INTEGRATED LARGE-SCALE ANALYSIS OF DISEASE SIGNATURES IN ORAL MICROBIOMES

by

OMPRAKASH SHETE

Under the Supervision of

DR. TARINI SHANKAR GHOSH

Assistant Professor,

Department of Computational Biology

Submitted

in partial fulfillment of the requirements for the degree of

Master of Technology

to

Indraprastha Institute of Information Technology Delhi

August, 2024

CERTIFICATE

This is to certify that the thesis titled "**Integrated Large-scale Analysis of Disease Signatures in Oral Microbiomes**" being submitted by **Omprakash Shete** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

August, 2024

Dr. Tarini Shankar Ghosh
Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

ACKNOWLEDGEMENT

Presentation, inspiration, and motivation have always been pivotal in the success of any venture. At the outset, I wish to express my profound and heartfelt gratitude to Prof. Dr. Tarini Shankar Ghosh for offering me the invaluable opportunity to work on the MetaOral project. His guidance and support have been instrumental in shaping my research journey, from the conception of the idea to the extraction of meaningful insights from the initial analyses of a large project.

I deeply appreciate his steadfast dedication and accessibility. Irrespective of the time or condition, he has consistently been available to respond to my queries, regardless of the lateness of the hour or the difficulty of the situation. The level of commitment he has shown towards my development has been truly extraordinary, frequently devoting more of his time to mentor me than he may have allocated for his own family. The remarkable level of support and encouragement I have received has not only motivated me but also compelled me to exert my utmost effort towards this project.

The initial analysis has provided essential knowledge that has facilitated further investigation and revelation. This progress would not have been achievable without his proficient guidance, mentorship, and substantial dedication of time and effort to my project.

I would like to express my sincere gratitude to my colleagues in the laboratory, namely Abhishek, Amit, Shivangi, Sourav, Alisha, Lavanya, Sakshi, and Vindhya. Their recommendations, assistance, and camaraderie fostered a sense of belonging within me, as if I were a member of their family. Their unwavering assistance throughout my project was extremely valuable, and I genuinely appreciate the time they devoted to supporting me in every conceivable manner.

I would like to express my appreciation to my classmates, namely Bhumika, Rudrakshi, and Siraj. They provided support, motivation, and companionship during my most difficult times. Their assistance played a crucial role in maintaining my motivation and concentration on my tasks.

Lastly, I would like to express my gratitude to my family for their consistent and unwavering support, both in terms of financial assistance and emotional encouragement, throughout my entire life. The unwavering faith they have in me has consistently provided me with the fortitude to confidently and resolutely pursue my academic and research aspirations.

Date: August 2024

Place: New Delhi

ABSTRACT

Oral health is essential for overall human well-being, with oral diseases impacting 3.5 billion people worldwide every year (WHO Report 2022). Oral microbiome is considered as the 2nd most abundant human body site with respect to microbial composition and diversity. Earlier studies have shown the variation in the oral microbiome in case of different oral and systemic diseases. This variation is associated with variety of endogenous and exogenous factors which include host lifestyle, drug use, host status, environmental factors etc. However, a complete understanding of the alterations of the oral microbiome across diseases (along with shared disease signatures, if any) is still lacking.

To investigate this, we have done an extensive meta-analysis of 17,031 Oral Microbiomes encompassing 48 different study cohorts and 29 diseases from 24 different countries across the globe, where we have analysed study-specific effects and the association of diseases with the oral microbiome. We observe that approximately 71% of studies have a significant association of oral alterations with different diseases. However, the age-specific variations in the oral microbial community play a key role as a confounder of these association patterns. We identify variabilities in the number of disease-gained and disease-lost across different diseases, and finally, we identify a shared signature of generic disease signature across oral microbiomes from different studies.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

INTRODUCTION:

Oral health is a vital component of overall well-being, playing a critical role in both systemic health and quality of life. The 2022 report by the World Health Organization emphasizes the alarming worldwide occurrence of oral diseases, impacting 3.5 billion individuals. Among all these diseases, oral cancer has global prevalence ranking as the 13th most common cancer due to its high rates of mortality and morbidity[1]. Periodontal diseases, which is the inflammation in the gums, impact approximately 19% of adults globally.

After the extensive research on the gut microbiome the oral microbiome has emerged as a vital area of research. It has come in to the notice that oral microbiome not only affects oral health but also general health. This diverse microbiome has a crucial function in maintaining oral health and preventing disease. Dysbiosis, defined as the disruption of the microbiome composition is linked to various diseases.

The oral microbiome refers to the overall collection of microorganisms that resides in the oral cavity, including bacteria, fungi, archaea, viruses, and other types of microorganisms. Oral microbiome is the initial site in the gastrointestinal tract that is directly exposed to the external environment, so it is highly influenced by various environmental factors. This makes it one of the most dynamic microbial ecosystems in the human body and leads to fluctuations in the makeup, quantity, and variety of these microorganisms. The oral microbiome is the second most populous and varied microbial community in the human body, after the gut microbiome. The diversity of microorganisms in this context refers not only to the abundance of different species, but also to the extensive array of functional abilities that these microorganisms possess. These capabilities play a significant role in promoting both oral and overall health. Gaining a comprehensive understanding of the complex details of the oral microbiome is essential in order to devise innovative approaches for the prevention and treatment of oral and systemic diseases[2].

The oral cavity has the complex anatomical structure which is composed of different sub-sites, and each of these sub-sites has distinct anatomy and physiology, which contribute to the diverse and abundant population of microorganisms present in the cavity. These sub-sites include buccal, tongue, teeth, gingiva, palate, and salivary glands. Each sub-site contributes to a unique and ideal environment. The tongue, with its papillae and grooves, has a large surface area and provide anaerobic conditions in its crevices, which facilitate the growth of bacteria like *Fusobacterium* and *Porphyromonas*[3]. Teeth and gingival sulcus, which have hard and rigid surfaces, accumulate biofilm, creating an ideal habitat for the growth of bacteria like Streptococcus and Actinomyces. Saliva, which is also the part of oral cavity has abundant bacteria - *Streptococcus* and *Neisseria*. In addition to this the oral cavity is exposure to food, drinks, air which introduces a continuous flow of new microorganisms, thereby increasing the diversity of the oral microbiome. The varied pH levels, oxygen availability, and nutrient sources in these sub-sites below the mouth support the growth of distinct and diverse microbial communities. The extensive different environment in mouth has a significant impact on the microbial diversity in oral microbiome[4].

CHAPTER 2

BACKGROUND AND OBJECTIVES:

2.1 Literature:

The very first study explaining about oral microbiome was published by Dal Bello et al, 2006, where they have explained briefly about *Lactobacilli* bacteria, which is prominent in gut microbiome. They were first to observe that oral cavity also contains a large abundance of *Lactobacilli*[5].

After this, many studies explored the Lactobacilli in oral cavity and one such study was published in year 2008, which explains the use of different strains of *Lactobacilli* species, collected from saliva and gingival sub-sites, as a probiotic to treat gastrointestinal disorders. Till certain limits the effect of *Lactobacilli* was observed and thus it was concluded that *Lactobacilli* in the oral cavity can be used as a probiotic for the treatment of diseases[6].

These studies became the basis of initiation of Human Oral Microbiome Database, an initiative of National Institute of Dental and Craniofacial Research (NIH). This database contains minor 16s datasets while most of it is consisted of genomes of oral species extracted from the oral cavity[7].

Further this database was utilized by many authors to get meaningful insights from the association of oral microbiome with oral diseases like Oral biofilms[8], Periodontitis[9], Dental caries[10] etc. Not only oral diseases but also systemic diseases like Type-1-diabetes[11], Polycystic ovary syndrome[12], etc. Along with these, some authors have also highlighted the studies showing the effects of exposures like tobacco[13,14], smoking[13,15], alcohol[14] consumption etc. on the oral microbiome.

However, many of these studies are on the small scale and there are many questions like we do not know if the alteration of Oral Microbiome is a cause or effect; how stable are these alterations in oral cavity; can these alterations be diagnosed and be used further for therapeutic purpose. For this purpose, we need to find what is the composition of oral microbiome; what are the alterations associated with; What are the signatures associated with different disease across multiple cohorts?

However, till now there is no any oral microbiome dataset repository containing sequence data along with metadata from different regions and for different diseases. And thus, establishment of it is the need for analysis said above. Thus, this study was meant to investigate above questions along with the establishment of a large data repository and to do so, we focused on the following objectives.

2.2 Objectives

A. Creating a large Oral Microbiome Repository from Publicly Available Data

B. Investigating Oral Microbiome Alterations in Different Diseases

C. Finding Association of Host Features with Alterations in Oral Microbiome

D. To Identify Shared Diseases-Associated Signatures in Oral Microbiome

CHAPTER 3

MATERIAL AND METHODOLOGY:

3.1. Creating a Data Repository:

Any Metagenomic analysis needs two basic things: Sequence data and Metadata. Sequence data is the samples that are taken from the patients and sequence and Metadata if the information of the patients with respect to age, gender, disease status, country, disease, diet etc.

Collection of Sequence data: Literature was collected using two different methods, the very first method was using the key word search on European Nucleotide Archive database and then literature search from the accession number. The second was using string search on PubMed database of National Institutes of Health[16]. Based on our use this literature was further filtered as shown in the *Figure 3.1.*

Collection of Metadata: Metadata was collected directly from the research papers and those whose metadata was not available were search on SRA Run Selector, a part of National Center for Biotechnology Information, National Institutes of Health. After collecting all the metadata, it was manually curated and was homogenized with respect to features of the patients.

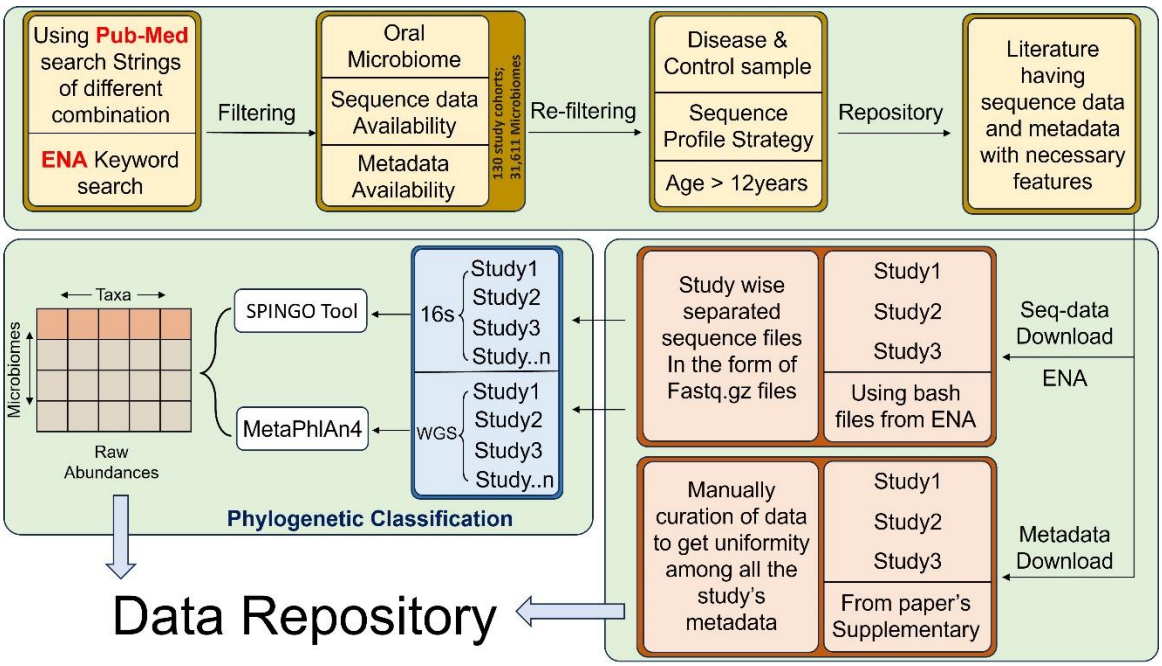

Figure 3.1: Data collection and Repository creation workflow

All the data collected was stored in the repository, irrespective of which type of data it is. Study names wise separation was done between the samples to know all the studies have their own disease and control matched samples.

3.2. Downloading and Pre-processing the Data:

Only sequence data from the filtered literatures was downloaded. Further based on the sequence strategy used (Amplicon or Shotgun), these sequence files were taken for phylogenetic classification. For Amplicon based sequences SPINGO tool[17] was used while for Shotgun sequenced data MetaPhlAn version 4 tool[18] was used.

The output of this tools were the species names with their abundance for each sample. The species abundance for each sample was stored in the form of table where columns define the species names and rows define sample names. This phylogenetic classification was done study cohort wise and was further proceeded to statistical analysis as shown in *Figure 3.1*

3.3.Statistical Analysis using R:

The species abundance for every study cohort generated after the phylogenetic classification was imported in R for statistical analysis. Every statistical analysis was done separately for every study cohort.

3.3.1. Extent of Disease Specific Variations:

In every microbiome study it is needed to know if the disease samples and control samples are distinguishable, which will let us know at what extent the samples are separated before going for the species level difference. To know this, we started with the PERMANOVA analysis which is non-parametric and multivariate statistical test[19]. The work flow of the PERMANOVA analysis is given in the *Figure. 3.3.1.* The row abundance was first row-sum normalized and then converted to distance matrix. The distances were Euclidean distances.

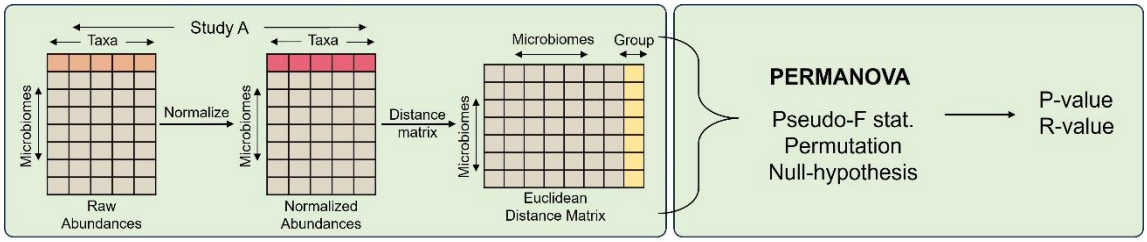

Figure 3.3.1: Workflow of PERMANOVA analysis

This distance matrix along with the metadata column stating disease status was given as a input in adonis2 function from vegan package for PERMANOVA. The output was P-value and R-value, where P value shows the significance in the separation of disease samples from the control one, while R-value shows the strength of separation.

3.3.2. Interaction of Host Demographics with Disease Status:

After PERMANOVA test, we were unknown about the features that control the separation of samples. In most of the studies disease status itself controls the

separation but there might be some other features that might control this separation. In order to know which features are associated with the separation of control and disease samples, we did ENV-FIT analysis. ENV-FIT analysis is the function that fits the environmental factors or vectors on the ordination plane and see the overlaps. It uses linear regression for this task. The R package vegan is used to do this using the function envfit[20].

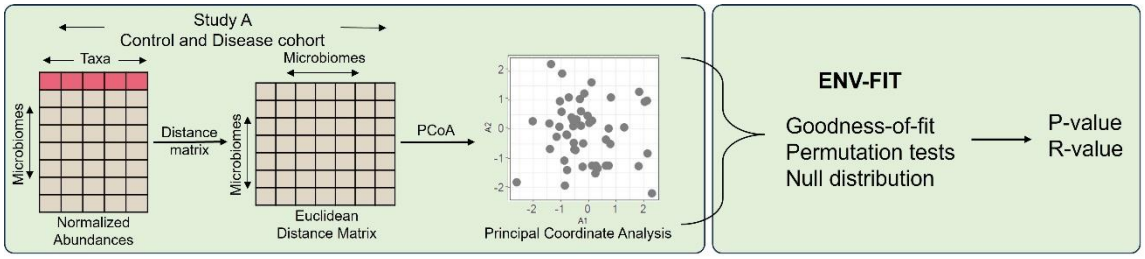

Figure 3.3.2: Workflow of ENV-FIT analysis

Where we used the normalized abundance to create a distance matrix and then dimensionally reduced this distance matrix to PCoA. The 2 components of this PCoA were further given as a input to ENV-FIT along with the environment variable which is age category and disease status in our case. See *Figure 3.3.2*

3.3.3. Confounding Factors in Disease Status:

Once the ENV-FIT analysis was done, some of the study cohorts shown the age category has an effect on the separation of the two groups. So, in order to confirm that this age category is independently interacting with the separation or acting as a confounding factor to disease status we did Multiple Linear Regression[19,21–24] where we took principle components as a dependent factor, disease status as a independent factor and age category as a confounding factor. See *Figure 3.3.3* which shows the workflow from normalized data to the input and then to the results of Multiple Linear Regression. If the P-value is significant for a study cohort, then the age category is acting as a confounder else only disease status is showing itself the separation.

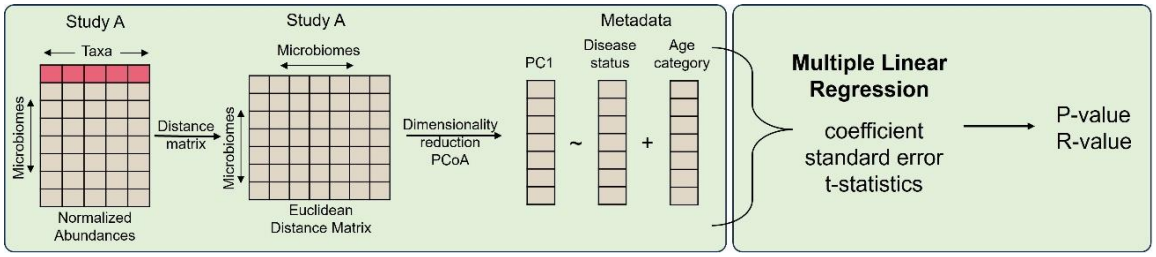

Figure 3.3.3: Workflow of the Multiple Linear Regression analysis

3.3.4. Disease and Health Associated Shared Signatures:

After finding the factors affecting the disease and control samples separation, we moved to see which species are associated with different diseases in every study and for this we used Wilcox test which is non-parametric and single variate statistical test. Wilcox test takes one species at a time to see the significant abundance difference of this species in two groups – disease and control. So this Wilcox test was run for every study cohort separately[25] as shown in *Figure 3.3.4.*

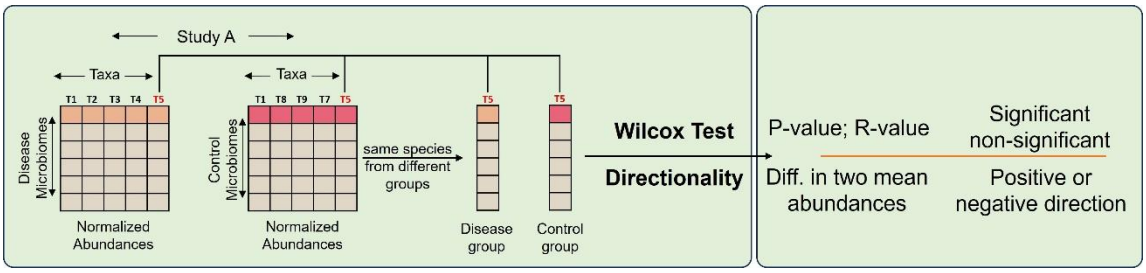

Figure 3.3.4: Workflow of the Wilcox Test and Directionality

Wilcox test yields P-value, estimate, etc. Along with this we calculate the difference of mean abundance of disease and control samples for every species. Based on P-value and mean we calculate the directionality of the species as positive or negative towards disease.

CHAPTER 4

RESULTS:

We collected total 17,031 samples from 48 study cohorts belonging to 24 different countries comprising of 29 diseases. Every analysis was done study wise. The result of each of them are as below:

4.1. Graphical Representation of Data Repository:

The data collected has been represented in *Figure 3.2*. Data collected is from total 24 countries where the largest collection is from the United States, covering about 29 different diseases. Of total 17,031 samples, only 706 samples belong to shotgun sequencing, while remaining 16,325 samples belong to amplicon sequencing which is also called as 16s sequencing. Control cohort samples were the most, accounting to 12,707 and diseased samples were only 4,324.

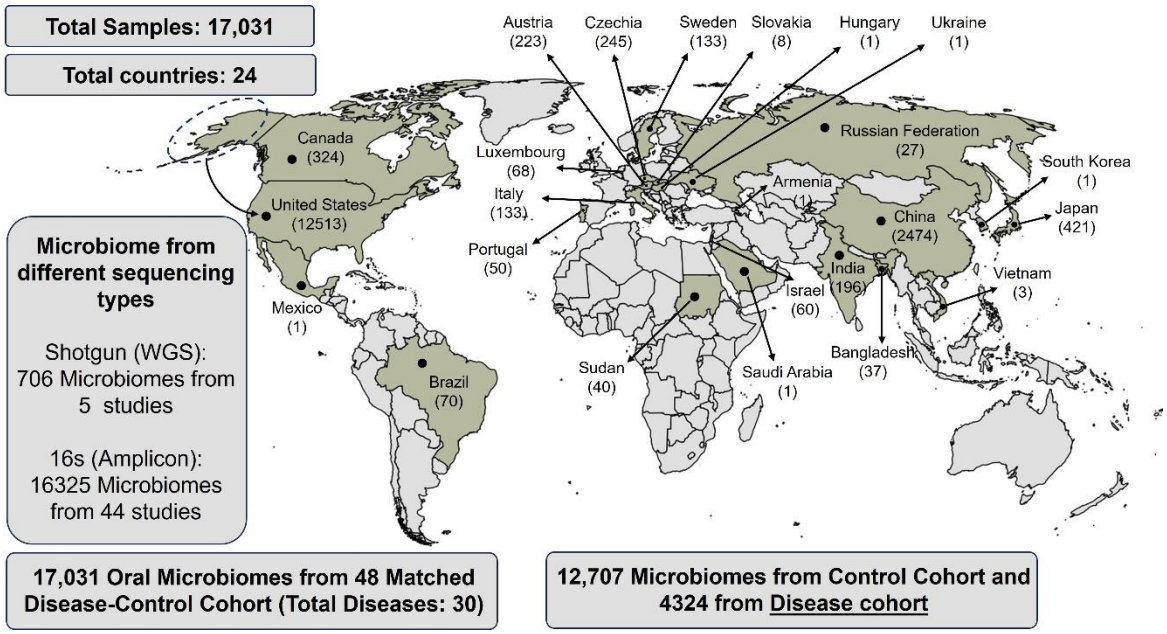

Figure 4.1: Data Representation of Collected Data

4.2. Graphical Representation from Statistical Analysis:

All the statistical results for each study cohort were analysed in R and their graphical representation are as below.

4.2.1. Extent of Disease Specific Variations using PERMANOVA:

Around 71% of studies (Pink colour) showing significant difference between Disease and Control Microbiomes of which number of studies with Periodontitis disease (85%) and Dental Caries (80%) are showing most significance. See Figure 4.2.1 The

y-axis of the plot is showing the strength of significance while x-axis are the study names and disease combination. This shows that disease samples and control samples are having difference at their abundance as well as ranking order of the species.

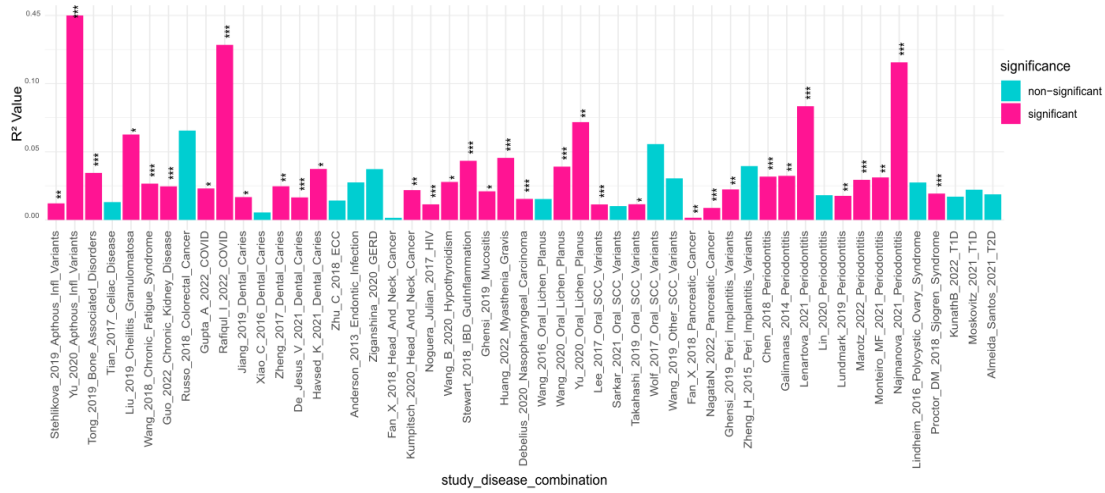

Figure 4.2.1: Study-wise PERMANOVA showing significance in around 71% studies

4.2.2. Interaction of Host Demographics with Disease Status using ENV-FIT:

We have used study condition and age category as two different environment variables separately.

See Figure 4.2.2a which shows the study names on the x-axis and y-axis is the R-value which shows the strength of significance. Here only 53% of studies (pink color) are showing significance which has reduces as compared to PERMANOVA analysis. Which gives a clue that any host feature has interaction with the microbiome separation. Percent of studies of Dental Caries showing significance also decreased to 20% but there was no any change in the percentage of studies showing significance for Periodontitis disease.

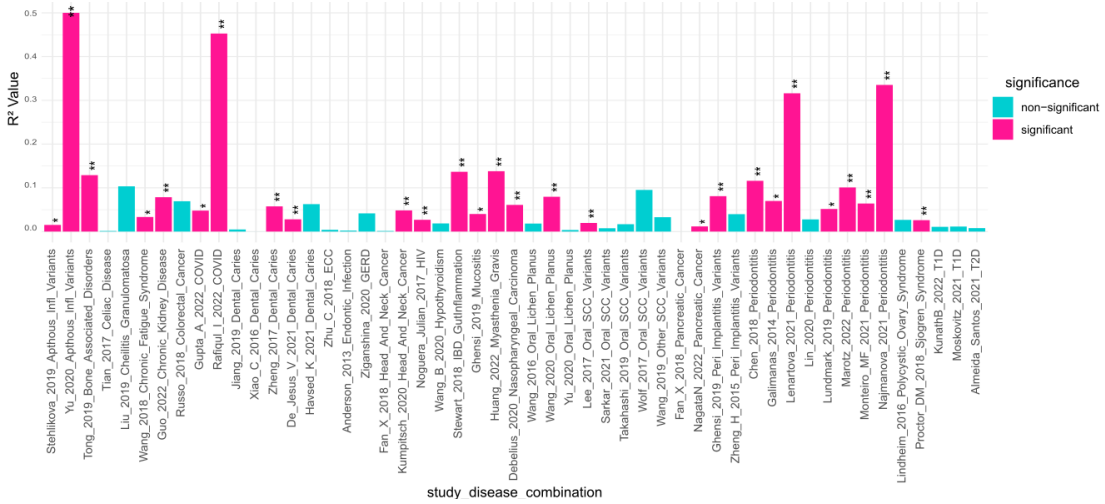

Figure 4.2.2a: Study-wise ENV-FIT using study condition as environment variable

When we used age category as an environmental variable for ENV-FIT analysis, around 18% of the studies shown significant interaction, see Figure 4.2.2b. Which

gives us a clue that age category is having interaction with the microbiome separation. Percent number of studies of Periodontitis disease and Dental Caries showing age category as a confounding factor were 38% and 20% respectively. Despite of no any change in the percent of studies showing significance in PERMANOVA and ENV-FIT for study condition for disease Periodontitis, still age category is showing significance for ENV-FIT in some studies.

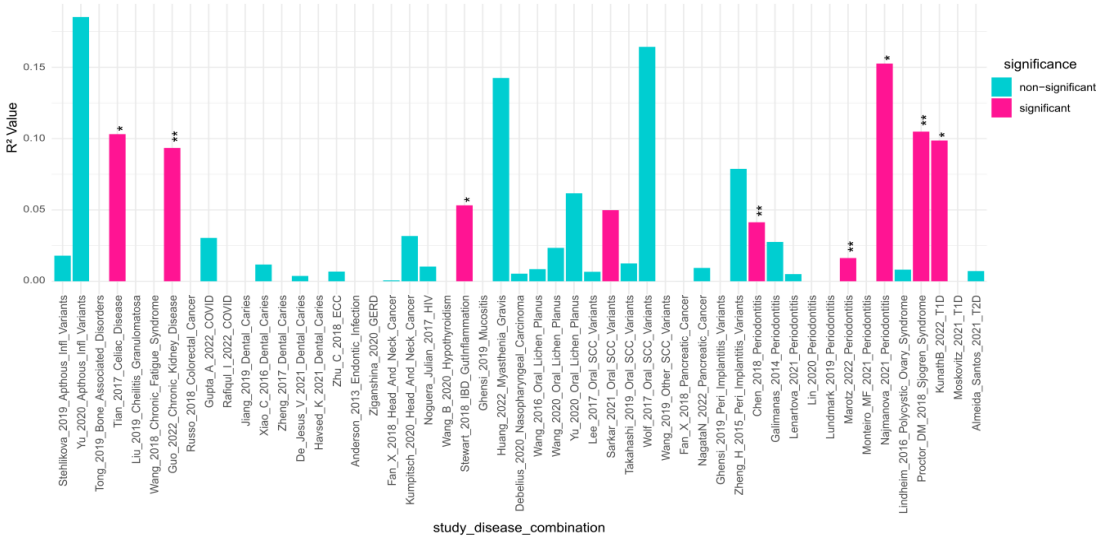

Figure 4.2.2b: Study-wise ENV-FIT using age category as environment variable

The decrease in the percentage of significant studies when study condition was used as an environmental variable is associated with the significance interaction of age category as shown above. Previously some of the studies on gut microbiome has shown the confounding factor associated with the study condition[19,22,26]. To know these, we did Multiple Linear Regression Model whose results are as below.

4.2.3. <u>Confounding Factors in Disease Status using Multiple Regression:</u>

To verify that age category as confounding factor, Multiple Linear Regression was carried for each study and 24% of studies shown significant interaction with age category as a confounder. See figure 4.2.3 which shows studies of Periodontitis and Dental caries are confounded by the age category which shows similar results as in ENV-FIT for age category as an environmental variable.

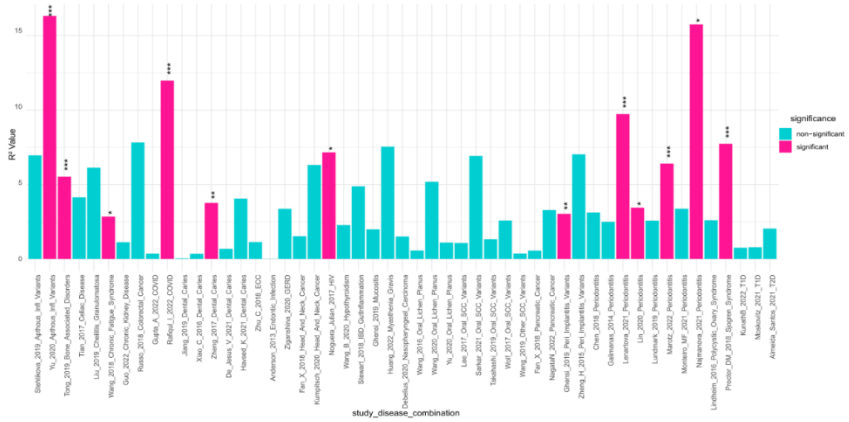

Figure 4.2.3: Studies showing age category as confounding factor

4.2.4. Disease Associated Shared Signatures using Wilcox Test:

To find the species associated with different disease we first used Wilcox and then based on mean abundance we got the directionality of each species in each study separately. Figure 4.2.4a shows the species associated with Periodontitis in different study cohorts. All the signatures are not consistent across all the study cohorts. The directionality shows the association with respect to the disease, if positive (2,3) they are associated with the disease else negative (-2,-3). Here Chen_2018, Marotz_2022 and Lenartova_2021 has maximum overlaps of signatures.

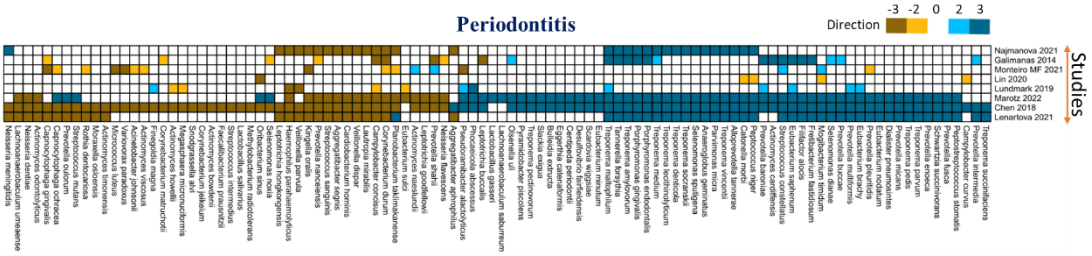

Figure 4.2.4a: Associated species with Periodontitis disease

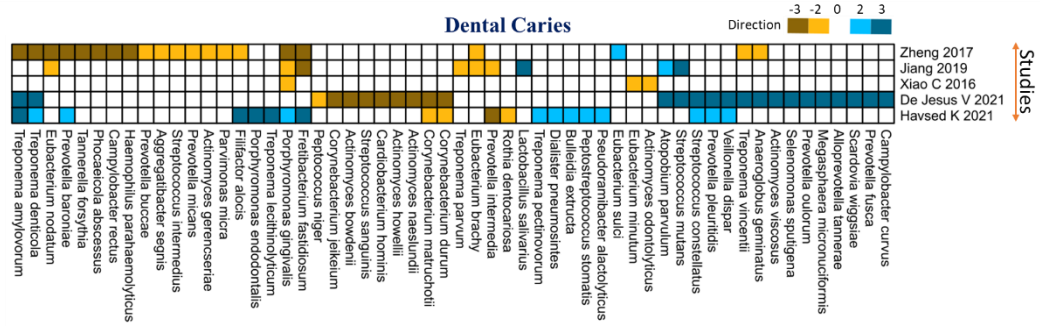

Figure 4.2.4b: Associated species with Dental Caries disease

As compared to Periodontitis, Dental Caries has less overlap and consistent signatures across studies, but there are study specific signatures associated with disease, see Figure 4.2.4b where Xiao_C_2016 study cohort has less signatures but every study cohort has their own set of signatures associated with disease.

When comparing these two figures – 4.2.4a and 4.2.4b with 4.2.4c, 4.2.4c has very less overlapped signatures for disease Oral Lichen Planus and Oral Squamous Cell Carcinoma Variants, other trends remain the same.

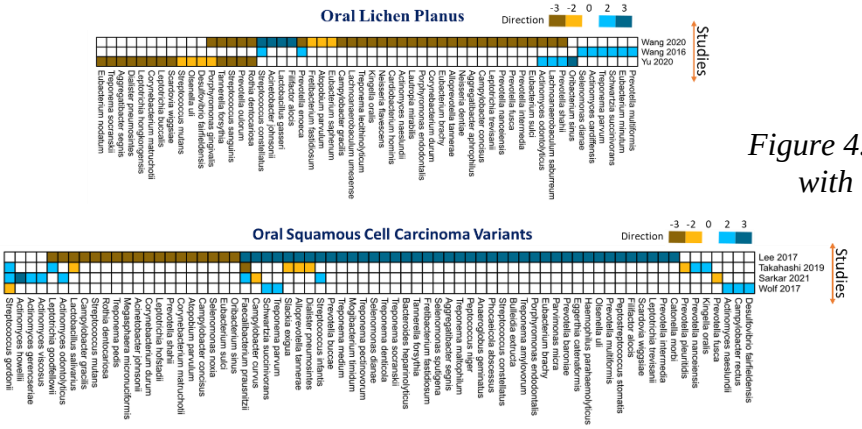

Figure 4.2.4c: Signatures associated with OLP and OSCC variants

Other diseases also have same pattern as for above diseases. Taking all these disease specific signatures together irrespective of studies we get the shared signatures as shown in the Figure 4.2.4d. Shared signatures are those which are consistent across multiple diseases, here we have assigned species to disease associated signatures which are positively associated with at least 5 or more diseases and negatively associated with maximum of 3 diseases (Right side of the plot with orange color line in graph). Opposite is for the non-disease associated shared signatures. This figure shows the separate signatures associated with disease and non-disease and thus can be used for further analysis.

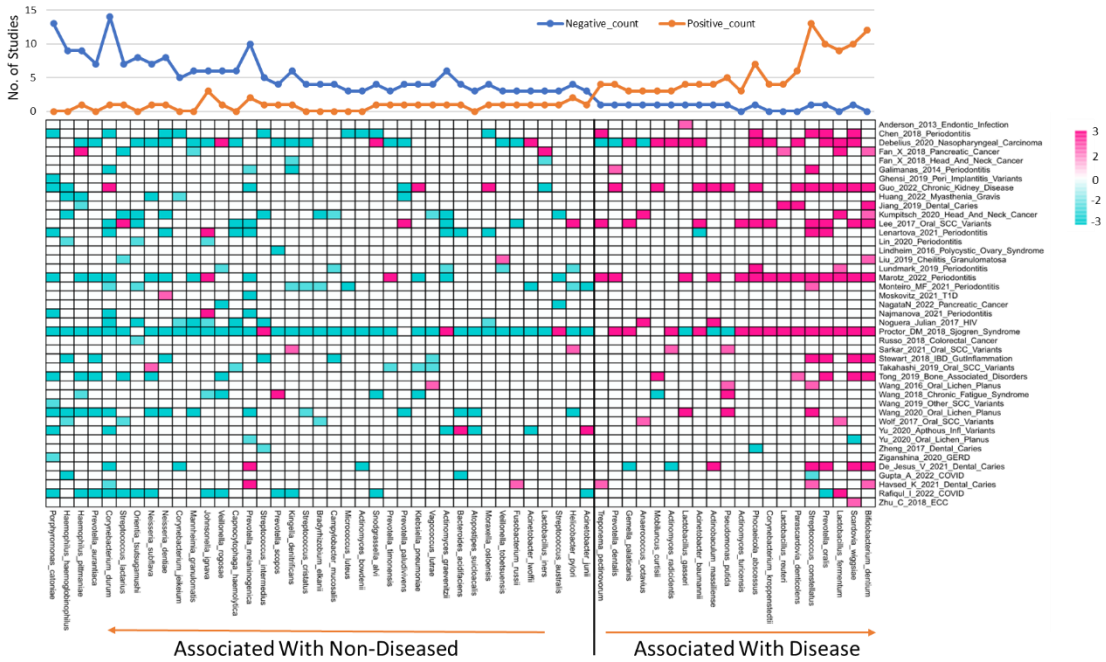

Figure 4.2.4d: Shared signatures associated with disease and non-disease (health)

CHAPTER 5

CONCLUSION AND LIMITATIONS:

PERMANOVA (Permutational Multivariate Analysis of Variance) analysis revealed statistically significant variations in the oral microbiome between disease and control group of individual samples. Around 71% of the study groups showed these notable differences, suggesting a strong connection between the presence of disease and changes in the oral microbiome. This indicates that the microbial composition in the mouth is responsive to different disease conditions, emphasizing its potential as a biomarker for identifying or tracking the development of diseases.

Although, age category was found to be a confounding factor in certain study cohorts, affecting the relationship between the oral microbiome and other factors. The initial analysis of this was carried through ENV-FIT model, a method that evaluates the alignment of environmental factor in a multidimensional space (PCoA). Subsequently, this finding was validated using multiple linear regression models. These findings suggest that age can influence the connection between the oral microbiome and health outcomes. It is important to consider age-related effects when analyzing microbiome data in order to prevent misleading associations.

Further the study wise disease association analysis shows different signatures associated with different diseases. The Wilcoxon test and directionality analysis of species associations demonstrated that the specific bacterial species associated with different diseases differed, suggesting a lack of consistency in the changes to the microbiome among patients with the same disease. This variability indicates that while changes in the oral microbiome are connected to disease states, the specific microbial profiles may vary, highlighting the intricate and diverse nature of shifts in the microbiome associated with diseases.

The analysis also detected specific bacterial species that were linked to multiple diseases called as shared signatures, indicating the existence of common microbial patterns across various diseases. These bacteria have the potential to be widely used as biomarkers or targets for therapeutic interventions due to their involvement in various diseases. The existence of these common indicators emphasizes the interdependence of the oral microbiome in different pathological conditions and emphasizes the possibility of conducting cross-disease microbiome research to reveal universal disease mechanisms.

Although we were able to show some features associated with the Oral microbiome, but there are limitations to the research which are stated below:

1. Geographic Bias: Approximately 20% of the study cohorts accounting to 73% of the oral microbiome data comes from the United States, which may limit the generalizability of the findings.
2. Low representation of whole genome sequenced samples: Only 5% of the samples from 17,031 samples are whole genome sequenced data accounting to 706 samples which will limit the detailed analysis

3. Uneven disease specific studies: As we were able to see the association of species with only few diseases which is due to uneven distribution of study cohorts across all the diseases.
4. Limited host features: In the current analysis we were able to see the age category feature interaction with microbiome but we were not able to see any other feature associated with microbiome due to limited number of feature available from the patients.

CHAPTER 6

REFERENCES:

1. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries - Bray - 2018 - CA: A Cancer Journal for Clinicians - Wiley Online Library. https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21492.
2. Kilian, M. *et al.* The oral microbiome – an update for oral healthcare professionals. *Br. Dent. J.* **221**, 657–666 (2016).
3. Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I. & Dewhirst, F. E. Defining the Normal Bacterial Flora of the Oral Cavity. *J. Clin. Microbiol.* **43**, 5721–5732 (2005).
4. Dewhirst, F. E. *et al.* The Human Oral Microbiome. *J. Bacteriol.* **192**, 5002–5017 (2010).
5. Dal Bello, F. & Hertel, C. Oral cavity as natural reservoir for intestinal lactobacilli. *Syst. Appl. Microbiol.* **29**, 69–76 (2006).
6. Kõll, P. *et al.* Characterization of oral lactobacilli as potential probiotics for oral health. *Oral Microbiol. Immunol.* **23**, 139–147 (2008).
7. Chen, T. *et al.* The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database* **2010**, baq013 (2010).
8. Langfeldt, D. *et al.* Composition of Microbial Oral Biofilms during Maturation in Young Healthy Adults. *PLOS ONE* **9**, e87449 (2014).
9. Chen, C. *et al.* Oral microbiota of periodontal health and disease and their changes after nonsurgical periodontal therapy. *ISME J.* **12**, 1210–1224 (2018).
10. Baker, J. L. *et al.* Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. *Genome Res.* **31**, 64–74 (2021).
11. Almeida-Santos, A., Martins-Mendes, D., Gayà-Vidal, M., Pérez-Pardal, L. & Beja-Pereira, A. Characterization of the Oral Microbiome of Medicated Type-2 Diabetes Patients. *Front. Microbiol.* **12**, (2021).
12. Lindheim, L. *et al.* The Salivary Microbiome in Polycystic Ovary Syndrome (PCOS) and Its Association with Disease-Related Parameters: A Pilot Study. *Front. Microbiol.* **7**, (2016).
13. Stewart, C. J. *et al.* Effects of tobacco smoke and electronic cigarette vapor exposure on the oral and gut microbiota in humans: a pilot study. *PeerJ* **6**, e4693 (2018).
14. Thomas, A. M. *et al.* Alcohol and tobacco consumption affects bacterial richness in oral cavity mucosa biofilms. *BMC Microbiol.* **14**, 250 (2014).
15. Halboub, E. *et al.* Tongue microbiome of smokeless tobacco users. *BMC Microbiol.* **20**, 201 (2020).

16. Vincent, B., Vincent, M. & Ferreira, C. G. Making PubMed Searching Simple: Learning to Retrieve Medical Literature Through Interactive Problem Solving. *The Oncologist* **11**, 243–251 (2006).

17. Allard, G., Ryan, F. J., Jeffery, I. B. & Claesson, M. J. SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics* **16**, 324 (2015).

18. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).

19. Ghosh, T. S., Das, M., Jeffery, I. B. & O'Toole, P. W. Adjusting for age improves identification of gut microbiome alterations in multiple diseases. *eLife* **9**, e50240 (2020).

20. Oksanen, J. *et al.* vegan: Community Ecology Package. (2022).

21. Ghosh, T. S. *et al.* Mediterranean diet intervention alters the gut microbiome in older people reducing frailty and improving health status: the NU-AGE 1-year dietary intervention across five European countries. *Gut* **69**, 1218–1228 (2020).

22. Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448–454 (2020).

23. Tito, R. Y. *et al.* Microbiome confounders and quantitative profiling challenge predicted microbial targets in colorectal cancer development. *Nat. Med.* **30**, 1339–1348 (2024).

24. Knight, R. *et al.* Best practices for analyzing microbiomes.

25. Qin, Y. *et al.* Consistent signatures in the human gut microbiome of old- and young-onset colorectal cancer. *Nat. Commun.* **15**, 3396 (2024).

26. He, Y. *et al.* Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* **24**, 1532–1535 (2018).