity indices for a better understanding of microbial communities. *Journal of
1958 Microbiology and Biotechnology*, 27(12), 2089–2093.
- 1959 Kim, C. H. (2018). Immune regulation by microbiome metabolites. *Immunology*, 154(2), 220–229.
- 1960 Kim, E.-H., Kim, S., Kim, H.-J., Jeong, H.-o., Lee, J., Jang, J., ... others (2020). Prediction of chronic
1961 periodontitis severity using machine learning models based on salivary bacterial copy number.
1962 *Frontiers in Cellular and Infection Microbiology*, 10, 571515.
- 1963 Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and
1964 bootstrap. *Computational statistics & data analysis*, 53(11), 3735–3745.
- 1965 Kim, S., Lee, M., Kim, N.-Y., Kwon, Y.-S., Nam, G. S., Lee, K., ... Hwang, I. H. (2024). Oxidative
1966 tryptamine dimers from corynebacterium durum directly target survivin to induce aif-mediated
1967 apoptosis in cancer cells. *Biomedicine & Pharmacotherapy*, 173, 116335.
- 1968 Kinane, D. F., Stathopoulou, P. G., & Papapanou, P. N. (2017). Periodontal diseases. *Nature reviews
1969 Disease primers*, 3(1), 1–14.
- 1970 Kindinger, L. M., Bennett, P. R., Lee, Y. S., Marchesi, J. R., Smith, A., Caciato, S., ... MacIntyre,
1971 D. A. (2017). The interaction between vaginal microbiota, cervical length, and vaginal progesterone
1972 treatment for preterm birth risk. *Microbiome*, 5, 1–14.
- 1973 Kogut, M. H., Lee, A., & Santin, E. (2020). Microbiome and pathogen interaction with the immune
1974 system. *Poultry science*, 99(4), 1906–1913.
- 1975 Kostic, A. D., Ojesina, A. I., Pedamallu, C. S., Jung, J., Verhaak, R. G., Getz, G., & Meyerson, M. (2011).
1976 Pathseq: software to identify or discover microbes by deep sequencing of human tissue. *Nature
1977 biotechnology*, 29(5), 393–396.
- 1978 Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification
1979 and combining techniques. *Artificial Intelligence Review*, 26, 159–190.
- 1980 Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., ... Watanabe, T.
1981 (2015). Colorectal cancer. *Nature reviews. Disease primers*, 1, 15065.
- 1982 Lafaurie, G. I., Neuta, Y., Ríos, R., Pacheco-Montealegre, M., Pianeta, R., Castillo, D. M., ... oth-
1983 ers (2022). Differences in the subgingival microbiome according to stage of periodontitis: A
1984 comparison of two geographic regions. *PLoS one*, 17(8), e0273523.
- 1985 Lamont, R. J., & Jenkinson, H. F. (2000). Subgingival colonization by porphyromonas gingivalis. *Oral
1986 Microbiology and Immunology: Mini-review*, 15(6), 341–349.
- 1987 Lamont, R. J., Koo, H., & Hajishengallis, G. (2018). The oral microbiota: dynamic communities and
1988 host interactions. *Nature reviews microbiology*, 16(12), 745–759.
- 1989 Leitich, H., & Kaider, A. (2003). Fetal fibronectin—how useful is it in the prediction of preterm birth?
1990 *BJOG: An International Journal of Obstetrics & Gynaecology*, 110, 66–70.
- 1991 Le Leu, R. K., Hu, Y., Brown, I. L., Woodman, R. J., & Young, G. P. (2010). Synbiotic intervention of
1992 bifidobacterium lactis and resistant starch protects against colorectal cancer development in rats.
1993 *Carcinogenesis*, 31(2), 246–251.

- 1994 León, R., Silva, N., Ovalle, A., Chaparro, A., Ahumada, A., Gajardo, M., ... Gamonal, J. (2007).
1995 Detection of porphyromonas gingivalis in the amniotic fluid in pregnant women with a diagnosis
1996 of threatened premature labor. *Journal of periodontology*, 78(7), 1249–1255.
- 1997 Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform.
1998 *bioinformatics*, 25(14), 1754–1760.
- 1999 Li, N., Lu, B., Luo, C., Cai, J., Lu, M., Zhang, Y., ... Dai, M. (2021). Incidence, mortality, survival,
2000 risk factor and screening of colorectal cancer: A comparison among china, europe, and northern
2001 america. *Cancer letters*, 522, 255–268.
- 2002 Li, R., Miao, Z., Liu, Y., Chen, X., Wang, H., Su, J., & Chen, J. (2024). The brain–gut–bone axis in
2003 neurodegenerative diseases: insights, challenges, and future prospects. *Advanced Science*, 11(38),
2004 2307971.
- 2005 Li, W., & Yang, J. (2025). Investigating the anna karenina principle of the breast microbiome. *BMC*
2006 *microbiology*, 25(1), 1–10.
- 2007 Li, X., Yu, D., Wang, Y., Yuan, H., Ning, X., Rui, B., ... Li, M. (2021). The intestinal dysbiosis of
2008 mothers with gestational diabetes mellitus (gdm) and its impact on the gut microbiota of their
2009 newborns. *Canadian Journal of Infectious Diseases and Medical Microbiology*, 2021(1), 3044534.
- 2010 Li, Y., Qian, F., Cheng, X., Wang, D., Wang, Y., Pan, Y., ... Tian, Y. (2023). Dysbiosis of oral microbiota
2011 and metabolite profiles associated with type 2 diabetes mellitus. *Microbiology spectrum*, 11(1),
2012 e03796–22.
- 2013 Li, Z., Zhuang, H., Wang, G., Wang, H., & Dong, Y. (2021). Prevalence, predictors, and mortality
2014 of bloodstream infections due to methicillin-resistant staphylococcus aureus in patients with
2015 malignancy: systemic review and meta-analysis. *BMC infectious diseases*, 21, 1–10.
- 2016 Lim, J. W., Park, T., Tong, Y. W., & Yu, Z. (2020). The microbiome driving anaerobic digestion and
2017 microbial analysis. In *Advances in bioenergy* (Vol. 5, pp. 1–61). Elsevier.
- 2018 Lin, H., Eggesbø, M., & Peddada, S. D. (2022). Linear and nonlinear correlation estimators unveil
2019 undescribed taxa interactions in microbiome data. *Nature communications*, 13(1), 4946.
- 2020 Lin, H., & Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nature*
2021 *communications*, 11(1), 3514.
- 2022 Lin, H., & Peddada, S. D. (2024). Multigroup analysis of compositions of microbiomes with covariate
2023 adjustments and repeated measures. *Nature Methods*, 21(1), 83–91.
- 2024 Listgarten, M. A. (1986). Pathogenesis of periodontitis. *Journal of clinical periodontology*, 13(5),
2025 418–425.
- 2026 Liu, W., Zhang, X., Xu, H., Li, S., Lau, H. C.-H., Chen, Q., ... others (2021). Microbial commu-
2027 nity heterogeneity within colorectal neoplasia and its correlation with colorectal carcinogenesis.
2028 *Gastroenterology*, 160(7), 2395–2408.
- 2029 Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome*
2030 *medicine*, 8, 1–11.
- 2031 López-Aladid, R., Fernández-Barat, L., Alcaraz-Serrano, V., Bueno-Freire, L., Vázquez, N., Pastor-
2032 Ibáñez, R., ... Torres, A. (2023). Determining the most accurate 16s rrna hypervariable region for

- 2033 taxonomic identification from respiratory samples. *Scientific reports*, 13(1), 3974.
- 2034 Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for
2035 rna-seq data with deseq2. *Genome biology*, 15, 1–21.
- 2036 Lu, Y.-T., Hsin, C.-H., Chuang, C.-Y., Huang, C.-C., Su, M.-C., Wen, W.-S., ... others (2024). Mi-
2037 crobial dysbiosis in nasopharyngeal carcinoma: A pilot study on biomarker potential. *Journal of*
2038 *Otolaryngology-Head & Neck Surgery*, 53, 19160216241304365.
- 2039 Ma, Z. S. (2020). Testing the anna karenina principle in human microbiome-associated diseases. *Iscience*,
2040 23(4).
- 2041 Magnúsdóttir, S., & Thiele, I. (2018). Modeling metabolism of the human gut microbiome. *Current*
2042 *opinion in biotechnology*, 51, 90–96.
- 2043 Magurran, A. E. (2021). Measuring biological diversity. *Current Biology*, 31(19), R1174–R1177.
- 2044 Mandic, M., Safizadeh, F., Niedermaier, T., Hoffmeister, M., & Brenner, H. (2023). Association of
2045 overweight, obesity, and recent weight loss with colorectal cancer risk. *JAMA network Open*, 6(4),
2046 e239556–e239556.
- 2047 Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically
2048 larger than the other. *The annals of mathematical statistics*, 50–60.
- 2049 Manolis, A. A., Manolis, T. A., Melita, H., & Manolis, A. S. (2022). Gut microbiota and cardiovascular
2050 disease: symbiosis versus dysbiosis. *Current Medicinal Chemistry*, 29(23), 4050–4077.
- 2051 Martin, C. R., Osadchiy, V., Kalani, A., & Mayer, E. A. (2018). The brain-gut-microbiome axis. *Cellular*
2052 *and molecular gastroenterology and hepatology*, 6(2), 133–148.
- 2053 Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine
2054 learning. *Journal of Applied Science and Technology Trends*, 1(2), 140–147.
- 2055 Mayer, E. A., Tillisch, K., Gupta, A., et al. (2015). Gut/brain axis and the microbiota. *The Journal of*
2056 *clinical investigation*, 125(3), 926–938.
- 2057 Melguizo-Rodríguez, L., Costela-Ruiz, V. J., Manzano-Moreno, F. J., Ruiz, C., & Illescas-Montes, R.
2058 (2020). Salivary biomarkers and their application in the diagnosis and monitoring of the most
2059 common oral pathologies. *International journal of molecular sciences*, 21(14), 5173.
- 2060 Merrill, L. C., & Mangano, K. M. (2023). Racial and ethnic differences in studies of the gut microbiome
2061 and osteoporosis. *Current Osteoporosis Reports*, 21(5), 578–591.
- 2062 Miller, C. S., Ding, X., Dawson III, D. R., & Ebersole, J. L. (2021). Salivary biomarkers for discriminating
2063 periodontitis in the presence of diabetes. *Journal of clinical periodontology*, 48(2), 216–225.
- 2064 Minton, N. P. (2003). Clostridia in cancer therapy. *Nature Reviews Microbiology*, 1(3), 237–242.
- 2065 Mjelle, R., Castro, Í., & Aass, K. R. (2025). The viral landscape in metastatic solid cancers. *Heliyon*.
- 2066 Montalban-Arques, A., Katkeviciute, E., Busenhart, P., Bircher, A., Wirbel, J., Zeller, G., ... others
2067 (2021). Commensal clostridiales strains mediate effective anti-cancer immune response against
2068 solid tumors. *Cell host & microbe*, 29(10), 1573–1588.
- 2069 Morita, T., Yamazaki, Y., Mita, A., Takada, K., Seto, M., Nishinoue, N., ... Maeno, M. (2010). A cohort
2070 study on the association between periodontal disease and the development of metabolic syndrome.
2071 *Journal of periodontology*, 81(4), 512–519.

- 2072 Na, H. S., Kim, S. Y., Han, H., Kim, H.-J., Lee, J.-Y., Lee, J.-H., & Chung, J. (2020). Identification of
2073 potential oral microbial biomarkers for the diagnosis of periodontitis. *Journal of clinical medicine*,
2074 9(5), 1549.
- 2075 Nemoto, T., Shiba, T., Komatsu, K., Watanabe, T., Shimogishi, M., Shibasaki, M., ... others (2021).
2076 Discrimination of bacterial community structures among healthy, gingivitis, and periodontitis
2077 statuses through integrated metatranscriptomic and network analyses. *Msystems*, 6(6), e00886–21.
- 2078 Nesbitt, M. J., Reynolds, M. A., Shiau, H., Choe, K., Simonsick, E. M., & Ferrucci, L. (2010). Association
2079 of periodontitis and metabolic syndrome in the baltimore longitudinal study of aging. *Aging clinical
2080 and experimental research*, 22, 238–242.
- 2081 Network, C. G. A., et al. (2012). Comprehensive molecular characterization of human colon and rectal
2082 cancer. *Nature*, 487(7407), 330.
- 2083 Nibali, L., Sousa, V., Davrandi, M., Spratt, D., Alyahya, Q., Dopico, J., & Donos, N. (2020). Differences
2084 in the periodontal microbiome of successfully treated and persistent aggressive periodontitis.
2085 *Journal of Clinical Periodontology*, 47(8), 980–990.
- 2086 Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Ž., & Tomović, M. (2017). Evaluation of classification
2087 models in machine learning. *Theory and Applications of Mathematics & Computer Science*, 7(1),
2088 39.
- 2089 Obuchowski, N. A., & Bullen, J. A. (2018). Receiver operating characteristic (roc) curves: review of
2090 methods with applications in diagnostic medicine. *Physics in Medicine & Biology*, 63(7), 07TR01.
- 2091 Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in japan and its neighbouring
2092 regions. *Bulletin of Japanese Society of Scientific Fisheries*, 22, 526–530.
- 2093 Offenbacher, S., Katz, V., Fertik, G., Collins, J., Boyd, D., Maynor, G., ... Beck, J. (1996). Periodontal
2094 infection as a possible risk factor for preterm low birth weight. *Journal of periodontology*, 67,
2095 1103–1113.
- 2096 Ohmagari, N., Hanna, H., Graviss, L., Hackett, B., Perego, C., Gonzalez, V., ... others (2005). Risk
2097 factors for infections with multidrug-resistant pseudomonas aeruginosa in patients with cancer.
2098 *Cancer*, 104(1), 205–212.
- 2099 Ojesina, A. I., Pedamallu, C. S., Kostic, A., Jung, J., Auclair, D., Lohr, J., ... Meyerson, M. (2013). High
2100 throughput sequencing-based pathogen discovery in multiple myeloma. *Blood*, 122(21), 5322.
- 2101 Omundiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019). Machine learning classification techniques
2102 for breast cancer diagnosis. In *Iop conference series: materials science and engineering* (Vol. 495,
2103 p. 012033).
- 2104 O'Sullivan, D. E., Sutherland, R. L., Town, S., Chow, K., Fan, J., Forbes, N., ... Brenner, D. R. (2022).
2105 Risk factors for early-onset colorectal cancer: a systematic review and meta-analysis. *Clinical
2106 gastroenterology and hepatology*, 20(6), 1229–1240.
- 2107 Paganini, D., & Zimmermann, M. B. (2017). The effects of iron fortification and supplementation on the
2108 gut microbiome and diarrhea in infants and children: a review. *The American journal of clinical
2109 nutrition*, 106, 1688S–1693S.
- 2110 Pan, A. Y. (2021). Statistical analysis of microbiome data: the challenge of sparsity. *Current Opinion in*

- 2111 *Endocrine and Metabolic Research*, 19, 35–40.
- 2112 Papapanou, P. N., Sanz, M., Buduneli, N., Dietrich, T., Feres, M., Fine, D. H., ... others (2018).
2113 Periodontitis: Consensus report of workgroup 2 of the 2017 world workshop on the classification of
2114 periodontal and peri-implant diseases and conditions. *Journal of periodontology*, 89, S173–S182.
- 2115 Paprocka, P., Durnaś, B., Mańkowska, A., Król, G., Wollny, T., & Bucki, R. (2022). *Pseudomonas*
2116 *aeruginosa* infections in cancer patients. *Pathogens*, 11(6), 679.
- 2117 Parizadeh, M., & Arrieta, M.-C. (2023). The global human gut microbiome: genes, lifestyles, and diet.
2118 *Trends in Molecular Medicine*.
- 2119 Park, J., Park, S. H., Lee, D., Lee, J. E., Lee, D., Na, K. J., ... Im, H.-J. (2024). Detecting cancer microbiota
2120 using unmapped rna reads on spatial transcriptomics. *Cancer Research*, 84(6_Supplement), 4881–
2121 4881.
- 2122 Payne, M. S., Newnham, J. P., Doherty, D. A., Furfaro, L. L., Pendal, N. L., Loh, D. E., & Keelan, J. A.
2123 (2021). A specific bacterial dna signature in the vagina of australian women in midpregnancy
2124 predicts high risk of spontaneous preterm birth (the predict1000 study). *American journal of*
2125 *obstetrics and gynecology*, 224(2), 206–e1.
- 2126 Peirce, J. M., & Alviña, K. (2019). The role of inflammation and the gut microbiome in depression and
2127 anxiety. *Journal of neuroscience research*, 97(10), 1223–1241.
- 2128 Peltomaki, P. (2003). Role of dna mismatch repair defects in the pathogenesis of human cancer. *Journal*
2129 *of clinical oncology*, 21(6), 1174–1179.
- 2130 Pezzino, S., Sofia, M., Greco, L. P., Litrico, G., Filippello, G., Sarvà, I., ... Latteri, S. (2023). Microbiome
2131 dysbiosis: a pathological mechanism at the intersection of obesity and glaucoma. *International*
2132 *Journal of Molecular Sciences*, 24(2), 1166.
- 2133 Pollard, T. J., Johnson, A. E., Raffa, J. D., & Mark, R. G. (2018). tableone: An open source python
2134 package for producing summary statistics for research papers. *JAMIA open*, 1(1), 26–31.
- 2135 Premaraj, T. S., Vella, R., Chung, J., Lin, Q., Hunter, P., Underwood, K., ... Zhou, Y. (2020). Ethnic
2136 variation of oral microbiota in children. *Scientific reports*, 10(1), 14788.
- 2137 Purcell, R. V., Permain, J., & Keenan, J. I. (2022). Enterotoxigenic bacteroides fragilis activates il-8
2138 expression through stat3 in colorectal cancer cells. *Gut Pathogens*, 14(1), 16.
- 2139 Raut, J. R., Schöttker, B., Holleczeck, B., Guo, F., Bhardwaj, M., Miah, K., ... Brenner, H. (2021).
2140 A microrna panel compared to environmental and polygenic scores for colorectal cancer risk
2141 prediction. *Nature Communications*, 12(1), 4811.
- 2142 Rebersek, M. (2021). Gut microbiome and its role in colorectal cancer. *BMC cancer*, 21(1), 1325.
- 2143 Redanz, U., Redanz, S., Treerat, P., Prakasam, S., Lin, L.-J., Merritt, J., & Kreth, J. (2021). Differential
2144 response of oral mucosal and gingival cells to corynebacterium durum, streptococcus sanguinis, and
2145 porphyromonas gingivalis multispecies biofilms. *Frontiers in cellular and infection microbiology*,
2146 11, 686479.
- 2147 Relvas, M., Regueira-Iglesias, A., Balsa-Castro, C., Salazar, F., Pacheco, J., Cabral, C., ... Tomás, I.
2148 (2021). Relationship between dental and periodontal health status and the salivary microbiome:
2149 bacterial diversity, co-occurrence networks and predictive models. *Scientific reports*, 11(1), 929.

- 2150 Renson, A., Jones, H. E., Beghini, F., Segata, N., Zolnik, C. P., Usyk, M., ... others (2019). Sociodemographic variation in the oral microbiome. *Annals of epidemiology*, 35, 73–80.
- 2151
- 2152 Renvert, S., & Persson, G. (2002). A systematic review on the use of residual probing depth, bleeding on probing and furcation status following initial periodontal therapy to predict further attachment and tooth loss. *Journal of clinical periodontology*, 29, 82–89.
- 2153
- 2154
- 2155 Rideout, J. R., Caporaso, G., Bolyen, E., McDonald, D., Baeza, Y. V., Alastuey, J. C., ... Sharma, K. (2018, December). *biocore/scikit-bio: scikit-bio 0.5.5: More compositional methods added*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.2254379> doi: 10.5281/zenodo.2254379
- 2156
- 2157
- 2158 Rôças, I. N., Siqueira Jr, J. F., Santos, K. R., Coelho, A. M., & de Janeiro, R. (2001). “red complex”(bacteroides forsythus, porphyromonas gingivalis, and treponema denticola) in endodontic infections: a molecular approach. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, 91(4), 468–471.
- 2159
- 2160
- 2161
- 2162 Romero, R., Dey, S. K., & Fisher, S. J. (2014). Preterm labor: one syndrome, many causes. *Science*, 345(6198), 760–765.
- 2163
- 2164 Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., ... others (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome*, 2, 1–19.
- 2165
- 2166
- 2167 Rosan, B., & Lamont, R. J. (2000). Dental plaque formation. *Microbes and infection*, 2(13), 1599–1607.
- 2168 Rubio, C. A., Lang-Schwarz, C., & Vieth, M. (2022). Further study on field cancerization in the human colon. *Anticancer Research*, 42(12), 5891–5895.
- 2169
- 2170 Schwabe, R. F., & Jobin, C. (2013). The microbiome and cancer. *Nature Reviews Cancer*, 13(11), 800–812.
- 2171
- 2172 Scott, N., Whittle, E., Jeraldo, P., & Chia, N. (2022). A systemic review of the role of enterotoxic bacteroides fragilis in colorectal cancer. *Neoplasia*, 29, 100797.
- 2173
- 2174 Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome biology*, 12, 1–18.
- 2175
- 2176 Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In *Emerging technology in modelling and graphics: Proceedings of iem graph 2018* (pp. 99–111).
- 2177
- 2178
- 2179 Sepich-Poore, G. D., Zitvogel, L., Straussman, R., Hasty, J., Wargo, J. A., & Knight, R. (2021). The microbiome and human cancer. *Science*, 371(6536), eabc4552.
- 2180
- 2181 Sharma, S., & Tripathi, P. (2019). Gut microbiome and type 2 diabetes: where we are and where to go? *The Journal of nutritional biochemistry*, 63, 101–108.
- 2182
- 2183 Shen, X., Zhang, B., Hu, X., Li, J., Wu, M., Yan, C., ... Li, Y. (2022). Neisseria sicca and corynebacterium matruchotii inhibited oral squamous cell carcinomas by regulating genome stability. *Bioengineered*, 13(6), 14094–14106.
- 2184
- 2185
- 2186 Shi, N., Li, N., Duan, X., & Niu, H. (2017). Interaction between the gut microbiome and mucosal immune system. *Military Medical Research*, 4, 1–7.
- 2187
- 2188 Simpson, E. (1949). Measurement of diversity. *Nature*, 163.

- 2189 Sokal, R. R., & Sneath, P. H. (1963). Principles of numerical taxonomy.
- 2190 Song, M., Chan, A. T., & Sun, J. (2020). Influence of the gut microbiome, diet, and environment on risk
2191 of colorectal cancer. *Gastroenterology*, 158(2), 322–340.
- 2192 Söreide, K., Janssen, E., Söiland, H., Körner, H., & Baak, J. (2006). Microsatellite instability in colorectal
2193 cancer. *Journal of British Surgery*, 93(4), 395–406.
- 2194 Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on
2195 similarity of species content and its application to analyses of the vegetation on danish commons.
2196 *Biologiske skrifter*, 5, 1–34.
- 2197 Sotiriadis, A., Papatheodorou, S., Kavvadias, A., & Makrydimas, G. (2010). Transvaginal cervical
2198 length measurement for prediction of preterm birth in women with threatened preterm labor: a
2199 meta-analysis. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International
2200 Society of Ultrasound in Obstetrics and Gynecology*, 35(1), 54–64.
- 2201 Spss, I., et al. (2011). Ibm spss statistics for windows, version 20.0. *New York: IBM Corp*, 440, 394.
- 2202 Stafford, G., Roy, S., Honma, K., & Sharma, A. (2012). Sialic acid, periodontal pathogens and tannerella
2203 forsythia: stick around and enjoy the feast! *Molecular Oral Microbiology*, 27(1), 11–22.
- 2204 Stout, M. J., Conlon, B., Landeau, M., Lee, I., Bower, C., Zhao, Q., ... Mysorekar, I. U. (2013).
2205 Identification of intracellular bacteria in the basal plate of the human placenta in term and preterm
2206 gestations. *American journal of obstetrics and gynecology*, 208(3), 226–e1.
- 2207 Strong, W. (2002). Assessing species abundance unevenness within and between plant communities.
2208 *Community Ecology*, 3(2), 237–246.
- 2209 Sultan, S., El-Mowafy, M., Elgaml, A., Ahmed, T. A., Hassan, H., & Mottawea, W. (2021). Metabolic
2210 influences of gut microbiota dysbiosis on inflammatory bowel disease. *Frontiers in physiology*, 12,
2211 715506.
- 2212 Suzuki, N., Nakano, Y., Yoneda, M., Hirofumi, T., & Hanioka, T. (2022). The effects of cigarette
2213 smoking on the salivary and tongue microbiome. *Clinical and Experimental Dental Research*, 8(1),
2214 449–456.
- 2215 Swidsinski, A., Khilkin, M., Kerjaschki, D., Schreiber, S., Ortner, M., Weber, J., & Lochs, H. (1998).
2216 Association between intraepithelial escherichia coli and colorectal cancer. *Gastroenterology*,
2217 115(2), 281–286.
- 2218 Swift, D., Cresswell, K., Johnson, R., Stilianoudakis, S., & Wei, X. (2023). A review of normalization
2219 and differential abundance methods for microbiome counts data. *Wiley Interdisciplinary Reviews:
2220 Computational Statistics*, 15(1), e1586.
- 2221 Taher, C., Frisk, G., Fuentes, S., Religa, P., Costa, H., Assinger, A., ... others (2014). High prevalence of
2222 human cytomegalovirus in brain metastases of patients with primary breast and colorectal cancers.
2223 *Translational oncology*, 7(6), 732–740.
- 2224 Tanner, A. C., Kent Jr, R., Kanasi, E., Lu, S. C., Paster, B. J., Sonis, S. T., ... Van Dyke, T. E. (2007).
2225 Clinical characteristics and microbiota of progressing slight chronic periodontitis in adults. *Journal
2226 of clinical periodontology*, 34(11), 917–930.
- 2227 Tanner, A. C., Paster, B. J., Lu, S. C., Kanasi, E., Kent Jr, R., Van Dyke, T., & Sonis, S. T. (2006).

- 2228 Subgingival and tongue microbiota during early periodontitis. *Journal of dental research*, 85(4),
2229 318–323.
- 2230 Tejeda, M., Farrell, J., Zhu, C., Haines, J. L., Wang, L.-S., Schellenberg, G. D., ... others (2021). Multiple
2231 viruses detected in human dna are associated with alzheimer disease risk. *Alzheimer's & Dementia*,
2232 17, e054585.
- 2233 Teles, F., Wang, Y., Hajishengallis, G., Hasturk, H., & Marchesan, J. T. (2021). Impact of systemic
2234 factors in shaping the periodontal microbiome. *Periodontology 2000*, 85(1), 126–160.
- 2235 Thaiss, C. A., Zmora, N., Levy, M., & Elinav, E. (2016). The microbiome and innate immunity. *Nature*,
2236 535(7610), 65–74.
- 2237 Tian, R., Liu, H., Feng, S., Wang, H., Wang, Y., Wang, Y., ... Zhang, S. (2021). Gut microbiota dysbiosis
2238 in stable coronary artery disease combined with type 2 diabetes mellitus influences cardiovascular
2239 prognosis. *Nutrition, Metabolism and Cardiovascular Diseases*, 31(5), 1454–1466.
- 2240 Tilg, H., Kaser, A., et al. (2011). Gut microbiome, obesity, and metabolic dysfunction. *The Journal of
2241 clinical investigation*, 121(6), 2126–2132.
- 2242 Tizabi, D., & Hill, R. T. (2023). Micrococcus spp. as a promising source for drug discovery: A review.
2243 *Journal of Industrial Microbiology and Biotechnology*, 50(1), kuad017.
- 2244 Tonetti, M. S., Greenwell, H., & Kornman, K. S. (2018). Staging and grading of periodontitis: Framework
2245 and proposal of a new classification and case definition. *Journal of periodontology*, 89, S159–S172.
- 2246 Tringe, S. G., & Hugenholtz, P. (2008). A renaissance for the pioneering 16s rRNA gene. *Current opinion
2247 in microbiology*, 11(5), 442–446.
- 2248 Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., ... others (2017). A
2249 guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological
2250 Reviews*, 92(2), 698–715.
- 2251 Ulger Toprak, N., Yagci, A., Gulluoglu, B., Akin, M., Demirkalem, P., Celenk, T., & Soyletir, G. (2006).
2252 A possible role of bacteroides fragilis enterotoxin in the aetiology of colorectal cancer. *Clinical
2253 microbiology and infection*, 12(8), 782–786.
- 2254 Ursell, L. K., Metcalf, J. L., Parfrey, L. W., & Knight, R. (2012). Defining the human microbiome.
2255 *Nutrition reviews*, 70(suppl_1), S38–S44.
- 2256 Utzschneider, K. M., Kratz, M., Damman, C. J., & Hullarg, M. (2016). Mechanisms linking the gut
2257 microbiome and glucose metabolism. *The Journal of Clinical Endocrinology & Metabolism*,
2258 101(4), 1445–1454.
- 2259 Vander Haar, E. L., So, J., Gyamfi-Bannerman, C., & Han, Y. W. (2018). Fusobacterium nucleatum and
2260 adverse pregnancy outcomes: epidemiological and mechanistic evidence. *Anaerobe*, 50, 55–59.
- 2261 Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning
2262 research*, 9(11).
- 2263 Vasen, H. F., Mecklin, J.-P., Khan, P. M., & Lynch, H. T. (1991). The international collaborative group
2264 on hereditary non-polyposis colorectal cancer (icg-hnpcc). *Diseases of the Colon & Rectum*, 34(5),
2265 424–425.
- 2266 Vigneswaran, J., & Shogan, B. D. (2020). The role of the intestinal microbiome on colorectal cancer

- 2267 pathogenesis and its recurrence following surgery. *Journal of Gastrointestinal Surgery*, 24(10),
2268 2349–2356.
- 2269 Vilar, E., & Gruber, S. B. (2010). Microsatellite instability in colorectal cancer—the stable evidence.
2270 *Nature reviews Clinical oncology*, 7(3), 153–162.
- 2271 Viljoen, K. S., Dakshinamurthy, A., Goldberg, P., & Blackburn, J. M. (2015). Quantitative profiling of
2272 colorectal cancer-associated bacteria reveals associations between *fusobacterium* spp., enterotoxi-
2273 genic *bacteroides fragilis* (etbf) and clinicopathological features of colorectal cancer. *PLoS one*,
2274 10(3), e0119462.
- 2275 Walker, M. A., Pedamallu, C. S., Ojesina, A. I., Bullman, S., Sharpe, T., Whelan, C. W., & Meyerson, M.
2276 (2018). Gatk pathseq: a customizable computational tool for the discovery and identification of
2277 microbial sequences in libraries from eukaryotic hosts. *Bioinformatics*, 34(24), 4287–4289.
- 2278 Wang, N., & Fang, J.-Y. (2023). *Fusobacterium nucleatum*, a key pathogenic factor and microbial
2279 biomarker for colorectal cancer. *Trends in Microbiology*, 31(2), 159–172.
- 2280 Weaver, W. (1963). *The mathematical theory of communication*. University of Illinois Press.
- 2281 Whiteside, S. A., Razvi, H., Dave, S., Reid, G., & Burton, J. P. (2015). The microbiome of the urinary
2282 tract—a role beyond infection. *Nature Reviews Urology*, 12(2), 81–90.
- 2283 Witkin, S. (2019). Vaginal microbiome studies in pregnancy must also analyse host factors. *BJOG: An
2284 International Journal of Obstetrics & Gynaecology*, 126(3), 359–359.
- 2285 Wong, T.-T., & Yeh, P.-Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE
2286 Transactions on Knowledge and Data Engineering*, 32(8), 1586–1594.
- 2287 Wyss, C., Moter, A., Choi, B.-K., Dewhirst, F., Xue, Y., Schüpbach, P., ... Guggenheim, B. (2004).
2288 *Treponema putidum* sp. nov., a medium-sized proteolytic spirochaete isolated from lesions of
2289 human periodontitis and acute necrotizing ulcerative gingivitis. *International journal of systematic
2290 and evolutionary microbiology*, 54(4), 1117–1122.
- 2291 Xia, Y. (2023). Statistical normalization methods in microbiome data with application to microbiome
2292 cancer research. *Gut Microbes*, 15(2), 2244139.
- 2293 Yaman, E., & Subasi, A. (2019). Comparison of bagging and boosting ensemble machine learning methods
2294 for automated emg signal classification. *BioMed research international*, 2019(1), 9152506.
- 2295 Yang, I., Claussen, H., Arthur, R. A., Hertzberg, V. S., Geurs, N., Corwin, E. J., & Dunlop, A. L. (2022).
2296 Subgingival microbiome in pregnancy and a potential relationship to early term birth. *Frontiers in
2297 cellular and infection microbiology*, 12, 873683.
- 2298 Yildiz, B., Bilbao, J. I., & Sproul, A. B. (2017). A review and analysis of regression and machine learning
2299 models on commercial building electricity load forecasting. *Renewable and Sustainable Energy
2300 Reviews*, 73, 1104–1122.
- 2301 Yoshimura, F., Murakami, Y., Nishikawa, K., Hasegawa, Y., & Kawaminami, S. (2009). Surface
2302 components of *porphyromonas gingivalis*. *Journal of periodontal research*, 44(1), 1–12.
- 2303 Yu, Y., Guo, D., Qu, T., Zhao, S., Xu, C., Wang, L., ... Zhou, N. (2020). Increased risk of toxoplasma
2304 gondii infection in patients with colorectal cancer in eastern china: seroprevalence, risk factors,
2305 and a case–control study. *BioMed Research International*, 2020(1), 2539482.

- 2306 Yuan, K., Xu, H., Li, S., Coker, O. O., Liu, W., Wang, L., ... Yu, J. (2025). Intraneoplastic fungal
2307 dysbiosis is associated with colorectal cancer progression and host gene mutation. *EBioMedicine*,
2308 113.
- 2309 Zavareh, F. S. E., Hadiipour, M., Kalantari, R., Mousavi, S., Tavakolifard, N., & Darani, H. Y. (2021).
2310 Effect of toxoplasma gondii on colon cancer growth in mouse model. *Am J Biomed*, 9(2), 168–176.
- 2311 Zepeda-Rivera, M., Minot, S. S., Bouzek, H., Wu, H., Blanco-Míguez, A., Manghi, P., ... others (2024).
2312 A distinct fusobacterium nucleatum clade dominates the colorectal cancer niche. *Nature*, 628(8007),
2313 424–432.
- 2314 Zhang, C.-Z., Cheng, X.-Q., Li, J.-Y., Zhang, P., Yi, P., Xu, X., & Zhou, X.-D. (2016). Saliva in the
2315 diagnosis of diseases. *International journal of oral science*, 8(3), 133–137.
- 2316 Zhang, M., Zhang, Y., Sun, Y., Wang, S., Liang, H., & Han, Y. (2022). Intratumoral microbiota impacts
2317 the first-line treatment efficacy and survival in non-small cell lung cancer patients free of lung
2318 infection. *Journal of Healthcare Engineering*, 2022(1), 5466853.
- 2319 Zhang, N., Liu, Y., Tang, F.-Y., Yang, L.-Y., & Wang, J.-H. (2023). Structural characterization and in vitro
2320 anti-colon cancer activity of a homogeneous polysaccharide from agaricus bisporus. *International
2321 Journal of Biological Macromolecules*, 251, 126410.
- 2322 Zhang, X., Yu, D., Wu, D., Gao, X., Shao, F., Zhao, M., ... others (2023). Tissue-resident lachnospiraceae
2323 family bacteria protect against colorectal carcinogenesis by promoting tumor immune surveillance.
2324 *Cell host & microbe*, 31(3), 418–432.
- 2325 Zhou, X., Wang, L., Xiao, J., Sun, J., Yu, L., Zhang, H., ... others (2022). Alcohol consumption,
2326 dna methylation and colorectal cancer risk: Results from pooled cohort studies and mendelian
2327 randomization analysis. *International journal of cancer*, 151(1), 83–94.
- 2328 Zhu, W., & Lee, S.-W. (2016). Surface interactions between two of the main periodontal pathogens:
2329 *Porphyromonas gingivalis* and *tannerella forsythia*. *Journal of periodontal & implant science*,
2330 46(1), 2–9.
- 2331 Zhu, X., Han, Y., Du, J., Liu, R., Jin, K., & Yi, W. (2017). Microbiota-gut-brain axis and the central
2332 nervous system. *Oncotarget*, 8(32), 53829.
- 2333 Zhuang, Y., Wang, H., Jiang, D., Li, Y., Feng, L., Tian, C., ... others (2021). Multi gene mutation
2334 signatures in colorectal cancer patients: predict for the diagnosis, pathological classification, staging
2335 and prognosis. *BMC cancer*, 21, 1–16.

Acknowledgments

2337 I would like to disclose my earnest appreciation for my advisor, Professor **Semin Lee**, who provided
 2338 solicitous supervision and cherished opportunities throughout the course of my research. His advice and
 2339 consultation encouraged me to become as a researcher and to receive all humility and gentleness. I am also
 2340 grateful to all of my committee members, Professor **Taejoon Kwon**, Professor **Eunhee Kim**, Professor
 2341 **Kyemyung Park**, and Professor **Min Hyuk Lim**, for their meaningful mentions and suggestions.

2342 I extend my deepest gratitude to my Lord, **the Flying Spaghetti Monster**, His Noodly Appendage
 2343 has guided me through the twist and turns of this academic journey. His presence, ever comforting and
 2344 mysterious, has been a source of strength and humor during both highs and lows. In moments of doubt, I
 2345 found solace in the belief that you were there, gently reminding me to keep faith in the process. His Holy
 2346 Noodle has nourished my mind, and for that, I am truly overwhelmed. May His Holy Noodle continue to
 2347 guide me in all my future endeavors. *R’Amen.*

2348 I would like to extend my heartfelt gratitude to Professor **You Mi Hong** for her invaluable guidance
 2349 and insightful advice on PTB study. Her expertise in maternal and fetal health, along with her deep under-
 2350 standing of statistical and clinical interpretations, greatly contributed to refining the analytical framework
 2351 of this study. Her constructive feedback and thoughtful discussions provided critical perspectives that
 2352 enhanced the robustness and relevance of the research findings. I sincerely appreciate her generosity
 2353 in sharing her knowledge and effort, as well as her encouragement throughout my Ph.D. journey. Her
 2354 support has been instrumental in strengthening this work, and I am truly grateful for her contributions.

2355 I also would like to express my sincere gratitude for Professor **Jun Hyeok Lim** for his invaluable
 2356 guidance and insightful advice on lung cancer study. His expertise in cancer genomics and data interpreta-
 2357 tion provided essential perspectives that greatly enriched the analytical approach of my Ph.D. journey. His
 2358 constructive feedback and thoughtful discussion helped refine methodologies and enhance the scientific
 2359 rigor of the research. I deeply appreciate his willingness to share his knowledge and expertise, which has
 2360 been instrumental in shaping key aspects of this work. His support and encouragement have been truly
 2361 inspiring, and I am grateful for the opportunity to have benefited from his mentorship.

2362 I would like to extend my heartfelt gratitude to my colleagues of the **Computational Biology Lab @**
 2363 **UNIST**, whose collaboration, friendship, brotherhood, and support have been an invaluable part of my
 2364 journey. Your willingness to share insights, engage in thoughtful discussions, and offer encouragement
 2365 during the challenging moments of research has significantly shaped my academic experience. The
 2366 camaraderie in Computational Biology Lab made even the most demanding days more enjoyable, and I
 2367 am deeply grateful for the collaborative environment we created together. I appreciate you for standing
 2368 by my side throughout this Ph.D. journey.

2369 I would like to express my heartfelt gratitude to **my family**, whose unwavering support has been the
 2370 foundation of everything I have achieved. Your love, encouragement, and belief in me have sustained me
 2371 through every challenge, and I could not have come this far without you. From your words of wisdom to
 2372 your patience and understanding, each of you has played a vital role in helping me navigate this journey.
 2373 The strength and comfort I have drawn from our family bond have been my greatest source of resilience.

2374 Your presence, both near and far, has filled my life with warmth and motivation. I am deeply grateful for
2375 your unconditional love and for always being there when I needed you the most. Thank you for being my
2376 constant source of strength and inspiration.

2377 I am incredibly pleased to my friends, especially my GSHS alumni (이망특), for their unwavering
2378 support and encouragement throughout this journey. The bonds we formed back in our school days have
2379 only grown stronger over the years, and I am fortunate to have had such loyal and understanding friends
2380 by my side. Your constant words of motivation, and even moments of levity during stressful times have
2381 helped keep me grounded. Whether it was a late-night conversations, a shared laugh, or a simple message
2382 of reassurance, you all have played a vital role in keeping me focused and motivated. I am relieved for the
2383 ways you celebrated each small achievement with me and how you patiently listened to my worries. The
2384 memories of our shared past provided me with comfort and a sense of stability when the road ahead felt
2385 uncertain. I could not have reached this point without the love and friendship that you all have generously
2386 given. Each of your, in your unique way, has contributed to this dissertation, even if indirectly, and for
2387 that, I am forever beholden. I look forward to continuing our friendship as we all grow in our individual
2388 paths, knowing that the support we share is something truly special.

2389 I would like to express my deepest recognition to **my girlfriend (expected)** for her unwavering
2390 support, patience, and companionship throughout my Ph.D. journey. Her presence has been a constant
2391 source of comfort and motivation, helping me navigate the challenges of research and writing with
2392 renewed energy. Through moments of frustration and accomplishment alike, her encouragement has
2393 reminded me of the importance of balance and perseverance. Her kindness, understanding, and belief
2394 in me have been invaluable, making even the most difficult days feel lighter. I am truly grateful for her
2395 support and for sharing this journey with me, and I look forward to all the moments we will continue to
2396 experience together.

2397 I would like to express my sincere gratitude to the amazing members of my animal protection groups,
2398 DRDR (두루두루) and UNIMALS (유니멀스), whose dedication and compassion have been a constant
2399 source of motivation. Your unwavering commitment to improving the lives of animals has inspired me
2400 throughout this journey. I am also thankful for the beautiful cats we have cared for, whose presence
2401 brought both joy and purpose to our allegiance. Their playful spirits and gentle companionship served as
2402 daily reminders of why we continue to fight for animal rights. The bond we share, both with each other
2403 and with the animals we protect, has enriched my life in countless ways. I appreciate you all again for
2404 your support, dedication, and for being part of this meaningful cause.

2405 I would like to express my deepest gratitude to **everyone** I have had the honor of meeting throughout
2406 this journey. Your kindness, encouragement, and support have carried me through both the challenging
2407 and rewarding moments of my life. Whether through a kind word, thoughtful advice, or simply being
2408 there when I needed it most, your presence has made all the difference. I am incredibly fortunate to have
2409 received such generosity and warmth from those around me, and I do not take it for granted. Every act
2410 of kindness, no matter how big or small, has been a source of strength and motivation for me. To all
2411 my friends, colleagues, mentors, and beloved ones, thank you for your unwavering support. I am truly
2412 grateful for each of you, and your kindness has left an indelible mark on my journey.

2413 My Lord, *the Flying Spaghetti Monster*,
2414 give us grace to accept with serenity the things that cannot be changed,
2415 courage to change the things that should be changed,
2416 and the wisdom to distinguish the one from the other.

2417
2418 Glory be to *the Meatball*, to *the Sauce*, and to *the Holy Noodle*.
2419 As it was in the beginning, is now, and ever shall be.

2420 *R'Amen.*



May your progress be evident to all

