Comparative Analysis of Human Oral Microbiome Communities in Health and Disease

Soleil Su[1], Jenny Mao[2], Daniel Ni[3], Steven Wei Chen[4], and Jason Pan[5]

[1]Biomedical Engineering, 2025, ls764
[2]Biological Engineering & Computer Science, 2025, jm2338
[3]Biomedical Engineering, 2025, jn397
[4]Computer Science, 2025, sc2342
[5]Chemistry & Statistical Science, 2024, jp952

ABSTRACT

The oral microbiome, a dynamic and intricate ecosystem encompassing bacteria, fungi, and viruses, is pivotal for maintaining oral health and has broader implications for overall well-being. Imbalances within the oral microbiota have been directly implicated in various pathologies, from tooth decay to oral cancer. This study employs advanced techniques, including shotgun sequencing and metagenomics, to conduct an in-depth analysis of 18 DNA samples sourced from individuals exhibiting diverse oral health statuses, implants, and diseases. Leveraging the mOTUs profiler, an advanced microbial profiling tool, this research aims to overcome inherent limitations associated with conventional profiling methods and offer a comprehensive understanding of the oral microbiome's bacterial composition.

The investigation reveals that healthier samples exhibit a diverse microbiome, while diseases and severe conditions correlate with the dominance of specific bacterial genera, including Streptococcus and Rothia. However, Principal Component Analysis (PCA) highlights outliers in severe implant cases. Future research should delve into the effects of unknown genera and identify contributors to PCA variance for a comprehensive understanding of oral microbiome dynamics. Nevertheless, the identification of specific bacterial contributors and exploration of unknown genera present crucial avenues for future research, promising a deeper comprehension of the intricate dynamics governing the oral microbiome.

Keywords (minumum 5): Short-gun sequencing, metagenomics, human microbiome, bioinformatics, taxonomy profiling

Project type: Original Research

Project repository: GitHub

1 Introduction

The oral cavity is a dynamic and intricate ecosystem that harbors a diverse community of microorganisms collectively known as the oral microbiome. It encompasses bacteria, fungi, viruses, and more. The composition and balance of the oral microbiome play a pivotal role in maintaining oral health, with repercussions that extend beyond the confines of the mouth, influencing overall health and well-being. Imbalanced oral microbiota has been directly implicated in the apparition of multiple oral pathologies, such as tooth decay and periodontitis [1]. Oral cancer is also linked to the presence of buccal dysbiosis [1]. In addition, emerging evidence suggests that the oral microbiome may serve as a reservoir of potential biomarkers, offering insights into the diagnosis of systemic disease [2].

Our project aims to understand the differences in bacterial composition of the oral microbiome between healthy and disease human oral samples. In our investigative pursuit of the human oral microbiome, our study is anchored in a comprehensive collection of 18 DNA samples, each sourced from distinct individuals exhibiting varying degrees of oral health, implants, and diseases. Augmenting this diverse set, our methodology incorporates essential controls - a Blank DNA sample serving as a negative control and an E.coli DNA sample as a positive control. These controls bolster the robustness of our analysis, ensuring fidelity and reliability in the subsequent examination of bacterial composition.

These DNA samples are obtained as metagenomic samples (collective DNA of a combination of DNA from multiple species). Then the DNA is extracted and gone through shotgun sequencing. The results are a set of metagenomic reads for each sample, forming the basis for our investigation.

Bacterial community profiling commonly relies on classifying PCR amplicon sequences derived from the small subunit ribosomal RNA gene, specifically the 16S rRNA gene of bacteria and archaea [3]. While this method is potent, it has been acknowledged for its inherent biases in composition estimates. These biases arise from factors such as variations in the copy numbers of the 16S rRNA gene per genome, disparities in the efficiency of PCR primers across different species, and the utilization of different sub-regions of the gene [3]. In addition, the high sequence conservation of the 16S rRNA gene creates limitations on its ability to effectively distinguish closely related organisms.

In our project, the tool we use for microbial profiling is called mOTUs. mOTUs consolidates data from >3100 metagenomic samples into an updated mOTU database to substantially improve the representation of human-associated and ocean microbial species [3]. The output of mOTUs is a taxonomic profile, providing insights into the differences in bacterial species among our diverse samples. This report will go over the pipeline of mOTUs in further detail.

2 Methods

2.1 Obtainning data

To investigate the distinctions between healthy and diseased oral microbiomes, we gather 18 metagenomic samples from volunteers at Harvard Dental School. In addition to these, we include two control samples: one blank (without any microbial DNA) and one containing Escherichia coli (E. coli). All these samples are acquired through shotgun sequencing and will be analyzed with mOTU profiler.

2.2 Quality Control

Ensuring the integrity of our data for mOTU profiler analysis, we rigorously perform quality control on our metagenomic data. This involves read filtering and base correction of FASTQ files, using the FASTP tool [4]. These preprocessing steps are crucial to eliminate errors and biases, providing clean, reliable data for accurate microbial profiling.

2.3 Base pipeline review

We profile the 20 filtered metagenomic samples using the MG-based operational taxonomic unit (mOTU) profiler version2, described in [3]. The mOTUs profiler works by identifying the specific gene marker in the sample and estimating bacteria abundance based on the number of gene markers presented in each sample. We follow the protocol described in [5]. In the following sections, we will walk through the built-in databases and the major algorithms used for taxonomic profiling in mOTU profiler.

2.3.1 Built-in Databases

The mOTUs2 profiler has a self-designed database of marker gene (MG) sequences extracted from reference genomes and metagenomic samples [3]. The original MGs were processed to form species-level mOTUs that were later compiled into a sequence database along with taxonomic annotations. The database is later used for short-read mapping with input data.

Marker gene selection

The systemic construction of mOTUs first started with 40 universal single-copy marker genes that had been shown to be able to accurately delineate prokaryotic species [6]. By performing multiple-sequence alignments (MSA) on the orthologous groups of the MGs that had been identified in more than a thousand prokaryotic genomes [6], homologous sequence positions were generated in a matrix format. Then, a profile hidden Markov model-based (HMM) tool was used to identify MGs in 3496 reference genome sequences downloaded from the US National Center for Biotechnology Information [6]. A profile HMM is used to capture patterns in sequences, and it typically has three states (match, insert, and delete states), a transition probability matrix, and an emission probability matrix. In this study, profile HMMs were derived from the results of MSAs and generated for the 40 MGs. A computational tool HMMER was used to create the models. For reference genome, the hmmsearch program of HMMER was used to identify approximately 40 MGs per genome, and each MG in each reference genome was found by choosing the highest-scoring target sequence. For metagenomic data, we calibrated bit-score cutoffs for each MG to achieve the highest accuracy of MG identification by using a training set of more than a thousand well-annotated genomes [6]. The set of calibrated cutoffs was then used to identify MGs in metagenomic data. After 40 MGs were identified in the reference genome and metagenomic data, they were filtered down to 10 best-performing genes based on the accuracy of MG identification and the accuracy in species abundance quantification.

MGCs generation

MGs are clustered to form MG clusters (MGCs), which can be divided into ref-MGCs and meta-MGCs. Ref-MGCs are built by clustering reference genomes. First, the pairwise global nucleotide identities, which is the number of congruent nucleotides divided by the gene length, were calculated for all genomes and each of the 40 MGs using vsearch [3]. Then, overall distances between genomes were acquired by finding the weighted average of pairwise global nucleotide identities of 40 MGs, using their gene lengths as weights. Then, genomes were clustered using a cutoff of 96.5% nucleotide identity, resulting in over 5000 clusters [3]. The ten MGs were extracted from the clusters, resulting in over 50,000 ref-MGCs, where each MGC corresponds to one MG. To build meta-MGCs, we first identified MGs from metagenomic data using the set of calibrated cutoffs specific to ten selected MGs. Then, pairwise nucleotide identities were calculated for all MGs, including ref-MGs and newly identified meta-MGs. Then, the open reference clustering method was used. It begins with comparing meta-MGs with a database of known sequences, which is the ref-MGCs, and those that match the ref-MGCs with at least 96.5% identity are clustered. Then, meta-MGs that do not find a match in ref-MGCs are clustered based on their similarity to each other, leading to the formation of meta-MGCs. In summary, ref-MGCs can contain both ref-MGs and meta-MGs, while meta-MGCs only contain meta-MGs.

mOTUs formation

Binning is an important process that combines MGCs originating from the same species. Ref-MGCs are binned into ref-mOTUs based on the original cluster affiliation during ref-MGCs formation. Because the clustering of meta-MGCs was performed independently for individual MGs, they cannot be combined to form meta-mOTUs the same way ref-mOTUs were formed. Instead, a different approach was used. The abundance of each of the genes from the same species is expected to correlate with each other, so we calculated the pairwise correlation of the abundance of all MGCs for each biome. The correlation measure and prevalence filtering were optimized for each biome by setting biome-specific parameters. Then, correlations for each biome were combined into association values in a false discovery rate (FDR) calibrated manner. For each pair of MGCs, the maximum value of FDR-calibrated association values was calculated across biomes [3]. For binning meta-MGCs, a version of the greedy algorithm was used. First, ref-MGCs were binned to form ref-mOTUs. Then, meta-MGCs were binned starting from the highest FDR-calibrated association values until a cutoff of 0.8 was reached [3]. Here, it is important to note that on top of the lowest association value requirement, an MGC can be added to ref-mOTUs or combined with another MGC only if the MG that corresponds to the MGC is not already added to the group. Meta-mOTUs can be defined as mOTUs only if they contain at least 6 MGCs. As a result, 2494 meta-mOTUs and 5232 ref-mOTUs were generated. Like MGCs, meta-mOTUs only contain meta-MGCs, while ref-mOTUs can contain both ref-MGCs and meta-MGCs.

Taxonomy annotation of mOTUs

Ref-mOTUs are annotated based on the original cluster affiliation as the clusters were annotated based on the taxonomy of member genomes. They can typically be homogenous (all member genomes have the same species name), heterogeneous (different species name), or undetermined (non-binomial species name). Only homogenous ref-mOTUs will be considered during profiling. For meta-mOTUs, the annotation is more complicated. First, the MGs were annotated using a reference protein sequence database called Uniprot's UniRef90 after the samples were translated into amino acid sequences. Then, similarities between the MG amino acid and reference protein sequences were calculated as bitscores, and proteins with a score higher than 90% of the highest bitscore were retained. For a meta-mOTU, the MG member with the highest score was selected, and the taxonomy annotation of the MG was then transferred to the associated MGC [3]. For each rank in the taxonomy tree, at least three MGCs must be annotated in order for the mOTU to be considered annotated, and at least half of the MGC taxonomy must be in agreement in order for the mOTU to be considered consistent. Unannotated or inconsistent mOTUs are not considered during profiling. In the end, each mOTU can only have a single annotation assigned to it to achieve species-level resolution.

mOTUs2 mapping database

The constructed mOTUs must be compiled into a sequence database that can later be used for matching short metagenomic sequences. Identical MG sequences were removed, and the MG sequences were extended at both ends by up to 100 nucleotides based on the origin of reference genomes or metagenomic data [3]. The resulting database is converted to sequence files in FASTA format with taxonomy annotations.

2.3.2 mOTUs2 profiling workflow

mOTU profiler identifies the bacteria species in our sample by aligning the quality-controlled sample sequences to the MGs of the mOTU built-in databases that are described above. The alignment is done using the BWA-MEM Genomic mapping algorithm.[7]

Alignment of metagenomic sequencing reads to MGs

Burrows-Wheeler Aligner

Burrows-Wheeler Aligner (BWA) is a software package for mapping sequences to against a large reference genome. The package consists of three algorithms: BWA-backtrack, BWA-SW, and BWA-MEM. mOTU profiler employs BWA Maximal Exact Matches (BWA-MEM), noted for its efficiency in aligning longer sequences and adeptly handling insertions and deletions. Compared to other genome aligners like Bowtie2 and HISAT2, a previous study [8] has also demonstrated that BWA achieves the highest alignment rate.

BWA-MEM, a local alignment algorithm, functions through three core computational stages: Seed Generation, Seed Extension, and Output Generation [9]. The process begins with the algorithm seeding a query sequence by locating exact matches (k-mers) within the reference genome, typically using a k-mer size of 19. BWA-MEM then extends these seeds up to a predetermined cutoff value, a threshold calculated by a function that tolerates large gaps in the alignment without imposing penalties.

During the Seed Generation phase, the algorithm identifies seed locations on the genome that exactly match subsequences between the read and the reference. Following this, Seed Extension is performed, where the algorithm attempts to elongate these seeds in both directions. This extension utilizes an affine-gap Smith-Waterman-like dynamic programming approach, allowing for inexact matches. Among all the extended seeds, the best-scoring alignment is selected.

The Seed Extension involves two distinct components: an outer function that iterates over all seeds, deciding whether to extend each one, and the actual Extend kernel. The outer function monitors all previously found extensions associated with a read. If a new seed significantly overlaps with a prior extension, it is disregarded to avoid redundancy. Optimal alignment is computed by filling a similarity matrix, accounting for all possible alignments. Each cell in this matrix relies on the values of its immediate top, top-left, and left neighbors as shown in the formula below, which take into account of affine gap penalty using the function γ.

$$F(i,j) = \max \begin{cases} F(i-1,j-1)+s(x_i,y_j), \\ \max\limits_{k=0,\dots,i-1}\{F(k,j)+\gamma(i-k)\}, \\ \max\limits_{k=0,\dots,j-1}\{F(i,k)+\gamma(j-k)\}. \end{cases}$$

Diverging from the standard Smith-Waterman algorithm, the initial values in this matrix are not zero. Instead, they depend on the score of the seed being extended, facilitated by an Initial Value block added to the systolic array. This modification enhances the accuracy of identifying optimal alignment sequences. Finally, BWA-MEM outputs the best alignments against our reference databases.

Estimation of read abundances for every marker gene cluster (MGC)

Since the marker genes are single-copy, the abundance of a particular type of bacteria in our sample can be reflected by the number of the alignments to the reference databases. The abundance for each MGC is first calculated by summing all the best unique alignments from the reads to reference, which result is a unique alignment profile. After, each read with multiple alignments should be distributed among its best-scoring MGCs. The distribution is not equal but based on the proportional abundance of each MGC as determined from the unique alignments. For example, if a read has a score of 1 (for simplicity), and it aligns to three MGCs with relative abundances of 40%, 30%, and 30% according to the unique alignment profile, then it contributes 0.4, 0.3, and 0.3 to each MGC respectively. After distributing all multiple alignment reads among their respective MGCs, the abundance counts for each MGC is updated. This count becomes a sum of the unique alignment counts and the distributed counts from multiple alignments.

Calculation of mOTU abundances

Base coverage is also a key to determine the abundance of bacteria species. mOTU profiler calculates the MGC base coverage by summing up the total number of bases aligning to each MGC and then dividing by the length of the respective gene in the MGC. Finally, the abundance of the mOTUs is calculated by taking the median of read abundances described above and the base coverages. The mOTU profiler generates a taxonomic profile for each sample, provided in a CSV file format. This file details not only the overall bacterial abundance but also the specific abundance of each bacterial type. The taxonomic profiles for our 20 samples are available in the supplementary materials for further examination and analysis.

2.4 Visualization

To visualize our result, we will utilize the Python packages matplotlib.pyplot [10] to generate a pie chart for the relative frequencies of different bacteria genera. We first calculated the relative frequency for each bacteria type because the human mouth contains not only bacteria but other organisms like viruses. The genera are ordered by abundance for easier reference and analysis.

2.5 Comparison

To analyze the results, we will leverage Principal Component Analysis to uncover underlying patterns and clusters behind our data and investigate what bacteria are the outliers in the study, indicating that some bacteria genuses are causing diseases in either implant or regular teeth. In addition, screen plots will also be utilized to ensure the quality of the PCA analysis, as well as serving as a dimensional reduction guideline.

3 Results

3.1 Bacterial Genera Composition Analysis

After examining the result from the corresponding pie charts from the 20 different samples of different individuals in various conditions, we came across that healthy samples tend to have more diverse microbiomes where different bacteria genera coexist with each other. On the other hand, in more severe cases, we do see there is a specific bacteria genus that dominates the microbiome community. Bacteria genera such as Streptococcus and Rothia, etc, which can cause mild to serious infection are the ones that we are specifically targeting for a closer look at each sample. In terms of Streptococcus, according to the CDC, it can cause Strep Throat, Cellulitis, Scarlet Fever, Impetigo, etc. [11] In terms of Rothia, according to the National Library of Medicine, it can cause a wide range of infections. [12]

Of our four healthy individuals with implanted teeth,as shown in Figure 1 to 4, three out of four (Figure 1,2, and 3) of them have a very diverse microbiome community such that each bacteria genus has a frequency of around 20%, which is an indicator that there is no strong presence of some bad bacteria genus that might affect the patient's health. On the other hand, in our two moderately healthy individuals with implant teeth, we found that one of the individuals (Figure 6) had around 38.9%, which is a strong presence out of all other bacteria genera. In addition, in the other moderate sample in Figure 5, we found there is a large presence of 25.3% of unknown bacteria genera that have not yet been previously identified in one of the individuals, which we might not know its effect on human health. In our severe samples, individuals were identified with a strong presence of streptococcus in two of the four samples (Figure 7 and 8) where 53% and 55.3% of bacteria genera were classified as streptococcus. Much less diverse microbes were observed such that the majority of the community is dominated by only two bacteria genera. In one of the samples (Figure 9), we also found a large presence of 24% of unknown bacteria genus similar to the sample of moderate individuals with implant teeth. However, for the severe implant sample 4 from Figure 10, we do see a diverse microbiome and less unknown genera which differs from the rest.

In our four healthy individuals with original teeth, we observed a similar trend as the healthy individuals with implanted teeth. However, there seems to be some disparity. In two of the four samples (Figure 13 and 14), pseudo propionibacterium dominates one individual's oral microbiome with 48.5% and streptococcus dominates another healthy individual's oral microbiome with 30.1%, indicating the difference in general trend as the four healthy individuals with implant teeth. On the other hand, the four diseased individuals with original teeth, all demonstrate some kind of abundance in one or two bacteria genera, such as Streptococcus, Rothia, Actinomyces, etc these types of bacteria genera can cause different kinds of infection. For example, in Figure 16 which is the diseased tooth sample 2, we see that three bacteria genera accounts for over three quarter of the oral microbiome. However, we could still see some diversity in the diseased samples as shown for diseased tooth sample 4 at Figure 18. Unknown bacteria genera are also present in these diseased individuals with original teeth.
Lastly, the control consists of two samples. One of them (Figure 19) is the positive control (E. coli and L. plantarum) and the other one (Figure 20) is the negative control (blank control that consists of all kinds of bacteria genera).

Our result shows there is a correlation between the abundance of some bacterial genera in the oral microbiome that might cause illness and the health condition of each individual. The abundance of some bacteria genus is mostly represented in severe cases and diseased samples. One of the limitation of using this strategy is that it is a little hard to tell the big difference between the tooth samples and implanted samples since our sample size is very small.

3.2 Principal Component Analysis

Principal Component Analysis is a powerful statistical and machine learning tool to realize data dimensional reduction and detect data clusters through PCA plots. It utilizes an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The steps encompass complex mathematical manipulations, and a general workflow entails the following stages: data standardization, covariance matrix calculation, eigendecomposition, linear projection, selection of principal components, and the interpretation and use of reduced data.

In our study, we leveraged PCA upon a dataset comprising 18 DNA samples, each obtained from individuals presenting a spectrum of oral health conditions, including those with implants and various oral diseases. Each feature, or dimension, is a bacteria genus in the sample. Through this, the primary objective is to understand the intrinsic data patterns present within the dataset. By doing so, we aim to alleviate the collinearity among features, and eventually realize data dimensional reduction so that future study and analysis are able to extract the reduced dataset to further investigate the effects of oral microbiome in oral diseases. Using python, we executed three distinct PCA corresponding to different subset samples, and plotted each PCA on the first two principal components. These sample groups include the entire collection of samples, implant samples, and tooth samples to detect clusters in various experimental groups. The rationale behind performing separate PCA for each group was to identify and compare potential clusters or patterns unique to each experimental group. This approach allowed us to gain a nuanced understanding of how the oral microbiome might differ not only in health and disease but also in relation to different oral environments such as implants versus natural teeth. The findings from these analyses are expected to provide suggestions that could guide future research in oral microbiology.

In the PCA of all samples (Figure 21), insightful observations about the distribution and relationships of the data points were depicted. Notably, a majority of the sample data points were found to be clustered around the origin (0,0) in the PCA plot. This clustering indicates a degree of similarity or commonality in the bacterial composition among these samples. However, some outliers, such as healthy implant, moderate implant, and severe implant, were observed, as these samples were located significantly distant from the central cluster, suggesting a distinct bacterial composition or diversity in these groups compared to the others. This variability could be inferred from the observation in the bacterial genera analysis, as it was concluded that the implant samples, especially in moderate and severe groups, were predominantly harbored by a Streptococcus population, while original tooth samples contained a more balanced microbiome community.

After decomposing all samples into implants and original tooth subsets, we executed PCA on these two subgroups to delve deeper into the data. In the PCA plot for implants (Figure 22), a similar pattern as the previous all sample PCA plot was observed. Once again the significant outlier happens to be moderate and severe implant samples. Through the genera analysis, the predominant Streptococcus population within these bacterial compositions explains the deviation of moderate and severe implant samples from the healthy implants. It is worth noting that one of the two moderate samples that remains within the cluster was found to have a more balanced bacteria community as opposed to being dominated by Streptococcus, which further consolidated our results that an increased Streptococcus presence is associated with oral disease in implant sites. Conversely, in the PCA plot for original teeth (Figure 23), two of the four healthy samples appeared to be distant from the data cluster. Analysis of their bacterial composition revealed a significant presence of Pseudopropionibacterium and Streptococcus, suggesting these genera's influence on the oral microbiome's overall structure and variability.

Lastly, screen plots (Figures 24-26) were produced for each PCA performed. These plots, which depict the proportion of variance explained by each principal component, indicated that the first two or three principal components accounted for approximately 50% of the variance in our dataset. This observation suggests that PCA was effective in identifying the underlying structure of the data since the first few principal components were able to capture a substantial amount of information in the data. Furthermore, this success reinforces the potential for dimensional reduction, which is especially beneficial for subsequent analyses, such as clustering or regression, as it allows for focusing on the most informative aspects of the data. In summary, the screen plots from our PCA analysis provide a strong basis for concluding that the first few principal components are sufficient to capture the key variance in our dataset, thereby validating the PCA approach and building the foundation for further dimensional reduction and machine learning analysis.

Note that one significant limitation in our usage of PCA is that the lack of sample points. This factor made it challenging to discern any well-structured clusters. In the study, we relied on approximations of cluster and outlier locations. In future studies, it would be beneficial to increase the sample size significantly for a more efficient use of PCA.

4 Discussion

Our investigation encompassed an analysis of the oral microbiome across diverse health conditions. Insights derived from the pie charts underscored a general trend: healthier samples exhibited a more diverse and balanced bacterial microbiome compared to diseased samples. Within the context of implant samples, an escalation in implant severity corresponded to the dominance of specific bacterial genera—such as streptococcus, rothia, and unidentified species—suggesting a potential link between the imbalance of these species and implant infections affecting oral health. In contrast, diseased original teeth samples displayed some diversity, yet certain bacterial genera, including rothia, streptococcus, and Actinomyces, remained abundant. Notably, outliers, such as one ostensibly healthy sample with strong streptococcus dominance, were observed. Nevertheless, the general trend still emphasizes the potential role of bacterial species imbalance in oral health.

Principal Component Analysis (PCA) revealed that the first two principal components collectively accounted for 42.3% of the dataset's variance, indicating the reasonableness of our PCA. However, distinct points, particularly those representing moderate and severe implant cases, deviated from the overall pattern.

For future directions, attention should be directed toward samples containing unknown genera in the pie charts, as these unknown entities could influence the conclusions drawn from our study. Further investigations into the effects of these unknown genera on human health are warranted. Additionally, probing into the specific bacterial genera contributing to the variance observed in the PCA analysis is essential for result validation and a deeper understanding of the underlying dynamics.

Author Contributions

This section is required to all project groups with more than a single member. Place group member's name under the appropriate contribution.

- Study design: Steven Wei Chen, Daniel Ni, Jenny Mao, Soleil Su, Jason Pan
- Coding: Steven Wei Chen
- Experiments:Steven Wei Chen, Daniel Ni, Jenny Mao, Soleil Su, Jason Pan
- Analyses:Steven Wei Chen, Daniel Ni, Jenny Mao, Soleil Su, Jason Pan
- Writing:
  - ○ *Introduction:* Jenny Mao
  - ○ *Methods:Daniel Ni, Soleil Su*
  - ○ *Results:* Steven Wei Chen, Jason Pan
  - ○ *Discussion:* Jenny Mao

References

1. Giordano-Kelhoffer B, Lorca C, March Llanes J, Rábano A, del Ser T, Serra A, Gallart-Palau X, 2022. Oral microbiota, its equilibrium and implications in the pathophysiology of human diseases: A systematic review. *Biomedicines*, .

2. Willis JR, Gabaldón T, 2020. The human oral microbiome in health and disease: From sequences to ecosystems. *Microorganisms*, .

3. Milanese A, Mende DR, Paoli L, et al., 2019. Microbial abundance, activity and population genomic profiling with motus2. *Nature Communications*, 10(1).

4. Chen S, Zhou Y, Chen Y, Gu J, 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890.

5. Ruscheweyh HJ, Milanese A, Paoli L, et al., 2021. motus: Profiling taxonomic composition, transcriptional activity and strain populations of microbial communities. *Current Protocols*, 1(8).

6. Sunagawa S, Mende DR, Zeller G, et al., 2013. Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods*, 10(12):1196–1199.

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

8. Musich R, 2020. *A recent (2020) comparative analysis of genome aligners shows HISAT2 and BWA are among the best tools*. PhD thesis, Rochester Institute of Technology. Order No. 28029122.

9. Houtgast E, Sima V, Bertels K, Al-Ars Z, 24 March 2016. Gpu-accelerated bwa-mem genomic mapping algorithm using adaptive load balancing. In *Architecture of Computing Systems – ARCS 2016*. Springer Cham.

10. Hunter JD, 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

11. Centers for Disease Control and Prevention, 2023. Group a streptococcal (gas) disease. Accessed: [2023].

12. Ramanan P, Barreto J, Osmon D, Tosh P, 2014. Rothia bacteremia: a 10-year experience at mayo clinic, rochester, minnesota. *J Clin Microbiol*, 52(9):3184–3189. Epub 2014 Jun 20.

Figures

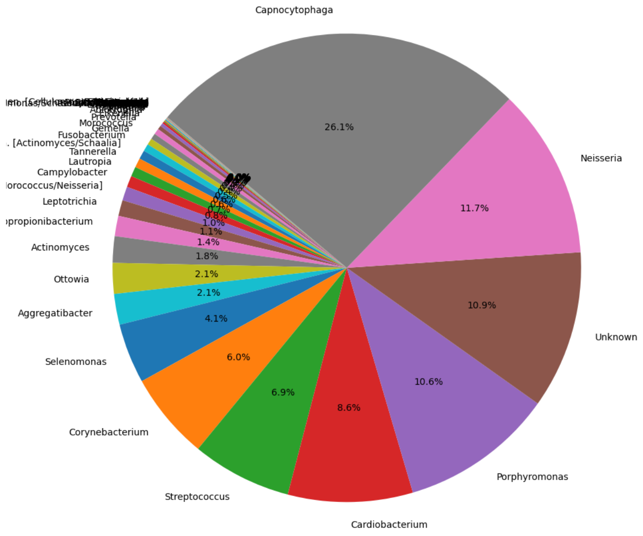


Figure 1. Healthy Implant Sample 1

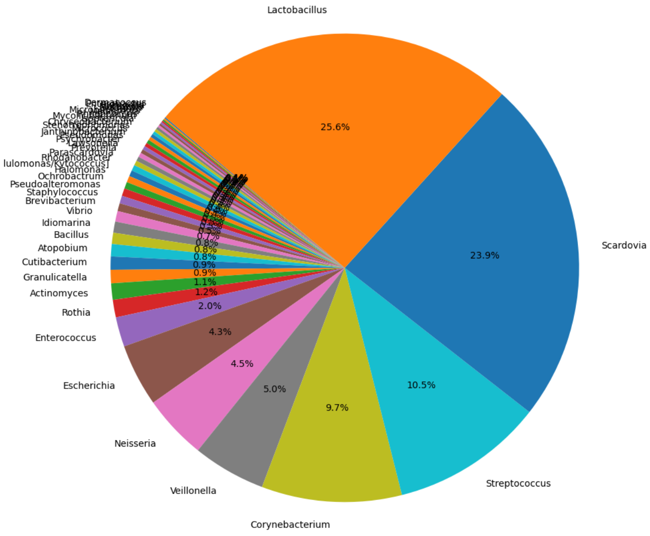


Figure 2. Healthy Implant Sample 2

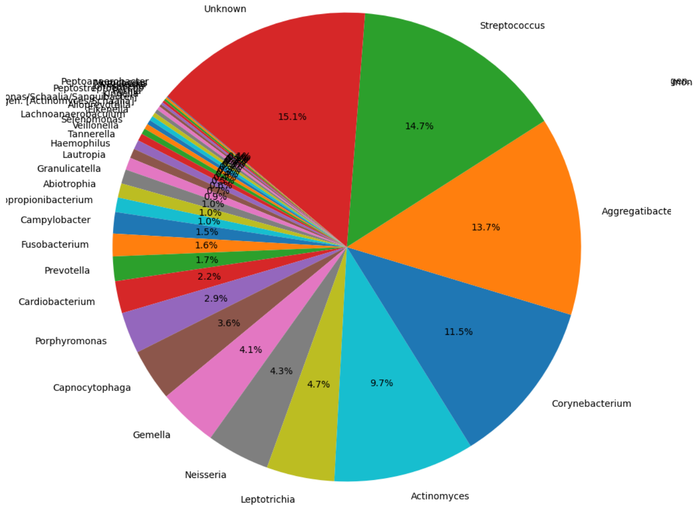


Figure 3. Healthy Implant Sample 3

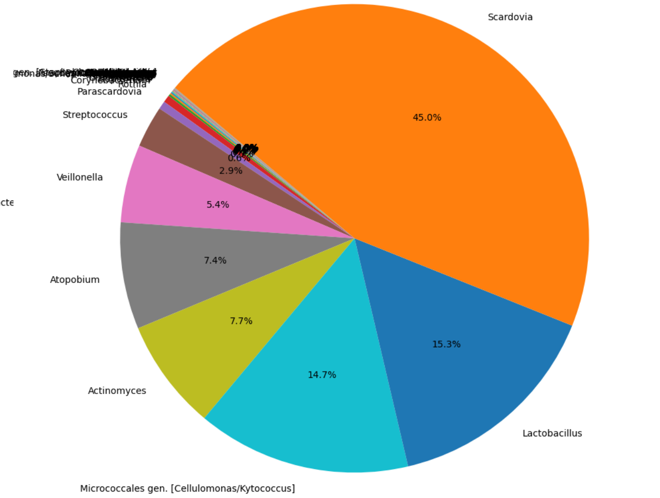


Figure 4. Healthy Implant Sample 4

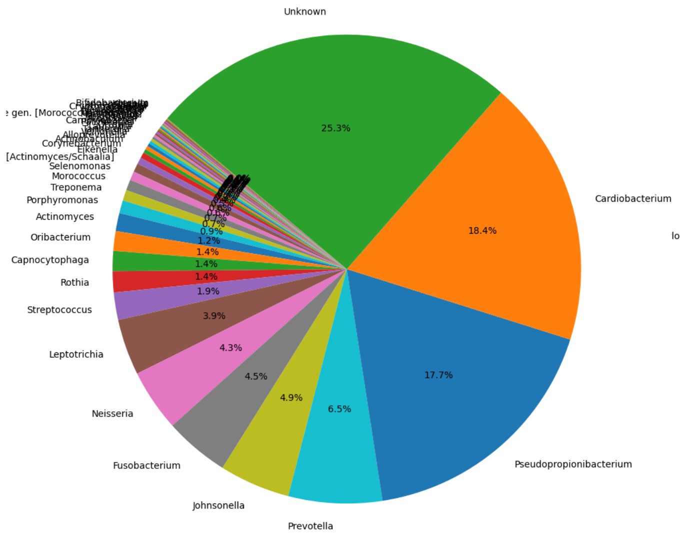


Figure 5. Moderate Implant Sample 1

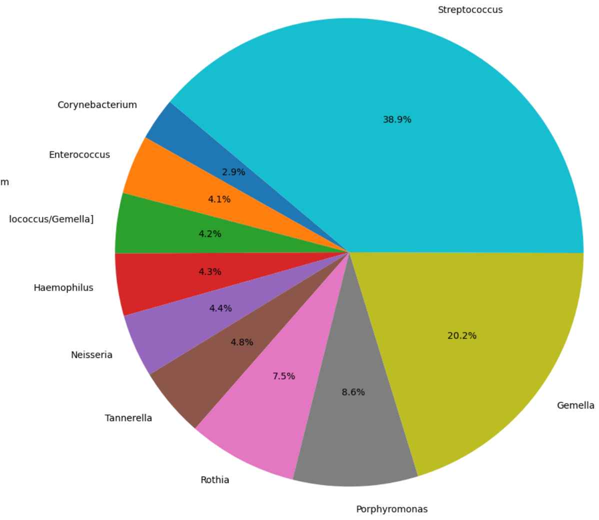


Figure 6. Moderate Implant Sample 2

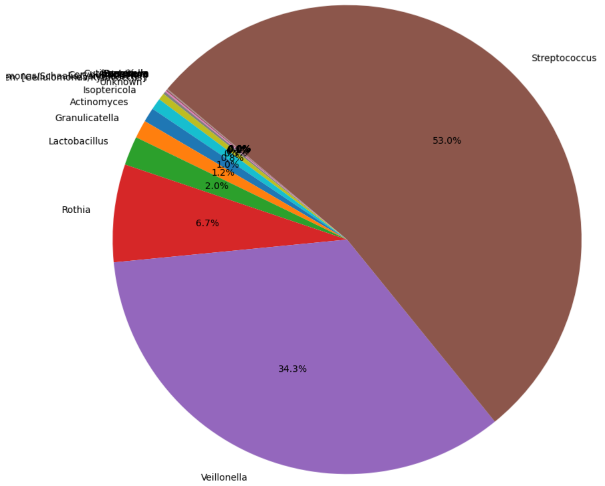


Figure 7. Severe Implant Sample 1

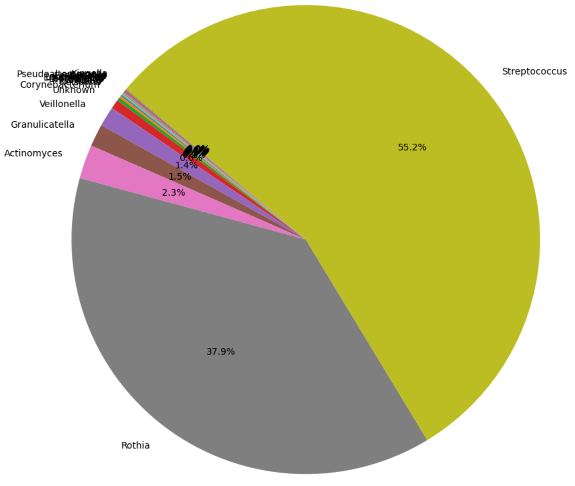


Figure 8. Severe Implant Sample 2

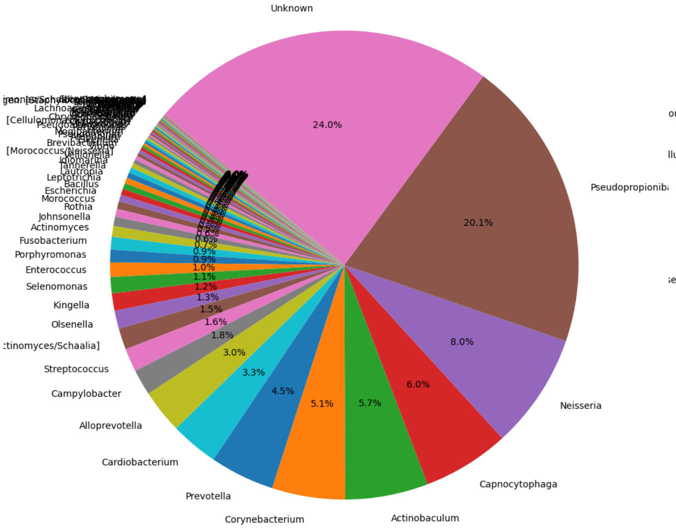


Figure 9. Severe Implant Sample 3

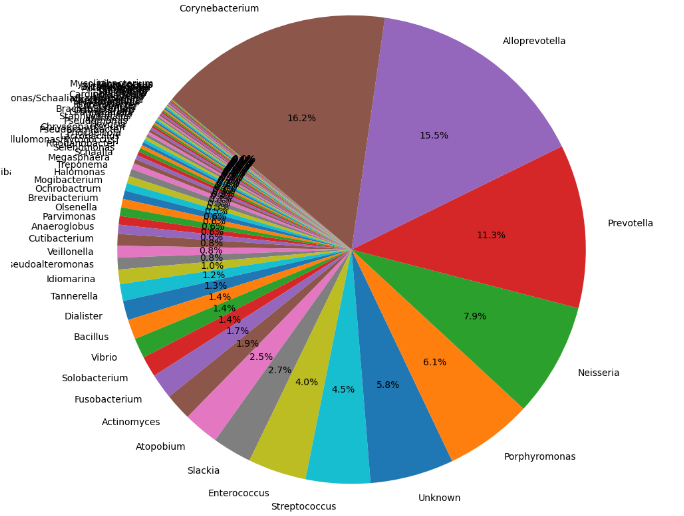


Figure 10. Severe Implant Sample 4

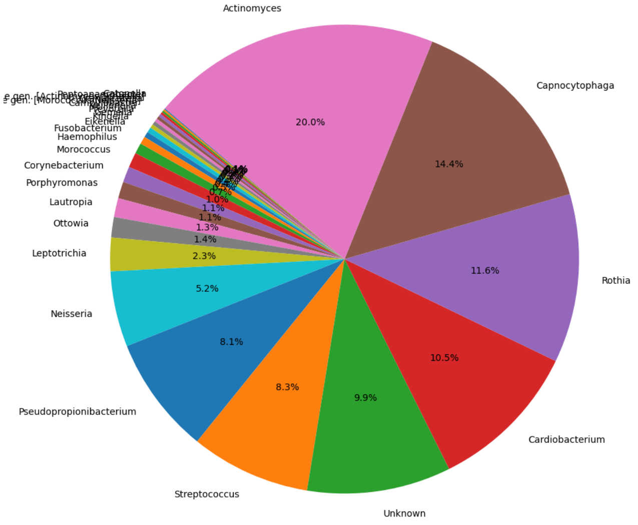


Figure 11. Healthy Tooth Sample 1

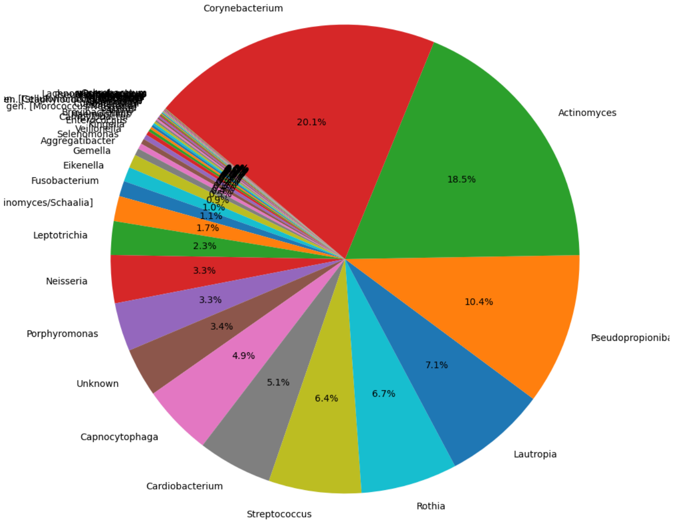


Figure 12. Healthy Tooth Sample 2

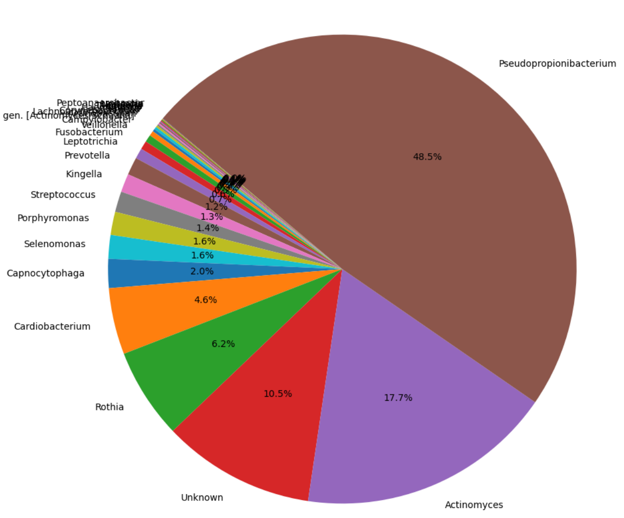


Figure 13. Healthy Tooth Sample 3

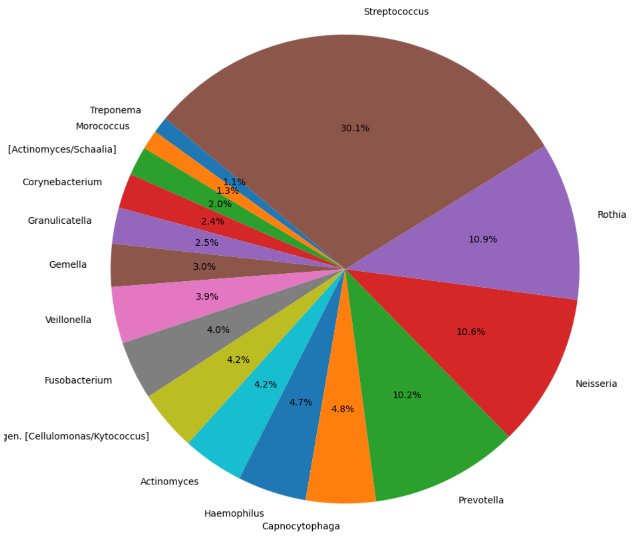


Figure 14. Healthy Tooth Sample 4

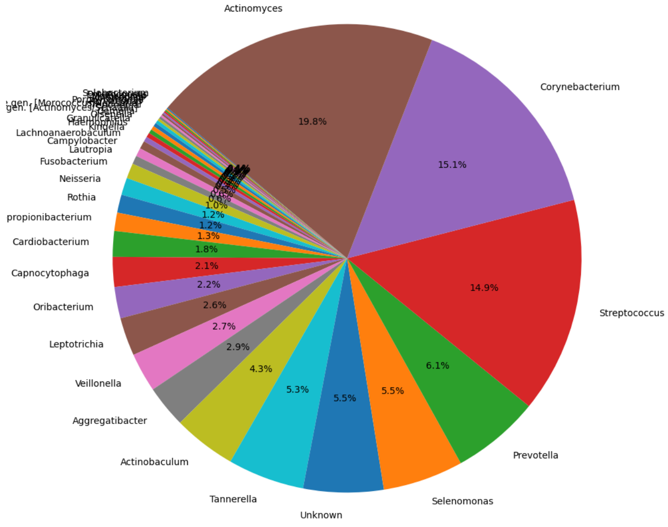


Figure 15. Diseased Tooth Sample 1

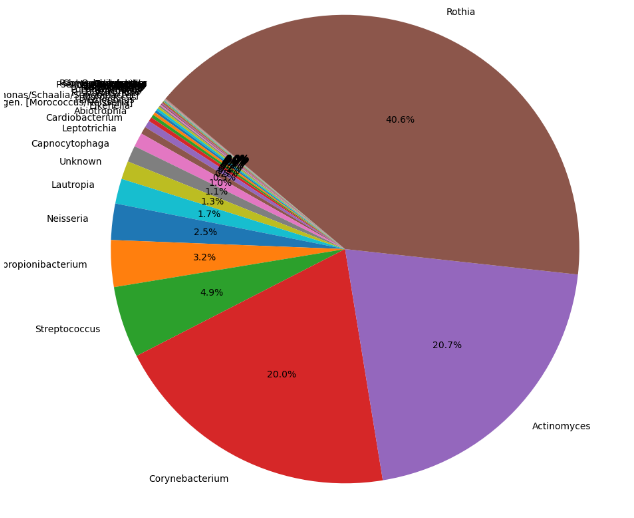


Figure 16. Diseased Tooth Sample 2

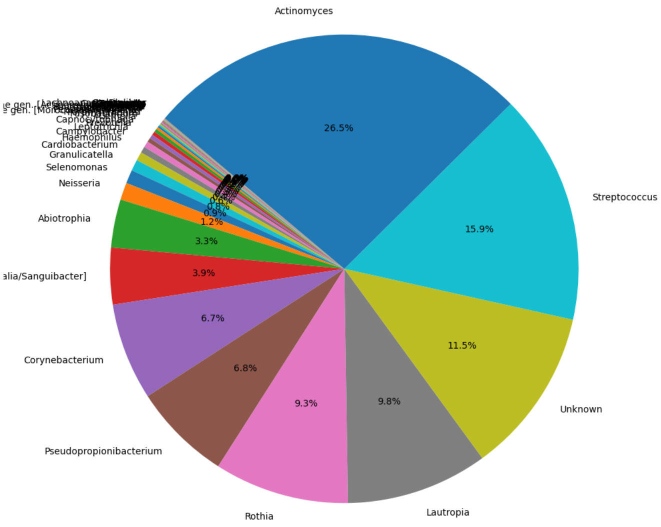


Figure 17. Diseased Tooth Sample 3

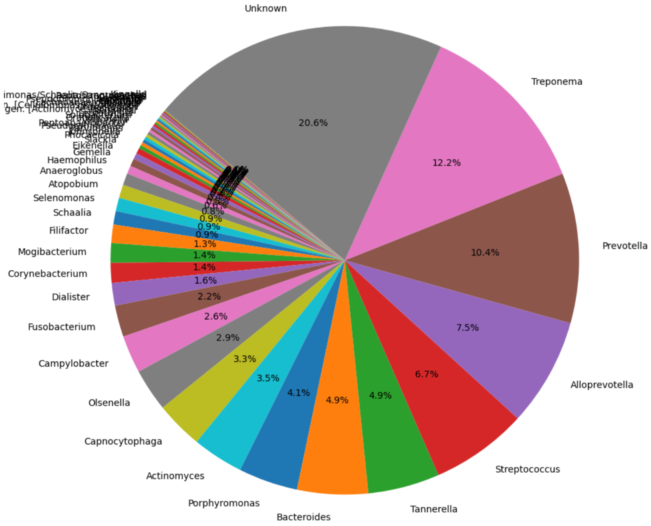


Figure 18. Diseased Tooth Sample 4

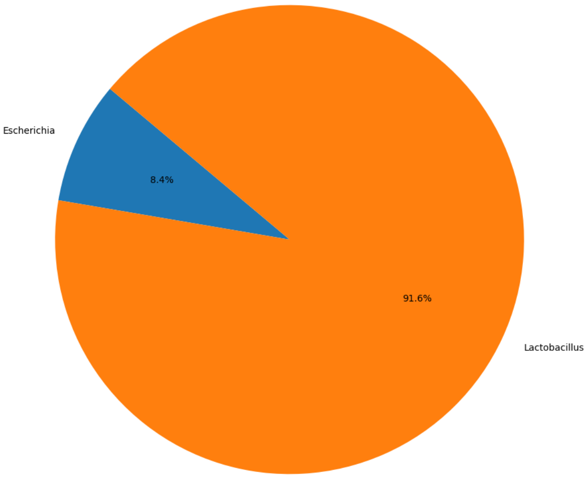


Figure 19. Positive Control

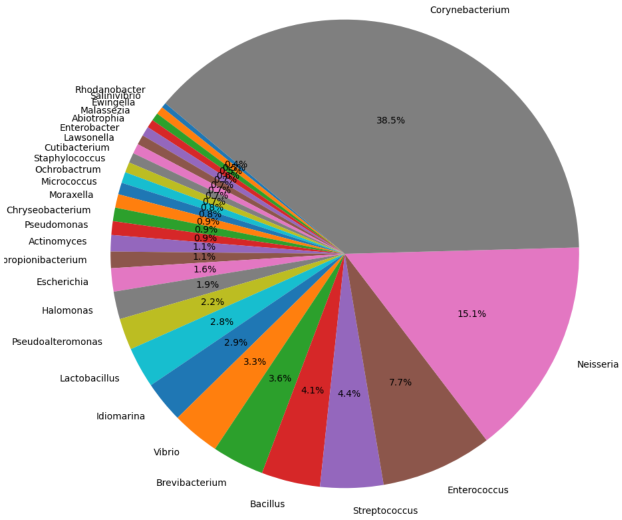


Figure 20. Negative Control

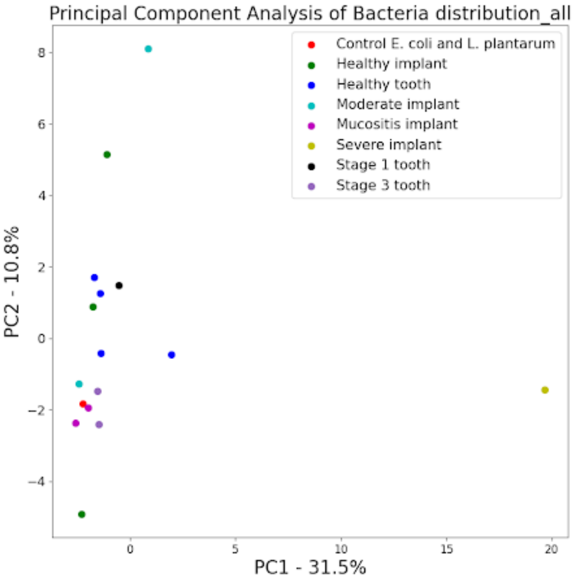


Figure 21. PCA All Samples

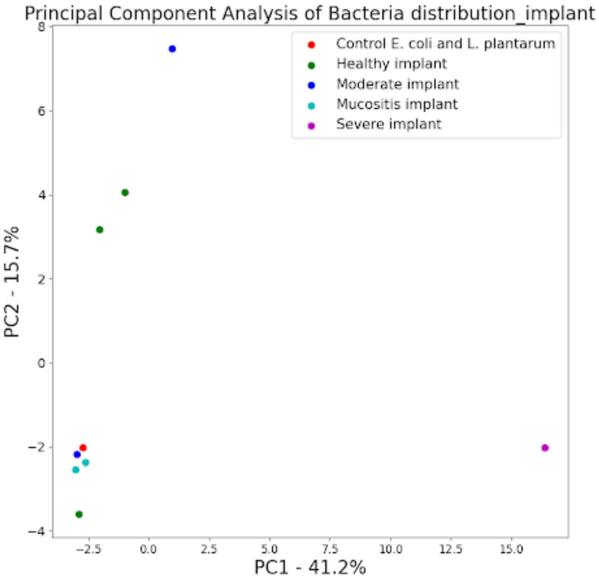


Figure 22. PCA Implant Samples

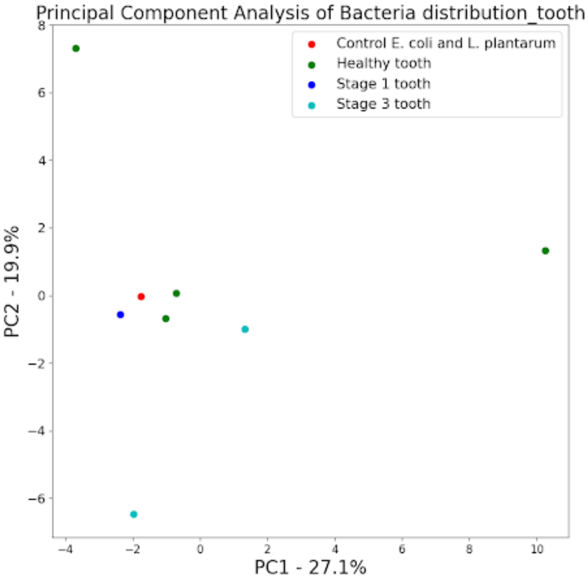


Figure 23. PCA Original Teeth Samples

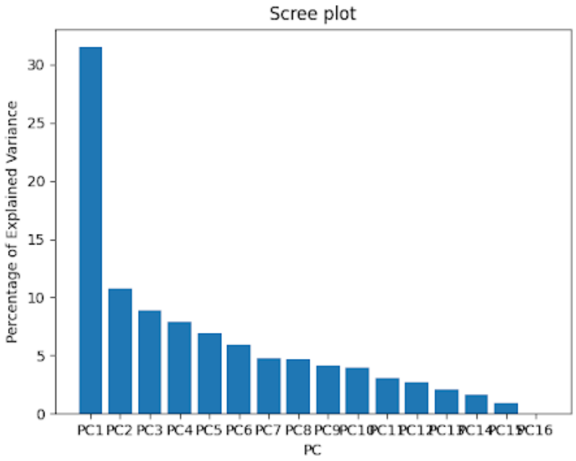


Figure 24. Screen Plot All Samples

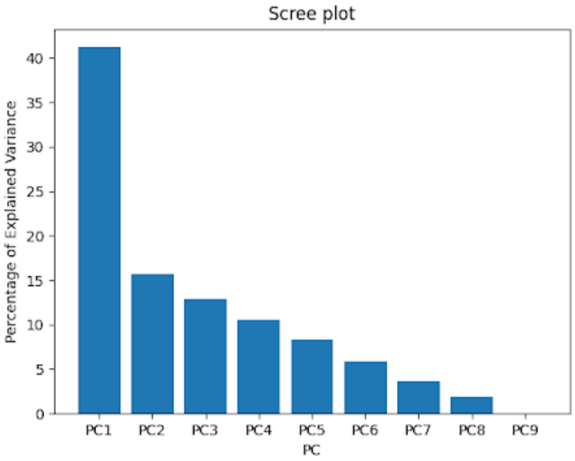


Figure 25. Screen Plot Implant Samples

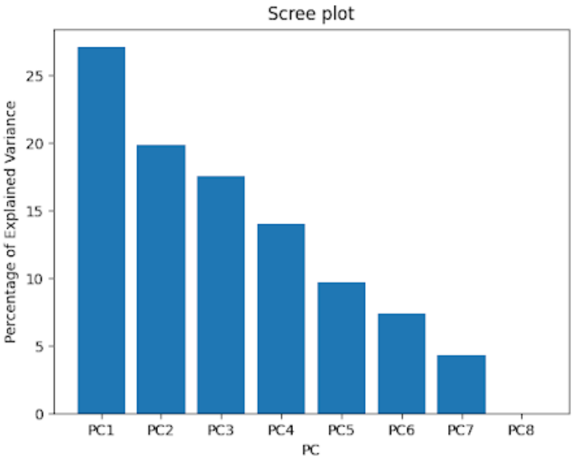
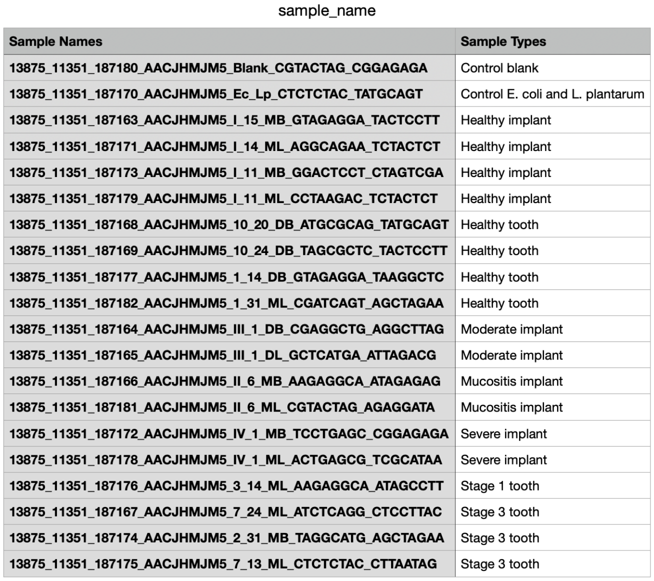


Figure 26. Screen Plot Original Teeth Samples

Tables



sample_name

Sample Names	Sample Types
13875_11351_187180_AACJHMJM5_Blank_CGTACTAG_CGGAGAGA	Control blank
13875_11351_187170_AACJHMJM5_Ec_Lp_CTCTCTAC_TATGCAGT	Control E. coli and L. plantarum
13875_11351_187163_AACJHMJM5_I_15_MB_GTAGAGGA_TACTCCTT	Healthy implant
13875_11351_187171_AACJHMJM5_I_14_ML_AGGCAGAA_TCTACTCT	Healthy implant
13875_11351_187173_AACJHMJM5_I_11_MB_GGACTCCT_CTAGTCGA	Healthy implant
13875_11351_187179_AACJHMJM5_I_11_ML_CCTAAGAC_TCTACTCT	Healthy implant
13875_11351_187168_AACJHMJM5_10_20_DB_ATGCGCAG_TATGCAGT	Healthy tooth
13875_11351_187169_AACJHMJM5_10_24_DB_TAGCGCTC_TACTCCTT	Healthy tooth
13875_11351_187177_AACJHMJM5_1_14_DB_GTAGAGGA_TAAGGCTC	Healthy tooth
13875_11351_187182_AACJHMJM5_1_31_ML_CGATCAGT_AGCTAGAA	Healthy tooth
13875_11351_187164_AACJHMJM5_III_1_DB_CGAGGCTG_AGGCTTAG	Moderate implant
13875_11351_187165_AACJHMJM5_III_1_DL_GCTCATGA_ATTAGACG	Moderate implant
13875_11351_187166_AACJHMJM5_II_6_MB_AAGAGGCA_ATAGAGAG	Mucositis implant
13875_11351_187181_AACJHMJM5_II_6_ML_CGTACTAG_AGAGGATA	Mucositis implant
13875_11351_187172_AACJHMJM5_IV_1_MB_TCCTGAGC_CGGAGAGA	Severe implant
13875_11351_187178_AACJHMJM5_IV_1_ML_ACTGAGCG_TCGCATAA	Severe implant
13875_11351_187176_AACJHMJM5_3_14_ML_AAGAGGCA_ATAGCCTT	Stage 1 tooth
13875_11351_187167_AACJHMJM5_7_24_ML_ATCTCAGG_CTCCTTAC	Stage 3 tooth
13875_11351_187174_AACJHMJM5_2_31_MB_TAGGCATG_AGCTAGAA	Stage 3 tooth
13875_11351_187175_AACJHMJM5_7_13_ML_CTCTCTAC_CTTAATAG	Stage 3 tooth

Figure 27. Samples