Doctoral Thesis

Metagenomic Profiling in Preterm Birth, Periodontitis, and Colorectal Cancer

Jaewoong Lee

Department of Biomedical Engineering

Ulsan National Institute of Science and Technology

2025

Metagenomic Profiling in Preterm Birth, Periodontitis, and Colorectal Cancer

Jaewoong Lee

Department of Biomedical Engineering

Ulsan National Institute of Science and Technology

Metagenomic Profiling in Preterm Birth, Periodontitis, and Colorectal Cancer

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

Metagenomic Profiling in Preterm Birth, Periodontitis, and Colorectal Cancer

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

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

Section 2 investigated the association between the prenatal salivary microbiome and preterm birth (PTB) using 16S ribosomal RNA (rRNA) gene sequencing and developed a random forest-based prediction model for risk of preterm birth. A total of 59 pregnant women were included as the study participants, with 30 in the PTB group and 29 in the full-term birth (FTB) group. Salivary microbiome samples were collected via mouthwash within 24 hours before delivery, and 16S rRNA gene sequencing was performed to analyze microbial taxonomic composition. Differentially abundant taxa (DAT) were identified by DESeq2, revealing the 25 significant taxa, including three PTB-enriched DAT and 22 FTB-enriched DAT, suggesting distinct microbial differences upon PTB. A random forest classifier was applied to predict PTB risk based on salivary microbiome composition, achieving the high balanced accuracy (0.765 ± 0.071, mean$\pm$SD) using the nine most important taxa. These findings indicate that salivary microbiome profiling may serve as a novel predictive tool for PTB risk assessment, complementing existing clinical predictors.

Section 3 characterized salivary microbiome compositions to classify peridontal health and different stages of periodontitis using 16S rRNA gene sequencing. A total of 250 study participants were included, comprising 100 periodontally healthy controls and 150 periodontitis patients equally classified into stage I, stage II, and stage III. Microbial diversity indices were calculated, and ANCOM was used to identify 20 differentially abundant taxa among the multiple periodontitis stages. A random forest machine learning model was developed to classify periodontitis stages based on the proportions of differentially abundant taxa, achieving an area-under-curve of 0.870 ± 0.079 (mean$\pm$SD). Among the identified differentially abundant taxa, *Porphyromonas gingivalis* and *Actinomyces* spp. were the most important features in distinguishing periodontitis stages. Random forest classifier also effectively distinguished healthy individuals from stage I periodontitis with an area-under-curve of 0.852 ± 0.103 (mean$\pm$SD) and detected periodontitis patients from healthy controls with an area-under-curve of 0.953 ± 0.049 (mean$\pm$SD). External validation with Spanish and Portuguese datasets showed a slight performance decrease, likely due

to ethnic variations in salivary microbiome composition, emphasizing the need for population-specific models. Finally, functional pathway enrichment analysis based on DAT was performed to derive potential microbial metabolic activities associated with periodontitis stages. These findings suggest that salivary microbiome composition profiling may serve as a non-invasive diagnostic technique for periodontitis, aiding in early detection and personalized dental care.

Section 4 conducted a comprehensive metagenomic analysis of colorectal cancer using PathSeq, focusing on key clinical outcomes, including recurrence history and overall survival duration. Significant differences in alpha-diversity and beta-diversity indices were observed between tumor and its adjacent normal tissues, with further stratification revealing distinct microbial diversity patterns associated with recurrence status and survival outcomes. Differentially abundant taxa were identified, highlighting microbial signatures may influence CRC progression and prognosis. To evaluate the predictive potential of these selected differentially abundant taxa, we developed a random forest-based machine learning model for CRC recurrence risk and survival duration. While the classification model for recurrence prediction achieved moderate balanced accuracy (0.570 ± 0.164, mean$\pm$SD), and the regression model of survival duration showed moderated mean-absolute errors (729.302 ± 179.940, mean$\pm$SD), these results suggest that gut microbiome composition alone may not be sufficient for personalized clinical predictions. These findings emphasize the need for multi-omics integration, combining host genomic alterations, *e.g.* somatic and germline mutations, with gut microbiome compositions, to improve CRC risk stratification and personalized medicine applications. This study highlights the potential role of gut microbiome for biomarkers in CRC diagnosis and prognosis while underscoring the complexity of host-microbiome interaction in CRC progression.

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

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

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

Contents

List of Figures

List of Tables

List of Abbreviations

ACC Accuracy

ACE Abundance-based coverage estimator

ASV Amplicon sequence variant

AUC Area-under-curve

BA Balanced accuracy

BMI Body mass index

C-section Cesarean section

CAL Clinical attachment level

DAT Differentially abundant taxa

F1 F1 score

Faith PD Faith's phylogenetic diversity

FC Fold change

FN False negative

FP False positive

FTB Full-term birth

GA Gestational age

MAE Mean absolute error

MSI Microsatellite instability

MSI-H MSI-High

MSI-L MSI-Low

MSS Microsatellite stable

MWU test Mann-Whitney U-test

OS Overall survival

PD Probing depth

PRE Precision

PROM Prelabor rupture of membrane

PTB Preterm birth

qPCR quantitative-PCR

RMSE Root mean squared error

ROC curve Receiver-operating characteristics curve

rRNA Ribosomal RNA

SD Standard deviation

SEN Sensitivity

SPE Specificity

t-SNE t-distributed stochastic neighbor embedding

TN True negative

TP True positive

1 Introduction

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

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

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

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

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

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

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 affecting 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.* DESeq2 (Love, Huber, & Anders, 2014) and ANCOM (Lin & Peddada, 2020), 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.

While identifying DAT provides important taxonomic insights into microbial shifts associated with disease, understanding the functional consequences of these changes is also crucial. Enriched pathway analysis based on DAT enables researchers to derive the metabolic and biological processes potentially altered by microbiome dysbiosis, offering a deeper understanding of how specific microbial communities may contribute to host physiology, immune modulation, and disease progression. By mapping DAT to known functional pathway using tools such as PICRUSt2 (Douglas et al., 2020; Ye & Doak, 2009; Louca & Doebeli, 2018), it is possible to predict metabolic capacities, biosynthetic functions, and inflammatory responses associated with microbial alterations. These functional insights complement taxonomic findings by bridging the gap between microbial presence and physiological relevance, thus providing a more comprehensive framework for exploring microbiome-disease associations. Incorporating pathway enrichment analysis into microbiome studies enhances the potential for discovering clinically actionable biomarkers and developing targeted microbiome-based therapeutic strategies.

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

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

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

The receiver-operating characteristics (ROC) curve is a graphical representation used to evaluate the performance of a classification model by plotting the sensitivity against $(1-\text{specificity})$ at multiple threshold setting (Gonçalves, Subtil, Oliveira, & de Zea Bermudez, 2014; Obuchowski & Bullen, 2018;

Centor, 1991). The ROC curve illustrates the trade-off between detecting true positives while minimizing false positives, suggesting determining the optimal decision threshold for classification. A key metric derived from the ROC curve is the area-under-curve (AUC), which quantifies overall ability of the classification model to discriminate between positive and negative predictions. An AUC value of 0.5 indicates a model performing no better than random chance, while value closer to 1.0 suggests high predictive accuracy. Thus, by analyzing the AUC value of the ROC curve, researchers can compare different models and select the better classification model that offers the best balance between sensitivity and specificity for a given application.

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

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

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

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}} \tag{1}$$

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

$$\mathrm{F1} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{3}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{4}$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}} \tag{5}$$

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{6}$$

$$\mathrm{MAE} = \sum_{i=1}^{n} |\mathrm{Prediction}_i - \mathrm{Real}_i| / n \tag{7}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} (\mathrm{Prediction}_i - \mathrm{Real}_i)^2 / n} \tag{8}$$

2 Predicting preterm birth using random forest classifier in salivary microbiome

This section includes the published contents:
Hong, Y. M., **Lee, Jaewoong**, Cho, D. H., Jeon, J. H., Kang, J., Kim, M. G., ... & Kim, J. K. (2023). Predicting preterm birth using machine learning techniques in oral microbiome. *Scientific Reports*, 13(1), 21105.

2.1 Introduction

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

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

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

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

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

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

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

2.2 Materials and methods

2.2.1 Study design and study participants

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

2.2.2 Clinical data collection and grouping

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

- maternal age at delivery
- diabetes mellitus
- hypertension
- overweight and obesity
- C-section
- history PROM or PTB
- gestational week on delivery
- birth weight
- sex

2.2.3 Salivary microbiome sample collection

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

2.2.4 16s rRNA gene sequencing

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

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

2.2.5 Bioinformatics analysis

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

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

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

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

2.2.6 Data and code availability

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

2.3 Results

2.3.1 Overview of clinical information

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

2.3.2 Comparison of salivary microbiomes composition

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

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

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

2.3.3 Random forest classification to predict PTB risk

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

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

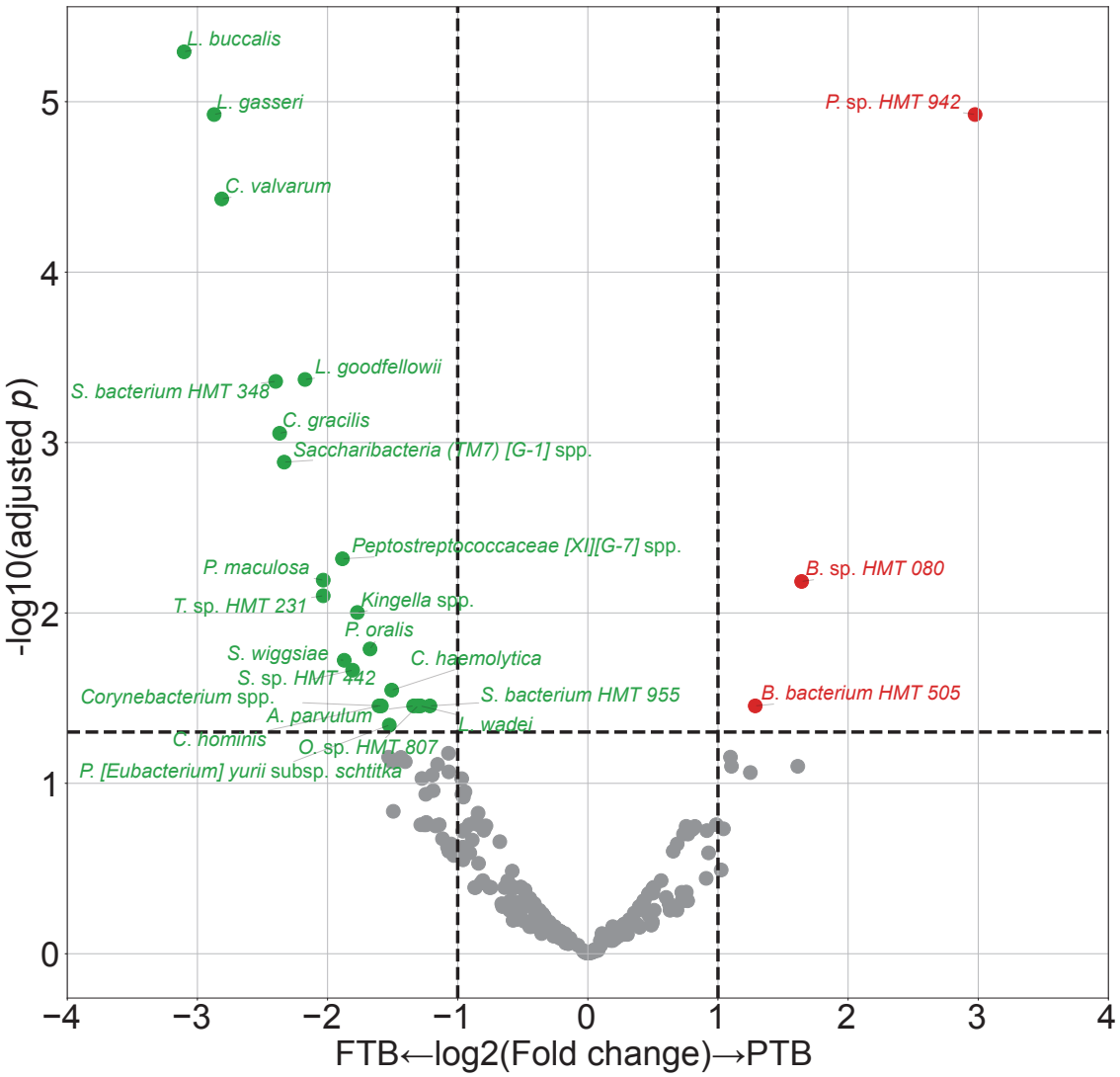

Figure 1: **DAT volcano plot for PTB prediction.**

Statistical threshold is: adjusted p-value < 0.05 and $|\log_2 \text{Fold Change}| > 1.0$. Red dots represent PTB-enriched DAT, while green dots represent FTB-enriched DAT.

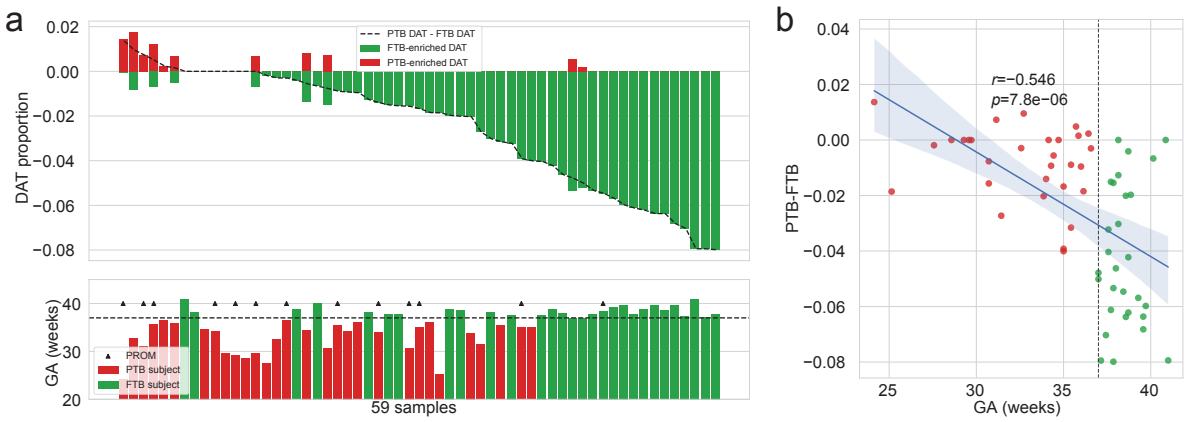

Figure 2: **Salivary microbiome compositions over DAT for PTB prediction.**
(a) Frequencies of DAT of PTB study subjects. The study participants are arranged in respect of (PTB-enriched DAT − FTB-enriched DAT). The study participants' GA is displayed in accordance with the upper panel's order (PTB: red 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 ($p = 7.8e-6$).

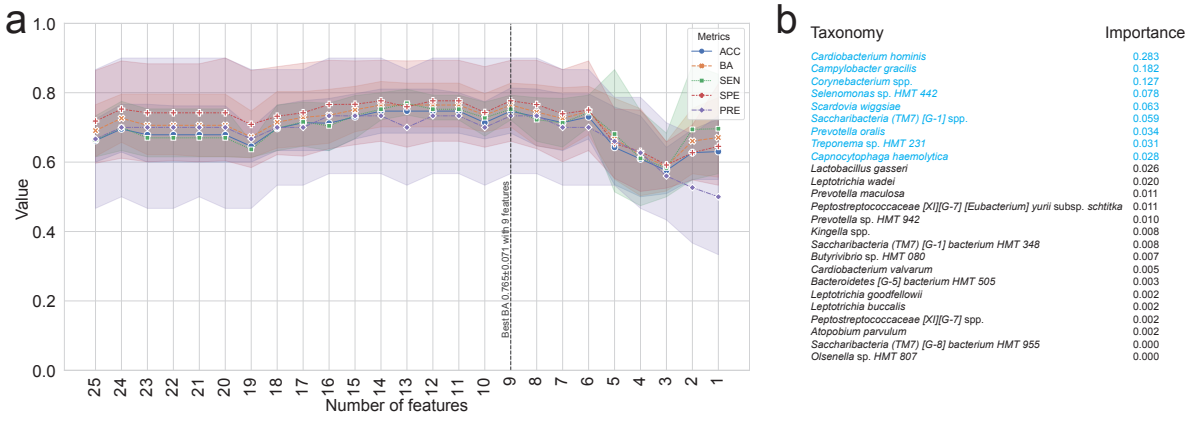

Figure 3: **Random forest-based PTB prediction model**.

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

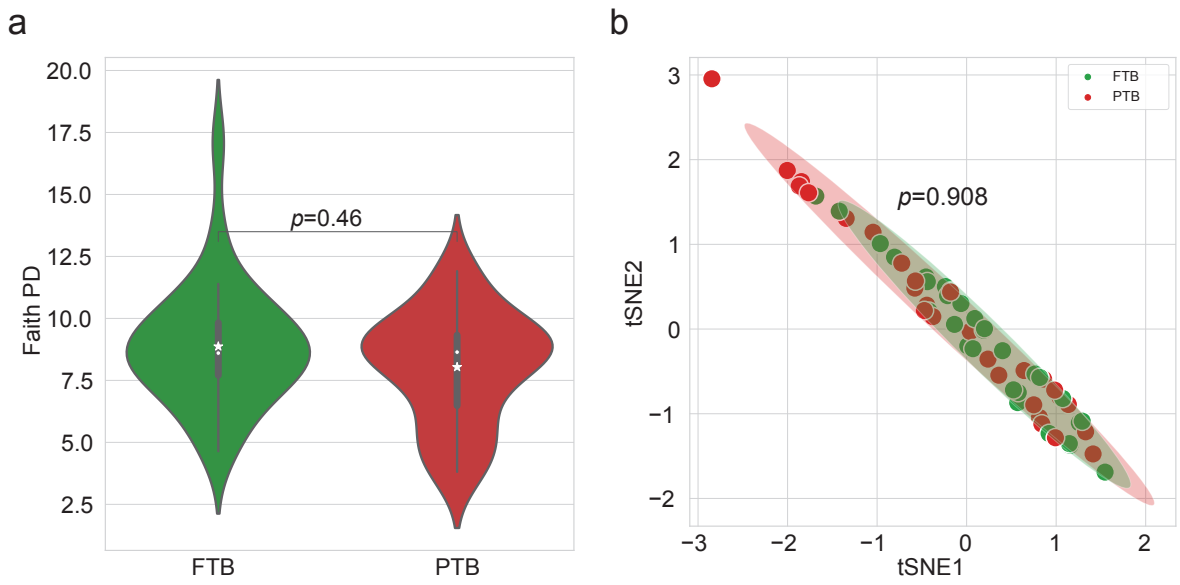

Figure 4: **Diversity indices about PTB study participants.**

(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 $p = 0.908$)

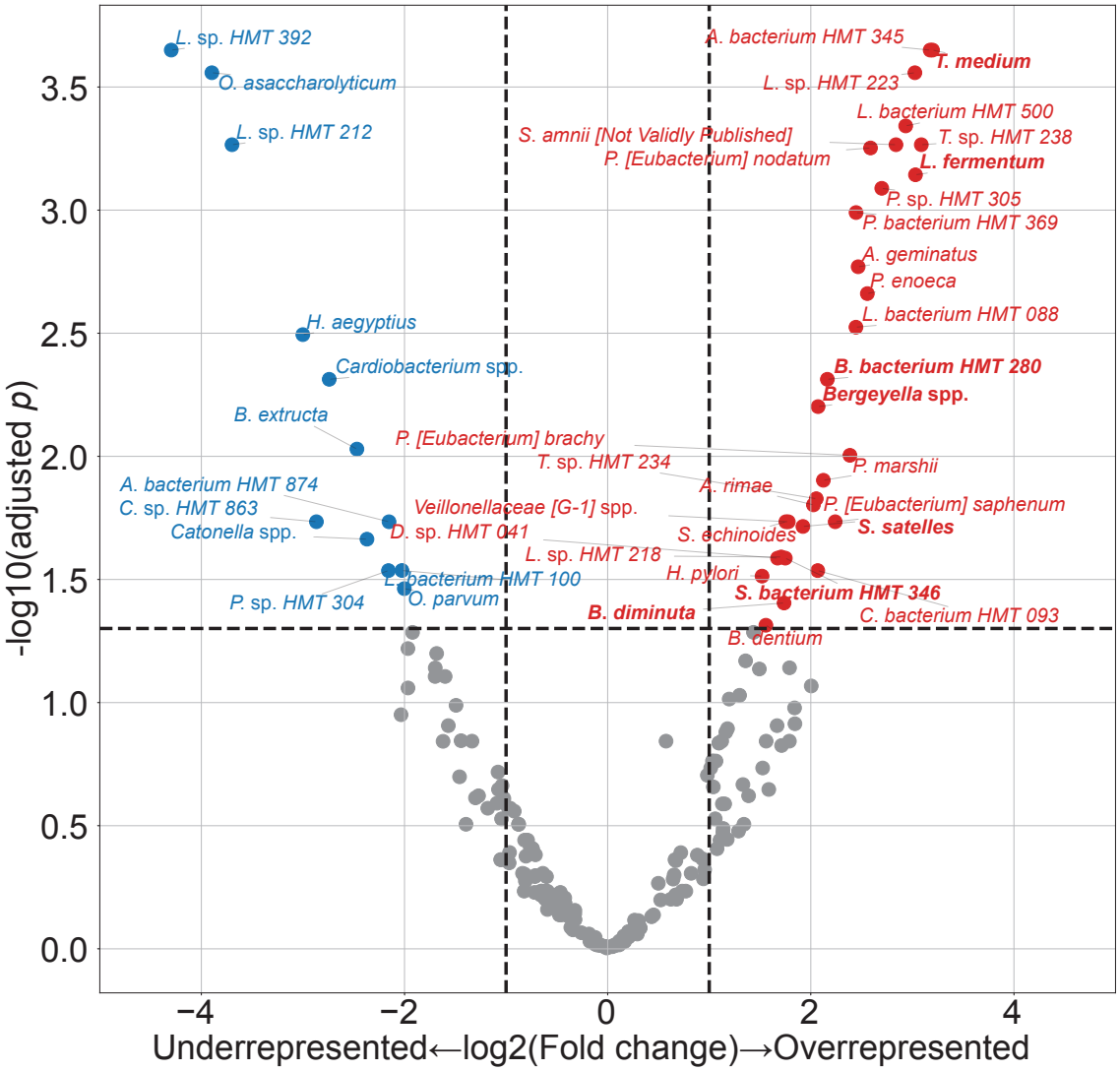

Figure 5: **PROM-related DAT between FTB and PTB.**

Statistical threshold is: adjusted p-value < 0.05 and $|\log_2 \text{Fold Change}| > 1.0$. 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.

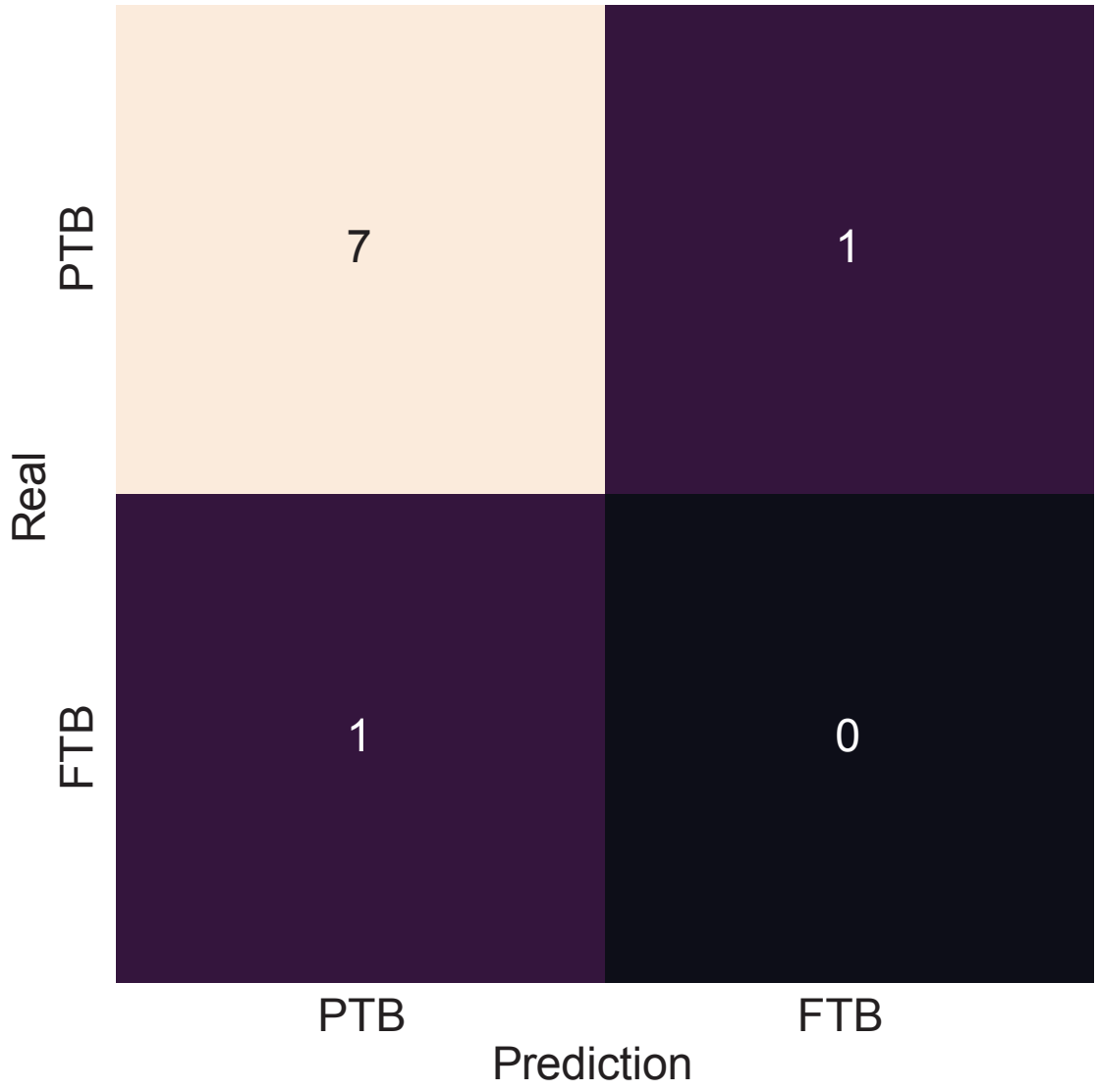

Figure 6: **Validation of random forest-based PTB prediction model**.
Nine twin pregnancies (eight PTB subjects and a FTB subject) that were excluded in the initial study subjects were subjected to a validation procedure. The random forest-based PTB prediction model shows 87.5% accuracy, comparable to the PTB classification evaluations on the singleton study subjects (0.714 ± 0.061. Mean$\pm$SD)

2.4 Discussion

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

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

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

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

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

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

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

3 Random forest prediction model for periodontitis stages based on the salivary microbiomes

3.1 Introduction

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

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

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

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

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

3.2 Materials and methods

3.2.1 Study participants enrollment

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

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

3.2.2 Periodontal clinical parameter diagnosis

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

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

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

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

0. Normal gingiva: without inflammation nor discoloration.

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

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

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

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

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

- Healthy:
    1. Bleeding sites $< 10\%$
    2. Probing depth: ≤ 3 mm
- Stage I:
    1. No tooth loss because of periodontitis.
    2. Inter-dental clinical attachment level at the site of the greatest loss: 1-2 mm
    3. Radiographic bone loss: $< 15\%$
- Stage II:
    1. No tooth loss because of periodontitis.
    2. Inter-dental clinical attachment level at the site of the greatest loss: 3-4 mm
    3. Radiographic bone loss: 15-33%
- Stage III:
    1. Teeth loss because of periodontitis: ≤ 3 teeth
    2. Inter-dental clinical attachment level at the site of the greatest loss: ≥ 5 mm
    3. Radiographic bone loss: $> 33\%$

3.2.3 Saliva sampling and DNA extraction procedure

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

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

- Forward: `5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3'`
- Reverse: `5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'`

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

1. Heat activation for 30 seconds at 95 °C.
2. 25 cycles for 30 seconds at 95 °C.
3. 30 seconds at 55 °C.
4. 30 seconds at 72 °C.

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

3.2.4 Bioinformatics analysis

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

- Abundance-based Coverage Estimator (ACE) (Chao & Lee, 1992)
- Chao1 (Chao, 1984)
- Fisher (Fisher, Corbet, & Williams, 1943)
- Margalef (Magurran, 2021)
- Observed ASVs (DeSantis et al., 2006)
- Berger-Parker d (Berger & Parker, 1970)
- Gini (Gini, 1912)

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

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

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

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

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

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

To derive the functional potential of the microbial communities, we performed pathway enrichment analysis based on DAT using PICRUSt2 (Douglas et al., 2020; Ye & Doak, 2009; Louca & Doebeli, 2018) according to the MetaCyc pathway database (Karp, Riley, Paley, & Pellegrini-Toole, 2002; Caspi et al., 2016, 2018). PICRUSt2 estimates the functional profile of microbial communities by predicting gene family abundances from 16S rRNA gene sequencing data using phylogenetic placement and ancestral-

state reconstruction. DAT identified by statistical methods were used as input, and the corresponding predicted metagenomic pathways in mutiple periodontitis stages. STAMP (version 2.1.3) was used to calculate the differentially abundant pathways between periodontitis stages (Parks, Tyson, Hugenholtz, & Beiko, 2014): 1) Welch's two-sided t-test $p \leq 0.01$ and 2) difference in mean proportion ≥ 0.05 or ratio of proportions ≥ 1.5. This approach enabled us to associate taxnomic changes with potential alterations in microbial functional pathways, offering insights into the biological mechanisms underlying host-microbiome interactions relevant to disease development and progression.

3.2.5 Data and code availability

All sequences from the 250 study participants have been published to the Sequence Read Archives (project ID `PRJNA976179`): `https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA976179`. Docker image that employed throughout this study is available in the DockerHub: `https://hub.docker.com/repository/docker/fumire/periodontitis_16s`. Every code used in this study can be found on GitHub: `https://github.com/CompbioLabUnist/Periodontitis_16S`.

3.3 Results

3.3.1 Summary of clinical information and sequencing data

Among clinical information of the study participants, clinical attachment level, probing depth, plaque index, and gingival index, were significantly increased with periodontitis severity (Kruskal-Wallis test $p < 0.001$), while sex were observed no significant difference (Table 2). Notably, clinical attachment level and probing depth have significant differences among the periodontitis severities (MWU test $p < 0.01$; Figure 16). Additionally, 71461.00$\pm$11792.30 and 45909.78$\pm$11404.65 (mean$\pm$SD) reads per sample were obtained before and after filtering low-quality reads and trimming extra-long tails, respectively (Figure 17). In 250 study subjects, we have found a total of 425 bacterial taxa (Figure 14).

3.3.2 Diversity indices reveal differences among the periodontitis severities

Rarefaction curves showed that the sequencing depth was sufficient (Figure 13). Alpha-diversity indices indicated significant differences between the healthy and the periodontitis stages (MWU test $p < 0.01$; Figure 7a-e); however, there were no significant differences between the periodontitis stages. This emphasizes how essential it is to classify the salivary microbiome compositions and distinguish between the stages of periodontitis using machine learning approaches.

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

3.3.3 DAT among multiple periodontitis severities and their correlation

Of the 425 total taxa that identified in the salivary microbiome composition (Figure 14), 20 DAT were identified (Table 5). Three separate subgroups were formed from the participants-level abundances of the DAT using a hierarchical clustering methodology (Figure 8a):

- Group 1
    1. *Treponema* spp.
    2. *Prevotella* sp. *HMT 304*
    3. *Prevotella* sp. *HMT 526*
    4. *Peptostreptococcaceae [XI][G-5] saphenum*
    5. *Treponema* sp. *HMT 260*
    6. *Mycoplasma faucium*
    7. *Peptostreptococcaceae [XI][G-9] brachy*
    8. *Lachnospiraceae [G-8] bacterium HMT 500*
    9. *Peptostreptococcaceae [XI][G-6] nodatum*
    10. *Fretibacterium* spp.

- Group 2
    1. *Porphyromonas gingivalis*
    2. *Campylobacter showae*
    3. *Filifactor alocis*
    4. *Treponema putidum*
    5. *Tannerella forsythia*
    6. *Prevotella intermedia*
    7. *Porphyromonas* sp. *HMT 285*
- Group 3
    1. *Actinomyces* spp.
    2. *Corynebacterium durum*
    3. *Actinomyces graevenitzii*

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

Correlation analysis from the DAT showed that DAT from Group 3 was negatively correlated with Group 1 and Group 2 (Figure 9), and strong correlations were observed the nine pairs of DAT (Figure 15).

3.3.4 Classification of periodontitis severities by random forest models

To confirm that using selected DAT bacterial profiles could have enhanced sequencing expenses without losing the classification evaluations, we built the random forest classification models based on DAT and full microbiome compositions (Figure 19). DAT based classifier showed non-significant different or better evaluations, by removing confounding taxa.

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

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

Based on the findings that the salivary microbiome composition in stage II is more comparable to those in stage III than to other severities (Figure 7f and Figure 7j), we combined stage II and stage III to perform a multi-label classification.

To examine alternative classification algorithms in comparison to random forest classification, we selected gradient boost algorithm because it is another algorithm of the few classification algorithms that can provide feature importances, which is essential for identifying key taxa contributing to the classification of periodontitis stages. Thus, we assessed gradient boosting algorithms (Figure 21). However, the classification evaluations obtained from gradient boosting have non-significant differences compared to random forest classification.

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

3.3.5 Over-represented fermentation pathways and under-represented glycolysis pathways along with periodontitis progression

We inferred the functions of stage-specific salivary microbiomes using PICRUSt2 to identify differentially represented pathways between healthy/stage I and stage II/III (Figure 12). Four pathways were predicted repeatedly to be significantly over-represented in stage II/III compared to healthy/stage I. These included pathways related to fermentation, including pyruvate fermentation to butanoate, as well as pathways associated with nucleotide biosynthesis. The most significant enriched pathway in stage II compared to healthy status was pyrimidine deoxyribonucleotide de novo biosynthesis III (Figure 12a), while stage III, it was pyruvate fermentation to acetone (Figure 12b). This pattern was also consistently observed in comparison with stage I (Figure 12c and Figure 12d). Interestingly, the glyoxylate cycle and glucose and glucose-1-phosphate degradation pathways were underrepresented in stage III compared to healthy status (Figure 12b). These data suggest that fermentation and nucleotide syntheses might play roles in the metabolic process as periodontitis progresses, similar to the findings of a previous report about aggressive periodontitis (Jorth et al., 2014).

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

Continuous variable: mean±SD. Categorical variable: count (proportion). 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 (mean±SD). *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	*P.gingivalis*	0.758±0.051	0.716±0.177	0.677±0.068	0.839±0.034	0.839±0.034	0.839±0.034	0.516±0.102
	P.gingivalis+*Act.*	0.792±0.043	0.822±0.105	0.723±0.057	0.861±0.029	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	0.889±0.015	0.668±0.033
Healthy vs. Stage I	*Act.*	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
	Act.+*P.gingivalis*	0.733±0.119	0.831±0.081	0.713±0.122	0.797±0.097	0.800±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	*P.gingivalis*	0.776±0.042	0.736±0.196	0.748±0.047	0.832±0.031	0.832±0.031	0.832±0.031	0.664±0.062
	P.gingivalis+*Act.*	0.843±0.035	0.876±0.109	0.823±0.039	0.882±0.026	0.882±0.026	0.882±0.026	0.764±0.052
	Top 6 taxa	0.885±0.036	0.914±0.027	0.871±0.038	0.914±0.027	0.914±0.025	0.914±0.025	0.828±0.051
Healthy vs. Stages I/II/III	*P.gingivalis*	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
	P.gingivalis+*Act.*	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**.
Statistical significance was determined by ANCOM W value.

No.	Taxonomy	ANCOM W score
1	*Porphyromonas gingivalis*	424
2	*Actinomyces* spp.	424
3	*Filifactor alocis*	421
4	*Prevotella intermedia*	419
5	*Treponema putidum*	418
6	*Tannerella forsythia*	415
7	*Porphyromonas* sp. *HMT 285*	412
8	*Peptostreptococcaceae [XI][G-6] nodatum*	412
9	*Fretibacterium* spp.	411
10	*Mycoplasma faucium*	411
11	*Prevotella* sp. *HMT 304*	411
12	*Lachnospiraceae [G-8] bacterium HMT 500*	409
13	*Treponema spp.*	408
14	*Prevotella* sp. *HMT 526*	401
15	*Peptostreptococcaceae [XI][G-9] brachy*	400
16	*Peptostreptococcaceae [XI][G-5] saphenum*	398
17	*Campylobacter showae*	395
18	*Treponema* sp. *HMT 260*	393
19	*Corynebacterium durum*	393
20	*Actinomyces graevenitzii*	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. Note that $\forall i, 0 \geq \text{importance}_i \geq 1$ and $\sum_i \text{importance}_i = 1$.

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
1	*Porphyromonas gingivalis*	0.297	*Actinomyces* spp.	0.360	*Porphyromonas gingivalis*	0.426	*Porphyromonas gingivalis*	0.461
2	*Actinomyces* spp.	0.195	*Porphyromonas gingivalis*	0.125	*Actinomyces* spp.	0.244	*Actinomyces* spp.	0.257
3	*Prevotella intermedia*	0.054	*Actinomyces graevenitzii*	0.095	*Actinomyces graevenitzii*	0.049	*Actinomyces graevenitzii*	0.059
4	*Actinomyces graevenitzii*	0.052	*Porphyromonas* sp. *HMT 285*	0.062	Corynebacterium durum	0.046	Corynebacterium durum	0.035
5	*Filifactor alocis*	0.050	*Lachnospiraceae [G-8] bacterium HMT 500*	0.052	*Filifactor alocis*	0.036	*Filifactor alocis*	0.032
6	*Campylobacter showae*	0.042	*Campylobacter showae*	0.050	*Prevotella intermedia*	0.033	*Campylobacter showae*	0.023
7	*Porphyromonas* sp. *HMT 285*	0.040	*Filifactor alocis*	0.039	*Tannerella forsythia*	0.025	*Porphyromonas* sp. *HMT 285*	0.022
8	Corynebacterium durum	0.032	Corynebacterium durum	0.038	*Campylobacter showae*	0.023	*Prevotella intermedia*	0.022
9	*Treponema* spp.	0.032	*Treponema* spp.	0.037	*Porphyromonas* sp. *HMT 285*	0.021	*Treponema* spp.	0.022
10	*Tannerella forsythia*	0.026	*Tannerella forsythia*	0.029	*Treponema* spp.	0.018	*Peptostreptococcaceae [XI][G-9] brachy*	0.015
11	*Treponema putidum*	0.025	*Prevotella intermedia*	0.026	*Peptostreptococcaceae [XI][G-9] brachy*	0.014	*Lachnospiraceae [G-8] bacterium HMT 500*	0.010
12	*Fretibacterium* spp.	0.023	*Fretibacterium* spp.	0.018	*Lachnospiraceae [G-8] bacterium HMT 500*	0.011	*Tannerella forsythia*	0.009
13	*Peptostreptococcaceae [XI][G-9] brachy*	0.021	*Peptostreptococcaceae [XI][G-9] brachy*	0.018	*Peptostreptococcaceae [XI][G-6] nodatum*	0.010	*Fretibacterium* spp.	0.009
14	*Treponema* sp. *HMT 260*	0.019	*Treponema putidum*	0.014	*Treponema putidum*	0.009	*Treponema putidum*	0.006
15	*Prevotella* sp. *HMT 526*	0.018	*Prevotella* sp. *HMT 526*	0.011	*Prevotella* sp. *HMT 526*	0.008	*Peptostreptococcaceae [XI][G-6] nodatum*	0.004
16	*Peptostreptococcaceae [XI][G-6] nodatum*	0.018	*Treponema* sp. *HMT 260*	0.008	*Fretibacterium* spp.	0.008	*Treponema* sp. *HMT 260*	0.004
17	*Prevotella* sp. *HMT 304*	0.017	*Peptostreptococcaceae [XI][G-6] nodatum*	0.008	*Treponema* sp. *HMT 260*	0.005	*Mycoplasma faucium*	0.004
18	*Mycoplasma faucium*	0.014	*Mycoplasma faucium*	0.004	*Prevotella* sp. *HMT 304*	0.005	*Prevotella* sp. *HMT 526*	0.003
19	*Peptostreptococcaceae [XI][G-5] saphenum*	0.014	*Prevotella* sp. *HMT 304*	0.003	*Mycoplasma faucium*	0.005	*Peptostreptococcaceae [XI][G-5] saphenum*	0.002
20	*Lachnospiraceae [G-8] bacterium HMT 500*	0.013	*Peptostreptococcaceae [XI][G-5] saphenum*	0.003	*Peptostreptococcaceae [XI][G-5] saphenum*	0.004	*Prevotella* sp. *HMT 304*	0.001

Table 7: **Beta-diversity pairwise comparisons on the periodontitis statuses**
Statistically significant (p-value) was determined by the PERMANOVA test.

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

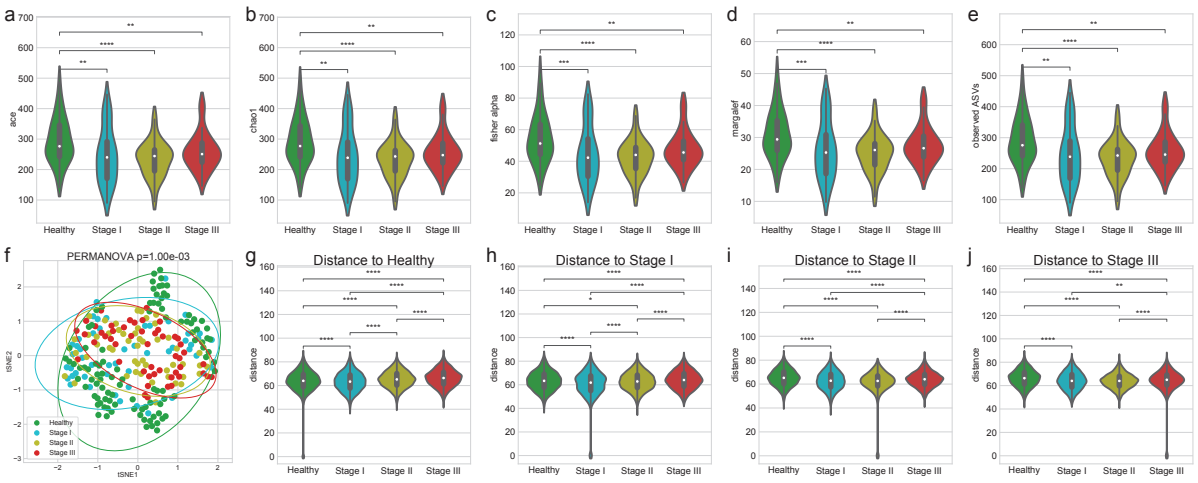

Figure 7: **Diversity indices for periodontitis.**

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

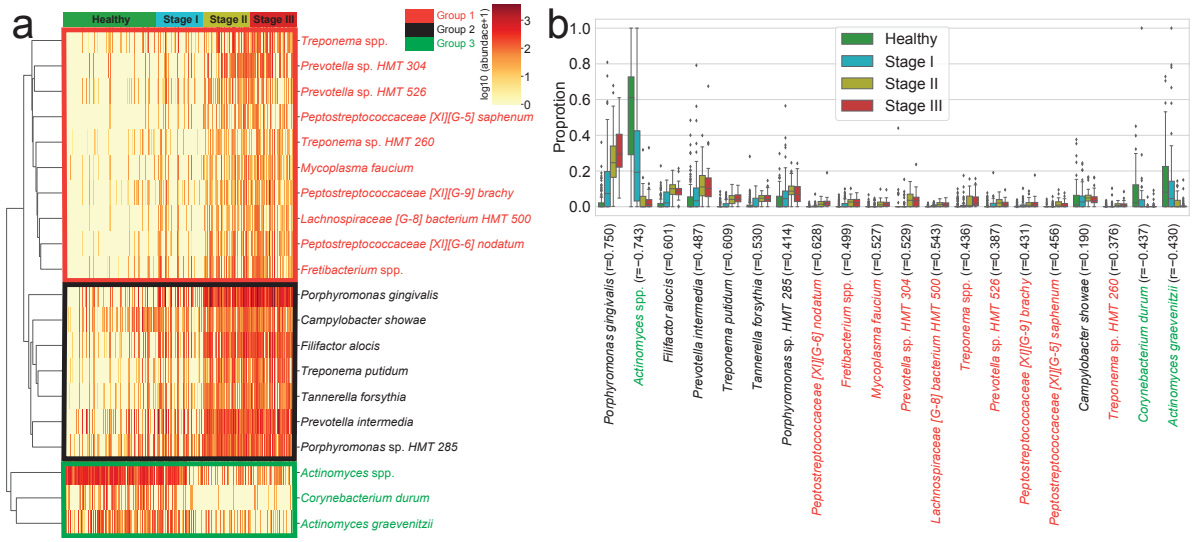

Figure 8: **DAT for periodontitis.**

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.

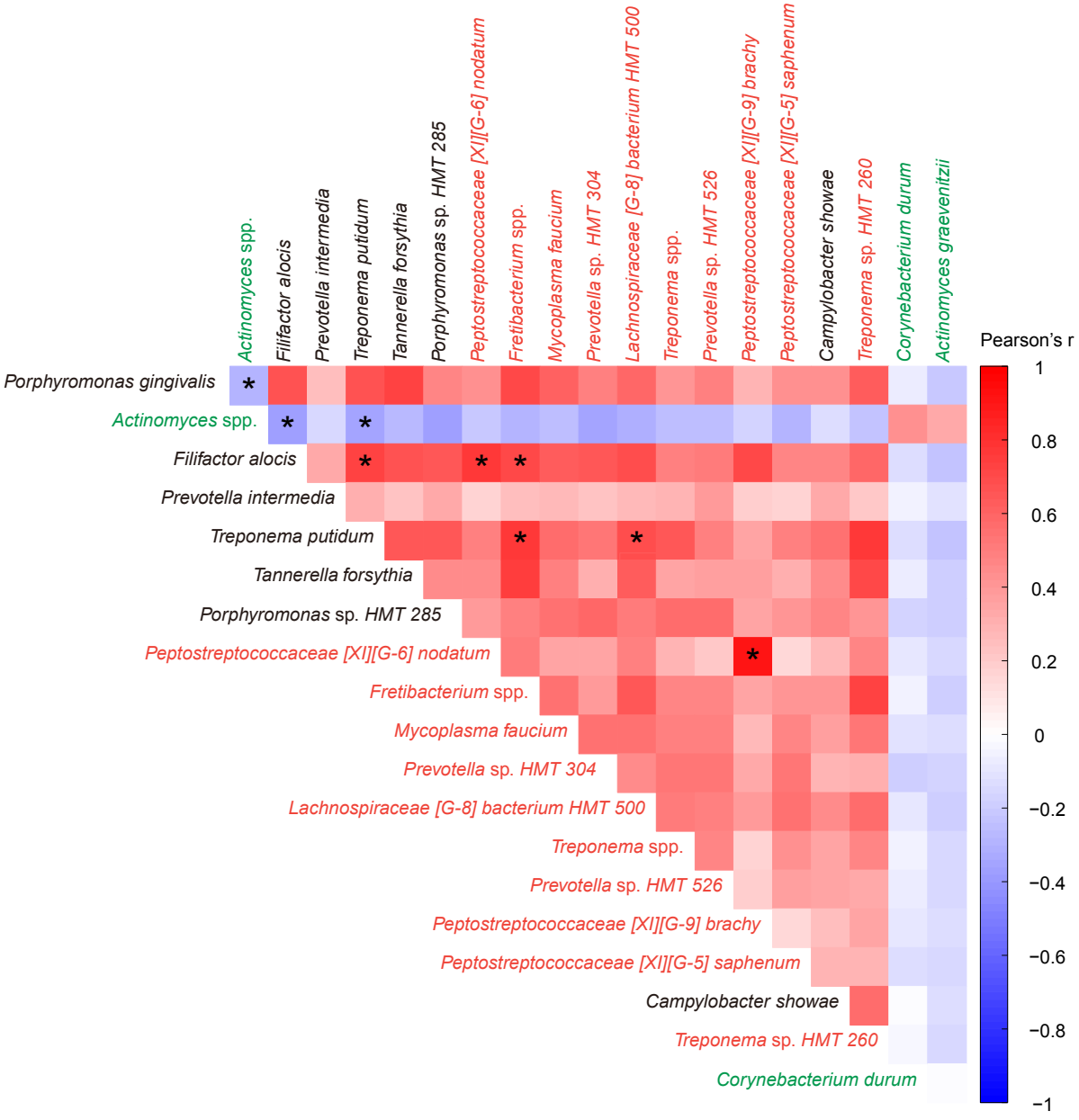

Figure 9: **Correlation heatmap between periodontitis DAT.**

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

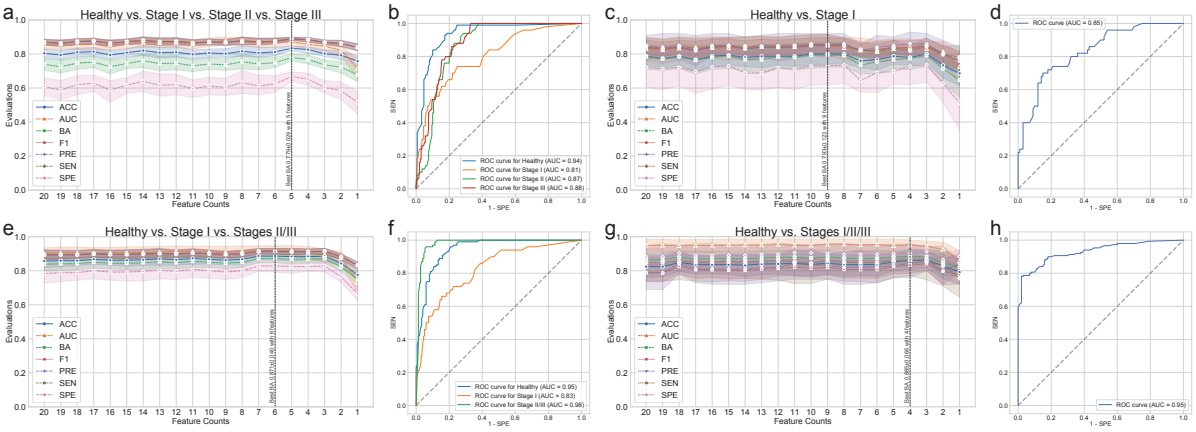

Figure 10: **Random forest classification metrics for periodontitis prediction.**
The classification metrics in the random forest classifications were as follows: ACC, AUC, BA, F1, PRE, SEN, and SPE. **(a)** Classification performance for healthy vs. stage I vs. stage II vs. stage III. **(b)** ROC curve 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).

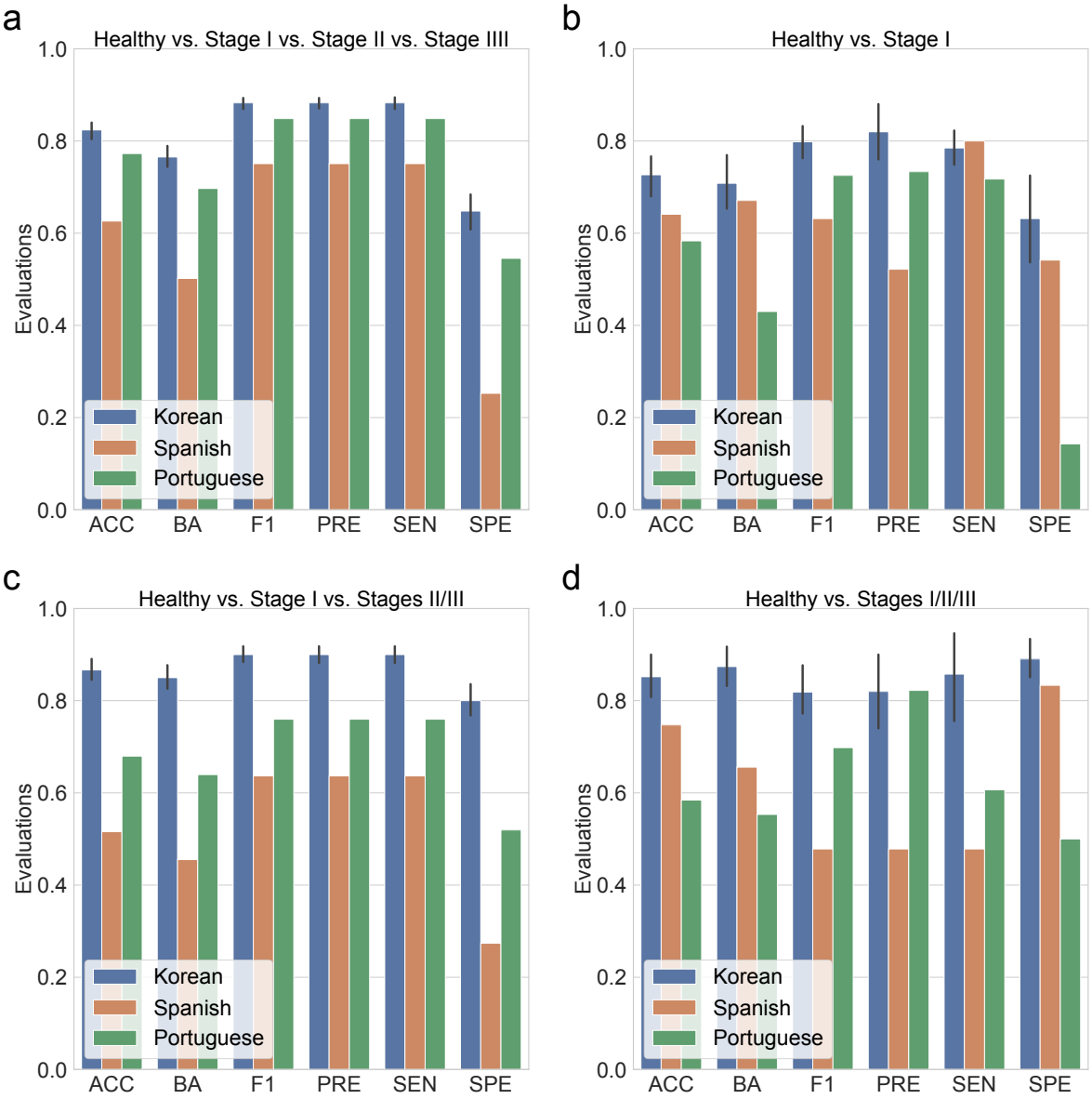

Figure 11: **Random forest classification metrics from external datasets**.

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.

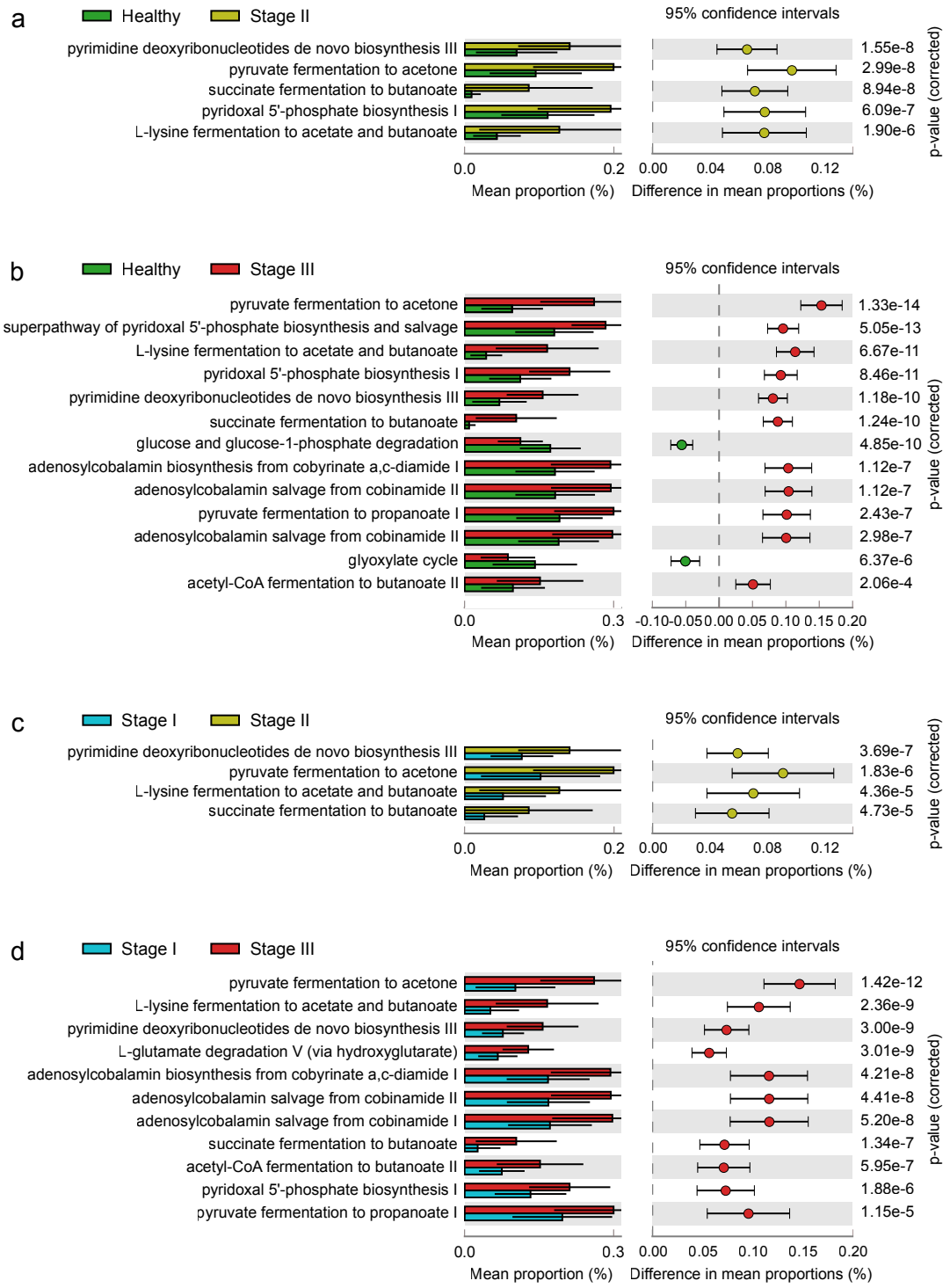

Figure 12: **Functional enrichment test**.

Significantly differentially enriched pathways between two periodontitis stages are shown, shorted by corrected *p*-value. **(a)** healthy vs. stage II. **(b)** healthy vs. stage III. **(c)** stage I vs. stage II. **(d)** stage I vs. stage III.

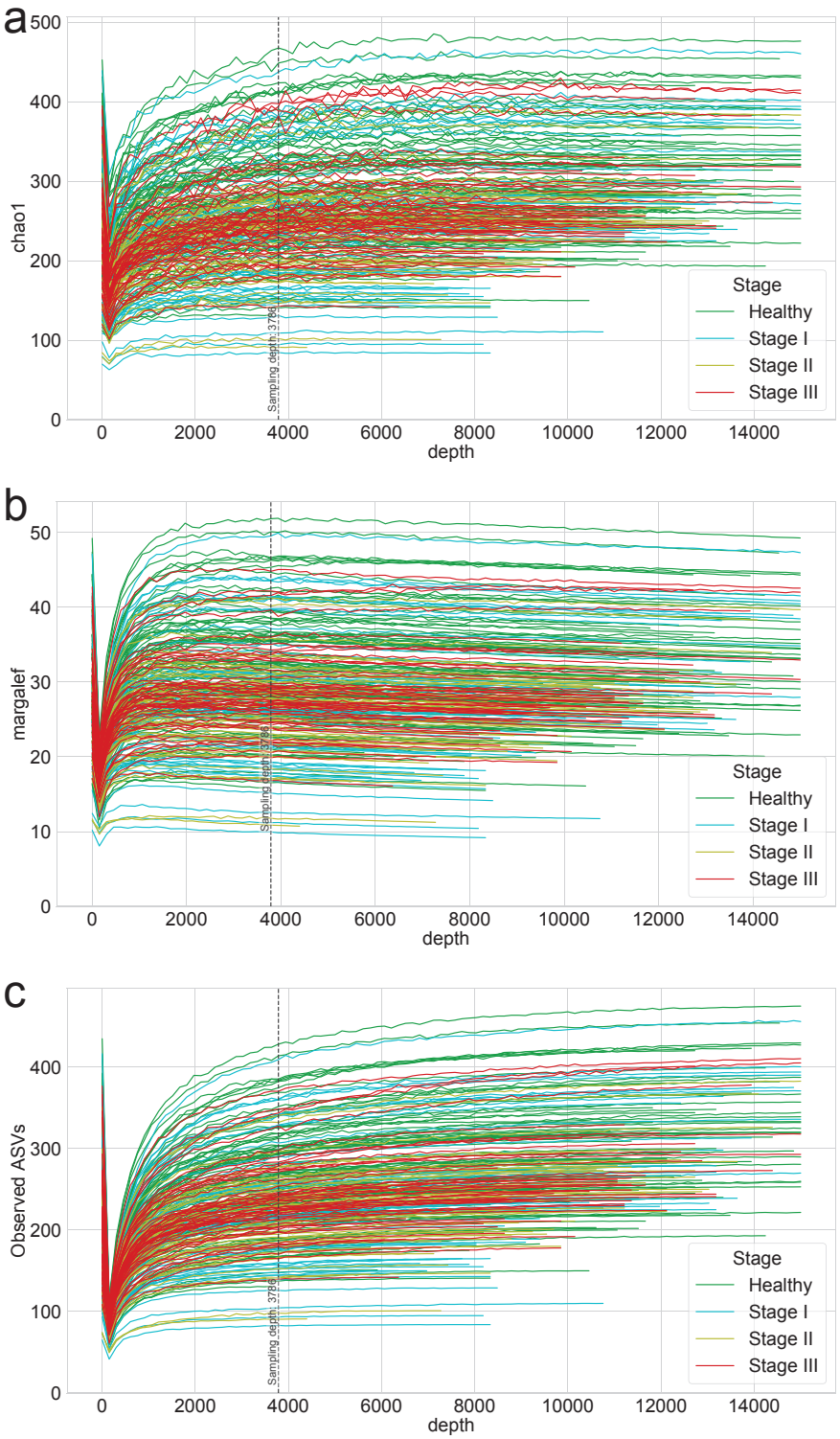

Figure 13: **Rarefaction curves for alpha-diversity indices**.
Rarefaction of **(a)** chao1 **(b)** margalef, and **(c)** observed ASVs were generated to measure species richness and determine the sampling depth of each sample.

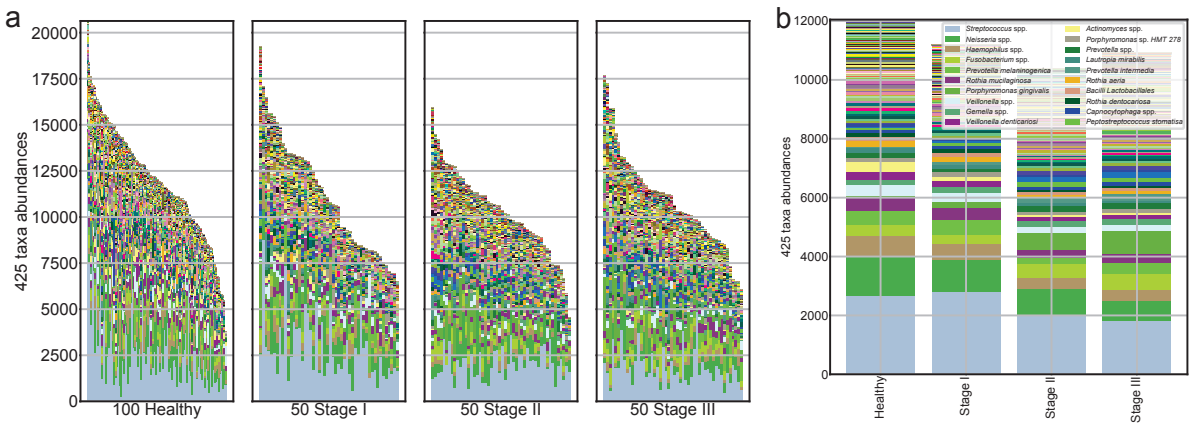

Figure 14: **Salivary microbiome compositions in the different periodontal stages**.
Stacked bar plot of the absolute abundance of bacterial species for all samples **(a)** and the mean absolute abundance of bacterial species in the healthy, stage I, stage II, and stage III groups **(b)**.

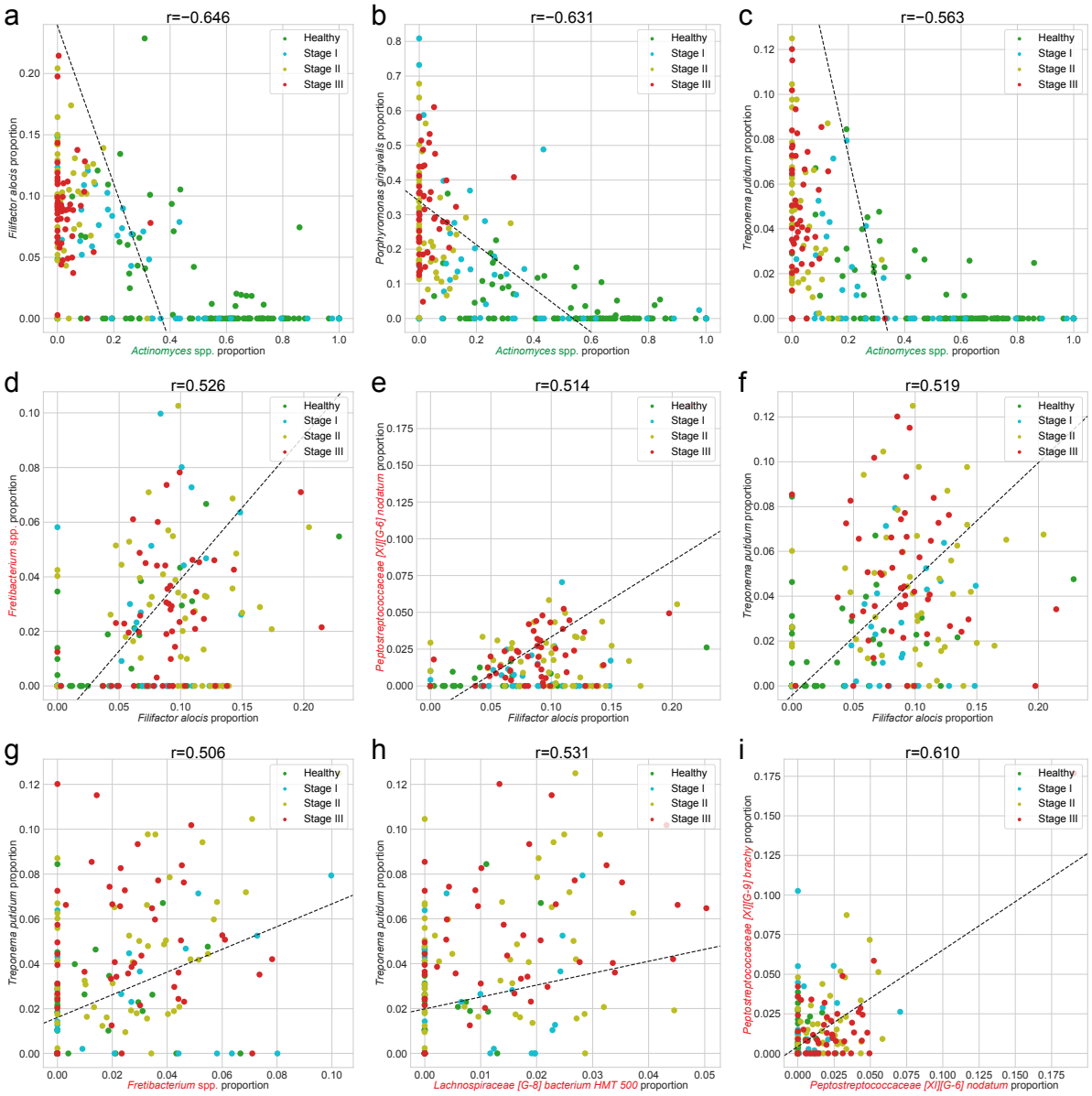

Figure 15: **Correlation plots for periodontitis DAT.**

We selected the combinations of DAT with absolute Spearman correlation coefficients greater than 0.5. The color represents periodontal healthy periodontal stages (green: healthy, cyan: stage I, yellow: stage II, and red: stage III).

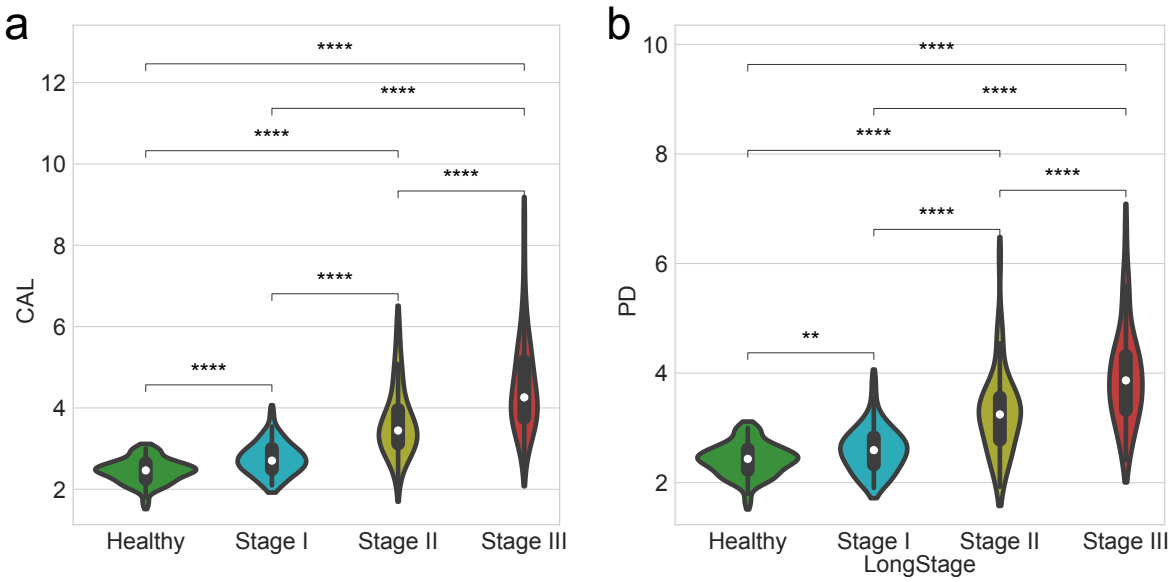

Figure 16: **Clinical measurements by the periodontitis stages**.
Comparisons of clinical measurement among healthy controls and patients with various periodontitis stages. **(a)** Clinical attachment level (CAL) **(b)** Probing depth (PD). Statistical significance determined by the MWU test: $p < 0.01$ (**) and $p < 0.0001$ (****).

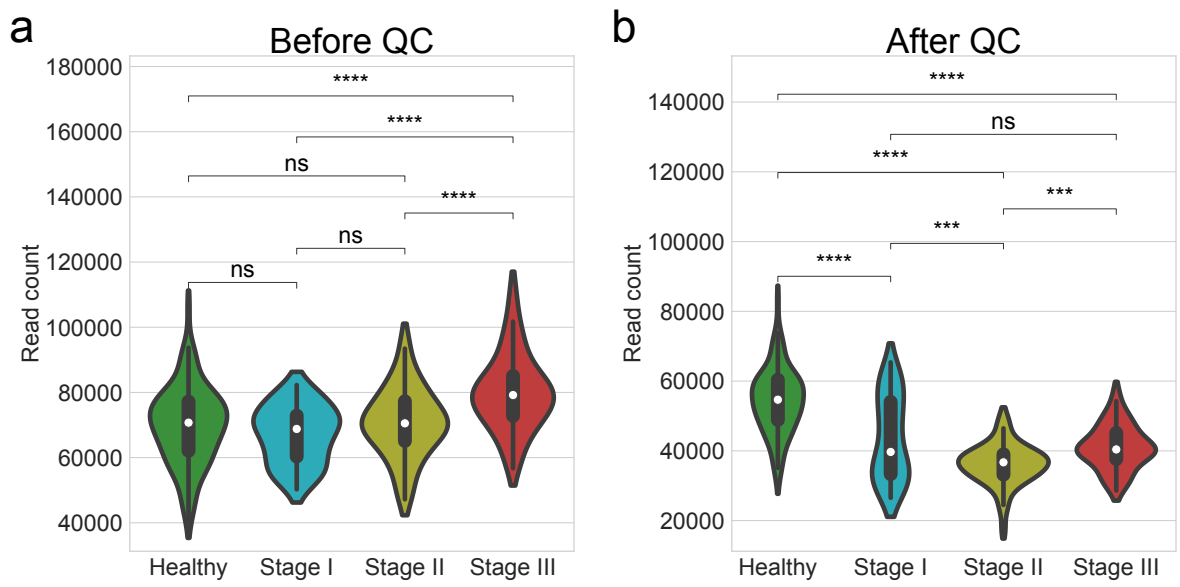

Figure 17: **Number of read counts by the periodontitis stages**.
Comparisons of the number of read counts among healthy controls and patients with various periodontitis stages. **(a)** Before quality check **(b)** After quality check. Statistical significance determined by the MWU test: $p \geq 0.05$ (ns), $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), and $p < 0.0001$ (****).

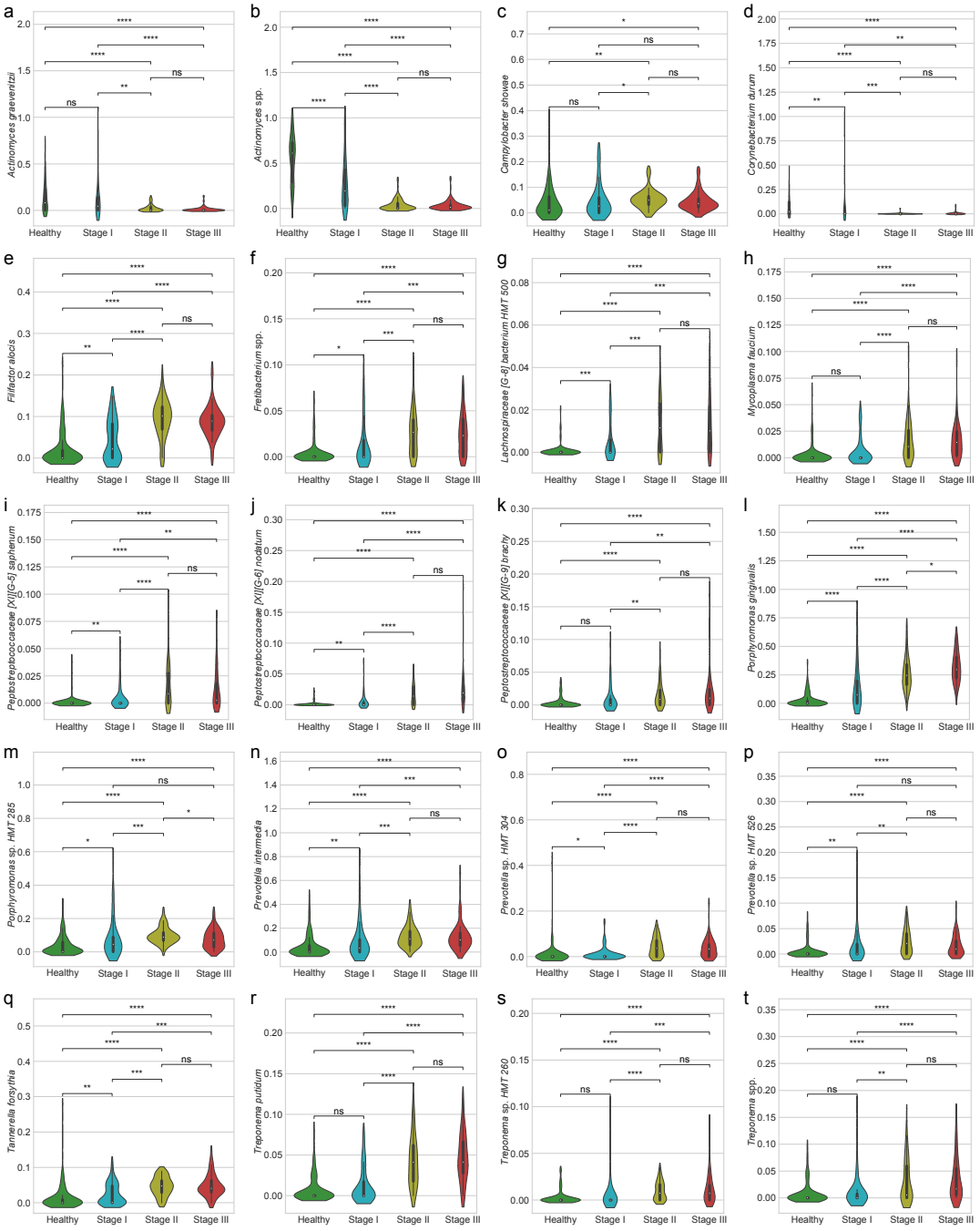

Figure 18: **Proportions of periodontitis DAT**.

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

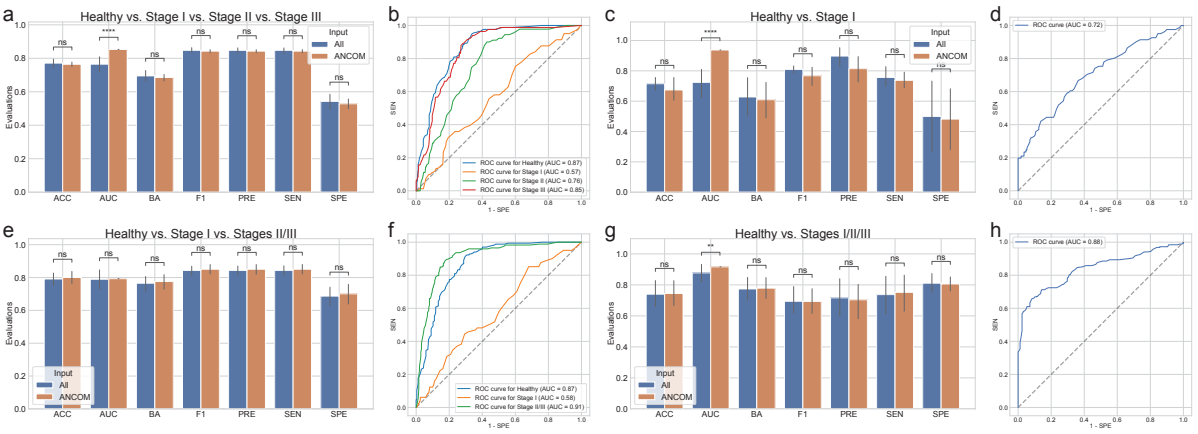

Figure 19: **Random forest classification metrics with the full microbiome compositions and ANCOM-selected DAT compositions**.

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

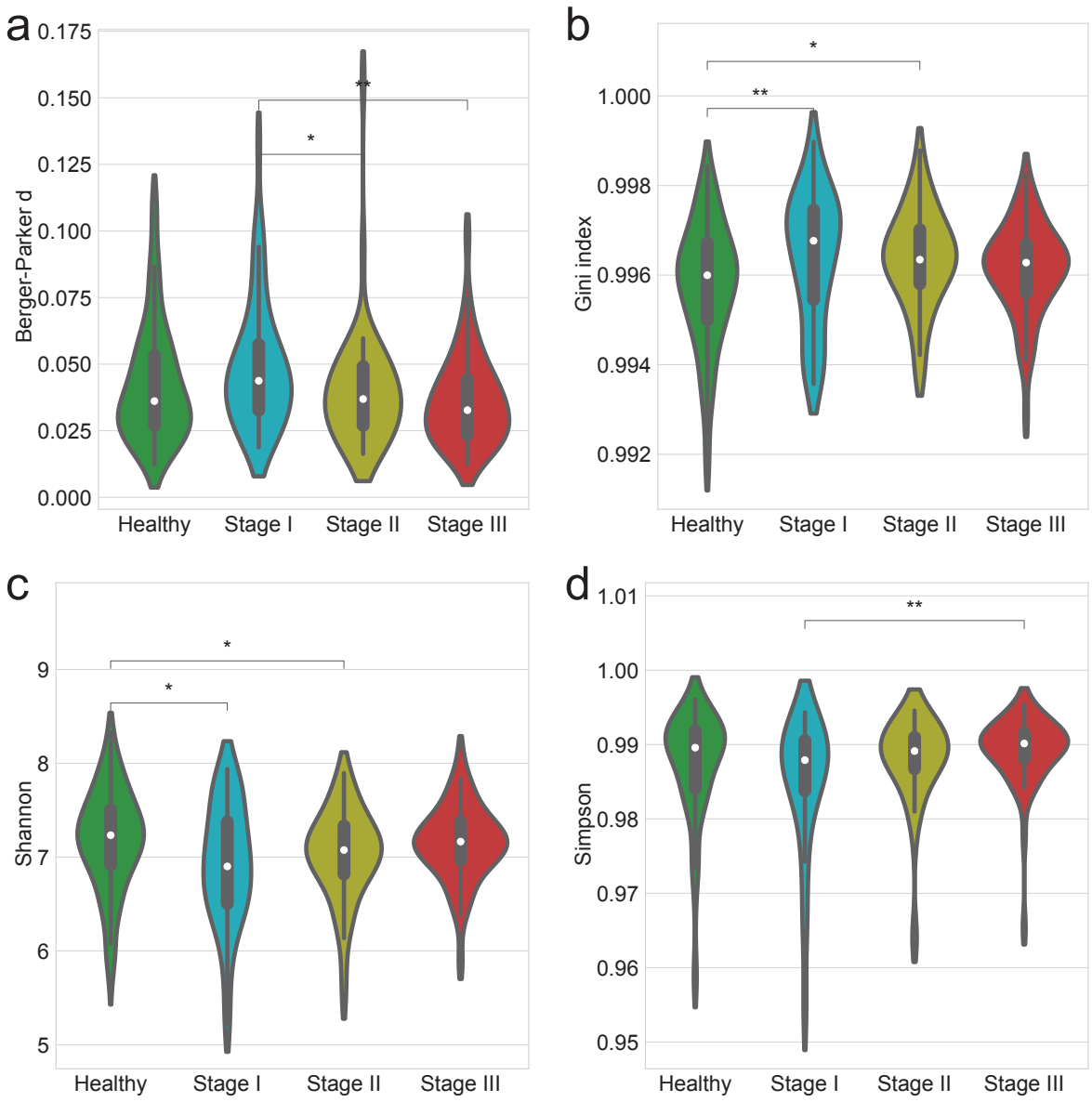

Figure 20: **Alpha-diversity indices account for evenness**.

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 $p < 0.01$ (**)

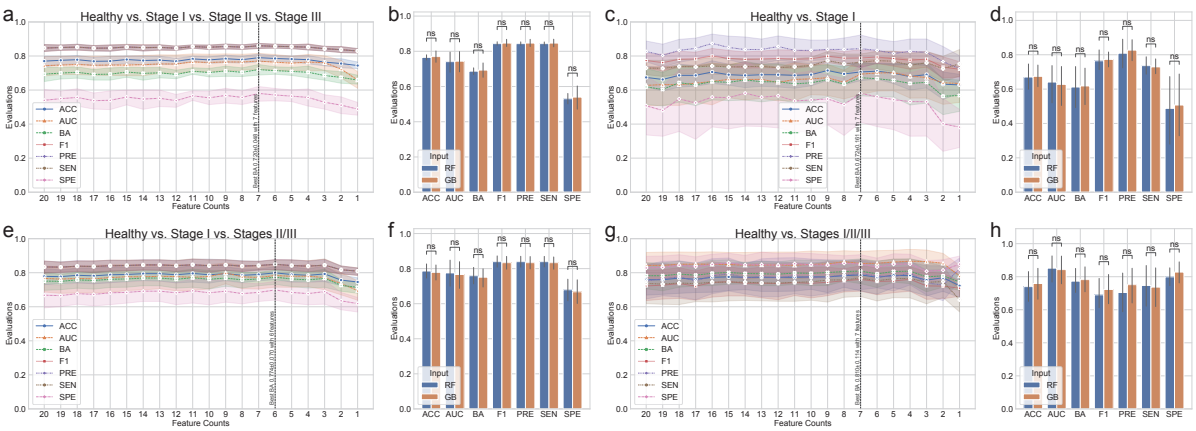

Figure 21: **Gradient Boosting classification metrics for periodontitis prediction.**
The classification metrics in the random forest classifications were as follows: ACC, AUC, BA, F1, PRE, SEN, and SPE. The feature counts mean that the classification model trained on the most important n features as 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)

3.4 Discussion

In order to investigate at potential alterations in the salivary microbiome compositions based on periodontal stages, including healthy, stage I, stage II, and stage III, we employed 16S rRNA gene sequencing to perform a cross-sectional periodontitis analysis. In this study, the 2018 periodontitis classification served as the basis for the classification of periodontitis stages (Papapanou et al., 2018). There were notable variations in the salivary microbiome composition among the multiple stages of periodontitis (Figure 14). Furthermore, our random forest classification model based on the proportions of DAT in the salivary microbiome compositions across study participants to predict multiple periodontitis statuses with high AUC of 0.870±0.079 (mean±SD) (Table 4).

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

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

According to our findings, the salivary microbiome compositions have 425 taxa (Figure 14). We computed the alpha-diversity indices to determine the variability within each salivary microbiome composition, including ace (Chao & Lee, 1992), chao1 (Chao, 1984), fisher alpha (Fisher et al., 1943), margalef (Magurran, 2021), observed ASVs (DeSantis et al., 2006), Berger-Parker *d* (Berger & Parker, 1970), Gini (Gini, 1912), Shannon (Weaver, 1963), and Simpson (Simpson, 1949) (Figure 7 and Figure 20). Alpha-diversity indices suggested that the microbial richness of periodontally healthy controls was higher than that of patients with periodontitis (Figure 7a-e and Figure 20). These results are in line with findings with that patients with advanced periodontitis, namely stage II and stage III, have less diversified communities than periodontally healthy controls (Jorth et al., 2014). Recognizing that the periodontitis severity increases the amount of *Porphyromonas gingivalis*, the salivary microbiome compositions from periodontally healthy controls conserved microbial networks dominated by *Streptococcus* spp. (Figure 14). *Porphyromonas gingivalis* is one of the known periodontal pathogen that could cause dysbiosys in the salivary microbiomes, suggesting in the pathophysiology of periodontitis. Despite this finding, earlier research found that subgingival microbiome of patients with periodontitis had a greater alpha-diversity index (observed ASVs) than that of healthy controls (Iniesta et al., 2023), might due to the different sampling sites between saliva and subgingival plaque. On the other hand, another research has addressed significant discrepancies in alpha-diversity indices from subgingival plaque, saliva, and tongue biofilms from healthy controls and periodontitis patients, resulting the highest alpha-diversity index in saliva collections (Belstrøm et al., 2021). Moreover, early-stage periodontitis, namely stage I,

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

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

Of the total of 425 taxa in the salivary microbiome composition that have been identified (Figure 14), ANCOM was applied to select 20 taxa as the DAT that indicated notable abundance variation among the periodontitis stages (Figure 8 and Table 5). Three sub-groups were formed from the DAT using hierarchical clustering (Figure 8a). Surprisingly, two of the red complex pathogens (Rôças, Siqueira Jr, Santos, Coelho, & de Janeiro, 2001), *Porphyromonas gingivalis* and *Tannerella forsythia*, were classified in Group 2 and were more prevalent in stage II and stage III periodontitis compared to healthy controls. *Campylobacter showae* was additionally placed in Group 2 of the orange complex pathogens (Gambin et al., 2021). Furthermoe, some of the DAT in Group 2 have reported their crucial roles in pathogenesis and development of periodontitis: *Filifactor alocis* (Aruni et al., 2015), *Treponema putidum* (Wyss et al., 2004), *Tannerella forsythia* (Stafford, Roy, Honma, & Sharma, 2012; W. Zhu & Lee, 2016), and *Prevotella intermedia* (Karched, Bhardwaj, Qudeimat, Al-Khabbaz, & Ellepola, 2022). Taken together, this indicates that DAT in Group 2 is essential to periodontitis. The portion of some Group 1 DAT, including *Peptostreptococcaceae[XI][G-5] saphenum*, *Peptostreptococcaceae[XI][G-6] nodatum*, and *Peptostreptococcaceae[XI][G-9] brachy*, in healthy controls and patients with periodontitis significantly differed, according to earlier research (Lafaurie et al., 2022). These outcomes support our research, implying that Group 1 DAT are also essential to the etiology and progression of periodontitis. However, in contrast to patients with periodontitis, Group 3 DAT, namely *Corynebacterium durum* and *Actinomyces graevenitzii*, were enriched in healthy controls, which is consistent with earlier research (Redanz et al., 2021; Nibali et al., 2020).

In our correlation analysis (Figure 9), we have discovered strongly negative correlations (coefficient ≤ -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 15). 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 15a), *Porphyromonas gingivalis* (Figure 15b), and *Treponema putidum* (Figure 15c). 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 15d-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 stages 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 patients, 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 stages 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%
- BA: 84%

- SEN: 94%
- SPE: 74%

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

- AUC: 95.3±4.9%
- BA: 88.5±6.6%
- SEN: 86.4±15.7%
- SPE: 90.5±7.0%

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

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

Finally, the functional enrichment test) indicated that four pathways were over-represented in stages II/III compared with healthy/stage I (Figure 12):

- L-lysine fermentation to acetate and butanoate
- Pyrimidine deoxyribonucleotides de novo biosynthesis III
- Pyruvate fermentation to acetone
- Succinate fermentation to butanoate

On the other hand, two pathways related to glucose and glucose-1-phosphate degradation and the glyoxylate cycle were underrepresented in stage III compared to healthy status (Figure 12b). These results are consistent with previous studies reporting that fermentation pathways are over-represented in periodontitis (Szafrański et al., 2015; Niederman, Buyle-Bodin, Lu, Robinson, & Naleway, 1997; Chang et al., 2020). In line with our results, previous study also found that the lysine fermentation to butyrate pathway and pyruvate fermentation pathway showed increased gene expression in diseased sites of patients with periodontitis (Jorth et al., 2014). Nevertheless, further studies are required to confirm the associations of the over-represented and under-represented pathways in periodontitis.

Regarding the clinical characteristics and potential confounders influencing the analysis of salivary microbiome compositions connected with periodontitis severity, this study had a number of limitations that were pointed out. We did not offer clinical information, such as the percentage of teeth, the percentage of bleeding on probing, nor dental furcation involvement (Table 3), even though we did gather information on attachment level, probing depth, plaque index, and gingival index (Renvert & Persson, 2002); this

might have it challenging to present thorough and in-depth data about periodontal health. Moreover, the broad age range may make it tougher to evaluate the relationship between age and periodontitis statuses, providing the necessity for future studies to consider into account more comprehensive clinical characteristics associated with periodontitis. Additionally, potential confounders–*e.g.* body mass index (Bombin, Yan, Bombin, Mosley, & Ferguson, 2022) and e-cigarette use (Suzuki, Nakano, Yoneda, Hirofuji, & Hanioka, 2022)–which might have affected dental health and salivary microbiome composition were disregarding consideration in addition to smoking status and systemic diseases. Thus, future research incorporating these components would offer a more thorough knowledge of how lifestyle factors interact and affect the salivary microbiome composition and periodontal health. Throughout, resolving these limitations will advance our understanding in pathogenesis and development of periodontitis, offering significant novel insights on the causal connection between systemic diseases and the salivary microbiome compositions.

4 Metagenomic signature analysis of Korean colorectal cancer

4.1 Introduction

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

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

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

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

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

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

4.2 Materials and methods

4.2.1 Study participants enrollment

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

4.2.2 DNA extraction procedure

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

4.2.3 Bioinformatics analysis

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

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

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

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

1. Berger-Parker *d* (Berger & Parker, 1970)
2. Chao1 (Chao, 1984)
3. Dominance
4. Doubles
5. Fisher (Fisher et al., 1943)
6. Good's coverage (Good, 1953)
7. Margalef (Magurran, 2021)
8. Mcintosh *e* (Heip, 1974)
9. Observed ASVs (DeSantis et al., 2006)
10. Simpson *d*
11. Singles
12. Strong (Strong, 2002)

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

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

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

$$\text{slope} = \frac{\Delta \text{OS}}{\Delta \text{DAT proportion}} \tag{9}$$

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

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

4.2.4 Data and code availability

All sequences from the 211 study participants have been published to the Korea Bioinformation Center (data ID `KGD10008857`): `https://kbds.re.kr/KGD10008857`. Docker image that employed throughout 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`.

4.3 Results

4.3.1 Summary of clinical characteristics

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

4.3.2 Gut microbiome compositions

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

1. Archaea kingdom: 31 genera
2. Bacteria kingdom: 1508 genera
3. Bamfordvirae kingdom: 1 genus
4. Eukaryota kingdom: 77 genera
5. Fungi kingdom: 137 genera
6. Loebvirae kingdom: 2 genera
7. Orthornavirae kingdom: 1 genus
8. Paranavirae kingdom: 3 genera
9. Shotokuvirae kingdom: 6 genera
10. Viruses kingdom: 76 genera

Among these kingdoms, the proportions of four major kingdoms, which have at least 50 genera, in the gut microbiome composition were displayed (Figure 22): bacteria kingdom, eukaryota kingdom, fungi kingdom, and viruses kingdom. In the bacteria kingdom (Figure 22a), *Bacteroides* genus is the most prevalent genus in the tumor tissue samples, followed by *Fusobacterium* and *Cutibacterium* genera. *Toxoplasma* and *Malassezia* genera were the dominant genus, which have over 90% of proprotions, in the eukaryota kingdom (Figure 22b) and the fungi kingdom (Figure 22c), respectively. On the other hand, *Roseolovirus* genus is the most popular genus of the viruses kingdom in the normal tissue samples (Figure 22d); contrarily, *Lymphocryptovirus* and *Cytomegalovirus* genera had been dominant genera in the tumor tissue samples. Taken together, these results suggest that the Anna Karenina principle (Ma, 2020; W. Li & Yang, 2025), *i.e.* in human microbiome-associated diseases, every disease-associated microbiome, including dysbiosis, is unique and patient-specific, whereas all healthy microbiomes are similar, also applies to CRC.

4.3.3 Diversity indices

In alpha-diversity analysis, which measures within-sample microbial community, revealed a significant increase in tumor samples compared to adjacent normal samples (Figure 23). Alpha-diversity indices, including Chao1, Fisher α, and observed features, were consistently higher in CRC tumor tissues (MWU test $p < 0.05$), indicating a more heterogeneous microbial community, *e.g.* the Anna Karenina principle, potentially influenced by tumor-associated dysbiosis.

To assess the microbial impact on CRC recurrence, alpha-diversity indices compared between normal and tumor tissue samples in accordance with recurrence information (Figure 24). In the recurrence patients, most alpha-diversity indices (11/12 indices; 92% indices), except Mcintosh index, exhibited increasing in tumor samples than normal samples (MWU test $p < 0.05$; Figure 24); In the non-recurrence patients, on the other hand, some alpha-diversity indices (8/12 indices; 67% indices) amplified in tumor samples than normal samples (MWU test $p < 0.05$; Figure 24). What is interesting about the alpha-diversity analysis in this figure is that a few indices, namely Fisher α (Figure 29e) and Margalef (Figure 24g), presented augmentation in normal sample of the recurrence patients than that of the non-recurrence patients (MWU test $p < 0.05$). Overall, these alpha-diversity results demonstrate that tumor samples have more diverse microbiome composition than normal samples. Furthermore, although only two indices significantly increased, the recurrence patients have diversified microbiome compositions than the non-recurrence patients in normal samples, not in tumor samples, indicating field cancerization by the gut microbiome leads to unfavorable prognosis such as recurrence (Curtius et al., 2018; Rubio et al., 2022).

To determine the microbial impact on OS of CRC patients, the Spearman correlation compared between alpha-diversity indices and OS duration (Figure 25). No significant Spearman correlation was found between every alpha-diversity indices and OS (Spearman correlation $p \geq 0.1$; Figure 25). However, a few alpha-diversity indices, *e.g.* Chao1 (Figure 25b), Good's coverage (Figure 25f), and observed features (Figure 25i), showed negative correlations with OS (Spearman correlation $p < 0.05$). Together these correlation results provide important insights into heterogeneous microbiome leads to shorter OS, suggesting the Anna Karenina principle and the filed cancerization.

In beta-diversity analysis, which calculates inter-sample microbial community, explain significant disparity between tumor samples and normal samples (Figure 26). Every six beta-diversity indices presented discrepancy between normal samples and tumor samples (PERMANOVA test $p < 0.001$), implying that tumor samples have distinct microbiome compositions from normal tissue samples.

Beta-diversity indices were evaluated between normal and tumor tissue samples along with recurrence history in order to evaluate the microbial influence on CRC recurrence (Figure 27). All six beta-diversity indices examined significant difference in microbial community structure between the recurrence patients and the non-recurrence patients (PERMANOVA test $p < 0.001$; Figure 27), indicating that tumor-associated gut microbiome composition varies resulting on recurrence status. tSNE-transformed plots further illustrated clear clustering patterns (Figure 27), suggesting again that the recurrence patients harbor dissimilar microbial communities compared to the non-recurrence patients. These observed differences in beta-diversity represent that microbial shifts, including dysbiosis, may be associated with

CRC progression and recurrence risk, possibly due to specific taxa contributing to a tumor-promoting microenvironment.

Moreover, beta-diversity analysis suggested a potential associated with OS duration in CRC patients. In all six beta-diversity indices, tSNE-transformed projection plots showed clear clustering patterns along OS duration (Figure 28), implying that possible microbiome composition shifts related to survival outcomes in CRC. However, since OS is a continuous variable, statistical significance testing could not be directly performed for these clustering patterns. Despite this limitation, the observed microbial community variations suggest that alterations in the gut microbiome composition may be associated to CRC prognosis and survival duration.

Together, diversity indices analyses revealed significant microbial community alterations between normal and tumor tissue samples, as well as between the recurrence and non-recurrence CRC patients. Alpha-diversity indices significantly increased in tumor tissue samples than normal tissue samples (MWU test $p < 0.05$; Figure 23). This increase was more pronounced in the recurrence patients (11/12 indices; 92% indices) compared to non-recurrence patients (8/12 indices; 67% indices) (Figure 24), indicating a potential link between microbial diversity and CRC recurrence. Additionally, negative correlation between OS and alpha-diversity indices were observed in normal samples (Spearman correlation $p < 0.05$; Figure 25), suggesting that lower microbial diversity may be associated with longer survival in CRC. On the other hand, beta-diversity indices analysis, showed significant separation between tumor and tumor tissue samples across all six beta-diversity indices (PERMANOVA test $p < 0.001$; Figure 26). Furthermore, the recurrence and non-recurrence patients displayed significantly discrete microbial compositions (PERMANOVA test $p < 0.001$; Figure 27), implying that microbial community shifts may reflect CRC progression and recurrence risk. These findings highlight the importance of microbiome diversity and gut microbiome composition in CRC prognosis and warrant further investigation into their potential as predictive biomarkers.

4.3.4 DAT selection

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

We identified 19 DAT associated with recurrence history across the total samples by ANCOM (Figure 29a), including 18 non-recurrence-enriched DAT and a recurrence-enriched DAT. When stratified by sample type, one DAT was enriched in normal samples of the non-recurrence patients (Figure 29b), whereas six DAT exhibited significant differential abundance in tumor samples (Figure 29c). These findings suggest that microbial composition variations in the tumor microenvironment are more pronounced in relation to recurrence status (Table 9), potentially indicating a microbial signature linked to CRC progression. These identified DAT may contribute to tumor-associated dysbiosis, influencing the likelihood of CRC recurrence through mechanisms such as inflammation, metabolic modulation, or

immune system interaction.

The non-recurrence-enriched DAT have decreased proportions both in normal and tumor samples of the recurrence patients than those in the non-recurrence patients (MWU test $p < 0.001$; Figure 29d-h). What is interesting about these non-recurrence-enriched DAT is that they belong to the *Micrococcus* genus. Among them, *Micrococcus aloeverae* was consistently identified in all three settings–total (Figure 29a), normal (Figure 29b), and tumor samples (Figure 29c)–indicating its stable presence regardless of tissue type. Variation in relative proportions of *Micrococcus aloeverae* (Figure 29d) suggests potential ecological adaptability within tumor microenvironment of CRC. The remaining *Micrococcus* genus DAT showed less variation between the recurrence and non-recurrence patients, reinforcing their limited associations with CRC recurrence. Moreover, only one taxon, *Pseudomonas* sp. *NBRC 111133*, was identified as recurrence-enriched DAT (Fiugre 29a). This suggests a potential association between *Pseudomonas* sp. *NBRC 111133* and CRC recurrence, indicating that its presence may contribute to a tumor-supportive microbial environment. *Pseudomonas* sp. *NBRC 111133* had higher relative proportions both in normal and tumor tissue samples of the recurrence patients than those of the non-recurrence patients (Figure 29i). Likewise, *Pseudomonas* sp. *NBRC 111133* were prevalent in tumor samples than normal samples of the non-recurrence patients (MWU test $p < 0.01$; Figure 29i); however, no significant difference between normal and tumor tissue samples of the recurrence patients.

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

Furthermore, correlation analysis between DAT abundance and OS duration identified a total of 16 over-represented DAT in the total samples (Figure 30a). When analyzed separately, 11 OS-correlated DAT, which consist of four under-represented and seven over-represented DAT showed significant correlations with OS in normal samples (Figure 30b), while four under-represented and 45 over-represented DAT were identified in tumor samples (Figure 30c), indicating that microbial composition shifts in tumor tissues may have a stronger association with survival outcomes. The higher number of survival-associated DAT in tumor tissue suggests that the tumor microbiome plays a more dynamic role in progression and prognosis of CRC. These findings highlight the potential of gut microbial composition as a prognostic indicator in CRC, warranting further investigation into the functional roles of these DAT in influencing clinical outcomes.

Among a total of 57 OS-correlated DAT (Table 10) with Spearman correlation and the slope (Equation 9). *Agaricus bisporus* (Figure 30d) and *Corynebacterium* sp. *KPL1824* (Figure 30h) are identified as over-represented DAT both in normal samples and tumor samples (Spearman correlation $p < 0.05$), whereas *Corynebacterium lowii* (Figure 30g) and *Paracoccus sphaerophysae* (Figure 30i) are selected as under-represented DAT both in normal samples and tumor samples (Spearman correlation $p < 0.05$). On the other hand, *Clostridiales bacterium* (Figure 30e) is classified as under-represented DAT only in normal samples (Spearman correlation $p < 0.01$), while *Corynebacterium kroppenstedtii* (Figure 30f) is described as over-represented DAT only in tumor samples (Spearman correlation $p < 0.001$).

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

4.3.5 Random forest prediction

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

To evaluate the predictive power of gut microbiome composition in CRC recurrence, we implemented a random forest classification model using two different input sets (Figure 31a-f): the entire gut microbiome composition and DAT-selected microbiome. Comparing these models allowed us to assess whether focusing on DAT-selected microbial features enhances classification performance. While the DAT-based classification models showed slightly improved classification metrics (MWU test $p \geq 0.05$), including ACC, AUC, and BA, over the entire microbiome-based model in the total sample (Figure 31a and Figure 30d), normal samples (Figure 31b and Figure 31e), and tumor samples (Figure 31c and Figure 31f), overall classification metrics remained around 60% (0.570 ± 0.164, mean$\pm$SD), suggesting moderated predictive capability. This relatively low metrics highlight the complexity of CRC recurrence, indicating that while dysbiosis may contribute to CRC progression, it is likely interwinded with host genetic factors such as germline and somatic mutations. Thus, the interplay between microbial shifts and tumor genomic alterations warrants further investigation, as integrating microbiome and genomic sequencing data may improve therapeutic strategies.

To assess the predictive capability of the gut microbiome composition in OS of CRC patients, we implemented a random forest regression model, comparing two different input sets (Figure 31g-i): the entire gut microbiome composition and DAT-selected microbiome. This comparison also aimed to determine whether focusing on key microbial features (DAT) enhances predictive accuracy. While DAT-based model showed a slight improvement over the entire microbiome-based model in normal samples (Figure 31h) and tumor samples (Figure 31i), the regression error remained high (729.302 ± 179.940, mean$\pm$SD), indicating substantial variability in survival outcomes that cannot be fully explained by gut

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

Table 8: **Clinical characteristics of CRC study participants.**

Continuous variable: mean±SD. Categorical variable: count (proportion). 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	186	7	18	
Recurrence, n (%)	False	132 (62.6%)	115 (61.8%)	4 (57.1%)	13 (72.2%)	0.657
	True	79 (37.4%)	71 (38.2%)	3 (42.9%)	5 (27.8%)	
Sex, n (%)	Male	137 (64.9%)	121 (65.1%)	6 (85.7%)	10 (55.6%)	0.357
	Female	74 (35.1%)	65 (34.9%)	1 (14.3%)	8 (44.4%)	
OS, mean±SD		1248.5±770.3	1256.7±766.4	1416.6±496.3	1097.7±903.2	0.580
Age, mean±SD		61.2±13.1	61.3±12.4	60.1±15.6	60.2±19.4	0.867

Table 9: **DAT list for CRC recurrence.**

Statistical significance was determined by ANCOM W. Significance threshold is $|\log_2 \mathrm{FC} > 1.0| \wedge W > 9600$. 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
Cutibacterium acnes	-1.878	10570				
Cutibacterium avidum	-1.383	10266				
Cutibacterium granulosum	-1.476	10271				
Micrococcus aloeverae	-2.280	10740	-1.821	10462	-2.481	10591
Micrococcus luteus	-2.216	10744				
Micrococcus sp. *CH3*	-2.323	10740			-2.493	10527
Micrococcus sp. *CH7*	-2.321	10740			-2.493	10542
Micrococcus sp. *HMSC31B01*	-2.282	10739			-2.458	10519
Micrococcus sp. *MS-ASIII-49*	-2.284	10740			-2.470	10527
Pseudomonas sp. *NBRC 111133*	1.139	9732				
Pseudonocardia sp. *P2*	-2.200	10736			-2.394	10253
Staphylococcus sp. *HMSC034A07*	-1.341	10050				
Staphylococcus sp. *HMSC063F03*	-1.322	10001				
Staphylococcus sp. *HMSC064E11*	-1.064	10163				
Staphylococcus sp. *HMSC067B04*	-1.343	9952				
Staphylococcus sp. *HMSC068G12*	-1.344	10173				
Staphylococcus sp. *HMSC072H01*	-1.298	10197				
Staphylococcus sp. *HMSC077C03*	-1.331	10115				
Treponema endosymbiont of Eucomonympha 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
Acinetobacter venetianus					3.087	0.203
Actinotalea ferrariae					2.574	0.200
Agaricus bisporus	2.329	0.287	2.925	0.276	2.258	0.306
Bifidobacterium boum					2.096	-0.216
Brevundimonas sp. *DS20*			2.180	0.279		
Clostridiales bacterium			2.631	-0.203		
Corynebacterium kroppenstedtii	2.117	0.220			2.117	0.302
Corynebacterium lipophiloflavum			2.137	0.227		
Corynebacterium lowii			2.006	-0.216		
Corynebacterium sp. *KPL1818*	2.101	0.209	2.487	0.220	2.044	0.215
Corynebacterium sp. *KPL1824*	2.057	0.207	2.511	0.212	2.003	0.226
Corynebacterium sp. *KPL1986*					2.205	0.202
Corynebacterium sp. *KPL1996*					2.205	0.202
Corynebacterium sp. *KPL1998*					2.205	0.202
Corynebacterium sp. *KPL2004*					2.205	0.202
Kocuria flava			2.729	0.214		
Kytococcus sedentarius					2.267	0.206
Lachnospiraceae bacterium AD3010			2.609	-0.203		
Lachnospiraceae bacterium NK4A136					2.538	-0.220
Methylorubrum extorquens					2.068	0.295
Microbacterium barkeri			2.071	0.389		
Paracoccus sphaerophysae					2.012	-0.209
Pontibacillus litoralis					2.580	-0.209
Porphyromonas macacae			2.476	-0.200		
Pseudomonas balearica					2.117	0.203
Pseudomonas monteilii					2.183	0.228
Rodentibacter myodis					2.444	0.245
Roseovarius tolerans					2.295	0.221
Staphylococcus epidermidis					2.243	0.214
Staphylococcus sp. *HMSC034A07*					2.183	0.209
Staphylococcus sp. *HMSC034D07*	2.278	0.206			2.252	0.253
Staphylococcus sp. *HMSC034G11*	2.362	0.208			2.357	0.261
Staphylococcus sp. *HMSC036A09*					2.308	0.239
Staphylococcus sp. *HMSC055A10*					2.168	0.222
Staphylococcus sp. *HMSC055B03*	2.134	0.202			2.134	0.266
Staphylococcus sp. *HMSC058E12*					2.106	0.216
Staphylococcus sp. *HMSC061C10*					2.882	0.207
Staphylococcus sp. *HMSC062B11*	2.391	0.203			2.377	0.253
Staphylococcus sp. *HMSC062D04*	2.278	0.202			2.274	0.259
Staphylococcus sp. *HMSC063F03*	2.376	0.201			2.367	0.251
Staphylococcus sp. *HMSC063F05*	2.387	0.210			2.381	0.266
Staphylococcus sp. *HMSC064E11*					2.276	0.218
Staphylococcus sp. *HMSC065D11*					2.329	0.245

Table 10 continued from previous page

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

Table 11: **Random forest classification and their evaluations.**

Metrics are shown as mean±SD.

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.**

Metrics are shown as mean±SD.

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

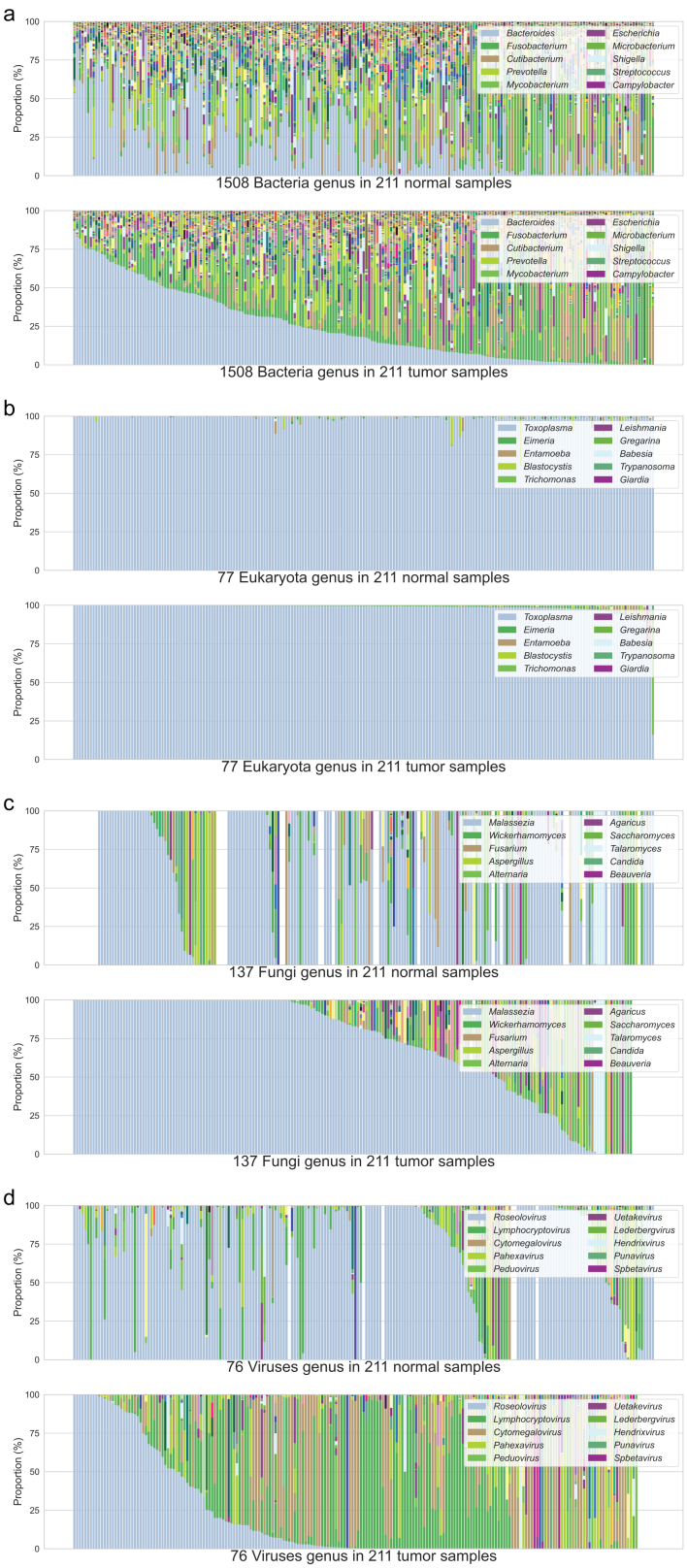

Figure 22: **Gut microbiome compositions in genus level.**

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

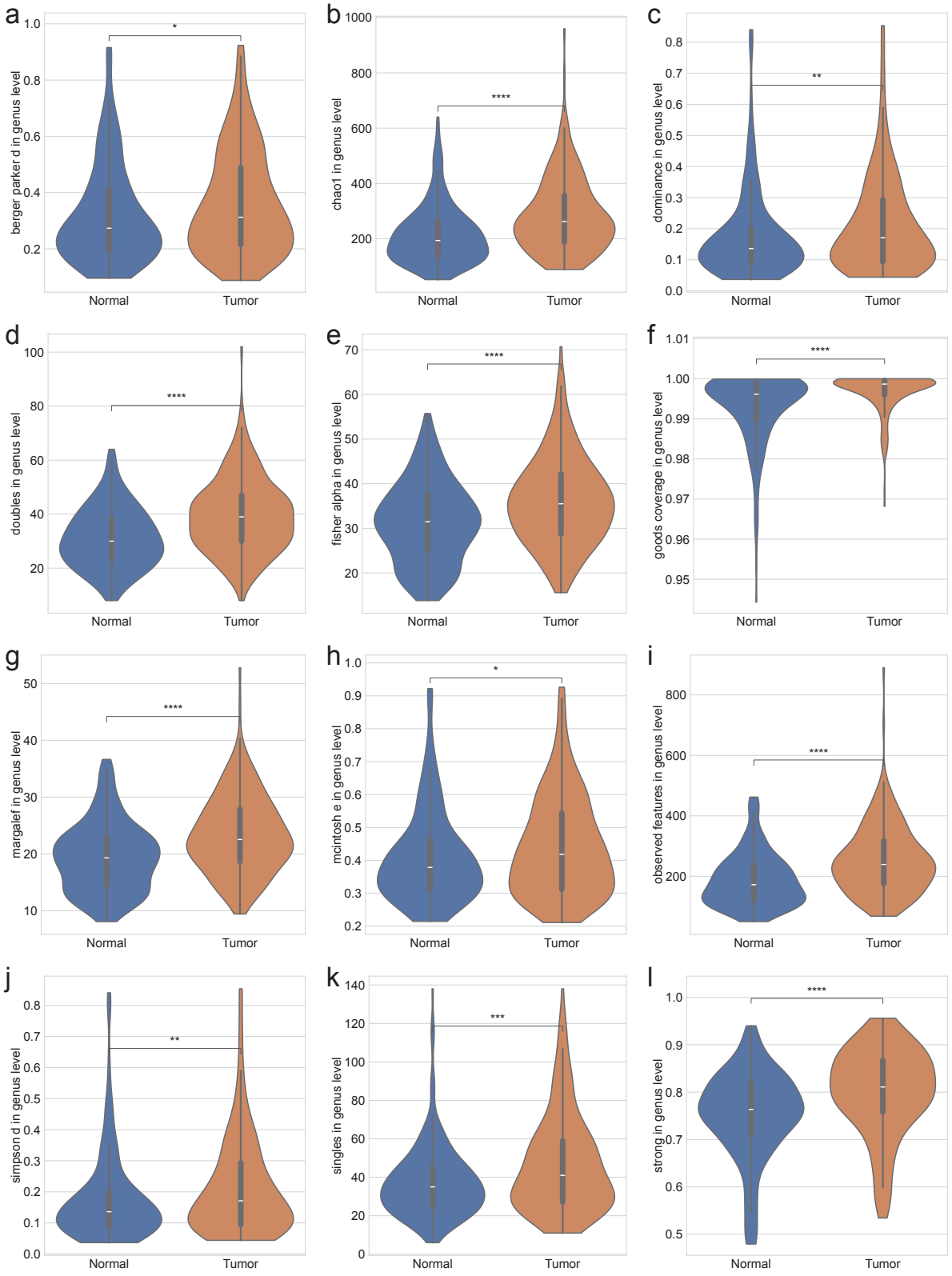

Figure 23: **Alpha-diversity indices in genus level.**

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

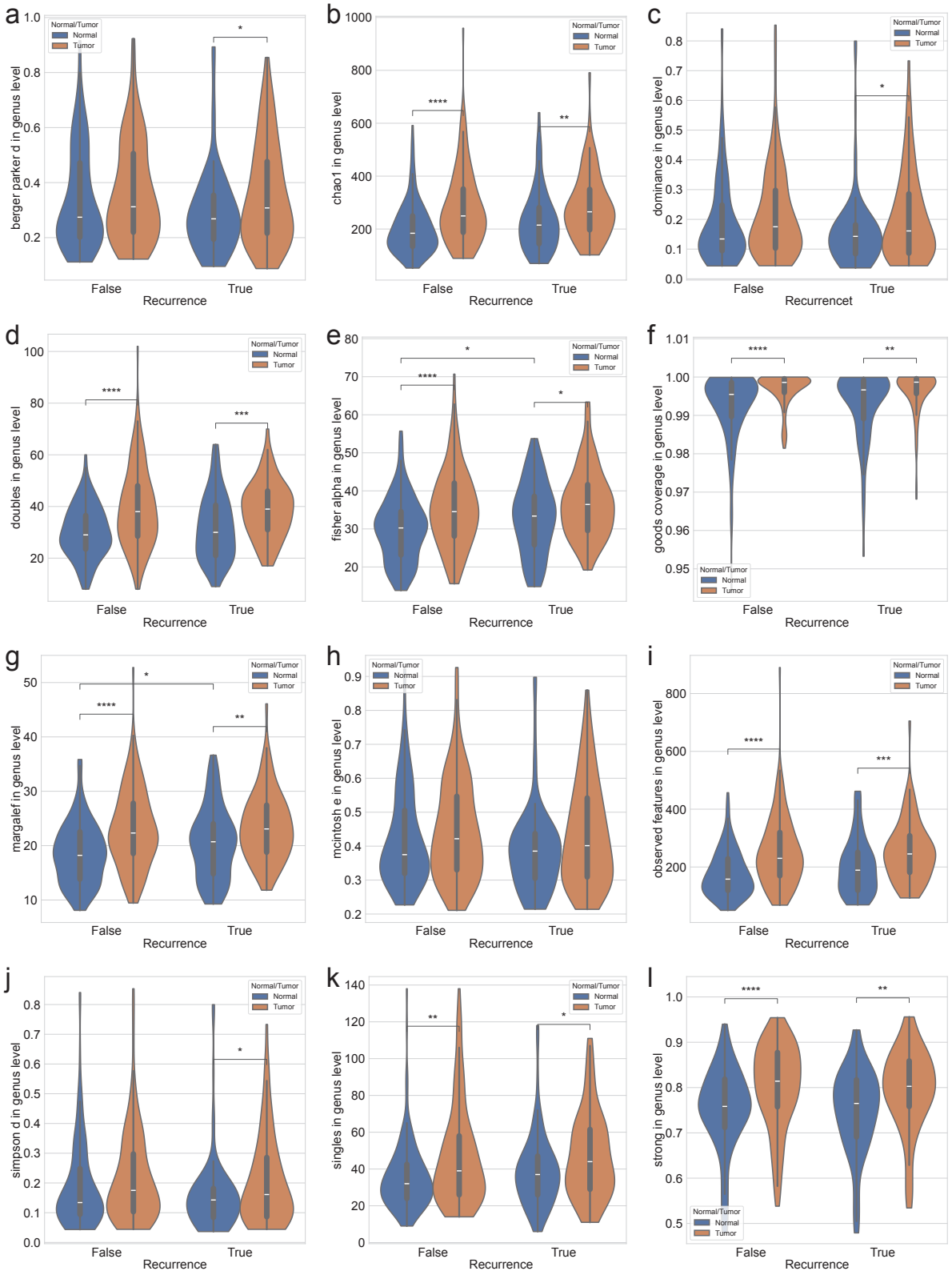

Figure 24: **Alpha-diversity indices with recurrence in genus level.**

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

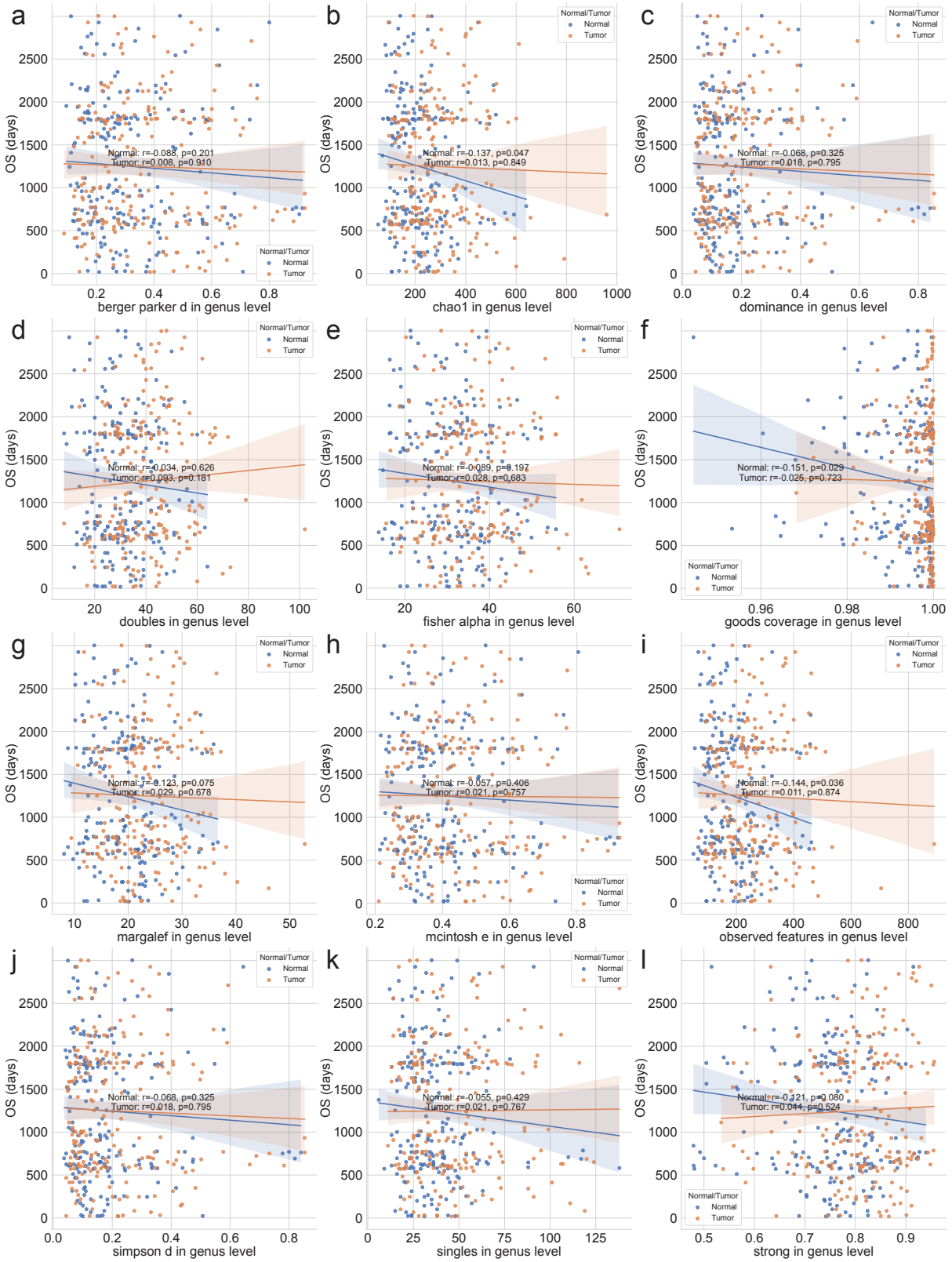

Figure 25: **Alpha-diversity indices with OS in genus level.**

(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.

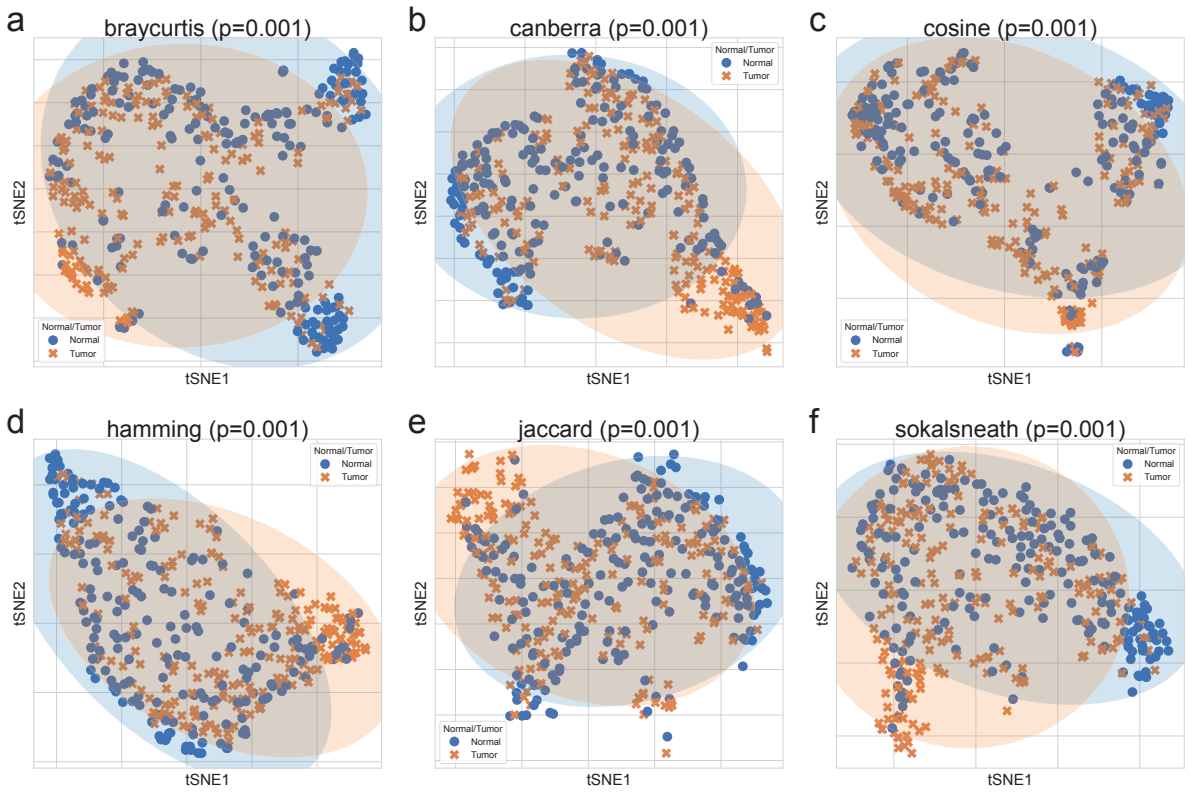

Figure 26: **Beta-diversity indices in genus level**.

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.

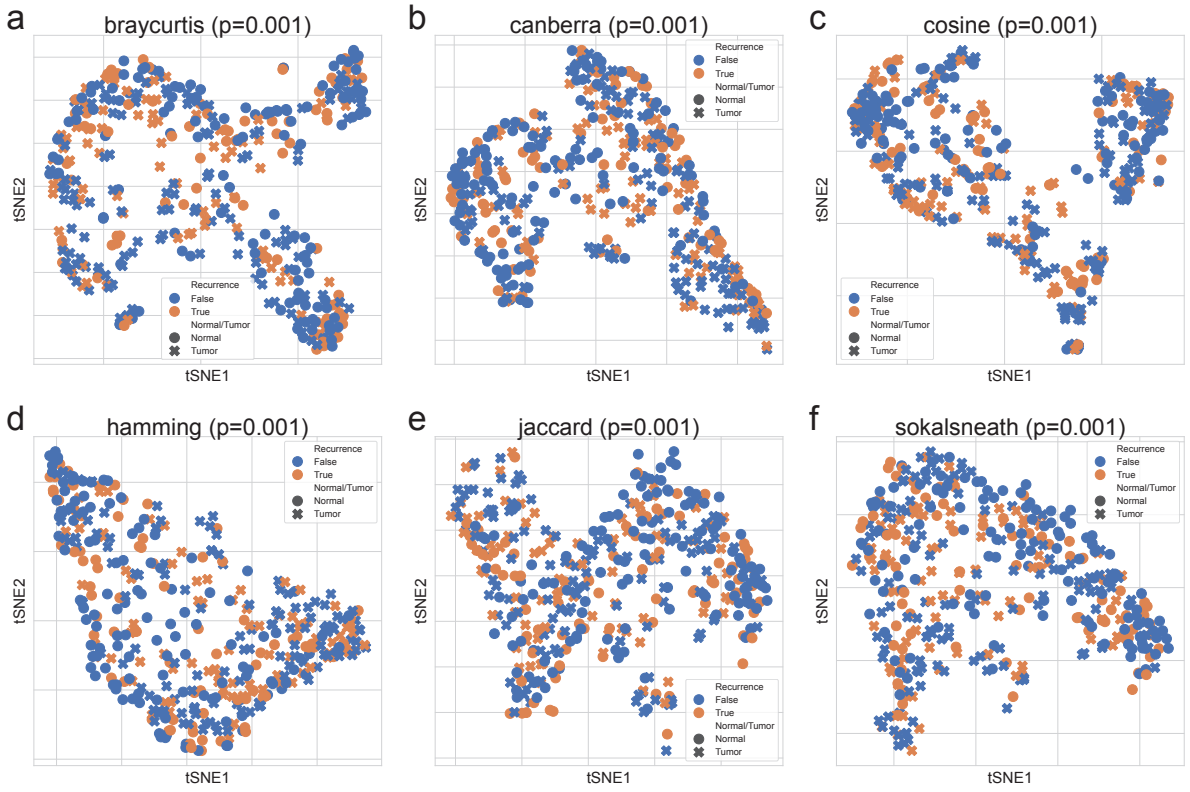

Figure 27: **Beta-diversity indices with recurrence in genus level**.

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.

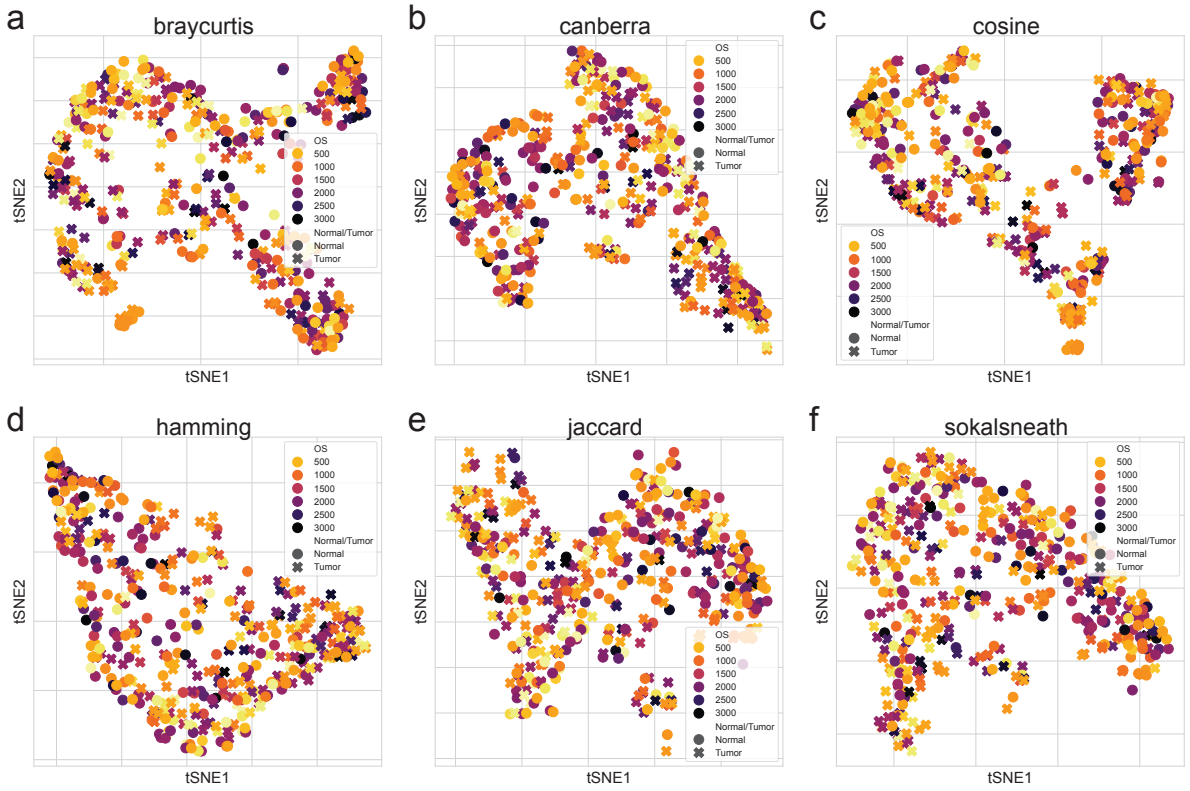

Figure 28: **Beta-diversity indices with OS in genus level.**

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

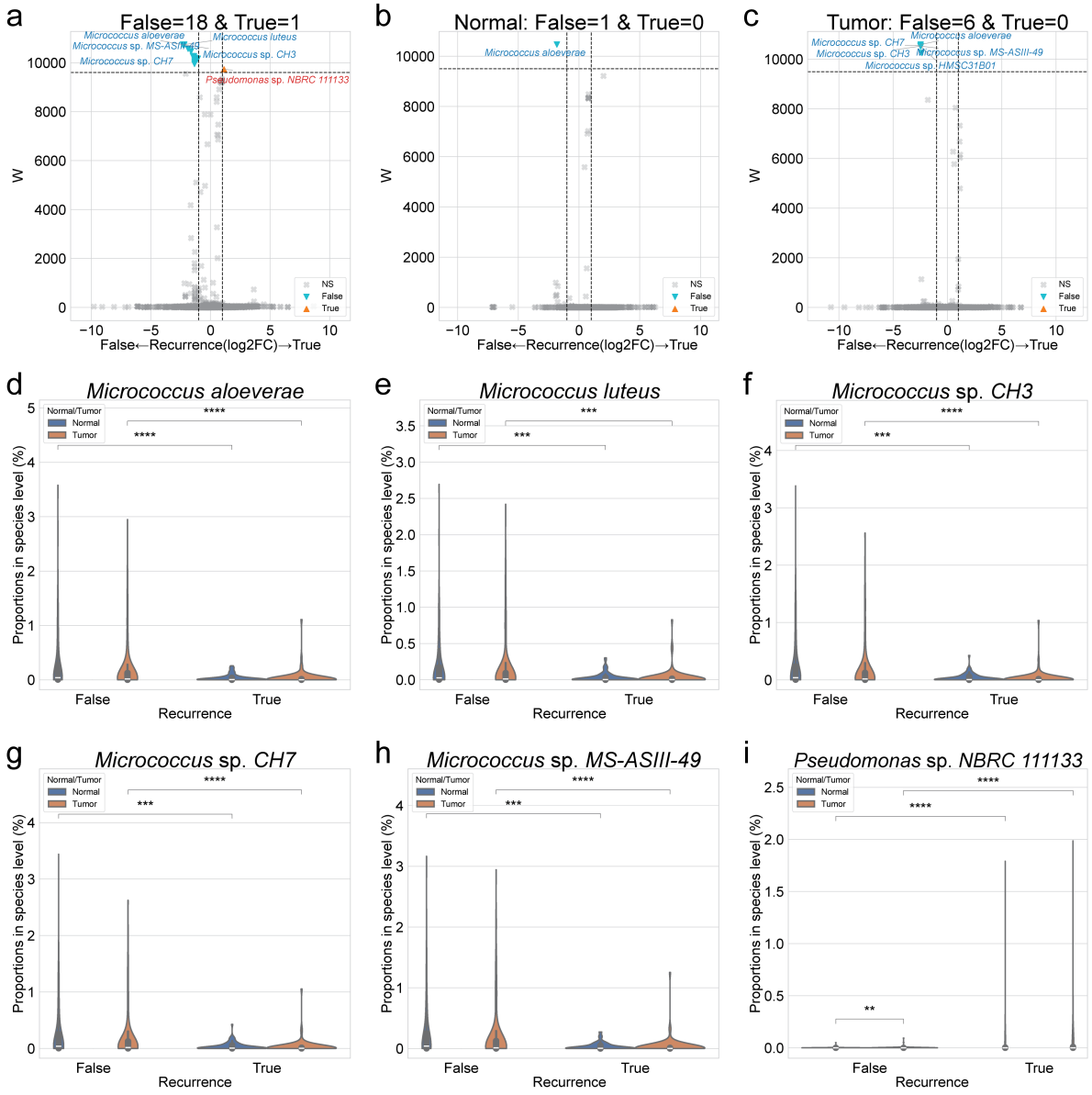

Figure 29: **DAT with recurrence in species level.**

(a-c) Volcano plots with recurrence. x-axis indicates $\log_2(\text{Fold Change})$ on recurrence, and y-axis indicates ANCOM significance (W). **(a)** Total **(b)** Normal samples **(c)** Tumor samples. **(d-i)** Violin plots of each taxon proportion with recurrence. **(d)** *Micrococcus aloeverae* **(e)** *Micrococcus luteus* **(f)** *Micrococcus* sp. *CH3* **(g)** *Micrococcus* sp. *CH7* **(h)** *Micrococcus* sp. *MS-ASIII-49* **(i)** *Pseudomonas* sp. *NBRC 111133*. Significant threshold is: $|\log_2 \text{Fold Change}| > 1.0$ and $W > 9600$. WU test: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), and $p < 0.0001$ (****)

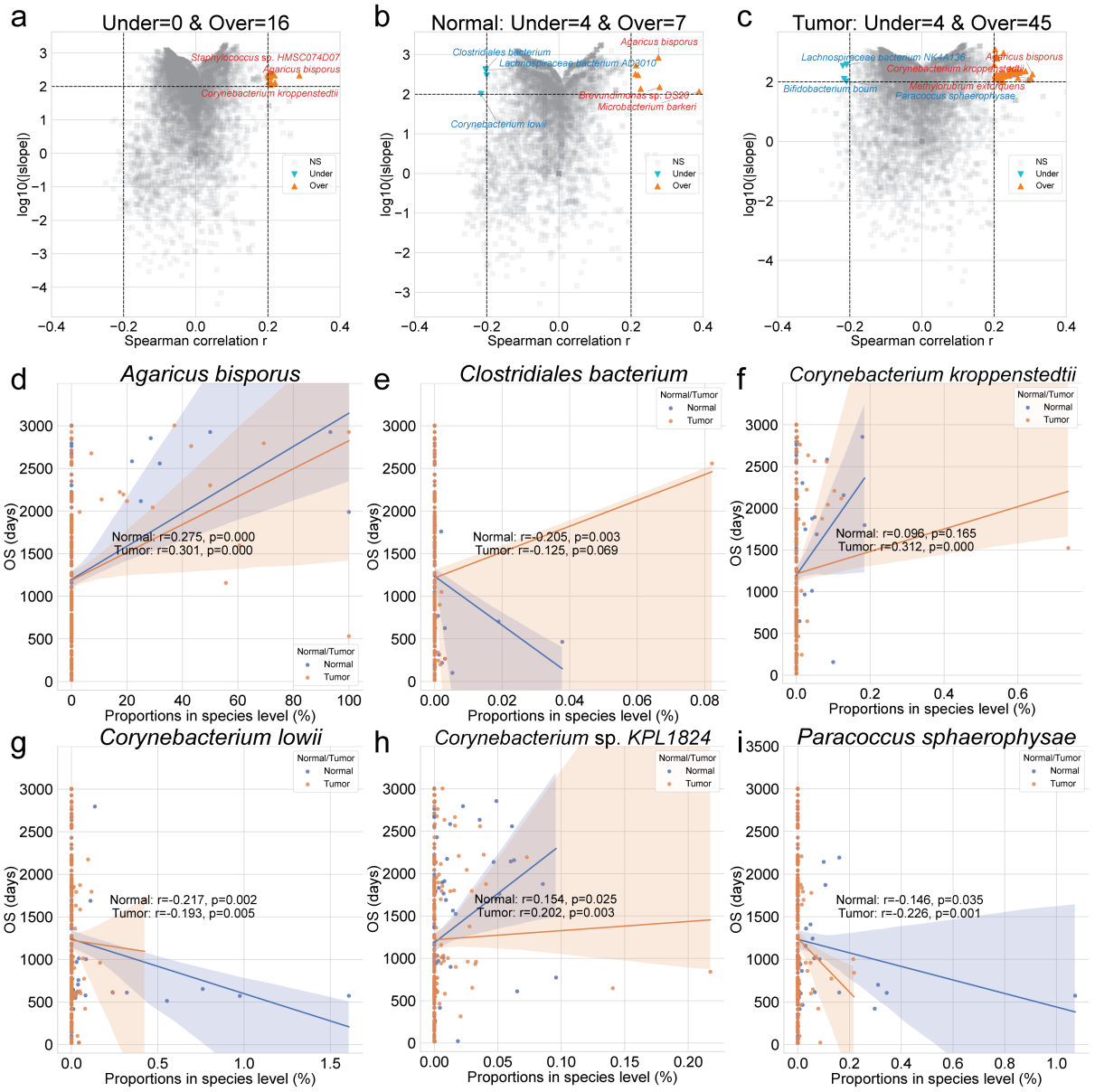

Figure 30: **DAT with OS in species level.**

(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): $|r| > 0.2$ and $\log_{10}\text{slope} > 2.0$.

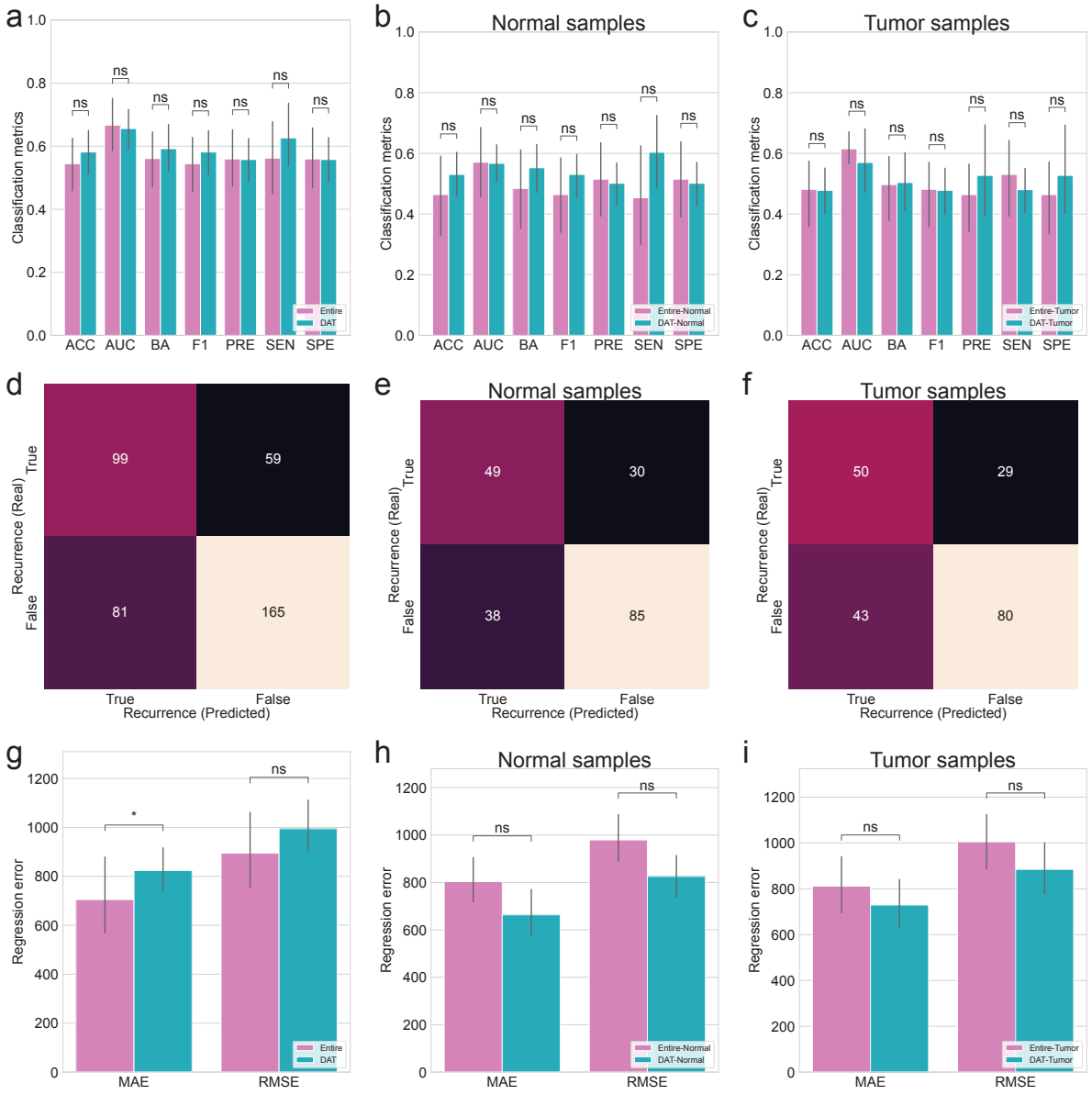

Figure 31: **Random forest classification and regression.**

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

4.4 Discussion

This study provides a comprehensive metagenomic signature analysis of Korean CRC patients by examining prevalent microbial taxa, diversity indices, DAT selection, and random forest-based predictions for recurrence and survival outcomes. Our analysis revealed distinct prevalent microbial communities in CRC patients (Figure 22), with significant difference between tumor tissues and adjacent matched normal tissues. Alpha-diversity indices analysis showed an overall shift in microbial diversity within tumor samples (Figure 23, Figure 24, and Figure 25), while beta-diversity analyses indicated significant changes in microbial composition associated with recurrence history and survival duration (Figure 26, Figure 27, and Figure 28). Through DAT selection by ANCOM and Spearman correlation, we identified key microbial taxa link to recurrence history (Table 9 and Figure 29) and OS duration (Table 10 and Figure 30), highlighting potential microbial biomarkers for CRC prognosis. To evaluate the predictive capacity of these microbial features, we implemented random forest-based machine learning models, where random forest classification demonstrated moderate accuracy (0.570 ± 0.164, mean$\pm$SD) for CRC recurrence prediction (Table 11 and Figure 31) and random forest regression showed slightly high error (729.302 ± 179.940, mean$\pm$SD) for OS prediction (Table 12 and Figure 31), suggesting that gut microbiome alterations alone are insufficient for precise prognosis and may interact with host genetic factors such as germline and somatic mutations. These findings underscore the potential of microbial biomarkers in CRC risk stratification, emphasizing the need for multi-omics integration to improve predictive models and personalized medicine strategies in CRC.

In the bacteria kingdom (Figure 22a), *Bacteroides* genus is the most frequent genus in tumor tissues, then came *Fusobacterium* and *Cutibacterium* genera. These results also accord with previous studies, which showed that *Bacteroides fragilis* (Scott, Whittle, Jeraldo, & Chia, 2022; Purcell, Permain, & Keenan, 2022), *Fusobacterium nucleatum* (Wang & Fang, 2023; Zepeda-Rivera et al., 2024), and *Cutibacterium acnes* (Benej et al., 2024) have significant roles in tumorigenesis and development of CRC. Further, not only those bacterium genera individually, the association between *Bacteroides* genus and *Fusobacterium* genus is reported (Viljoen, Dakshinamurthy, Goldberg, & Blackburn, 2015; Joo et al., 2024; Duy et al., 2024; Conde-Pérez et al., 2024), suggesting possible contribution to CRC pathogenesis through mechanisms such as biofilm formation, immune evasion, and/or metabolic interactions with other dysbiotic taxa. Given that *Fusobacterium* genus has been shown to co-aggregate with *Bacteroides* genus, it is plausible that *Cutibacterium* genus might interact with these genera to influence inflammation, epithelial barrier integrity, and tumor progression. Thus, further studies integrating functional metagenomics, metabolomics, and host-microbiome interactions are warranted to elucidate the precise role of *Cutibacterium* genus and its relationship with CRC-associated microbial networks.

Analysis of eukaryotic and fungal microbial compositions revealed that the *Toxoplasma* genus was prevalent in both normal and tumor samples (Figure 22b), while the *Malassezia* genus was more prevalent in tumor samples (Figure 22c). The consistent presence of *Toxoplasma* genus across both sample types suggests that this intracellular pathogen may be a stable component of the gut microbiome, although its role in CRC pathogenesis remains unclear (Yu et al., 2020; Zavareh et al., 2021). In contrast, the increase

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

In normal tissue samples, *Roseolovirus* genus was the most prevalent viral taxon (Figure 22d), indicating its stable presence in the gut virome of healthy colonic tissues. However, in tumor tissue samples, *Lymphocryptovirus* and *Cytomegalovirus* genera were more prevalent viral taxa, suggesting an alteration in viral community structure associated with CRC progression. This viral compositional shift aligns with the Anna Karenina principle (Ma, 2020; W. Li & Yang, 2025), implying that microbial communities in diseased states exhibit greater instability and variability compare to their adjacent normal tissues. The emergence of *Lymphocryptovirus* (Mjelle, Castro, & Aass, 2025; De Flora & Bonanni, 2011) and *Cytomegalovirus* (Harkins et al., 2002; Taher et al., 2014; Bender et al., 2009) genera in tumor samples raises the possibility that oncogenic viruses may contribute to CRC carcinogenesis by promoting chronic inflammation, immune modulation, and/or direct viral-host interactions affecting cellular transformation. Moreover, the detection of tumor-associated viral alterations in adjacent normal tissues supports the concept of field cancerization (Curtius et al., 2018; Rubio et al., 2022), where viral dysbiosis may extend beyond the tumor itself, creating a pro-tumorigenic microenvironment even before malignant transformation occurs. These findings underscore the potential impact of viral communities in CRC and highlight the requirement for further research into their functional roles in carcinogenesis of CRC.

Alpha-diversity indices revealed a significant increase in microbial diversity in tumor samples compared to its adjacent normal tissues (MWU test $p < 0.05$; Figure 23), suggesting CRC is associated with a more heterogeneous gut microbiome (Liu et al., 2021). The increase in alpha-diversity indices within tumor tissues may support the Anna Karenina principle and/or the concept of field cancerization, where microbial alterations extend beyond the tumor site and contribute to a pre-malignant microenvironment. The enrichment of distinct bacterial, eukaryotic, fungal, and viral taxa within tumor samples suggest that microbial dysbiosis in CRC is not limited to a single pathogenic genus or species but rather involves complex community-level changes.

Furthermore, alpha-diversity indices in relation to recurrence history revealed distinct microbial diversity patterns between normal and tumor tissue samples (Figure 24). In recurrence patients, tumor samples exhibited a greater increase in alpha-diversity indices compared to their adjacent normal tissues (11/12 indices, 92% indices; Figure 24), suggesting that a more heterogeneous microbial community may be linked to tumor aggressiveness and recurrence potential (Huo et al., 2022; Vigneswaran & Shogan, 2020). This trend aligns with a highly diverse but dysregulated microbiome in tumor samples may contribute to immune evasion, chronic inflammation, and tumor-promoting metabolic changes. In non-recurrence patients, although tumor samples still exhibited increased alpha-diversity indices compared to normal tissues, the difference was less pronounced (8/12 indices, 67% indices; Figure 24), suggesting that a relatively more stable microbiome in tumor tissues may be associated with favorable

survival outcomes (Avuthu & Guda, 2022). These findings reinforce the concept that tumor microbiome changes are inconsistent across CRC patients, supporting the Anna Karenina principle. Additionally, the differences in alpha-diversity indices of normal tissues between recurrence and non-recurrence patients further suggest (Figure 24e and Figure 24g) that specific microbial communities may influence post-treatment disease progression.

Moreover, alpha-diversity indices and OS duration in CRC patients revealed distinct patterns between normal and tumor tissues (Figure 25), suggesting that microbial diversity in non-cancerous lesions may play a role in cancer prognosis (Galeano Niño et al., 2022). While no significant correlation was found between tumor-associated microbiome and OS duration, three of the 12 alpha-diversity indices exhibited negative correlations with OS in normal tissues (Figure 30b, Figure 30f, and Figure 30i), indicating that lower microbial heterogeneity in normal lesions was associated with longer survival. This finding suggests that a more heterogeneous microbial community in normal colon tissues may contribute to a microenvironment that fosters tumor progression, aligning with the field cancerization. Therefore, the negative correlations observed only in normal tissues suggests that pre-onset dysbiosis in non-cancerous regions could influence prognosis of CRC, potentially serving as an early indicator of cancer progression risk.

Beta-diversity indices revealed significant differences in gut microbiome compositions between tumor and normal tissues (Figure 26), aligning with the alpha-diversity indices and further confirming the presence of dysbiosis in gut microbiome of CRC. The distinct clustering of tumor and normal samples in beta-diversity indices (PERMANOVA $p < 0.001$) suggests that CRC is associated with a major alteration in microbial structure. This transformation may be driven by the expansion of tumor-associated taxa and the shrinkage of protective taxa, resulting in a tumor-supportive microenvironment. This clear separation in beta-diversity indices between tumor and normal tissues supports again the field cancerization, where microbial alterations extend beyond tumor lesions and affect surrounding non-cancerous lesions.

Furthermore, beta-diversity indices demonstrated significant microbial composition shifts between normal and tumor tissues in accordance with recurrence status (Figure 27), suggesting that dysbiosis in the gut microbiome may play an essential role in CRC progression and post-treatment recurrence. By the beta-diversity indices, the observed recurrence-associated microbial shifts highlight the potential of beta-diversity index as predictive markers for recurrence risk of CRC, warranting further studies to explore their functional significance and potential integration into microbiome-based prognostic models.

Moreover, beta-diversity indices suggested a potential association between the gut microbiome composition and OS in CRC patients (Figure 28), as distinct clustering were observed in relation to survival duration. However, due to the continuous nature of survival duration, direct statistical comparison using PERMANOVA test could be not performed, limiting the ability to formally quantify these differences. Despite this limitation, the observed separation of microbial communities along OS suggests that the gut microbiome composition may play a major role in CRC prognosis, potentially influencing immune response, tumor progression, and treatment outcomes. This lack of statistical validation highlights the need for alternative approaches to better assess the relationship between microbiome structure and survival outcome. Further investigation is required to determine whether specific microbial taxa drive these

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

To identify recurrence-related DAT in CRC, we applied ANCOM to compare the gut microbiome compositions between recurrence and non-recurrence patients (Table 9 and Figure 29). By applying ANCOM separately to total samples (Figure 29a), normal samples (Figure 29b), and tumor samples (Figure 29c), we identified both global and tissue-specific microbial shifts linked to CRC recurrence. Among these 19 recurrence-related DAT (Table 9), several DAT belonging to the *Micrococcus* and *Staphylococcus* genera were nominated as non-recurrence-enriched DAT. *Micrococcus* genus has been reported with anti-bacterial, anti-fungal, and anti-inflammatory activities (Tizabi & Hill, 2023), and another study has found that the production of carotenoid pigments from *Micrococcus luteus* (Figure 29e) exhibited promising antibiotics agents (Hegazy, Abu-Hussien, Elsenosy, El-Sayed, & Abo El-Naga, 2024). In this CRC study participants, *Cutibacterium acnes* was selected one of the non-recurrence-enriched DAT (Table 9). This finding is consistent with previous studies which have suggested that *Cutibacterium acnes* inhibits the activities of pathogens, such as *Staphylococcus aureus*, and suppresses tumor growth (Benej et al., 2024; Ding, Lian, Tam, & Oh, 2024). On the other hand, in this CRC study participants, many *Staphylococcus* species have chosen as non-recurrence-enriched DAT (Table 9); however, this outcome is contrary to previous studies which have described that cancer-promoting activity of *Staphylococcus aureus* (Z. Li, Zhuang, Wang, Wang, & Dong, 2021; Cuervo et al., 2010), suggesting opposite behaviors between *Staphylococcus aureus* and other *Staphylococcus* species. Last but not least, *Pseudomonas* sp. *NBRC 111133* has been found as the only recurrence-enriched DAT (Figure 29i). This also accords with earlier studies, which showed that *Pseudomonas aeruginosa* infections in cancer patients (Ohmagari et al., 2005; Paprocka et al., 2022).

To determine the OS-correlated DAT in CRC, we applied Spearman correlation to measure effects of the gut microbiome composition with OS (Table 10 and Figure 30). By implementing Spearman correlation to total samples (Figure 30a), normal samples (Figure 30b), and tumor samples (Figure 30c), we found that CRC survival is associated with both tissue type-specific and global microbial alterations. Among these 57 OS-correlated DAT (Table 30), several OS-correlated DAT from the *Corynebacterium* and *Staphylococcus* genera have significant correlations with survival duration of CRC. *Agaricus bisporus* has positive correlation with OS both in normal and tumor samples (Figure 30d). In accordance with this finding, previous studies have demonstrated that a polysaccharide produced from *Agaricus bisporus* exhibited anti-cancerous activity in colon cancer (Dong, Wang, Tang, Liu, & Gao, 2024; El-Deeb et al., 2022; N. Zhang, Liu, Tang, Yang, & Wang, 2023). Furthermore, most of *Corynebacterium* genus, including *Corynebacterium kroppenstedtii* (Figure 30f) and *Corynebacterium* sp. *KPL1824* (Figure 30h), have positive correlations with OS; however, *Corynebacterium lowii* (Figure 30g) has negative correlation with OS both in normal and tumor samples. Comparison of the findings with those of other studies confirms a breast cancer risk factor of *Corynebacterium afermentans* (J. An, Kwon, Oh, & Kim, 2025), an increasing of *Corynebacterium appendicis* in CRC (Hasan et al., 2022), an inhibition role of *Corynebacterium matruchotii* of cancer growth in oral squamous cell carcinoma (Shen et al., 2022), and a promoting cancer cell apoptosis of *Corynebacterium durum* (S. Kim et al., 2024), warranting future

investigations to selecting pro-tumoric and anti-tumoric species of *Corynebacterium* genus. *Clostridiales bacterium* has negative correlation with OS in normal samples (Figure 30e). However, this result does not support previous researches which have demonstrated that anti-cancer activities with immune modulation of *Clostridiales* genus (Montalban-Arques et al., 2021; Minton, 2003), suggesting different roles from *Clostridiales* species for immune response against cancer. Many species from *Staphylococcus* genus have positive correlations with OS (Table 10). Although, these results differ from some published studies which indicated cancer prevention and treatment via reduction of *Staphylococcus epidermidis* (Bernardo et al., 2023; Kepp, Zitvogel, & Kroemer, 2023), these results are consistent with other published researches which suggested that other species of *Staphylococcus* genus exhibited anti-cancer activities (Hassan, Mustafa, Rahim, & Isa, 2016; M. Zhang et al., 2022). Moreover, *Lachnospiraceae bacterium AD3010* and *Porphyromonas macacae* have negative correlations with OS in normal samples, while *Lachnospiraceae bacterium NK4A136* and *Paracoccus sphaerophysae* have negative correlations with OS in tumor samples (Table 10). Previous studies have addressed that high abundance of *Lachnospiraceae* genus in the gut microbiome showed anti-tumor roles in the CRC (Hexun et al., 2023; X. Zhang et al., 2023), indicating that more comprehensive investigation of species from *Lachnospiraceae* genus might be required. *Porphyromonas gingivalis*, a well-known periodontitis pathogens from *Porphyromonas* genus, was also reported promoting cancer resistance and development on CRC, lung cancer, and oesophageal cancer (León et al., 2007; Katz et al., 2009; S. Gao et al., 2021), providing a warrant to elucidate cancer-related roles of not only *Porphyromonas gingivalis* but also other *Porphyromonas* genus. *Paracoccus sphaerophysae* displayed negative correlation with OS in tumor samples and insignificantly negative correlation with OS (Spearman $|r| \leq 0.2$) in normal samples (Figure 30i), it is consistent with the literature which have shown *Paracoccus* genus is more prevalent in nasopharyngeal carcinoma group than healthy individuals (Lu et al., 2024).

One limitation of this study is the reliance on correlation analysis to evaluate association between microbiome features and overall survival, rather than utilizing more robust survival analysis methods such as the Kaplan-Meier estimator. While correlation analysis offers a simple approach to assess relationship between microbial abundance and survival duration as a continuous variable, it does not account for time-to-event distribution or censoring, both of which are fundamental in clinical survival data. In contrast, Kaplan-Meier analysis, along with log-rank testing, enables the comparison of survival curves between groups, providing clear interpretations of how specific microbiome profiles may stratify CRC patients by survival risk. The absence of Kaplan-Meier-based sub-grouping limits the ability to determine threshold-based clinical relevance of microbial biomarkers. Future studies should incorporate Kaplan-Meier plats, Cox proportional hazards models, and other survival-specific methods to strengthen the predictive and prognostic utility of microbiome data in survival outcomes.

To assess the predictive potential of recurrence-related DAT in CRC recurrence risk, we implemented a random forest classification model (Table 11 and Figure 31). The classification model achieved moderated classification performance (0.570 ± 0.164, mean$\pm$SD), indicating that while gut microbial provides some predictive value, it is likely insufficient as a standalone biomarker for recurrence risk of CRC. This limited predictive accuracy may be attributed to the complex and dynamic nature of gut microbiome

network, where epigenetic modifications and immune modulation collectively influence development and progression of CRC. Additionally, host-microbiome interactions, including metabolic pathways, may further contribute to recurrence of cancer, warranting a more integrative multi-omics appraoch. Therefore, future studies incorporating genomic sequencing data, *e.g.* somatic mutations and host immune signatures, could provide a more comprehensive understanding of how microbial dysbiosis interacts with tumor biology.

To evaluate the predictive potential of OS-related DAT in survival duration of CRC, we employed a random forest regression model (Table 12 and Figure 31). The regression model exhibited moderate regression error (729.302$\pm$179.940, mean$\pm$SD), suggesting that while gut microbiome composition provides some predictive values for cancer patient survival, it is likely influenced by additional host-specific and environmental factors. The complex interplay between the gut microbiome and CRC progression involves sophisticated microbial networks, metabolic interactions, and immune response, making it difficult to capture survival outcomes solely based on microbiome features. Furthermore, host-microbiome interactions, including MSI, tumor mutational burden, and epigenetic modifications, likely play a crucial role in determine favorable or unfavorable survival. These findings highlight the need for multi-omics integration, combining genomic sequencing data and metagenomic functional analysis, to gain deeper insights into how microbial dysbiosis interacts with tumor biology and clinical outcomes. Future studies incorporating machine learning models with multi-layered biological data may improve the accuracy of survival prediction and contribute to personalized medicine appraoches for CRC therapeutics.

5 Conclusion

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

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

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

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

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

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

References

Aagaard, K., Ma, J., Antony, K. M., Ganu, R., Petrosino, J., & Versalovic, J. (2014). The placenta harbors a unique microbiome. *Science translational medicine*, *6*(237), 237ra65–237ra65.

Abu-Ghazaleh, N., Chua, W. J., & Gopalan, V. (2021). Intestinal microbiota and its association with colon cancer and red/processed meat consumption. *Journal of gastroenterology and hepatology*, *36*(1), 75–88.

Abusleme, L., Hoare, A., Hong, B.-Y., & Diaz, P. I. (2021). Microbial signatures of health, gingivitis, and periodontitis. *Periodontology 2000*, *86*(1), 57–78.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., & Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Mathematical geology*, *32*, 271–275.

Aja, E., Mangar, M., Fletcher, H., & Mishra, A. (2021). Filifactor alocis: recent insights and advances. *Journal of dental research*, *100*(8), 790–797.

Alelyani, S. (2021). Stable bagging feature selection on medical data. *Journal of Big Data*, *8*(1), 11.

Altabtbaei, K., Maney, P., Ganesan, S. M., Dabdoub, S. M., Nagaraja, H. N., & Kumar, P. S. (2021). Anna karenina and the subgingival microbiome associated with periodontitis. *Microbiome*, *9*, 1–15.

Altıngöz, S. M., Kurgan, Ş., Önder, C., Serdar, M. A., Ünlütürk, U., Uyanık, M., … Günhan, M. (2021). Salivary and serum oxidative stress biomarkers and advanced glycation end products in periodontitis patients with or without diabetes: A cross-sectional study. *Journal of periodontology*, *92*(9), 1274–1285.

Alverdy, J., Hyoju, S., Weigerinck, M., & Gilbert, J. (2017). The gut microbiome and the mechanism of surgical infection. *Journal of British Surgery*, *104*(2), e14–e23.

An, J., Kwon, H., Oh, S.-Y., & Kim, Y. J. (2025). Association between breast cancer risk factors and blood microbiome in patients with breast cancer. *Scientific Reports*, *15*(1), 6115.

An, S., & Park, S. (2022). Association of physical activity and sedentary behavior with the risk of colorectal cancer. *Journal of Korean Medical Science*, *37*(19).

Anderson, M. J. (2014). Permutational multivariate analysis of variance (permanova). *Wiley statsref: statistics reference online*, 1–15.

Aruni, A. W., Mishra, A., Dou, Y., Chioma, O., Hamilton, B. N., & Fletcher, H. M. (2015). Filifactor alocis–a new emerging periodontal pathogen. *Microbes and infection*, *17*(7), 517–530.

Avuthu, N., & Guda, C. (2022). Meta-analysis of altered gut microbiota reveals microbial and metabolic biomarkers for colorectal cancer. *Microbiology Spectrum*, *10*(4), e00013–22.

Aziz, Q., & Thompson, D. G. (1998). Brain-gut axis in health and disease. *Gastroenterology*, *114*(3), 559–578.

Bai, X., Wei, H., Liu, W., Coker, O. O., Gou, H., Liu, C., ... others (2022). Cigarette smoke promotes colorectal cancer through modulation of gut microbiota and related metabolites. *Gut*, *71*(12), 2439–2450.

Baldelli, V., Scaldaferri, F., Putignani, L., & Del Chierico, F. (2021). The role of enterobacteriaceae in gut microbiota dysbiosis in inflammatory bowel diseases. *Microorganisms*, *9*(4), 697.

Bardou, M., Rouland, A., Martel, M., Loffroy, R., Barkun, A. N., & Chapelle, N. (2022). Obesity and colorectal cancer. *Alimentary Pharmacology & Therapeutics*, *56*(3), 407–418.

Barlow, G. M., Yu, A., & Mathur, R. (2015). Role of the gut microbiome in obesity and diabetes mellitus. *Nutrition in clinical practice*, *30*(6), 787–797.

Basavaprabhu, H., Sonu, K., & Prabha, R. (2020). Mechanistic insights into the action of probiotics against bacterial vaginosis and its mediated preterm birth: An overview. *Microbial pathogenesis*, *141*, 104029.

Belstrøm, D., Constancias, F., Drautz-Moses, D. I., Schuster, S. C., Veleba, M., Mahé, F., & Givskov, M. (2021). Periodontitis associates with species-specific gene expression of the oral microbiota. *npj Biofilms and Microbiomes*, *7*(1), 76.

Bender, C., Zipeto, D., Bidoia, C., Costantini, S., Zamò, A., Menestrina, F., & Bertazzoni, U. (2009). Analysis of colorectal cancers for human cytomegalovirus presence. *Infectious agents and cancer*, *4*, 1–6.

Benej, M., Hoyd, R., Kreamer, M., Wheeler, C. E., Grencewicz, D. J., Choueiry, F., ... others (2024). The tumor microbiome reacts to hypoxia and can influence response to radiation treatment in colorectal cancer. *Cancer research communications*, *4*(7), 1690–1701.

Berger, W. H., & Parker, F. L. (1970). Diversity of planktonic foraminifera in deep-sea sediments. *Science*, *168*(3937), 1345–1347.

Berghella, V. (2012). Universal cervical length screening for prediction and prevention of preterm birth. *Obstetrical & gynecological survey*, *67*(10), 653–657.

Bernardo, G., Le Noci, V., Ottaviano, E., De Cecco, L., Camisaschi, C., Guglielmetti, S., ... others (2023). Reduction of staphylococcus epidermidis in the mammary tumor microbiota induces antitumor immunity and decreases breast cancer aggressiveness. *Cancer Letters*, *555*, 216041.

Blencowe, H., Cousens, S., Oestergaard, M. Z., Chou, D., Moller, A.-B., Narwal, R., ... others (2012). National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The lancet*, *379*(9832), 2162–2172.

Boland, C. R., & Goel, A. (2010). Microsatellite instability in colorectal cancer. *Gastroenterology*, *138*(6), 2073–2087.

Boleij, A., Hechenbleikner, E. M., Goodwin, A. C., Badani, R., Stein, E. M., Lazarev, M. G., ... others (2015). The bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clinical Infectious Diseases*, *60*(2), 208–215.

Bolstad, A., Jensen, H. B., & Bakken, V. (1996). Taxonomy, biology, and periodontal aspects of fusobacterium nucleatum. *Clinical microbiology reviews*, *9*(1), 55–71.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... others (2019). Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature biotechnology*, *37*(8), 852–857.

Bombin, A., Yan, S., Bombin, S., Mosley, J. D., & Ferguson, J. F. (2022). Obesity influences composition of salivary and fecal microbiota and impacts the interactions between bacterial taxa. *Physiological reports*, *10*(7), e15254.

Bonnet, M., Buc, E., Sauvanet, P., Darcha, C., Dubois, D., Pereira, B., ... Darfeuille-Michaud, A. (2014). Colonization of the human gut by e. coli and colorectal cancer risk. *Clinical Cancer Research*, *20*(4), 859–867.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Brennan, C. A., & Garrett, W. S. (2019). Fusobacterium nucleatum—symbiont, opportunist and oncobacterium. *Nature Reviews Microbiology*, *17*(3), 156–166.

Broom, L. J., & Kogut, M. H. (2018). The role of the gut microbiome in shaping the immune system of chickens. *Veterinary immunology and immunopathology*, *204*, 44–51.

Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern recognition*, *36*(6), 1291–1302.

Bullman, S., Pedamallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., ... others (2017). Analysis of fusobacterium persistence and antibiotic response in colorectal cancer. *Science*, *358*(6369), 1443–1448.

Burt, R. W., Leppert, M. F., Slattery, M. L., Samowitz, W. S., Spirio, L. N., Kerber, R. A., ... others (2004). Genetic testing and phenotype in a large kindred with attenuated familial adenomatous polyposis. *Gastroenterology*, *127*(2), 444–451.

Cai, Y., Li, Y., Xiong, Y., Geng, X., Kang, Y., & Yang, Y. (2024). Diabetic foot exacerbates gut mycobiome dysbiosis in adult patients with type 2 diabetes mellitus: revealing diagnostic markers. *Nutrition & Diabetes*, *14*(1), 71.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods*, *13*(7), 581–583.

Canakci, V., & Canakci, C. F. (2007). Pain levels in patients during periodontal probing and mechanical non-surgical therapy. *Clinical oral investigations*, *11*, 377–383.

Cappellato, M., Baruzzo, G., & Di Camillo, B. (2022). Investigating differential abundance methods in microbiome data: A benchmark study. *PLoS computational biology*, *18*(9), e1010467.

Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., ... others (2016). The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic acids research*, *44*(D1), D471–D480.

Caspi, R., Billington, R., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., ... others (2018). The metacyc database of metabolic pathways and enzymes. *Nucleic acids research*, *46*(D1),

D633–D639.

Castaner, O., Goday, A., Park, Y.-M., Lee, S.-H., Magkos, F., Shiow, S.-A. T. E., & Schröder, H. (2018). The gut microbiome profile in obesity: a systematic review. *International journal of endocrinology*, *2018*(1), 4095789.

Center, M. M., Jemal, A., Smith, R. A., & Ward, E. (2009). Worldwide variations in colorectal cancer. *CA: a cancer journal for clinicians*, *59*(6), 366–378.

Centor, R. M. (1991). Signal detectability: the use of roc curves and their analyses. *Medical decision making*, *11*(2), 102–106.

Cerqueira, F. M., Photenhauer, A. L., Pollet, R. M., Brown, H. A., & Koropatkin, N. M. (2020). Starch digestion by gut bacteria: crowdsourcing for carbs. *Trends in Microbiology*, *28*(2), 95–108.

Champagne, C., McNairn, H., Daneshfar, B., & Shang, J. (2014). A bootstrap method for assessing classification accuracy and confidence for agricultural land use mapping in canada. *International Journal of Applied Earth Observation and Geoinformation*, *29*, 44–52.

Chang, H.-J., Lee, S.-J., Yong, T.-H., Shin, N.-Y., Jang, B.-G., Kim, J.-E., ... others (2020). Deep learning hybrid method to automatically diagnose periodontal bone loss and stage periodontitis. *Scientific reports*, *10*(1), 7531.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, 265–270.

Chao, A., & Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the American statistical Association*, *87*(417), 210–217.

Chapple, I. L., Mealey, B. L., Van Dyke, T. E., Bartold, P. M., Dommisch, H., Eickholz, P., ... others (2018). Periodontal health and gingival diseases and conditions on an intact and a reduced periodontium: Consensus report of workgroup 1 of the 2017 world workshop on the classification of periodontal and peri-implant diseases and conditions. *Journal of periodontology*, *89*, S74–S84.

Chen, T., Marsh, P., & Al-Hebshi, N. (2022). Smdi: an index for measuring subgingival microbial dysbiosis. *Journal of dental research*, *101*(3), 331–338.

Chen, T., Yu, W.-H., Izard, J., Baranova, O. V., Lakshmanan, A., & Dewhirst, F. E. (2010). The human oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database*, *2010*.

Chen, X., D'Souza, R., & Hong, S.-T. (2013). The role of gut microbiota in the gut-brain axis: current challenges and perspectives. *Protein & cell*, *4*, 403–414.

Chen, X., Jansen, L., Guo, F., Hoffmeister, M., Chang-Claude, J., & Brenner, H. (2021). Smoking, genetic predisposition, and colorectal cancer risk. *Clinical and translational gastroenterology*, *12*(3), e00317.

Chen, X., Li, H., Guo, F., Hoffmeister, M., & Brenner, H. (2022). Alcohol consumption, polygenic risk score, and early-and late-onset colorectal cancer risk. *EClinicalMedicine*, *49*.

Chew, R. J. J., Tan, K. S., Chen, T., Al-Hebshi, N. N., & Goh, C. E. (2024). Quantifying periodontitis-associated oral dysbiosis in tongue and saliva microbiomes—an integrated data analysis. *Journal of Periodontology*.

Čižmárová, B., Tomečková, V., Hubková, B., Hurajtová, A., Ohlasová, J., & Birková, A. (2022). Salivary redox homeostasis in human health and disease. *International Journal of Molecular Sciences*, *23*(17), 10076.

Conde-Pérez, K., Aja-Macaya, P., Buetas, E., Trigo-Tasende, N., Nasser-Ali, M., Rumbo-Feal, S., ... others (2024). The multispecies microbial cluster of fusobacterium, parvimonas, bacteroides and faecalibacterium as a precision biomarker for colorectal cancer diagnosis. *Molecular Oncology*, *18*(5), 1093–1122.

Cuervo, S. I., Cortés, J. A., Sánchez, R., Rodríguez, J. Y., Silva, E., Tibavizco, D., & Arroyo, P. (2010). Risk factors for mortality caused by staphylococcus aureus bacteremia in cancer patients. *Enfermedades infecciosas y microbiologia clinica*, *28*(6), 349–354.

Cullin, N., Antunes, C. A., Straussman, R., Stein-Thoeringer, C. K., & Elinav, E. (2021). Microbiome and cancer. *Cancer Cell*, *39*(10), 1317–1341.

Curtius, K., Wright, N. A., & Graham, T. A. (2018). An evolutionary perspective on field cancerization. *Nature Reviews Cancer*, *18*(1), 19–32.

Dabke, K., Hendrick, G., Devkota, S., et al. (2019). The gut microbiome and metabolic syndrome. *The Journal of clinical investigation*, *129*(10), 4050–4057.

De Flora, S., & Bonanni, P. (2011). The prevention of infection-associated cancers. *Carcinogenesis*, *32*(6), 787–795.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... Andersen, G. L. (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, *72*(7), 5069–5072.

Ding, R., Lian, S. B., Tam, Y. C., & Oh, C. C. (2024). The cutaneous microbiome in skin cancer–a systematic review. *JDDG: Journal der Deutschen Dermatologischen Gesellschaft*, *22*(2), 177–184.

Dong, K., Wang, J., Tang, F., Liu, Y., & Gao, L. (2024). A polysaccharide with a triple helix structure from agaricus bisporus: Characterization and anti-colon cancer activity. *International Journal of Biological Macromolecules*, *281*, 136521.

Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., ... Langille, M. G. (2020). Picrust2 for prediction of metagenome functions. *Nature biotechnology*, *38*(6), 685–688.

Doyle, R., Alber, D., Jones, H., Harris, K., Fitzgerald, F., Peebles, D., & Klein, N. (2014). Term and preterm labour are associated with distinct microbial community structures in placental membranes which are independent of mode of delivery. *Placenta*, *35*(12), 1099–1101.

Duy, T. N., Le Huy, H., Thanh, Q. Đ., Thi, H. N., Minh, H. N. T., Dang, M. N., ... Tat, T. N. (2024). Association between bacteroides fragilis and fusobacterium nucleatum infection and colorectal cancer in vietnamese patients. *Anaerobe*, *88*, 102880.

El-Deeb, N. M., Ibrahim, O. M., Mohamed, M. A., Farag, M. M., Farrag, A. A., & El-Aassar, M. (2022). Alginate/κ-carrageenan oral microcapsules loaded with agaricus bisporus polysaccharides mh751906 for natural killer cells mediated colon cancer immunotherapy. *International Journal of Biological Macromolecules*, *205*, 385–395.

Fahmy, C. A., Gamal-Eldeen, A. M., El-Hussieny, E. A., Raafat, B. M., Mehanna, N. S., Talaat, R. M., & Shaaban, M. T. (2019). Bifidobacterium longum suppresses murine colorectal cancer through the modulation of oncomirs and tumor suppressor mirnas. *Nutrition and cancer*, *71*(4), 688–700.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological conservation*, *61*(1), 1–10.

Fettweis, J. M., Serrano, M. G., Brooks, J. P., Edwards, D. J., Girerd, P. H., Parikh, H. I., ... others (2019). The vaginal microbiome and preterm birth. *Nature medicine*, *25*(6), 1012–1021.

Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 42–58.

Flanagan, L., Schmid, J., Ebert, M., Soucek, P., Kunicka, T., Liska, V., ... others (2014). Fusobacterium nucleatum associates with stages of colorectal neoplasia development, colorectal cancer and disease outcome. *European journal of clinical microbiology & infectious diseases*, *33*, 1381–1390.

Fortenberry, J. D. (2013). The uses of race and ethnicity in human microbiome research. *Trends in microbiology*, *21*(4), 165–166.

Francescone, R., Hou, V., & Grivennikov, S. I. (2014). Microbiome, inflammation, and cancer. *The Cancer Journal*, *20*(3), 181–189.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367–378.

Fushiki, T. (2011). Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, *21*, 137–146.

Galeano Niño, J. L., Wu, H., LaCourse, K. D., Kempchinsky, A. G., Baryiames, A., Barber, B., ... others (2022). Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer. *Nature*, *611*(7937), 810–817.

Gambin, D. J., Vitali, F. C., De Carli, J. P., Mazzon, R. R., Gomes, B. P., Duque, T. M., & Trentin, M. S. (2021). Prevalence of red and orange microbial complexes in endodontic-periodontal lesions: a systematic review and meta-analysis. *Clinical Oral Investigations*, 1–14.

Gao, J., Yin, J., Xu, K., Li, T., & Yin, Y. (2019). What is the impact of diet on nutritional diarrhea associated with gut microbiota in weaning piglets: a system review. *BioMed research international*, *2019*(1), 6916189.

Gao, R., Kong, C., Li, H., Huang, L., Qu, X., Qin, N., & Qin, H. (2017). Dysbiosis signature of mycobiota in colon polyp and colorectal cancer. *European Journal of Clinical Microbiology & Infectious Diseases*, *36*, 2457–2468.

Gao, S., Liu, Y., Duan, X., Liu, K., Mohammed, M., Gu, Z., ... others (2021). Porphyromonas gingivalis infection exacerbates oesophageal cancer and promotes resistance to neoadjuvant chemotherapy. *British Journal of Cancer*, *125*(3), 433–444.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*, 3–42.

Ghanavati, R., Akbari, A., Mohammadi, F., Asadollahi, P., Javadi, A., Talebi, M., & Rohani, M. (2020). Lactobacillus species inhibitory effect on colorectal cancer progression through modulating the

wnt/β-catenin signaling pathway. *Molecular and Cellular Biochemistry*, *470*, 1–13.

Ghojogh, B., & Crowley, M. (2019). The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. *arXiv preprint arXiv:1905.12787*.

Ghorbani, E., Avan, A., Ryzhikov, M., Ferns, G., Khazaei, M., & Soleimanpour, S. (2022). Role of lactobacillus strains in the management of colorectal cancer: An overview of recent advances. *Nutrition*, *103*, 111828.

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current understanding of the human microbiome. *Nature medicine*, *24*(4), 392–400.

Gini, C. (1912). Variabilità e mutabilità (variability and mutability). *Tipografia di Paolo Cuppini, Bologna, Italy*, *156*.

Goldenberg, R. L., Culhane, J. F., Iams, J. D., & Romero, R. (2008). Epidemiology and causes of preterm birth. *The lancet*, *371*(9606), 75–84.

Gonçalves, L., Subtil, A., Oliveira, M. R., & de Zea Bermudez, P. (2014). Roc curve estimation: An overview. *REVSTAT-Statistical journal*, *12*(1), 1–20.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*(3-4), 237–264.

Goodyear, M. D., Krleza-Jeric, K., & Lemmens, T. (2007). *The declaration of helsinki* (Vol. 335) (No. 7621). British Medical Journal Publishing Group.

Haffajee, A., Teles, R., & Socransky, S. (2006). Association of eubacterium nodatum and treponema denticola with human periodontitis lesions. *Oral microbiology and immunology*, *21*(5), 269–282.

Hajishengallis, G. (2015). Periodontitis: from microbial immune subversion to systemic inflammation. *Nature reviews immunology*, *15*(1), 30–44.

Hamjane, N., Mechita, M. B., Nourouti, N. G., & Barakat, A. (2024). Gut microbiota dysbiosis-associated obesity and its involvement in cardiovascular diseases and type 2 diabetes. a systematic review. *Microvascular Research*, *151*, 104601.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*, *29*(2), 147–160.

Hampel, H., Frankel, W. L., Martin, E., Arnold, M., Khanduja, K., Kuebler, P., ... others (2008). Feasibility of screening for lynch syndrome among patients with colorectal cancer. *Journal of Clinical Oncology*, *26*(35), 5783–5788.

Han, Y. W. (2015). Fusobacterium nucleatum: a commensal-turned pathogen. *Current opinion in microbiology*, *23*, 141–147.

Han, Y. W., & Wang, X. (2013). Mobile microbiome: oral bacteria in extra-oral infections and inflammation. *Journal of dental research*, *92*(6), 485–491.

Hand, D. J. (2012). Assessing the performance of classification methods. *International Statistical Review*, *80*(3), 400–414.

Harkins, L., Volk, A. L., Samanta, M., Mikolaenko, I., Britt, W. J., Bland, K. I., & Cobbs, C. S. (2002). Specific localisation of human cytomegalovirus nucleic acids and proteins in human colorectal cancer. *The Lancet*, *360*(9345), 1557–1563.

Hartstra, A. V., Bouter, K. E., Bäckhed, F., & Nieuwdorp, M. (2015). Insights into the role of the microbiome in obesity and type 2 diabetes. *Diabetes care*, *38*(1), 159–165.

Hasan, R., Bose, S., Roy, R., Paul, D., Rawat, S., Nilwe, P., ... Choudhury, S. (2022). Tumor tissue-specific bacterial biomarker panel for colorectal cancer: Bacteroides massiliensis, alistipes species, alistipes onderdonkii, bifidobacterium pseudocatenulatum, corynebacterium appendicis. *Archives of microbiology*, *204*(6), 348.

Hashemi Goradel, N., Heidarzadeh, S., Jahangiri, S., Farhood, B., Mortezaee, K., Khanlarkhani, N., & Negahdari, B. (2019). Fusobacterium nucleatum and colorectal cancer: A mechanistic overview. *Journal of Cellular Physiology*, *234*(3), 2337–2344.

Hassan, Z., Mustafa, S., Rahim, R. A., & Isa, N. M. (2016). Anti-breast cancer effects of live, heat-killed and cytoplasmic fractions of enterococcus faecalis and staphylococcus hominis isolated from human breast milk. *In Vitro Cellular & Developmental Biology-Animal*, *52*, 337–348.

Hegazy, A. A., Abu-Hussien, S. H., Elsenosy, N. K., El-Sayed, S. M., & Abo El-Naga, M. Y. (2024). Optimization, characterization and biosafety of carotenoids produced from whey using micrococcus luteus. *BMC biotechnology*, *24*(1), 74.

Heip, C. (1974). A new index measuring evenness. *Journal of the Marine Biological Association of the United Kingdom*, *54*(3), 555–557.

Helmink, B. A., Khan, M. W., Hermann, A., Gopalakrishnan, V., & Wargo, J. A. (2019). The microbiome, cancer, and cancer therapy. *Nature medicine*, *25*(3), 377–388.

Hexun, Z., Miyake, T., Maekawa, T., Mori, H., Yasukawa, D., Ohno, M., ... Tani, M. (2023). High abundance of lachnospiraceae in the human gut microbiome is related to high immunoscores in advanced colorectal cancer. *Cancer Immunology, Immunotherapy*, *72*(2), 315–326.

Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, *54*(2), 427–432.

Hiranmayi, K. V., Sirisha, K., Rao, M. R., & Sudhakar, P. (2017). Novel pathogens in periodontal microbiology. *Journal of Pharmacy and Bioallied Sciences*, *9*(3), 155–163.

Honda, K., & Littman, D. R. (2012). The microbiome in infectious disease and inflammation. *Annual review of immunology*, *30*(1), 759–795.

Honest, H., Forbes, C., Durée, K., Norman, G., Duffy, S., Tsourapas, A., ... others (2009). Screening to prevent spontaneous preterm birth: systematic reviews of accuracy and effectiveness literature with economic modelling. *Health Technol Assess*, *13*(43), 1–627.

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

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, *5*(2), 1.

Huang, R.-Y., Lin, C.-D., Lee, M.-S., Yeh, C.-L., Shen, E.-C., Chiang, C.-Y., ... Fu, E. (2007). Mandibular disto-lingual root: a consideration in periodontal therapy. *Journal of periodontology*, *78*(8), 1485–1490.

Huo, R.-X., Wang, Y.-J., Hou, S.-B., Wang, W., Zhang, C.-Z., & Wan, X.-H. (2022). Gut mucosal

microbiota profiles linked to colorectal cancer recurrence. *World journal of gastroenterology*, *28*(18), 1946.

Iams, J. D., & Berghella, V. (2010). Care for women with prior preterm birth. *American journal of obstetrics and gynecology*, *203*(2), 89–100.

Ide, M., & Papapanou, P. N. (2013). Epidemiology of association between maternal periodontal disease and adverse pregnancy outcomes–systematic review. *Journal of clinical periodontology*, *40*, S181–S194.

Iniesta, M., Chamorro, C., Ambrosio, N., Marín, M. J., Sanz, M., & Herrera, D. (2023). Subgingival microbiome in periodontal health, gingivitis and different stages of periodontitis. *Journal of Clinical Periodontology*, *50*(7), 905–920.

Inra, J. A., Steyerberg, E. W., Grover, S., McFarland, A., Syngal, S., & Kastrinos, F. (2015). Racial variation in frequency and phenotypes of apc and mutyh mutations in 6,169 individuals undergoing genetic testing. *Genetics in Medicine*, *17*(10), 815–821.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, *44*, 223–270.

Janda, J. M., & Abbott, S. L. (2007). 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, *45*(9), 2761–2764.

Jiang, W., & Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. *Statistics in medicine*, *26*(29), 5320–5334.

John, G. K., & Mullin, G. E. (2016). The gut microbiome and obesity. *Current oncology reports*, *18*, 1–7.

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., ... others (2019). Evaluation of 16s rrna gene sequencing for species and strain-level microbiome analysis. *Nature communications*, *10*(1), 5029.

Joo, J. E., Chu, Y. L., Georgeson, P., Walker, R., Mahmood, K., Clendenning, M., ... others (2024). Intratumoral presence of the genotoxic gut bacteria pks+ e. coli, enterotoxigenic bacteroides fragilis, and fusobacterium nucleatum and their association with clinicopathological and molecular features of colorectal cancer. *British Journal of Cancer*, *130*(5), 728–740.

Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N., & Whiteley, M. (2014). Metatranscriptomics of the human oral microbiome during health and disease. *MBio*, *5*(2), 10–1128.

Joscelyn, J., & Kasper, L. H. (2014). Digesting the emerging role for the gut microbiome in central nervous system demyelination. *Multiple Sclerosis Journal*, *20*(12), 1553–1559.

Kang, Y., Kang, X., Yang, H., Liu, H., Yang, X., Liu, Q., ... others (2022). Lactobacillus acidophilus ameliorates obesity in mice through modulation of gut microbiota dysbiosis and intestinal permeability. *Pharmacological research*, *175*, 106020.

Karched, M., Bhardwaj, R. G., Qudeimat, M., Al-Khabbaz, A., & Ellepola, A. (2022). Proteomic analysis of the periodontal pathogen prevotella intermedia secretomes in biofilm and planktonic lifestyles. *Scientific Reports*, *12*(1), 5636.

Karp, P. D., Riley, M., Paley, S. M., & Pellegrini-Toole, A. (2002). The metacyc database. *Nucleic acids research*, *30*(1), 59–61.

Katz, J., Chegini, N., Shiverick, K., & Lamont, R. (2009). Localization of p. gingivalis in preterm delivery placenta. *Journal of dental research*, *88*(6), 575–578.

Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., & Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature*, *474*(7351), 327–336.

Kelly, B. J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J. D., Collman, R. G., ... Li, H. (2015). Power and sample-size estimation for microbiome studies using pairwise distances and permanova. *Bioinformatics*, *31*(15), 2461–2468.

Kennedy, J., Alexander, P., Taillie, L. S., & Jaacks, L. M. (2024). Estimated effects of reductions in processed meat consumption and unprocessed red meat consumption on occurrences of type 2 diabetes, cardiovascular disease, colorectal cancer, and mortality in the usa: a microsimulation study. *The Lancet Planetary Health*, *8*(7), e441–e451.

Kepp, O., Zitvogel, L., & Kroemer, G. (2023). *Prevention and treatment of cancers by tumor antigen-expressing staphylococcus epidermidis* (Vol. 12) (No. 1). Taylor & Francis.

Kim, B.-R., Shin, J., Guevarra, R. B., Lee, J. H., Kim, D. W., Seol, K.-H., ... Isaacson, R. E. (2017). Deciphering diversity indices for a better understanding of microbial communities. *Journal of Microbiology and Biotechnology*, *27*(12), 2089–2093.

Kim, C. H. (2018). Immune regulation by microbiome metabolites. *Immunology*, *154*(2), 220–229.

Kim, E.-H., Kim, S., Kim, H.-J., Jeong, H.-o., Lee, J., Jang, J., ... others (2020). Prediction of chronic periodontitis severity using machine learning models based on salivary bacterial copy number. *Frontiers in Cellular and Infection Microbiology*, *10*, 571515.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, *53*(11), 3735–3745.

Kim, S., Lee, M., Kim, N.-Y., Kwon, Y.-S., Nam, G. S., Lee, K., ... Hwang, I. H. (2024). Oxidative tryptamine dimers from corynebacterium durum directly target survivin to induce aif-mediated apoptosis in cancer cells. *Biomedicine & Pharmacotherapy*, *173*, 116335.

Kinane, D. F., Stathopoulou, P. G., & Papapanou, P. N. (2017). Periodontal diseases. *Nature reviews Disease primers*, *3*(1), 1–14.

Kindinger, L. M., Bennett, P. R., Lee, Y. S., Marchesi, J. R., Smith, A., Cacciatore, S., ... MacIntyre, D. A. (2017). The interaction between vaginal microbiota, cervical length, and vaginal progesterone treatment for preterm birth risk. *Microbiome*, *5*, 1–14.

Kogut, M. H., Lee, A., & Santin, E. (2020). Microbiome and pathogen interaction with the immune system. *Poultry science*, *99*(4), 1906–1913.

Kostic, A. D., Ojesina, A. I., Pedamallu, C. S., Jung, J., Verhaak, R. G., Getz, G., & Meyerson, M. (2011). Pathseq: software to identify or discover microbes by deep sequencing of human tissue. *Nature biotechnology*, *29*(5), 393–396.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, *26*, 159–190.

Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., ... Watanabe, T. (2015). Colorectal cancer. *Nature reviews. Disease primers*, *1*, 15065.

Lafaurie, G. I., Neuta, Y., Ríos, R., Pacheco-Montealegre, M., Pianeta, R., Castillo, D. M., ... others (2022). Differences in the subgingival microbiome according to stage of periodontitis: A comparison of two geographic regions. *PloS one*, *17*(8), e0273523.

Lamont, R. J., & Jenkinson, H. F. (2000). Subgingival colonization by porphyromonas gingivalis. *Oral Microbiology and Immunology: Mini-review*, *15*(6), 341–349.

Lamont, R. J., Koo, H., & Hajishengallis, G. (2018). The oral microbiota: dynamic communities and host interactions. *Nature reviews microbiology*, *16*(12), 745–759.

Leitich, H., & Kaider, A. (2003). Fetal fibronectin—how useful is it in the prediction of preterm birth? *BJOG: An International Journal of Obstetrics & Gynaecology*, *110*, 66–70.

Le Leu, R. K., Hu, Y., Brown, I. L., Woodman, R. J., & Young, G. P. (2010). Synbiotic intervention of bifidobacterium lactis and resistant starch protects against colorectal cancer development in rats. *Carcinogenesis*, *31*(2), 246–251.

León, R., Silva, N., Ovalle, A., Chaparro, A., Ahumada, A., Gajardo, M., ... Gamonal, J. (2007). Detection of porphyromonas gingivalis in the amniotic fluid in pregnant women with a diagnosis of threatened premature labor. *Journal of periodontology*, *78*(7), 1249–1255.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, *25*(14), 1754–1760.

Li, N., Lu, B., Luo, C., Cai, J., Lu, M., Zhang, Y., ... Dai, M. (2021). Incidence, mortality, survival, risk factor and screening of colorectal cancer: A comparison among china, europe, and northern america. *Cancer letters*, *522*, 255–268.

Li, R., Miao, Z., Liu, Y., Chen, X., Wang, H., Su, J., & Chen, J. (2024). The brain–gut–bone axis in neurodegenerative diseases: insights, challenges, and future prospects. *Advanced Science*, *11*(38), 2307971.

Li, W., & Yang, J. (2025). Investigating the anna karenina principle of the breast microbiome. *BMC microbiology*, *25*(1), 1–10.

Li, X., Yu, D., Wang, Y., Yuan, H., Ning, X., Rui, B., ... Li, M. (2021). The intestinal dysbiosis of mothers with gestational diabetes mellitus (gdm) and its impact on the gut microbiota of their newborns. *Canadian Journal of Infectious Diseases and Medical Microbiology*, *2021*(1), 3044534.

Li, Y., Qian, F., Cheng, X., Wang, D., Wang, Y., Pan, Y., ... Tian, Y. (2023). Dysbiosis of oral microbiota and metabolite profiles associated with type 2 diabetes mellitus. *Microbiology spectrum*, *11*(1), e03796–22.

Li, Z., Zhuang, H., Wang, G., Wang, H., & Dong, Y. (2021). Prevalence, predictors, and mortality of bloodstream infections due to methicillin-resistant staphylococcus aureus in patients with malignancy: systemic review and meta-analysis. *BMC infectious diseases*, *21*, 1–10.

Lim, J. W., Park, T., Tong, Y. W., & Yu, Z. (2020). The microbiome driving anaerobic digestion and microbial analysis. In *Advances in bioenergy* (Vol. 5, pp. 1–61). Elsevier.

Lin, H., & Peddada, S. D. (2020). Analysis of compositions of microbiomes with bias correction. *Nature*

communications, *11*(1), 3514.

Listgarten, M. A. (1986). Pathogenesis of periodontitis. *Journal of clinical periodontology*, *13*(5), 418–425.

Liu, W., Zhang, X., Xu, H., Li, S., Lau, H. C.-H., Chen, Q., ... others (2021). Microbial community heterogeneity within colorectal neoplasia and its correlation with colorectal carcinogenesis. *Gastroenterology*, *160*(7), 2395–2408.

Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome medicine*, *8*, 1–11.

López-Aladid, R., Fernández-Barat, L., Alcaraz-Serrano, V., Bueno-Freire, L., Vázquez, N., Pastor-Ibáñez, R., ... Torres, A. (2023). Determining the most accurate 16s rrna hypervariable region for taxonomic identification from respiratory samples. *Scientific reports*, *13*(1), 3974.

Louca, S., & Doebeli, M. (2018). Efficient comparative phylogenetics on large trees. *Bioinformatics*, *34*(6), 1053–1055.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, *15*, 1–21.

Lu, Y.-T., Hsin, C.-H., Chuang, C.-Y., Huang, C.-C., Su, M.-C., Wen, W.-S., ... others (2024). Microbial dysbiosis in nasopharyngeal carcinoma: A pilot study on biomarker potential. *Journal of Otolaryngology-Head & Neck Surgery*, *53*, 19160216241304365.

Ma, Z. S. (2020). Testing the anna karenina principle in human microbiome-associated diseases. *Iscience*, *23*(4).

Magnúsdóttir, S., & Thiele, I. (2018). Modeling metabolism of the human gut microbiome. *Current opinion in biotechnology*, *51*, 90–96.

Magurran, A. E. (2021). Measuring biological diversity. *Current Biology*, *31*(19), R1174–R1177.

Mandic, M., Safizadeh, F., Niedermaier, T., Hoffmeister, M., & Brenner, H. (2023). Association of overweight, obesity, and recent weight loss with colorectal cancer risk. *JAMA network Open*, *6*(4), e239556–e239556.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

Manolis, A. A., Manolis, T. A., Melita, H., & Manolis, A. S. (2022). Gut microbiota and cardiovascular disease: symbiosis versus dysbiosis. *Current Medicinal Chemistry*, *29*(23), 4050–4077.

Martin, C. R., Osadchiy, V., Kalani, A., & Mayer, E. A. (2018). The brain-gut-microbiome axis. *Cellular and molecular gastroenterology and hepatology*, *6*(2), 133–148.

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, *1*(2), 140–147.

Mayer, E. A., Tillisch, K., Gupta, A., et al. (2015). Gut/brain axis and the microbiota. *The Journal of clinical investigation*, *125*(3), 926–938.

Melguizo-Rodríguez, L., Costela-Ruiz, V. J., Manzano-Moreno, F. J., Ruiz, C., & Illescas-Montes, R. (2020). Salivary biomarkers and their application in the diagnosis and monitoring of the most common oral pathologies. *International journal of molecular sciences*, *21*(14), 5173.

Merrill, L. C., & Mangano, K. M. (2023). Racial and ethnic differences in studies of the gut microbiome and osteoporosis. *Current Osteoporosis Reports*, *21*(5), 578–591.

Miller, C. S., Ding, X., Dawson III, D. R., & Ebersole, J. L. (2021). Salivary biomarkers for discriminating periodontitis in the presence of diabetes. *Journal of clinical periodontology*, *48*(2), 216–225.

Minton, N. P. (2003). Clostridia in cancer therapy. *Nature Reviews Microbiology*, *1*(3), 237–242.

Mjelle, R., Castro, Í., & Aass, K. R. (2025). The viral landscape in metastatic solid cancers. *Heliyon*.

Montalban-Arques, A., Katkeviciute, E., Busenhart, P., Bircher, A., Wirbel, J., Zeller, G., ... others (2021). Commensal clostridiales strains mediate effective anti-cancer immune response against solid tumors. *Cell host & microbe*, *29*(10), 1573–1588.

Morita, T., Yamazaki, Y., Mita, A., Takada, K., Seto, M., Nishinoue, N., ... Maeno, M. (2010). A cohort study on the association between periodontal disease and the development of metabolic syndrome. *Journal of periodontology*, *81*(4), 512–519.

Na, H. S., Kim, S. Y., Han, H., Kim, H.-J., Lee, J.-Y., Lee, J.-H., & Chung, J. (2020). Identification of potential oral microbial biomarkers for the diagnosis of periodontitis. *Journal of clinical medicine*, *9*(5), 1549.

Nemoto, T., Shiba, T., Komatsu, K., Watanabe, T., Shimogishi, M., Shibasaki, M., ... others (2021). Discrimination of bacterial community structures among healthy, gingivitis, and periodontitis statuses through integrated metatranscriptomic and network analyses. *Msystems*, *6*(6), e00886–21.

Nesbitt, M. J., Reynolds, M. A., Shiau, H., Choe, K., Simonsick, E. M., & Ferrucci, L. (2010). Association of periodontitis and metabolic syndrome in the baltimore longitudinal study of aging. *Aging clinical and experimental research*, *22*, 238–242.

Network, C. G. A., et al. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*(7407), 330.

Nibali, L., Sousa, V., Davrandi, M., Spratt, D., Alyahya, Q., Dopico, J., & Donos, N. (2020). Differences in the periodontal microbiome of successfully treated and persistent aggressive periodontitis. *Journal of Clinical Periodontology*, *47*(8), 980–990.

Niederman, R., Buyle-Bodin, Y., Lu, B.-Y., Robinson, P., & Naleway, C. (1997). Short-chain carboxylic acid concentration in human gingival crevicular fluid. *Journal of dental research*, *76*(1), 575–579.

Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Ž., & Tomović, M. (2017). Evaluation of classification models in machine learning. *Theory and Applications of Mathematics & Computer Science*, *7*(1), 39.

Obuchowski, N. A., & Bullen, J. A. (2018). Receiver operating characteristic (roc) curves: review of methods with applications in diagnostic medicine. *Physics in Medicine & Biology*, *63*(7), 07TR01.

Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in japan and its neighbouring regions. *Bulletin of Japanese Society of Scientific Fisheries*, *22*, 526–530.

Offenbacher, S., Katz, V., Fertik, G., Collins, J., Boyd, D., Maynor, G., ... Beck, J. (1996). Periodontal infection as a possible risk factor for preterm low birth weight. *Journal of periodontology*, *67*, 1103–1113.

Ohmagari, N., Hanna, H., Graviss, L., Hackett, B., Perego, C., Gonzalez, V., ... others (2005). Risk

factors for infections with multidrug-resistant pseudomonas aeruginosa in patients with cancer. *Cancer*, *104*(1), 205–212.

Ojesina, A. I., Pedamallu, C. S., Kostic, A., Jung, J., Auclair, D., Lohr, J., ... Meyerson, M. (2013). High throughput sequencing-based pathogen discovery in multiple myeloma. *Blood*, *122*(21), 5322.

Omondiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019). Machine learning classification techniques for breast cancer diagnosis. In *Iop conference series: materials science and engineering* (Vol. 495, p. 012033).

O'Sullivan, D. E., Sutherland, R. L., Town, S., Chow, K., Fan, J., Forbes, N., ... Brenner, D. R. (2022). Risk factors for early-onset colorectal cancer: a systematic review and meta-analysis. *Clinical gastroenterology and hepatology*, *20*(6), 1229–1240.

Paganini, D., & Zimmermann, M. B. (2017). The effects of iron fortification and supplementation on the gut microbiome and diarrhea in infants and children: a review. *The American journal of clinical nutrition*, *106*, 1688S–1693S.

Pan, A. Y. (2021). Statistical analysis of microbiome data: the challenge of sparsity. *Current Opinion in Endocrine and Metabolic Research*, *19*, 35–40.

Papapanou, P. N., Sanz, M., Buduneli, N., Dietrich, T., Feres, M., Fine, D. H., ... others (2018). Periodontitis: Consensus report of workgroup 2 of the 2017 world workshop on the classification of periodontal and peri-implant diseases and conditions. *Journal of periodontology*, *89*, S173–S182.

Paprocka, P., Durnaś, B., Mańkowska, A., Król, G., Wollny, T., & Bucki, R. (2022). Pseudomonas aeruginosa infections in cancer patients. *Pathogens*, *11*(6), 679.

Parizadeh, M., & Arrieta, M.-C. (2023). The global human gut microbiome: genes, lifestyles, and diet. *Trends in Molecular Medicine*.

Park, J., Park, S. H., Lee, D., Lee, J. E., Lee, D., Na, K. J., ... Im, H.-J. (2024). Detecting cancer microbiota using unmapped rna reads on spatial transcriptomics. *Cancer Research*, *84*(6_Supplement), 4881–4881.

Parks, D. H., Tyson, G. W., Hugenholtz, P., & Beiko, R. G. (2014). Stamp: statistical analysis of taxonomic and functional profiles. *Bioinformatics*, *30*(21), 3123–3124.

Payne, M. S., Newnham, J. P., Doherty, D. A., Furfaro, L. L., Pendal, N. L., Loh, D. E., & Keelan, J. A. (2021). A specific bacterial dna signature in the vagina of australian women in midpregnancy predicts high risk of spontaneous preterm birth (the predict1000 study). *American journal of obstetrics and gynecology*, *224*(2), 206–e1.

Peirce, J. M., & Alviña, K. (2019). The role of inflammation and the gut microbiome in depression and anxiety. *Journal of neuroscience research*, *97*(10), 1223–1241.

Peltomaki, P. (2003). Role of dna mismatch repair defects in the pathogenesis of human cancer. *Journal of clinical oncology*, *21*(6), 1174–1179.

Pezzino, S., Sofia, M., Greco, L. P., Litrico, G., Filippello, G., Sarvà, I., ... Latteri, S. (2023). Microbiome dysbiosis: a pathological mechanism at the intersection of obesity and glaucoma. *International Journal of Molecular Sciences*, *24*(2), 1166.

Pollard, T. J., Johnson, A. E., Raffa, J. D., & Mark, R. G. (2018). tableone: An open source python

package for producing summary statistics for research papers. *JAMIA open*, *1*(1), 26–31.

Premaraj, T. S., Vella, R., Chung, J., Lin, Q., Hunter, P., Underwood, K., ... Zhou, Y. (2020). Ethnic variation of oral microbiota in children. *Scientific reports*, *10*(1), 14788.

Purcell, R. V., Permain, J., & Keenan, J. I. (2022). Enterotoxigenic bacteroides fragilis activates il-8 expression through stat3 in colorectal cancer cells. *Gut Pathogens*, *14*(1), 16.

Raut, J. R., Schöttker, B., Holleczek, B., Guo, F., Bhardwaj, M., Miah, K., ... Brenner, H. (2021). A microrna panel compared to environmental and polygenic scores for colorectal cancer risk prediction. *Nature Communications*, *12*(1), 4811.

Rebersek, M. (2021). Gut microbiome and its role in colorectal cancer. *BMC cancer*, *21*(1), 1325.

Redanz, U., Redanz, S., Treerat, P., Prakasam, S., Lin, L.-J., Merritt, J., & Kreth, J. (2021). Differential response of oral mucosal and gingival cells to corynebacterium durum, streptococcus sanguinis, and porphyromonas gingivalis multispecies biofilms. *Frontiers in cellular and infection microbiology*, *11*, 686479.

Relvas, M., Regueira-Iglesias, A., Balsa-Castro, C., Salazar, F., Pacheco, J., Cabral, C., ... Tomás, I. (2021). Relationship between dental and periodontal health status and the salivary microbiome: bacterial diversity, co-occurrence networks and predictive models. *Scientific reports*, *11*(1), 929.

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.

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.

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

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.

Romero, R., Dey, S. K., & Fisher, S. J. (2014). Preterm labor: one syndrome, many causes. *Science*, *345*(6198), 760–765.

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.

Rosan, B., & Lamont, R. J. (2000). Dental plaque formation. *Microbes and infection*, *2*(13), 1599–1607.

Rubio, C. A., Lang-Schwarz, C., & Vieth, M. (2022). Further study on field cancerization in the human colon. *Anticancer Research*, *42*(12), 5891–5895.

Schwabe, R. F., & Jobin, C. (2013). The microbiome and cancer. *Nature Reviews Cancer*, *13*(11), 800–812.

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.

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).

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.

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.

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.

Shi, N., Li, N., Duan, X., & Niu, H. (2017). Interaction between the gut microbiome and mucosal immune system. *Military Medical Research*, *4*, 1–7.

Simpson, E. (1949). Measurement of diversity. *Nature*, *163*.

Sokal, R. R., & Sneath, P. H. (1963). Principles of numerical taxonomy.

Song, M., Chan, A. T., & Sun, J. (2020). Influence of the gut microbiome, diet, and environment on risk of colorectal cancer. *Gastroenterology*, *158*(2), 322–340.

Söreide, K., Janssen, E., Söiland, H., Körner, H., & Baak, J. (2006). Microsatellite instability in colorectal cancer. *Journal of British Surgery*, *93*(4), 395–406.

Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, *5*, 1–34.

Sotiriadis, A., Papatheodorou, S., Kavvadias, A., & Makrydimas, G. (2010). Transvaginal cervical length measurement for prediction of preterm birth in women with threatened preterm labor: a meta-analysis. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, *35*(1), 54–64.

Spss, I., et al. (2011). Ibm spss statistics for windows, version 20.0. *New York: IBM Corp*, *440*, 394.

Stafford, G., Roy, S., Honma, K., & Sharma, A. (2012). Sialic acid, periodontal pathogens and tannerella forsythia: stick around and enjoy the feast! *Molecular Oral Microbiology*, *27*(1), 11–22.

Stout, M. J., Conlon, B., Landeau, M., Lee, I., Bower, C., Zhao, Q., ... Mysorekar, I. U. (2013). Identification of intracellular bacteria in the basal plate of the human placenta in term and preterm gestations. *American journal of obstetrics and gynecology*, *208*(3), 226–e1.

Strong, W. (2002). Assessing species abundance unevenness within and between plant communities. *Community Ecology*, *3*(2), 237–246.

Sultan, S., El-Mowafy, M., Elgaml, A., Ahmed, T. A., Hassan, H., & Mottawea, W. (2021). Metabolic influences of gut microbiota dysbiosis on inflammatory bowel disease. *Frontiers in physiology*, *12*, 715506.

Suzuki, N., Nakano, Y., Yoneda, M., Hirofuji, T., & Hanioka, T. (2022). The effects of cigarette smoking on the salivary and tongue microbiome. *Clinical and Experimental Dental Research*, *8*(1),

449–456.

Swidsinski, A., Khilkin, M., Kerjaschki, D., Schreiber, S., Ortner, M., Weber, J., & Lochs, H. (1998). Association between intraepithelial escherichia coli and colorectal cancer. *Gastroenterology*, *115*(2), 281–286.

Swift, D., Cresswell, K., Johnson, R., Stilianoudakis, S., & Wei, X. (2023). A review of normalization and differential abundance methods for microbiome counts data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *15*(1), e1586.

Szafrański, S. P., Deng, Z.-L., Tomasch, J., Jarek, M., Bhuju, S., Meisinger, C., ... Wagner-Döbler, I. (2015). Functional biomarkers for chronic periodontitis and insights into the roles of prevotella nigrescens and fusobacterium nucleatum; a metatranscriptome analysis. *npj Biofilms and Microbiomes*, *1*(1), 1–13.

Taher, C., Frisk, G., Fuentes, S., Religa, P., Costa, H., Assinger, A., ... others (2014). High prevalence of human cytomegalovirus in brain metastases of patients with primary breast and colorectal cancers. *Translational oncology*, *7*(6), 732–740.

Tanner, A. C., Kent Jr, R., Kanasi, E., Lu, S. C., Paster, B. J., Sonis, S. T., ... Van Dyke, T. E. (2007). Clinical characteristics and microbiota of progressing slight chronic periodontitis in adults. *Journal of clinical periodontology*, *34*(11), 917–930.

Tanner, A. C., Paster, B. J., Lu, S. C., Kanasi, E., Kent Jr, R., Van Dyke, T., & Sonis, S. T. (2006). Subgingival and tongue microbiota during early periodontitis. *Journal of dental research*, *85*(4), 318–323.

Tejeda, M., Farrell, J., Zhu, C., Haines, J. L., Wang, L.-S., Schellenberg, G. D., ... others (2021). Multiple viruses detected in human dna are associated with alzheimer disease risk. *Alzheimer's & Dementia*, *17*, e054585.

Teles, F., Wang, Y., Hajishengallis, G., Hasturk, H., & Marchesan, J. T. (2021). Impact of systemic factors in shaping the periodontal microbiome. *Periodontology 2000*, *85*(1), 126–160.

Thaiss, C. A., Zmora, N., Levy, M., & Elinav, E. (2016). The microbiome and innate immunity. *Nature*, *535*(7610), 65–74.

Tian, R., Liu, H., Feng, S., Wang, H., Wang, Y., Wang, Y., ... Zhang, S. (2021). Gut microbiota dysbiosis in stable coronary artery disease combined with type 2 diabetes mellitus influences cardiovascular prognosis. *Nutrition, Metabolism and Cardiovascular Diseases*, *31*(5), 1454–1466.

Tilg, H., Kaser, A., et al. (2011). Gut microbiome, obesity, and metabolic dysfunction. *The Journal of clinical investigation*, *121*(6), 2126–2132.

Tizabi, D., & Hill, R. T. (2023). Micrococcus spp. as a promising source for drug discovery: A review. *Journal of Industrial Microbiology and Biotechnology*, *50*(1), kuad017.

Tonetti, M. S., Greenwell, H., & Kornman, K. S. (2018). Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. *Journal of periodontology*, *89*, S159–S172.

Tringe, S. G., & Hugenholtz, P. (2008). A renaissance for the pioneering 16s rrna gene. *Current opinion in microbiology*, *11*(5), 442–446.

Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., ... others (2017). A

guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews*, *92*(2), 698–715.

Ulger Toprak, N., Yagci, A., Gulluoglu, B., Akin, M., Demirkalem, P., Celenk, T., & Soyletir, G. (2006). A possible role of bacteroides fragilis enterotoxin in the aetiology of colorectal cancer. *Clinical microbiology and infection*, *12*(8), 782–786.

Ursell, L. K., Metcalf, J. L., Parfrey, L. W., & Knight, R. (2012). Defining the human microbiome. *Nutrition reviews*, *70*(suppl_1), S38–S44.

Utzschneider, K. M., Kratz, M., Damman, C. J., & Hullarg, M. (2016). Mechanisms linking the gut microbiome and glucose metabolism. *The Journal of Clinical Endocrinology & Metabolism*, *101*(4), 1445–1454.

Vander Haar, E. L., So, J., Gyamfi-Bannerman, C., & Han, Y. W. (2018). Fusobacterium nucleatum and adverse pregnancy outcomes: epidemiological and mechanistic evidence. *Anaerobe*, *50*, 55–59.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Vasen, H. F., Mecklin, J.-P., Khan, P. M., & Lynch, H. T. (1991). The international collaborative group on hereditary non-polyposis colorectal cancer (icg-hnpcc). *Diseases of the Colon & Rectum*, *34*(5), 424–425.

Vigneswaran, J., & Shogan, B. D. (2020). The role of the intestinal microbiome on colorectal cancer pathogenesis and its recurrence following surgery. *Journal of Gastrointestinal Surgery*, *24*(10), 2349–2356.

Vilar, E., & Gruber, S. B. (2010). Microsatellite instability in colorectal cancer—the stable evidence. *Nature reviews Clinical oncology*, *7*(3), 153–162.

Viljoen, K. S., Dakshinamurthy, A., Goldberg, P., & Blackburn, J. M. (2015). Quantitative profiling of colorectal cancer-associated bacteria reveals associations between fusobacterium spp., enterotoxigenic bacteroides fragilis (etbf) and clinicopathological features of colorectal cancer. *PloS one*, *10*(3), e0119462.

Walker, M. A., Pedamallu, C. S., Ojesina, A. I., Bullman, S., Sharpe, T., Whelan, C. W., & Meyerson, M. (2018). Gatk pathseq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. *Bioinformatics*, *34*(24), 4287–4289.

Wang, N., & Fang, J.-Y. (2023). Fusobacterium nucleatum, a key pathogenic factor and microbial biomarker for colorectal cancer. *Trends in Microbiology*, *31*(2), 159–172.

Weaver, W. (1963). *The mathematical theory of communication*. University of Illinois Press.

Whiteside, S. A., Razvi, H., Dave, S., Reid, G., & Burton, J. P. (2015). The microbiome of the urinary tract—a role beyond infection. *Nature Reviews Urology*, *12*(2), 81–90.

Witkin, S. (2019). Vaginal microbiome studies in pregnancy must also analyse host factors. *BJOG: An International Journal of Obstetrics & Gynaecology*, *126*(3), 359–359.

Wong, T.-T., & Yeh, P.-Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, *32*(8), 1586–1594.

Wyss, C., Moter, A., Choi, B.-K., Dewhirst, F., Xue, Y., Schüpbach, P., ... Guggenheim, B. (2004).

Treponema putidum sp. nov., a medium-sized proteolytic spirochaete isolated from lesions of human periodontitis and acute necrotizing ulcerative gingivitis. *International journal of systematic and evolutionary microbiology*, *54*(4), 1117–1122.

Xia, Y. (2023). Statistical normalization methods in microbiome data with application to microbiome cancer research. *Gut Microbes*, *15*(2), 2244139.

Yaman, E., & Subasi, A. (2019). Comparison of bagging and boosting ensemble machine learning methods for automated emg signal classification. *BioMed research international*, *2019*(1), 9152506.

Yang, I., Claussen, H., Arthur, R. A., Hertzberg, V. S., Geurs, N., Corwin, E. J., & Dunlop, A. L. (2022). Subgingival microbiome in pregnancy and a potential relationship to early term birth. *Frontiers in cellular and infection microbiology*, *12*, 873683.

Ye, Y., & Doak, T. G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS computational biology*, *5*(8), e1000465.

Yildiz, B., Bilbao, J. I., & Sproul, A. B. (2017). A review and analysis of regression and machine learning models on commercial building electricity load forecasting. *Renewable and Sustainable Energy Reviews*, *73*, 1104–1122.

Yoshimura, F., Murakami, Y., Nishikawa, K., Hasegawa, Y., & Kawaminami, S. (2009). Surface components of porphyromonas gingivalis. *Journal of periodontal research*, *44*(1), 1–12.

Yu, Y., Guo, D., Qu, T., Zhao, S., Xu, C., Wang, L., ... Zhou, N. (2020). Increased risk of toxoplasma gondii infection in patients with colorectal cancer in eastern china: seroprevalence, risk factors, and a case–control study. *BioMed Research International*, *2020*(1), 2539482.

Yuan, K., Xu, H., Li, S., Coker, O. O., Liu, W., Wang, L., ... Yu, J. (2025). Intraneoplastic fungal dysbiosis is associated with colorectal cancer progression and host gene mutation. *EBioMedicine*, *113*.

Zavareh, F. S. E., Hadiipour, M., Kalantari, R., Mousavi, S., Tavakolifard, N., & Darani, H. Y. (2021). Effect of toxoplasma gondii on colon cancer growth in mouse model. *Am J Biomed*, *9*(2), 168–176.

Zepeda-Rivera, M., Minot, S. S., Bouzek, H., Wu, H., Blanco-Míguez, A., Manghi, P., ... others (2024). A distinct fusobacterium nucleatum clade dominates the colorectal cancer niche. *Nature*, *628*(8007), 424–432.

Zhang, C.-Z., Cheng, X.-Q., Li, J.-Y., Zhang, P., Yi, P., Xu, X., & Zhou, X.-D. (2016). Saliva in the diagnosis of diseases. *International journal of oral science*, *8*(3), 133–137.

Zhang, M., Zhang, Y., Sun, Y., Wang, S., Liang, H., & Han, Y. (2022). Intratumoral microbiota impacts the first-line treatment efficacy and survival in non-small cell lung cancer patients free of lung infection. *Journal of Healthcare Engineering*, *2022*(1), 5466853.

Zhang, N., Liu, Y., Tang, F.-Y., Yang, L.-Y., & Wang, J.-H. (2023). Structural characterization and in vitro anti-colon cancer activity of a homogeneous polysaccharide from agaricus bisporus. *International Journal of Biological Macromolecules*, *251*, 126410.

Zhang, X., Yu, D., Wu, D., Gao, X., Shao, F., Zhao, M., ... others (2023). Tissue-resident lachnospiraceae family bacteria protect against colorectal carcinogenesis by promoting tumor immune surveillance. *Cell host & microbe*, *31*(3), 418–432.

Zhou, X., Wang, L., Xiao, J., Sun, J., Yu, L., Zhang, H., ... others (2022). Alcohol consumption, dna methylation and colorectal cancer risk: Results from pooled cohort studies and mendelian randomization analysis. *International journal of cancer*, *151*(1), 83–94.

Zhu, W., & Lee, S.-W. (2016). Surface interactions between two of the main periodontal pathogens: Porphyromonas gingivalis and tannerella forsythia. *Journal of periodontal & implant science*, *46*(1), 2–9.

Zhu, X., Han, Y., Du, J., Liu, R., Jin, K., & Yi, W. (2017). Microbiota-gut-brain axis and the central nervous system. *Oncotarget*, *8*(32), 53829.

Zhuang, Y., Wang, H., Jiang, D., Li, Y., Feng, L., Tian, C., ... others (2021). Multi gene mutation signatures in colorectal cancer patients: predict for the diagnosis, pathological classification, staging and prognosis. *BMC cancer*, *21*, 1–16.

Acknowledgments

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

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

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

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

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

- Dr. Hyoung-oh Jeong
- Dr. Jinho Jang
- Dr. Seunghoon Kim
- Dr. Yeonsong Choi
- Dr. Taejoo Hwang

- Byeongjun Park
- Suhyun Park
- Ilsun Yun
- Kyoung Jun Lee
- Sabin Park
- Hyo Kyung Lee
- David Whee-Young Choi
- Changhoe Kim
- Seeun Choi

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

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

- 강승윤
- 구인용
- 권오준
- 권혁재
- 김정훈
- 김종민
- 류호준
- 문지석
- 박철홍
- 이병우

- 이선구
- 임병찬
- 조동혁
- 진창훈
- 최성재
- 최원준

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

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

My Lord, *the Flying Spaghetti Monster*,
give me grace to accept with serenity the things that cannot be changed,
courage to change the things that should be changed,
and the wisdom to distinguish the one from the other.
R'Amen.

May your progress be evident to all