Final project: Metagenomic analysis for the Characterization of a uSGB

Alberto Lupatin Andrea Policano Letizia Girardi

May 9, 2024

1 Introduction

1.1 Metagenomic in Microbial Research

Metagenomic can be defined as the study of uncultured microorganisms from the environment, which can include humans or other living hosts with focus on taxonomic and functional characteristics of the total collection of microorganisms within a community. Studying the taxonomical and functional characteristics of microbial communities is getting easier and easier thanks to metagenomic tools available nowadays.

Given how widespread and abundant microbes are, studying the metagenome provides us with plenty of information both on human and non-human microbiome and environment.

1.2 Oral Cavity Diseases

The wide-spread prosthetic restorations based on dental implants enables optimal oral rehabilitations of totally and partially edentulous patients. In fact, the prevalence of dental implants in the global population is estimated to reach up to 23% by the year 2026. The growing number of patients translates to a higher number of potential peri-implant complications, which include peri-implantitis. Peri-implantitis is defined as progressive, irreversible disease affecting both hard (alveolar bone) and soft tissues (supracrestal tissues and mucosa) surrounding dental implants. [1] Even though this disease has a precise definition, peri-implantitis is often a controversial term in medicine, as it is typically used to describe any "less-than-ideal" condition surrounding a dental implant fixture. [3]

1.3 Microbiome Unknowns linked with Peri-Implantitis

Bacteria proliferation plays a key role in the development of peri-implantitis. In fact, previous studies have demonstrated that specific microbial communities could exclusively, or at least predominantly, reside in the biofilm of implants with peri-implantitis.[5] Nevertheless, several bacterial strains and species are yet to be identified.

Metagenomics plays a fundamental role in this sense, as its tools could help us identify the species present in the collected samples to analyze whether a correlation between bacterial species and patients' metadata can be found.

1.4 Theoretical Principles

MAGs (Metagenome-Assembled Genomes): microbial genomes reconstructed from metagenome data. MAGs are fundamental to better understand microbial populations and their interactions with the environment they live in. Moreover, most MAGs belong to novel species, therefore helping to decrease the amount of unknown bacteria. [6] MAGs are typically represented in FASTA format, with each MAG represented as a single file containing the nucleotide sequences of contigs assembled from metagenomic reads.

SGBs (Species Genome Bins): clusters of MAGs binned together as they could possibly represent the genome of a particular species. SGBs are obtained by computing the distance scores between couples of putative genomes. Based on the reference genome, SGBs are then divided into:

- Known (kSGB): a known bacterium is found in the reference database
- Unknown (uSGB): no known bacteria has been found
- Non Human

OTU (Operational Taxonomic Unit) are used in numerical taxonomy. These units may refer to an individual, species, genus or class which are grouped according to their DNA sequence similarity. Usually, two sequences belong to the same OTU when their similarity is above 97%.They are used as pragmatic substituted terms for microbial individuals at different taxonomic levels. [2]

1.5 Motivation and General Workflow

The aim of this report is to obtain as much information as possible about a dental plaque SGB, starting from a set of 30 genomes (MAGs). More specifically, our main goal is to find whether peri-implantitis could be characterized by the bacteria present in different plaque samples. Furthermore, by comparing the results obtained from our analysis with the patients' metadata, we aim to identify some external factors, such as smoking, obesity and sex, that could influence the disease's development. These studies will be based on the analysis of the assembly and of the bacterial genomes.

Regarding the workflow, genomes have firstly been collected from the oral cavity of Italian patients after the dental implants were applied.

Then, bacterial genomes have been sequenced by a high-throughput machine. Finally, the output MAGs were further analyzed.

In order to get the most out of a set of MAGs, we proceeded with the following pipeline:

1. Taxonomic Assignment
2. Genome Annotation
3. Pangenome Analysis
4. Phylogenetic Analysis
5. Association with Host Metadata

2 Methods

2.1 CheckM

CheckM is a software tool used in computational microbial genomics for assessing the quality of microbial genomes recovered from isolates, single cells, or metagenomes. It provides a robust estimation of the quality of genomic data, leveraging lineage-specific marker sets to estimate genome completeness, contamination and heterogeneity. This process is essential for ensuring the reliability of genomic analyses. [4] We used this tool in order to assess the quality of the given samples.

Putting it into practice, for each step, we used a specific conda environment to manage dependencies and isolate its execution environment from the other software tools.

For this analysis we run the following script:

```
conda deactivate && conda activate checkm
checkm taxonomy_wf domain Bacteria Samples checkm_output -t 4
```

where:

- `taxonomy_wf`: command used to perform a taxonomy-specific workflow, which is particularly useful for assigning taxa labels to MAGs. This command requires specifying the taxonomic rank (domain), name (Bacteria), input directory (Samples), and output directory (checkmoutput)
- `-t`: parameter specifies the number of threads: the higher the value, the lower the time needed for the analysis

2.2 PhyloPhIAn

PhyloPhlAn is a computational tool meticulously designed for conducting phylogenetic analyses on genomes and metagenomes, aiming to allocate taxonomic labels to genomes and MAGs through comprehensive whole-genome data. PhyloPhlAn strategically leverages the positioning of MAGs on the phylogenetic tree to deduce their taxonomic labels. The closer a MAG is to the reference genomes of known species or genus on the tree, the more likely it is to be assigned to that taxon. Knowing the taxonomy of a MAG can help predict its functional capabilities. Understanding the taxonomy of a MAG facilitates predictions regarding its functional capabilities. For execution, the following commands have been utilized:

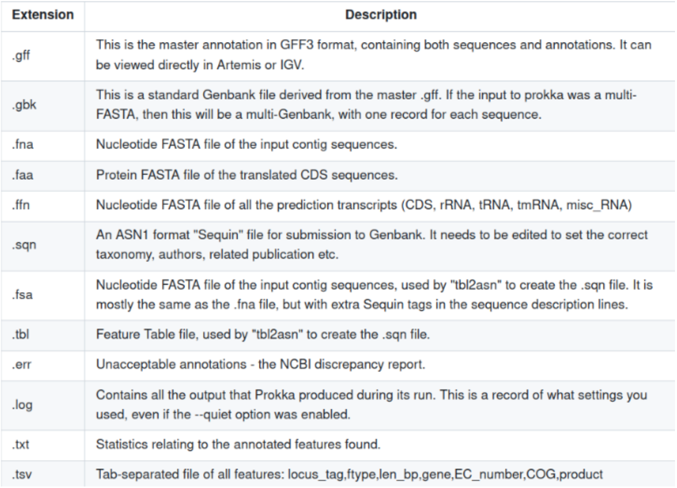
```
conda deactivate && conda activate ppa
phylophlan_metagenomic -i SGB1566/Samples -e ".fna" -d CMG2324 --database_folder
ppa_db --nproc 4 --verbose -n 1
```

Where:

- `-nproc`: argument for expediting the analysis by engaging 4 processing cores
- `-n`: argument for specifying the number of resulting SGB
- `-database_folder ppa_db`: argument specifying the path to the folder containing the downloaded reference database

2.3 Prokka

Prokka is a versatile command line tool designed for rapid gene annotation and feature identification within prokaryotic genome sequences. It operates with preassembled genomic DNA sequences in FASTA format, ideally devoid of gaps. The sequence file serves as the sole mandatory parameter for the software. Prokka uses external feature prediction tools such as Prodigal and Aragorn to precisely define the coordinates of genomic features within contigs. The process unfolds in two primary steps: initially, protein-coding regions are identified via Prodiga; subsequently, encoded protein functions are predicted by comparing them to single proteins, or domains, databases. Where no matches are found, annotations are designated as *hypothetical protein*. Prokka generates 13 files within the specified output directory Figure 1



Extension	Description
.gff	This is the master annotation in GFF3 format, containing both sequences and annotations. It can be viewed directly in Artemis or IGV.
.gbk	This is a standard Genbank file derived from the master .gff. If the input to prokka was a multi-FASTA, then this will be a multi-Genbank, with one record for each sequence.
.fna	Nucleotide FASTA file of the input contig sequences.
.faa	Protein FASTA file of the translated CDS sequences.
.ffn	Nucleotide FASTA file of all the prediction transcripts (CDS, rRNA, tRNA, tmRNA, misc_RNA)
.sqn	An ASN1 format "Sequin" file for submission to Genbank. It needs to be edited to set the correct taxonomy, authors, related publication etc.
.fsa	Nucleotide FASTA file of the input contig sequences, used by "tbl2asn" to create the .sqn file. It is mostly the same as the .fna file, but with extra Sequin tags in the sequence description lines.
.tbl	Feature Table file, used by "tbl2asn" to create the .sqn file.
.err	Unacceptable annotations - the NCBI discrepancy report.
.log	Contains all the output that Prokka produced during its run. This is a record of what settings you used, even if the --quiet option was enabled.
.txt	Statistics relating to the annotated features found.
.tsv	Tab-separated file of all features: locus_tag,ftype,len_bp,gene,EC_number,COG,product

Figure 1: Prokka's output files

2.4 Roary

A rapid standalone pan genome pipeline, swiftly constructs the pan genome across numerous prokaryotic samples. This analysis defines the genetic architecture of prokaryotic genomes, enhancing comprehension of critical processes like selection and evolution.

When inputting the annotated assemblies per sample (MAGs) in .GFF format -generated by Prokka- the tool extracts coding regions and converts them to protein sequences. These sequences undergo a filtration process to eliminate partial sequences. Then, an all-against-all comparison ensues via BLASTP on the refined sequences, employing a user-defined percentage sequence identity (typically set at 95%). Lastly, a graph is built to illustrate cluster relationships based on their order of occurrence in the input sequences. Isolates are clustered based on gene presence in the accessory genome, with the weight of isolated contributions to the graph determined by cluster size.

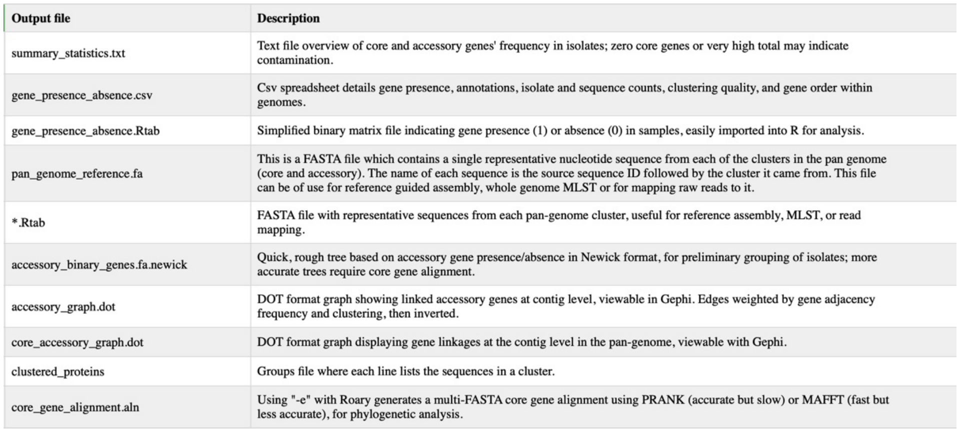
To run the software and create a pan genome we used the roary script.

```
conda deactivate && conda activate roary
roary */*.gff -f roary_output_w_aln -cd 90 -p 8 -e -n
```

where:

- `-i 95%`: minimum percentage identity for blastp
- `-cd`: percentage of isolates a gene must be in to be core. Two thresholds (95% and 90%) have been set in order to study the behavior changes
- `-e`: flag for creating a multiFASTA alignment of core genes using PRANK
- `-n`: flag for using the faster MAFFT algorithm for creating the FASTA alignment

In order to visualize the output data 2, we used the python script roary_plots.py available on GitHub and two different online tools: Alignment and Tree.



Output file	Description
summary_statistics.txt	Text file overview of core and accessory genes' frequency in isolates; zero core genes or very high total may indicate contamination.
gene_presence_absence.csv	Csv spreadsheet details gene presence, annotations, isolate and sequence counts, clustering quality, and gene order within genomes.
gene_presence_absence.Rtab	Simplified binary matrix file indicating gene presence (1) or absence (0) in samples, easily imported into R for analysis.
pan_genome_reference.fa	This is a FASTA file which contains a single representative nucleotide sequence from each of the clusters in the pan genome (core and accessory). The name of each sequence is the source sequence ID followed by the cluster it came from. This file can be of use for reference guided assembly, whole genome MLST or for mapping raw reads to it.
*.Rtab	FASTA file with representative sequences from each pan-genome cluster, useful for reference assembly, MLST, or read mapping.
accessory_binary_genes.fa.newick	Quick, rough tree based on accessory gene presence/absence in Newick format, for preliminary grouping of isolates; more accurate trees require core gene alignment.
accessory_graph.dot	DOT format graph showing linked accessory genes at contig level, viewable in Gephi. Edges weighted by gene adjacency frequency and clustering, then inverted.
core_accessory_graph.dot	DOT format graph displaying gene linkages at the contig level in the pan-genome, viewable with Gephi.
clustered_proteins	Groups file where each line lists the sequences in a cluster.
core_gene_alignment.aln	Using "-e" with Roary generates a multi-FASTA core gene alignment using PRANK (accurate but slow) or MAFFT (fast but less accurate), for phylogenetic analysis.

Figure 2: Roary's output files

2.5 FastTree

This tool is able to infer an approximately-maximum-likelihood phylogenetic tree using nucleotide or protein sequences. It is a fast open-source software that can handle alignments with up to a million of sequences in a reasonable amount of time and memory.
Its pipeline can be divided into four steps:

- Heuristic Neighbour-Joining to get a rough topology
- Reducing the Length of the Tree
- Maximizing the tree's Likelihood
- Estimate the Reliability of each split in the tree

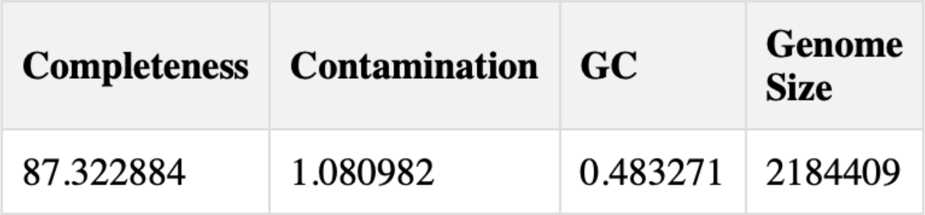
From a practical point of view, we installed the ETE Toolkit and used the function Phylo from the Biopython package.

3 Results

3.1 Description of the set of genomes

The dataset contains a selection of individuals with detailed records regarding sex, BMI, age, smoking status, and health status. Through the use of PhyloPhlAn it was possible to recognise that the samples belongs to the same SGB of an unknown element of the Bacteroidota phylum, known to be Gram- and opportunistic pathogens, even though in normal conditions they present a stable symbiotic relationship with their host.
In addition, the second nearest known element was the known bacterium *Prevotella pleuritidis*: known to be obligate anaerobic and non-motile, they also are non-spore forming and Gram-negative. Aside from that, *P.pleuritidis* is especially associated with disease connected to the respiratory apparatus, also typically present in cases of pleura inflammation (pleuritis) and abscesses, presenting, as for the previous case, an antagonistic behavior against the host.

Examining the MAGs quality from the *bin_ state_ ext* Checkm's output file, we found the dataset to be of high quality, with an average completeness of 87.32% and a mean contamination of 1.08%, as shown in Figure 3. The GC content ranged from 47.42% to 49.73%, slightly below typical values and the genome size the range goes from 1277960 up to 2522737 nucleotides.



Completeness	Contamination	GC	Genome Size
87.322884	1.080982	0.483271	2184409

Figure 3: MAGs quality check - average values

3.2 Genome annotation

MAGs analysis was conducted using Prokka's TSV files, with each file detailing the genome of a sample. For every protein-coded gene, information includes locus tag, type (CDS or tRNA), length (bp), gene name, Enzyme Commission (EC) number for catalysis classification, COG, and involved reactions.

Numerical analysis in Bash and R revealed 55,353 proteins, with 44% known and 53% hypothetical, plus a small fraction of tRNAs, rRNAs, tmRNAs, and repeated regions.

Following that, we verified what the core genes were, based on its intrinsic definition: by searching what genes were present in every sample. This would help us to see whether these genes' functions could be connected to oral cavity's diseases thus providing some insight regarding the disease development.

Firstly, we filtered by the EC_number resulting in 48 core genes. However, EC_number is not an optimal parameter in this case study, as different genes may have the same EC_number.

In order to correct this bias, we filtered by gene name. In this case, only 11 core genes were found. Those 11 elements were further analyzed through an enrichment analysis, using the EnrichR software available online. The main section taken in consideration were Elsevier pathway collection, GWAS Catalog 2023 and KOMP2 Mouse Phenotype. 4

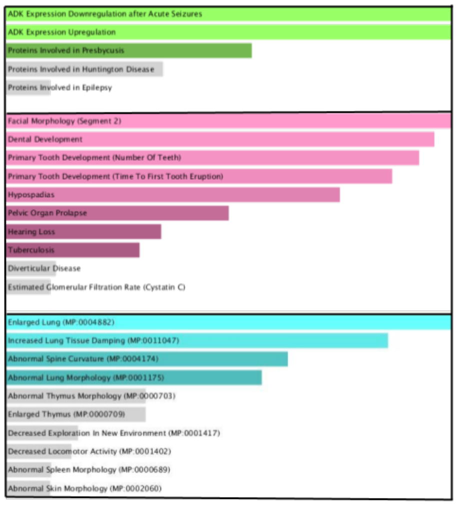


Figure 4: Enrichment Sections of Interest

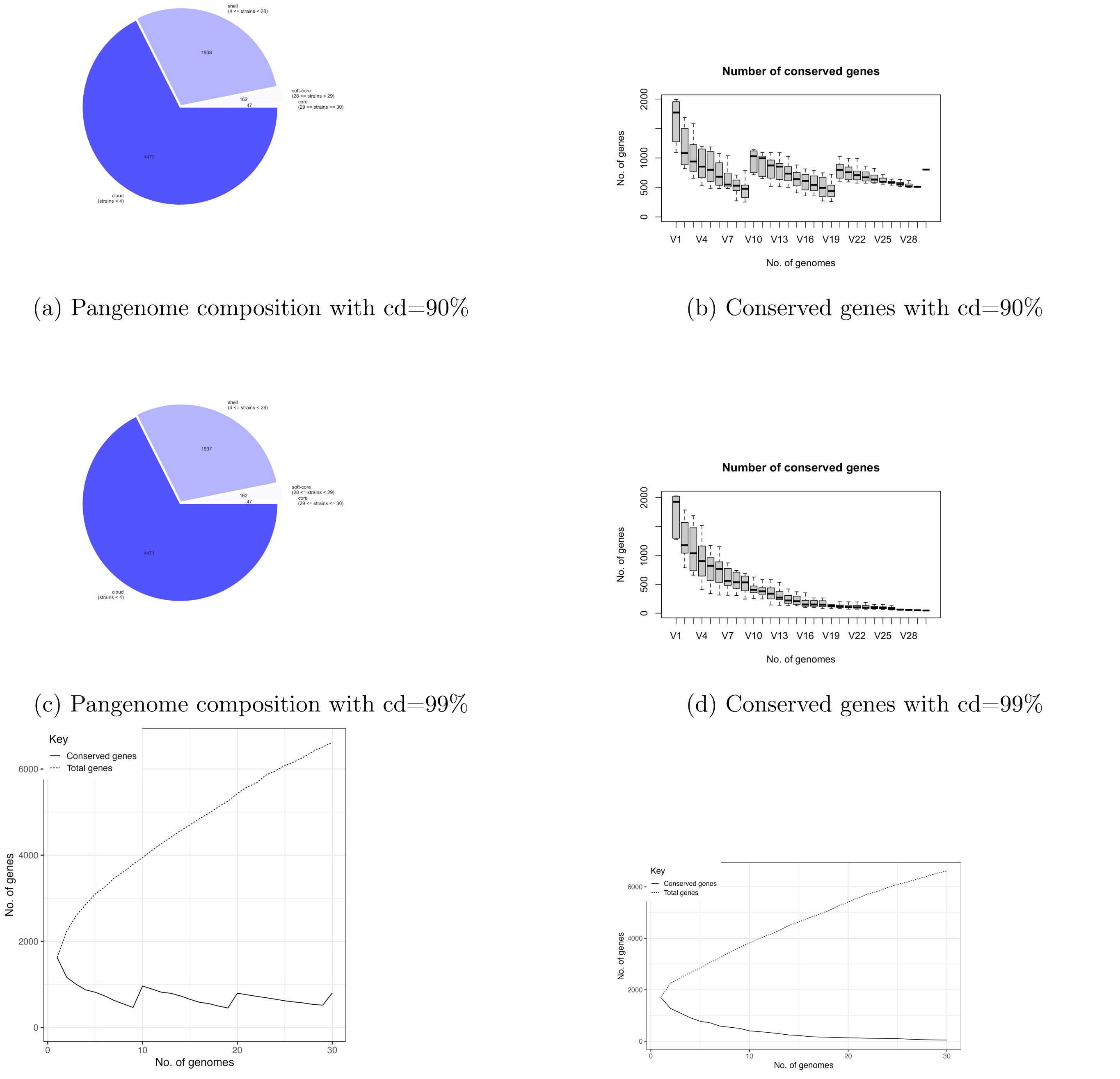
The results showed a high association between the 11 genes presumably defined as core and ADK differential expression in specific conditions (usually connected with cancer, type 1 and 2 diabetes and other inflammatory related and epilepsy). Aside from that, ADK plays a fundamental role in many enzymatic reactions linked to regulating the levels of Adenosine in many cellular mechanisms, especially those involved in epigenetic and metabolic/bioenergetic functions, therefore being an efficient target for opportunistic pathogens such as those belonging to the Bacteroidetes phylum. Moreover, this pathway is correlated with facial morphology in humans, especially concerning the dental apparatus, giving us more certainties regarding the possibility of having found a possible, perhaps different strain highly similar with the two already mentioned in the result section "Description of the set of Genomes".

3.3 Pangenome Analysis

Input from Prokka's genome annotation yielded a pangenome of 6621 genes at 90% CD: 209 core (soft-core = 162; core = 47) and 6411 accessory (shell = 1938; cloud = 4473) genes. However, a limit of setting a not-so-strict parameter is shown in Figure 5b, representing the number of conserved genes. This pattern is a known mathematical issue of Roary.

As a consequence, we increases this parameter to 99. The pangenome doesn't change severely: is composed by 6621 genes, 209 of which represent the core genome (Soft-core = 162 ; Core = 47), while the remaining 6410 belonged to the accessory genome (Shell = 1937 ; Cloud = 4471). 5c On the other hand a significant difference is observed in the graph representing the number of conserved genes.

In both cases, we found the pangenomes are open, so new genes can be added without expanding the core genome, as shown by the increasing total gene count versus stable conserved genes in Figure 5e.



(a) Pangenome composition with cd=90%

(b) Conserved genes with cd=90%

(c) Pangenome composition with cd=99%

(d) Conserved genes with cd=99%

(e) Conserved genes vs total genes with cd=90%

(f) Conserved genes vs total genes with cd=99%

Figure 5: Pangenome Analysis

3.4 Phylogenetic Analysis with Host Metadata

After building the phylogenetic tree 6, we proceeded to do further analysis by merging the samples with the respective metadata. Even though no statistical analysis were performed, some inferences can still be made. In particular, by looking at the clusters, it's possible to define whether a correlation can be found between some specific strains and some particular patients' characteristics.

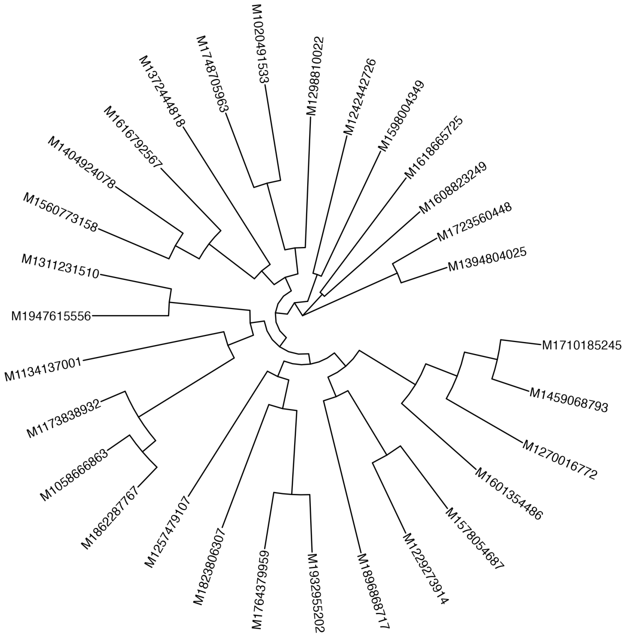


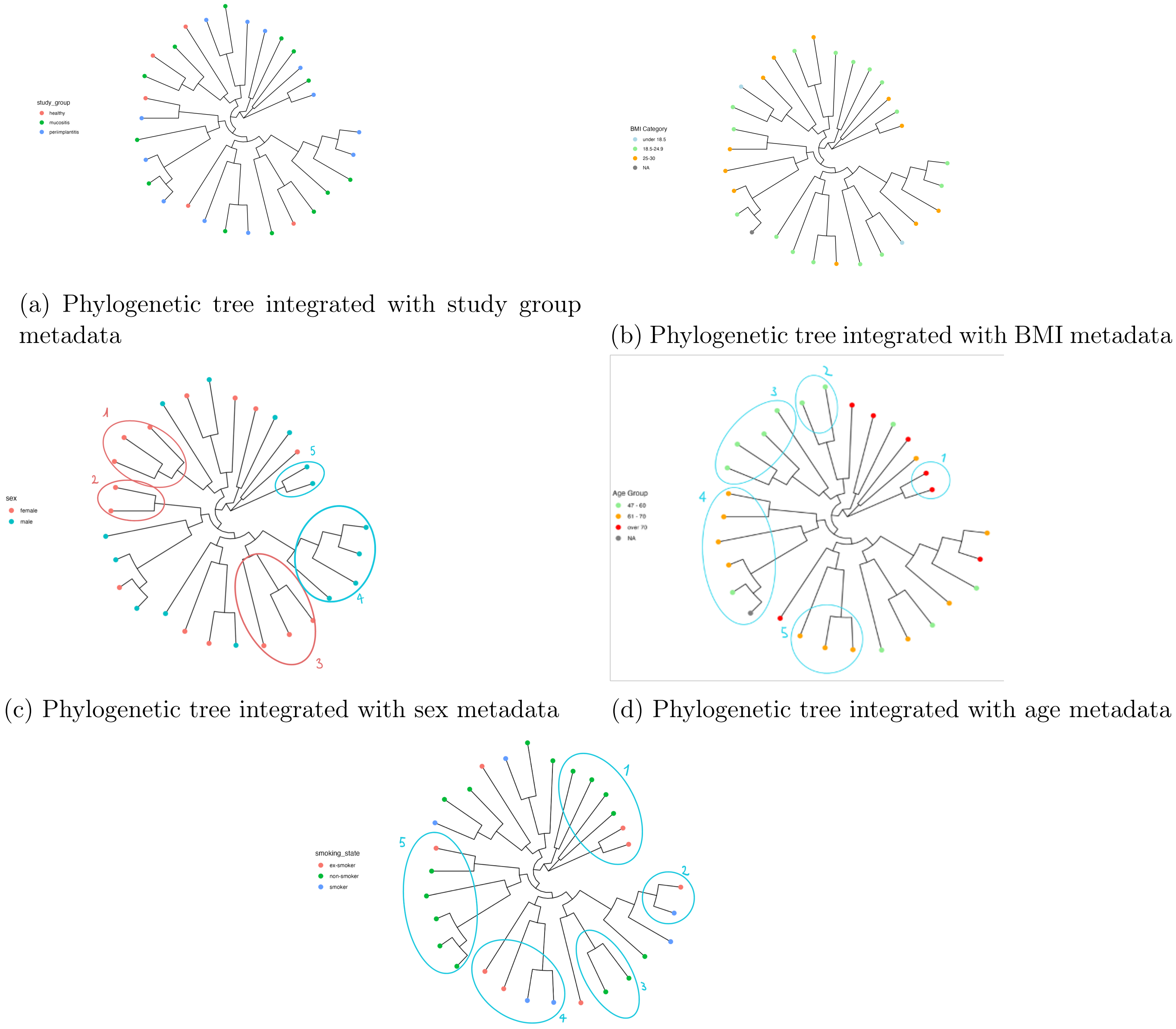
Figure 6: Phylogenetic tree with labels

We tested all the five metadata types: sex, BMI, age, smoking state, study group.

Firstly, we integrated metadata regarding the patients' sexes. As shown in Figure 7c, we can isolate 5 clusters. In particular, clusters 1 and 2 seem to be correlated with males, while clusters 2 and 3 seem to be correlated with females. As a consequence, we can establish that a correlation between sex and some bacterial strains can be found.

Secondly, we analyzed BMI correlations by categorizing patients into the five WHO's recommended categorization: underweight (BMI < 18.5), healthy weight ($18.5 <$ BMI > 25), overweight ($25 <$ BMI > 30), obese (≥ 30). Despite some strains including healthy weight individuals can be isolated, no significant correlation emerged.

After that, we integrated ages metadata. As shown in Figure 7d, we can identify 5 clusters. In particular, three clusters seem to be correlated with an age minor to 60 years old while two clusters seem to be correlated with an age higher than 70 years old. As a consequence, we can establish that a correlation between age and some bacterial strains can be found.



(a) Phylogenetic tree integrated with study group metadata

(b) Phylogenetic tree integrated with BMI metadata

(c) Phylogenetic tree integrated with sex metadata

(d) Phylogenetic tree integrated with age metadata

(e) Phylogenetic tree integrated with smoking state metadata

Figure 7: Phylogenetic Analysis with Host Metadata

Regarding the smoking state, we divided patients into three categories: those who smoke, those who don't smoke and those who stopped smoking. Since detailed metadata regarding the latter is not provided, some speculation can be made. 7e

For instance, in cluster number 5 we can assume that the only ex-smoker present hasn't smoked so severely in its life, so it can be clustered with other non-smoking samples.
Unlike the previous case, looking at cluster 1, a separation between ex-smokers and non-smokers can be seen; however, there is still some similarity between these two sample types. Hence, here we can assume that smoking has partly affected the presence of a specific bacterial clade. Similarly, in cluster 4 we suppose that patients who quitted smoking have a slightly different bacterial clade than patients that still smoke.
Bearing in mind all the above, these hypotheses remain as is without more detailed metadata and statistical analysis. Still, considering all biases, one possible statement is that maybe smoking still affects the oral cavity microbiome even after quitting.

Lastly, we checked a correlation with the study group. The samples were divided into 3 categories: healthy, with peri-implantitis and with mucositis. Even though this metadata was expected to be the most relevant, as it is known that the microbiome is severely influenced by the disease's status, no cluster can clearly be isolated.

4 Conclusions

This study aimed to unravel the microbial complexities associated with peri-implantitis by leveraging the power of metagenomics. Through the examination of 30 MAGs, we sought to identify specific bacterial communities that could be linked to peri-implantitis, understanding the potential influence of various host factors like smoking, obesity and sex. Our study revealed that, even if our MAGs don't match any already known SGB, the samples predominantly belonged to an uncharacterized species genome bin within the Bacteroidota phylum. Notably, the genera Capnocythophaga and Sphingobacterium, along with elements of the Prevotella genus, have been identified as potential pathogens. Moreover, the pangenome analysis indicated an open pangenome structure, suggesting a high degree of genetic diversity of it. Finally, the integration of host metadata, including sex, age and smoking status, suggests correlations between certain microbial strains and host characteristics. Nonetheless, additional studies could be useful to further clarify the complex interactions between the oral microbiome and host factors in peri-implantitis.

References

[1] M. Chmielewski and A. Pilloni. Current Molecular, Cellular and Genetic Aspects of Peri-Implantitis Disease: A Narrative Review. *Dent J (Basel)*, 11(5), May 2023.

[2] Y. He, J. G. Caporaso, X. T. Jiang, H. F. Sheng, S. M. Huse, J. R. Rideout, R. C. Edgar, E. Kopylova, W. A. Walters, R. Knight, and H. W. Zhou. Erratum to: Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome*, 3:34, 2015.

[3] Raza A. Hussain, Michael Miloro, and Jennifer B. Cohen. An update on the treatment of peri-implantitis. *Dental Clinics of North America*, 65(1):43–56, 2021. Implant Surgery.

[4] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*, 25(7):1043–1055, Jul 2015.

[5] P. Sahrmann, F. Gilli, D. B. Wiedemeier, T. Attin, P. R. Schmidlin, and L. Karygianni. The Microbiome of Peri-Implantitis: A Systematic Review and Meta-Analysis. *Microorganisms*, 8(5), May 2020.

[6] J. C. Setubal. Metagenome-assembled genomes: concepts, analogies, and challenges. *Biophys Rev*, 13(6):905–909, Dec 2021.