UNIVERSITÀ DI TRENTO

Computational Microbial Genomics

Project Report
Zehra Korkusuz & Eevi Sipponen
Code/Output repository: github.com/zehrakorkusuz/Microbial_Genomics

A. Introduction

The oral microbiome, a community of microbial organisms in oral cavity plays an important role in maintaining oral health by aiding digestion, inhibiting the growth and dominance of pathogenic microorganisms, regulating pH to protect the tooth enamel from erosion and even modulating the immune response locally and systemically.

Oral microbiome is a preventive barrier but some organisms in this community can be detrimental to oral health. Some species are found in dental plaques and are associated with dental decay and gum disease. [1] Plaques are naturally formed by bacteria, and they can be managed with adequate oral hygiene, somehow when oral hygiene is compromised and this plaque is left to accumulate it can create tartar, a hardened plaque on the base of the teeth which eventually can cause bad breath, tooth decay, gum disease.

Dental implants are an effective treatment for replacing teeth and provide favorable outcomes in restoring aesthetics and functionality as well as preserving bone tissue and improving oral health but also susceptible to plaque formation like natural teeth. After implants, plaque can create an inflammatory response in the tissues surrounding the implant and cause peri-implant mucositis. Mucositis is the inflammation of the mucosal tissue without bone loss and is a reversible condition. If left untreated, mucositis may advance to peri-implantitis and can cause bone loss. Different stages are presented in Figure 1.

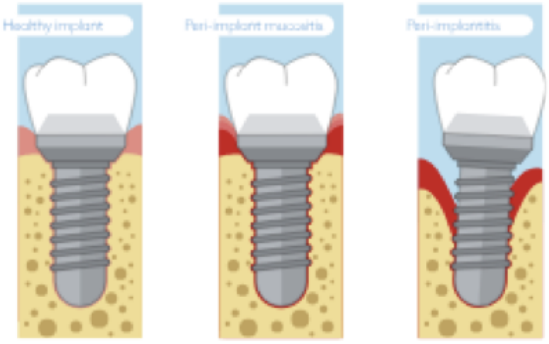

Figure 1: Comparative Representation of Dental Implant Post-Treatment Conditions [2].

In this project, we study unknown Species-Level Genome bins (uSGBs) prevalent in the dental plaque of 30 people with different health conditions including healthy, mucositis or peri-implantitis. The samples were taken from patients after varying times post-implant placement. Therefore samples reflect the clinical state of an implant at the sampling moment, not after a fixed period. All samples were sequenced with Illumina NovaSeq 6000 System, using shotgun metagenomics approach. This allows genomic analysis at species level and

analysis of the functional profile of the microbial communities without a priori information of the species. Considering that the reads may also involve unknown species, all samples were assembled using the same de Novo assembly approach to create contigs and then binning to create metagenome-assembled genomes (MAGs) and cluster the SGBs. Our dataset has an unknown SGB labeled as uSGB985.

Aiming to get the most out of the dataset, we employ several steps and use bioinformatics libraries to understand the taxonomy, characteristics and functionality of the SGB and analyze it along with the metadata to see if there is a common trait in unhealthy groups which bacteria strains might showcase some pathogenicity.

B. Methods

B.1 Quality control with CheckM

From sequencing to assembly techniques, methods introduce biases which can lead to misinterpretation of the results. CheckM is an automated quality control tool that estimates the completeness, contamination by analyzing the presence and lack of the universal single copy genes in prokaryotes. Additional information such as coding density, coverage, GC content are also provided in the outputs. [3] These statistical properties vary among organisms and can be a quality measure of the assembly. For instance, GC content distribution shows us the consistency and variations of the GC content between different bins as well as the potential contamination when we see bimodal distributions. Example of the distribution on Figure 2.

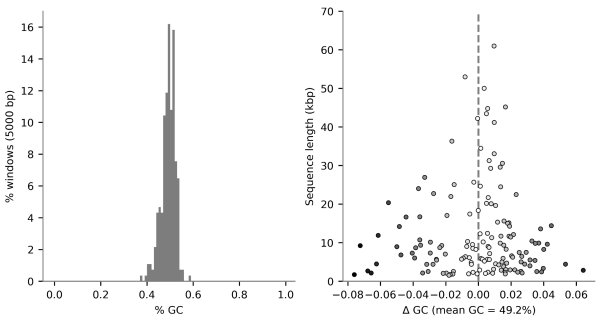

Figure 2: Example on GC content plot run on a MAG.

In the novel genomic content and lack of specific lineage markers, CheckM has some limitations. Loss of genes in bacteria can overestimate incompleteness. CheckM is able to provide refined estimates for well-characterized lineages for reduced genomes. But to improve quality estimates for genomes recovered from novel lineage, a manual assessment of duplication or gene loss is required. [3]

In order to overcome this CheckM2 uses machine learning to capture more complex relationships, particularly in metagenomic analysis [4]. We opt for CheckM as a more conventional approach.

B.2 Taxonomic Assignment with PhyloPhlAn 3.0

Taxonomic assignment is the classification of the Operational Taxonomic Units (OTUs) to the sequences and the reads of the genome. Identification of certain microbial organisms and the diversity allow us to study the role of these microbial communities in maintaining health, differentiate commensal strains from pathogenic strains and develop strategies to overcome disease states such as developing antibiotics.

Phylophlan is a phylogenetic analysis tool, including a taxonomic assignment pipeline for assigning mags/genomes to the closest species. Phylophlan integrates public datasets and known species to assign the closest species to the MAGs. For different resolution levels in the phylogenetic tree, Phylophlan uses a different set of markers. For higher levels of the tree, 400 universal markers are used while assigning species and strains, UniRef90 proteins are selected as phylogenetic markers. UniRef90 refers to sequences clustered if 90% identical in the proteome database, UniProt. [5] This approach utilizes multi-resolution phylogenetic reconstruction to approach assigning taxonomic labels from phylum to species level to input genomes or MAGs. PhyloPhlAn 3.0 database has over 150 000 MAGs and 80 000 reference genomes integrated from which it uses. [4]

We used the phylophlan_metagenomic pipeline to assign uSGBs to the closest taxonomy.

Customized parameters: n=1, database=CMG2324, nproc=4

n: number of SGBs to reported based on the sorted distance, nproc: thread number

For each bin of the dataset, also a mash distance is calculated to estimate the distance from the closest taxonomy. If the distance is less than 5%, the bin is assigned to the taxonomy. The threshold allows discovery of the new species while considering the genomic diversity within the species. [6]

One of the limitations of the tool is the reliance on the external dataset and it is updated periodically.

B.3 Genome Annotation with Prokka

In order to understand the relationship between the health conditions and how the bacteria in the environment in these varying states, we need to understand the metabolism, virulence factors contributing to the pathogenicity, resistance mechanism to antibiotics.

Prokka is a genome annotation tool, particularly developed for prokaryotes, detecting the coding regions in the genome. It uses a hierarchical methodology to map the functions of the protein, starting with a small and more reliable database and domain specific databases to curated protein families models. [7]

Functional Annotation Steps:

1. User-curated protein annotations, searching with BLAST+ (Basic Local Alignment Search tool)
2. UniProt bacterial proteins database, covering more than 50% of the genome, searching with BLAST
3. RefSeq, in specific genus or domain, searching with BLAST
4. HMM Profile Databases such as Pfam, TIGRFAM (profiles developed with Markov Models) using hmmscan from HMMR 3 library
5. If no annotation is found among the previous searches, the genes are annotated as "hypothetical protein" [7]

In inferring the different genes such as tRNA, rRNA Prokka uses references. Table 1 shows a list of the reference Prokka relies on to perform functional annotation.

Table 1: Feature Prediction References , adapted from [7].

Tools (reference)	**Features predicted**
RNAmmer (Lagesen et al. , 2007)	Ribosomal RNA genes
Prodigal (Hyatt 2010)	Coding sequence
Aragorn (Laslett and Canback, 2004)	Transfer RNA genes
SignalP (Petersen et al. , 2011)	Signal leader peptides
Infernal (Kolbe and Eddy, 2011)	Non-coding RNA

Identifying CDS (coding sequences) provides insights about the putative roles of the genes and is fundamental to conduct pangenomic analysis to observe conversed genes and varying genes between different strains. This understanding can aid us in identifying drug targets in pathogens. For instance, various drugs can target the highly conserved regions to target the strains despite the mutations, some target the virulent regions to inhibit the bacteria to disable pathogenicity or target the genomes unique to certain strains harmful to host health.

A performance evaluation review article concludes that different tools perform better in the annotation of the CDS and the tool's performance was dependent on the organism. Prodigal; Prokka's annotation reference was performing best in *E.coli* and *S.aureus* [8]

Customized parameters: domain = Bacteria

B.4 Pangenome Analysis with Roary

Pangenome analysis was performed using Roary, a high speed pangenome pipeline that calculates the pangenome from annotated assemblies in GFF3 format. The input files can be produced in Prokka analysis. Roary's pipeline first extracts the coding regions from input

sequences and converts them into protein sequences, filters out any incomplete or partial sequences, and then pre-clusters the complete sequences iteratively with CD-HIT, resulting in a reduced set of protein sequences. Then BLASTP performs an all-against-all comparison on these sequences with a set percentage sequence identity [-i], default 95%. Clustering is done with Markov cluster algorithm (MCL) and merged with CD-HIT results. [9]

Basic usage of Roary is: `roary [options] *.gff`. From default parameters we changed the percentage of isolates a gene must be in to be considered part of the core genome [-cd] from 99% to 90% and number of threads [-p] from 1 to 4.

Roary provides a variety of output files as a result. We used additional scripts to create visualizations of the results. Rscript create_pan_genome_plots.R was used to create a set of plots describing the behavior of different parameters, such as number of conserved genes, and new genes, when adding new genomes. A Python script roary_plots.py by Marco Galardini was used to visualize the composition of the pangenome. [9], [10] However, to make pangenome analysis consistent with roary's results, we changed the roary_plots.py scripts criteria for core genome and accessory genome to match roary's parameters. A gene was considered part of the core genome when it was found in at least 90% of the genomes, and otherwise part of the accessory genome.

B.5. Phylogenetic Analysis with Roary and FastTree

Sequence based phylogenies are crucial knowledge in analyzing functions of genes and in understanding evolutionary relationships. We performed phylogenetic analysis using the FastTree tool. It is a tool that infers approximately-maximum-likelihood phylogenetic trees from nucleotide or protein sequence alignments. It uses approximations and heuristics to make it more computationally feasible, which makes it suitable for large amounts of data but may lead to some inaccuracies. [11]

As an input FastTree takes an alignment of nucleotide sequences as FASTA files as an input. Output is a phylogenetic tree in Newick format. Newick format file can be visualized using R package ggtree [12]. Results also provide local support values, ranging from 0 to 1, for estimating the reliability of each split, computed using the Shimodaira-Hasegawa test [11].

Tuned parameters used in the analysis:
`-pseudo`: to enable pseudocounts
`-spr 4`: increase the number of rounds of subtree-prune-regraft moves
`-mlacc 2` and `-slownni`: to make the maximum-likelihood nearest-neighbor interchange more comprehensive
`-fastest`: speed up the neighbor-joining phase
and `-no2nd`: improve accuracy by slowing the neighbor-joining phase
`-mlnni 4`: reduce the rounds of maximum-likelihood nearest-neighbor interchanges
`-gtr`: to use GTR-CAT model instead or Jukes-Cantor+CAT model for alignment
`-nt`: nucleotide alignment [11]

C. Results & Discussion

C.1 Metadata, Taxonomy and Quality

We encoded the study_group as healthy/unhealthy, and smoking history to yes/no also including ex-smokers with people who smoke to simplify our analysis. There was no statistically significant relationship between smoking and health conditions, as well as smoking habits and the subgroups in the unhealthy group. ($p > 0{,}40$) Yet, we observed that 6 out of 8 smokers had mucositis while the distribution of the smoking groups seemed to be randomly distributed in the other study groups and aging is associated with the disease but this could be due to a variety of people such as people seeking implant treatment as they get older. Average BMI in healthy, mucositis and peri-implantitis groups were 23.9, 23.1 22.6 and not a statistically significant difference.

Based on the taxonomic assignment with PhyloPhlAn all MAGs were assigned in the same kSGB985 with the following taxonomy:

Kingdom: Bacteria
└── Phylum: *Actinobacteria*
└── Class: *Coriobacteriia*
└── Order: *Coriobacteriales*
└── Family: *Atopobiaceae*
└── Genus: *Lancefieldella*
└── Species: *Lancefieldella rimae.*

L. rimae is a gram-positive, oval shaped, and mobile bacteria. It is a mesophilic anaerobe that was first isolated from human gingival crevices. Its full scientific name is *Lancefieldella rimae* (Olsen et al. 1991) Nouioui et al. 2018. [13] L. rimae is found in the microbiome of the healthy human saliva and oesophagus. A study suggests that *L. rimae* is strongly correlated with periodontal disease [14]. Periodontal diseases and peri-implantitis, despite the perception of being identical in etiology, have different microbial profiles [15]. Also, in our dataset, only 5 out of 30 people were in a healthy state.

Average genomic distance from kSGB985 was 0,041 with maximum distance 0,050 and minimum 0,036. We obtained information about the quality of our genomes. The used criteria and results for quality assessment are presented in Table 2.

Table 2: Quality criteria and results.

Quality Criteria				
Completeness	**Contamination**	**Quality of Genomes**	**Number of Genomes**	**Percentage**
50-90%	<5%	**Medium Quality**	13	43%
>90%	<5%	**High Quality**	17	57%

To better understand the results we obtained, we compared them to the corresponding values of *L. rimaes* type strain [16]. Values are reported in Table 3.

Table 3: Comparison of our data and L. rimae type strain.

Parameter	**Our data**	**Reference strain** [15]
Average Genome Size	1,460 MB	1,6 MB
Average GC content	49,7%	49,5%
Average number of contigs	110,4	9
Strain heterogeneity	0% in 25/30 samples and 100% on 5/30	

From Table 3 we can see that our results are in line with the reference strains values. GC content was found to be consistent through samples. As expected, we found a negative correlation of -0,86 between the number of contigs and genomes completeness. Correlation between contamination and number of contigs was moderate 0,44. Even though we had samples with 100% heterogeneity in this project we decided to continue with all the data we had but a more in-depth study should take this into account to include strain-specific differences.

C.2 Genome annotation

Table 4: Gene Annotation Summary.

	Average	**Minimum value**	**Maximum value**
tRNA	39	17	50
rRNA	2	1	3
tmRNA	1	1	1
CDS	1305	806	1603
Known Proteins	2143	1285	2608
Hypothetical Proteins	548	362	703

The presence of the tRNA and rRNA genes are within the expected range for bacterial genome indicating a well preserved translational machinery. The CDS count reflects a moderate genome size typical for bacteria and range of coding genes from 806 to 1603 shows the genetic diversity within the bins belonging to *Lancefieldella rimae* specie.

The hypothetical proteins may be related to the functions unique to the presence of this bacteria in the oral microbiome and the novel genes.

Table 5: Assembly Summary

	Average	**Minimum value**	**Maximum value**
Contigs	110	10	315
N50 Contigs	73238	3033	282375

Assembly summary shows the variety between the quality and completeness which is important in interpreting the results of the prokka output. A smaller number of contigs and higher N50 are indicating a higher quality of assembly.

For example, we can hypothesize that the different number of rRNA and tRNA genes in healthy groups can give insight about the association with the disease [17].

We noticed that samples connected to peri-implantitis lack rRNA compared to other groups, this can be seen in Figure 3. 0 indicates that no rRNA detected and assigned for visualization purposes.

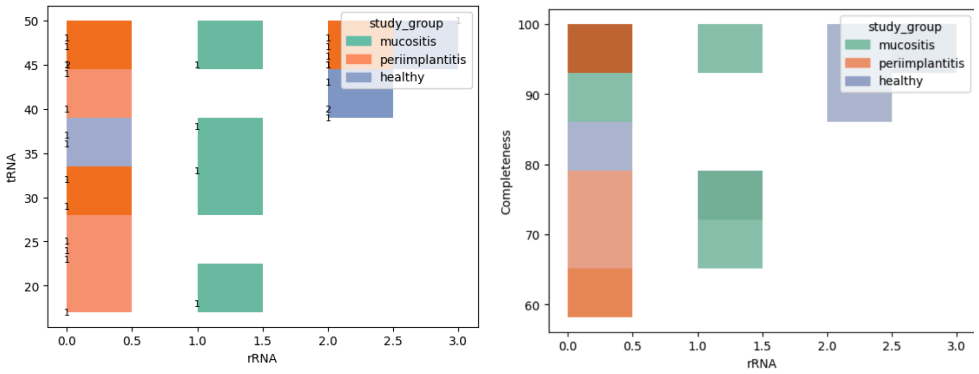

Figure 3: Comparison of the rRNA and tRNA presence vs Completeness and rRNA presence.

This could suggest that the bacteria has more active metabolism in healthy individuals and might be important in sustaining the healthy state of the oral microbiome. However, this kind of assumption cannot be made directly from the data we have. It is not known whether the difference in RNAs reflect the biological properties or the completeness of the samples. To investigate this, we looked into the correlation between completeness and tRNA in each health group. Correlation in the healthy group was 0,957, while in mucositis group 0,080 and peri-implantitis 0,4106.

In peri-implant disease, dominance of few bacteria and lower abundance of remaining strain may be the reason that we couldn't annotate the rRNA presence in the peri-implantitis group. Further insights could be derived with transcriptomics. [18]

C.3 Pangenome analysis

In pangenome analysis we discovered that the core genome consists of 500 core genes. Accessory genome had 4544 genes. The composition of the pangenome is visualized in Figure 4.

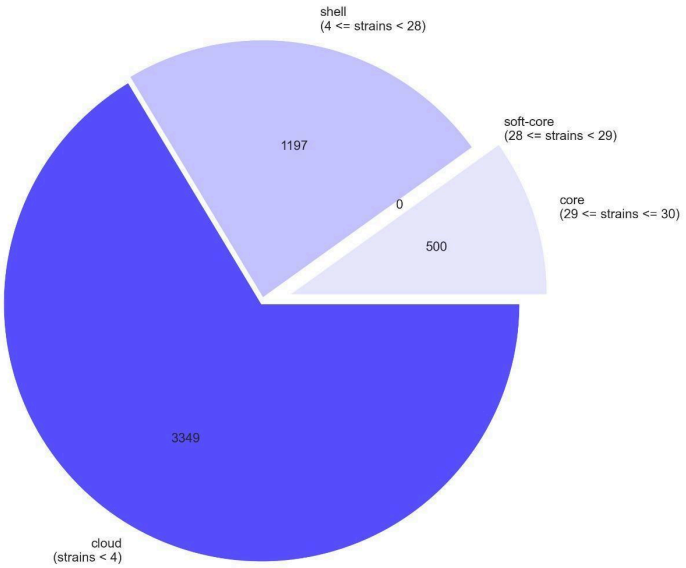

Figure 4 : Visualization of the pangenome of our MAGs by Gene Presence.

Pangenome as a whole has 5046 genes. In Figure 4 accessory genome is divided into shell and cloud. We also visualized the accessory genomes' presence-absence matrix against a phylogenetic tree. This is shown in Figure 5 below.

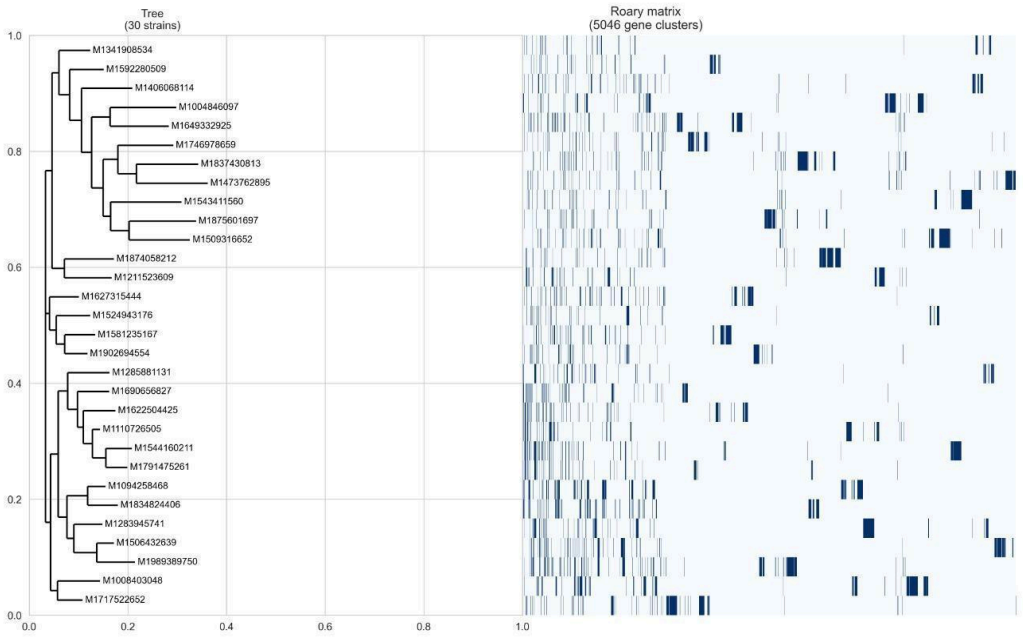

Figure 5: Phylogenetic tree against a presence-absence matrix of accessory genes.

From the presence-absence matrix we cannot see significant clustering or association between strains.

From Roary analysis we were able to determine whether our SGB has an open or closed pangenome. This can be observed by looking at how the total number of genes behave when new genomes are added to the pangenome. This has been visualized in Figure 6.

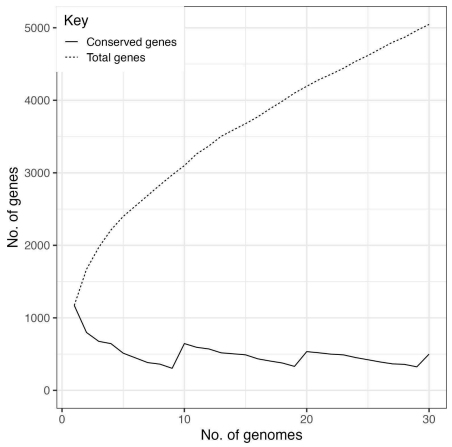

Figure 6: Total number of genes as a function of the number of genomes.

The number of genes keeps increasing when new genomes are added which indicates that the pangenome is open. Open pangenome suggest high genetic variability and diversity within species.

C.4 Phylogenetic analysis and association with metadata

The result from phylogenetic analysis was visualized and each sample annotated with study group information provided in the metadata. Tree is presented in Figure 7.

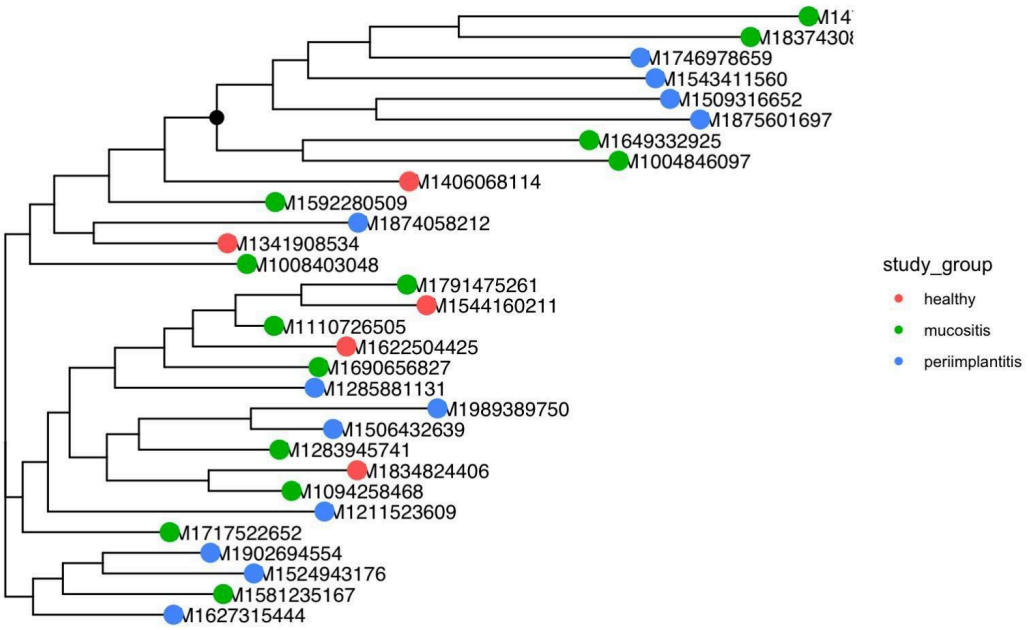

Figure 7: Phylogenetic tree (Accessory binary genes) annotated with metadata.

The phylogenetic tree is divided into three main branches. One branch at the bottom of the tree contains only samples associated with peri-implantitis or mucositis. One internal node (marked with a black dot) can be found which roots a clade with no samples from healthy individuals. When comparing Figure 6 tree to Figure 7 tree, the trees are identical, but the layout is different for samples M1874058212, M1341908534, M1008403048 and M1717522652. This is expected since both trees are built based on accessory genome data. Using accessory genome data instead of core genome data provides insight into newly acquired genes rather than the conserved genes. Both of these trees suggest that the pathogenic species may have a common origin. However proper conclusions cannot be made based on this with our limited data and analysis.

To gain more insight on this bacteria and our kSGBs association with mucositis and peri-implantitis additional analyses are needed. Possible next steps in the analysis could include the search for resistance and virulence genes as well as antigens. Virulence genes could help understand pathogenic potential and behavior and which genes or gene clusters contribute to that. Identification of antibiotic resistance genes could contribute to better understanding of the prevalence and diversity of the pathogen. [19] Antigens on the other hand give information on bacteria characterization and bacteria-host cell interactions. All these aspects could contribute to the understanding of pathogenicity, genetic diversity and thus be valuable, for example, in clinical applications.

D. Conclusion

Our dataset had 30 *Lancefieldella rimae* metagenome-assembled genome associated with oral microbiomes of 5 healthy and 25 people with mucositis and peri-implantitis disease after implant treatment. Because all of our samples share *L. rimae*, we can only make inferences at the strain level.

In quality check, we realized that peri-implantitis was correlated with the missing count of rRNA genes due to an incomplete genome. The incompleteness in the peri-implantitis samples can be due to the low abundance of *L. rimae* and the dominance of other pathogens in a sample. Another study supports the idea that the diversity is lower in mucositis and peri-implant disease compared to healthy oral microbiome profiles.

We found the pangenome of *L.rimae* open, having significantly more new genomes, which can reflect the diversity in the species and the evolutionary capacity to adapt to the environment. It also partially explains the portion of the hypothetical protein. In the phylogenetic tree, 2 clans seem to emerge containing only strains in unhealthy samples.

Analyzing metabolic pathways, and searching for virulence genes and antibiotic resistance can provide additional insights into the impact of the strains on the host health. Also comparing the groups in the lack and presence of the species L.rimae could explain the species-level impact on the host health.

List of Tables and Figures

References:

[1] J. A. Aas *et al.*, 'Bacteria of dental caries in primary and permanent teeth in children and young adults', *J Clin Microbiol*, vol. 46, no. 4, pp. 1407–1417, Apr. 2008, doi: 10.1128/JCM.01410-07.

[2] 'Peri-implant diseases', European Federation of Periodontology. Accessed: Apr. 07, 2024. [Online]. Available: https://www.efp.org/for-patients/dental-implants/peri-implant-diseases/

[3] 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*, vol. 25, no. 7, pp. 1043–1055, Jul. 2015, doi: 10.1101/gr.186072.114.

[4] A. Chklovski, D. H. Parks, B. J. Woodcroft, and G. W. Tyson, 'CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning'. bioRxiv, p. 2022.07.11.499243, Jul. 11, 2022. doi: 10.1101/2022.07.11.499243.

[5] F. Asnicar *et al.*, 'Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0', *Nat Commun*, vol. 11, no. 1, p. 2500, May 2020, doi: 10.1038/s41467-020-16366-7.

[6] B. D. Ondov *et al.*, 'Mash: fast genome and metagenome distance estimation using MinHash', *Genome Biology*, vol. 17, no. 1, p. 132, Jun. 2016, doi: 10.1186/s13059-016-0997-x.

[7] 'Prokka: rapid prokaryotic genome annotation | Bioinformatics | Oxford Academic'. Accessed: Apr. 07, 2024. [Online]. Available: https://academic.oup.com/bioinformatics/article/30/14/2068/2390517

[8] N. J. Dimonaco, W. Aubrey, K. Kenobi, A. Clare, and C. J. Creevey, 'No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study', *Bioinformatics*, vol. 38, no. 5, pp. 1198–1207, Mar. 2022, doi: 10.1093/bioinformatics/btab827.

[9] A. J. Page *et al.*, 'Roary: rapid large-scale prokaryote pan genome analysis', *Bioinformatics*, vol. 31, no. 22, pp. 3691–3693, Nov. 2015, doi: 10.1093/bioinformatics/btv421.

[10] 'sanger-pathogens/Roary'. Pathogen Informatics, Wellcome Sanger Institute, Mar. 24, 2024. Accessed: Apr. 07, 2024. [Online]. Available: https://github.com/sanger-pathogens/Roary

[11] M. N. Price, P. S. Dehal, and A. P. Arkin, 'FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments', *PLoS ONE*, vol. 5, no. 3, p. e9490, Mar. 2010, doi: 10.1371/journal.pone.0009490.

[12] 'ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data - Yu - 2017 - Methods in Ecology and Evolution - Wiley Online Library'. Accessed: Apr. 07, 2024. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.12628

[13] L. C. Reimer, J. Sarda Carbasse, J. Koblitz, A. Podstawka, and J. Overmann, ‘Lancefieldella rimae (Olsen et al. 1991) Nouioui et al. 2018’. [object Object], Dec. 21, 2023. doi: 10.13145/BACDIVE3041.20230509.8.1.

[14] E. L. Veras *et al.*, ‘Newly identified pathogens in periodontitis: evidence from an association and an elimination study’, *J Oral Microbiol*, vol. 15, no. 1, p. 2213111, doi: 10.1080/20002297.2023.2213111.

[15] ‘Periodontal and peri‐implant diseases: identical or fraternal infections? - Robitaille - 2016 - Molecular Oral Microbiology - Wiley Online Library’. Accessed: Apr. 07, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/omi.12124

[16] ‘Lancefieldella rimae ATCC 49626 genome assembly ASM17401v1’, NCBI. Accessed: Apr. 07, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_000174015.1/

[17] J. A. Klappenbach, J. M. Dunbar, and T. M. Schmidt, ‘rRNA Operon Copy Number Reflects Ecological Strategies of Bacteria’, *Appl Environ Microbiol*, vol. 66, no. 4, pp. 1328–1333, Apr. 2000.

[18] P. Ghensi *et al.*, ‘Strong oral plaque microbiome signatures for dental implant diseases identified by strain-resolution metagenomics’, *npj Biofilms Microbiomes*, vol. 6, no. 1, pp. 1–12, Oct. 2020, doi: 10.1038/s41522-020-00155-7.

[19] J. Wilson, M. Schurr, C. LeBlanc, R. Ramamurthy, K. Buchanan, and C. Nickerson, ‘Mechanisms of bacterial pathogenicity’, *Postgrad Med J*, vol. 78, no. 918, pp. 216–224, Apr. 2002, doi: 10.1136/pmj.78.918.216.