Università degli Studi di Trento: Department of Cellular, Computational and Integrative Biology , 1–7
Date: 06/04/2025

UNIVERSITÀ DI TRENTO

Pangenome Architecture and Phylogenetic Diversity in SGB3588

Sofia Pietrini, Anna Salsano and Caterina Sanetti

Introduction

The human oral cavity represents a complex and dynamic ecosystem that hosts a diverse microbial community, which, given the easy sampling procedure, is one of the most studied microbiomes (1). These microorganisms colonize both hard and soft surfaces, forming structured biofilms that play a crucial role in oral health, forming structured biofilms that potentially drive diseases when disrupted (2). Recent advances in molecular biology, particularly high-throughput sequencing technologies such as shotgun metagenomics and Metagenome-Assembled Genome (MAG) analysis, have revolutionized our understanding of oral microbial communities. In this study, we focus on the characterization of a SGB (Species-level Genome Bin) sampled from individuals displaying diverse conditions: peri-implantitis, mucolitis and a control group. Through integrated pangenomic and phylogenetic analysis, we aim to discover the open nature of the pangenome and phylogenetic relationships among strains, particularly examining correlations between MAGs from the same disease state, while comparing these disease-associated profiles with healthy controls. We hypothesize that MAGs from distinct peri-implant conditions may exhibit differential quality metrics, coding sequence abundance, and phylogenetic clustering patterns, which could serve as biomarkers for these clinically significant conditions.

Materials and Methods

Data

The dataset used included 30 MAGs collected from the oral cavity of patients with dental implants, all classified under SGB3588. Each MAG was annotated with the patient's age, smoking status (current, former, or non-smoker), and clinical diagnosis (mucositis or peri-implantitis). The MAGs are labeled with unique identifiers (*magID* and *sampleID*), allowing for precise tracking of microbial genomic data alongside patient characteristics. The inclusion of both disease states and healthy controls allows for comparative analyses to identify microbial signatures associated with implant-related pathologies.

Quality Checking

To assess the quality of metagenome-assembled genomes (MAGs), we utilized CheckM v1.2.3 (3) with a taxonomy workflow.

```
checkM taxonomy_wf domain Bacteria \
mags checkm_output -t 4
```

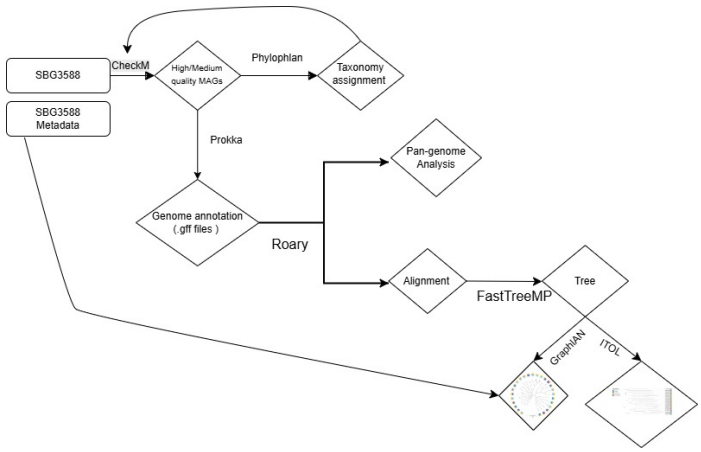


Fig. 1: Workflow of the analysis

MAGs were classified into three quality tiers: low-quality (completeness <50% or contamination >5%), medium-quality (completeness ≥50% and <90% with contamination ≤5%), and high-quality (completeness ≥90% and contamination ≤5%).

Taxonomic Assignment

Once we discarded the five low quality MAGs, we proceeded with taxonomic assignment using PhyloPhlAn v3.0.67 (4), in combination with MetaRefSGB database, a collection of microbial genomes, clustered based on species-level distance. The phylophlan_metagenomic pipeline was employed to compare each MAG against SGBs (species genome bin) present in the database. The output obtained from this analysis presented the distances between our samples and the closest SGB.

Genome Annotation

An important step to continue the MAGs analysis, is the performance of the Genome Annotation, a process to identify and label genomic features present in the contigs. In this study, we performed genome annotation using Prokka 1.14.6 (5), which processes .fna files from the MAGs as input and generates a set of 12 output files per MAG, including .gff, .fna, .faa, and .txt files. These output files contain essential information regarding protein sequences and gene classifications, distinguishing between coding sequences (CDS), hypothetical proteins and known proteins.
We wrote a loop for Prokka to iterate over each of the MAG folders

containing the .fna files, for the only input that Prokka accepts is one .fna (FASTA format DNA sequence alignment file) file at the time.

Pangenome Analysis

Pan-genome analysis was performed using Roary v3.13.0 (6), which constructs a bacterial pan-genome from annotated genome assemblies. The tool processes Prokka-generated GFF3 files by first extracting protein-coding sequences, filtering out partial genes. An all-against-all BLASTP comparison (set at 95% identity threshold using parameter -i 95) identifies homologous sequences, which are then clustered into orthologous groups. The pipeline distinguishes core genes (conserved in $\geq$ 90% of isolates, specified by -cd 90) from accessory and strain-specific genes, while handling paralogs through conserved gene neighborhood analysis. The command :

```
roary prokka_output/*/*.gff -f roary_output \
-i 95 -cd 90 -p 4
```

directed this analysis, where:

- -f defined the output directory;
- -i set the output directory;
- -cd determined the core gene threshold;
- -p enabled parallel processing with 4 threads.

Resulting gene clusters were analyzed for presence/absence patterns and visualized using Roary's companion scripts (create_pan_genome_plots.R and roary_plots.py) to generate pan-genome accumulation curves and phylogenetic heatmaps.

Phylogenetic Analysis

To infer the phylogenetic relationships among the isolates, we performed a pan-genome analysis using Roary.

```
roary prokka_output/*/*.gff -f roary_output \
-i 95 -cd 90 -p 4 -n -e
```

where -e -n set a fast core gene alignment usign Mafft (7).
The resulting core gene alignment was used as input for FastTreeMP (v2.1.11) (8) to construct a maximum-likelihood phylogeny under the GTR model (-gtr -nt). To optimize tree accuracy, we applied subtree-prune-regraft (SPR) moves (-spr 4), limited nearest-neighbor interchanges (-mlnni 4), and enabled additional likelihood accuracy checks (-mlacc 2). The analysis was set to prioritize computational efficiency (-slownni -fastest -no2nd) while maintaining robustness. The final tree was exported in Newick format.
For visualization and annotation, the phylogenetic tree was imported into Interactive Tree of Life (iTOL) (9) and GraPhlAn (v1.1.3) (10), where metadata was integrated to enhance interpretability.

Results and Discussion

Initial quality assessment of the 30 metagenome-assembled genomes (MAGs) was performed using CheckM's taxonomy workflow , with domain as taxon rank and *Bacteria* as taxon, which identified five low-quality MAGs. One MAG (M1927064548) was excluded for failing both completeness and contamination thresholds, while the remaining four were discarded due to excessive contamination (Figure 2). Analysis of GC content distribution across samples revealed stable values, with a median of 47.2% (range: 0.465–0.480) and low variability (SD = 0.44%), indicating high consistency across genomes. Genome sizes ranged from 1.35 to 2.87 Mbp. Visual inspection of MAG quality categories (High/Medium vs. Low) across patient metadata revealed no apparent association with smoking status (smoker, non-smoker, ex-smoker) or clinical groups (mucositis, healthy, peri-implantitis), as shown in Figure 2.

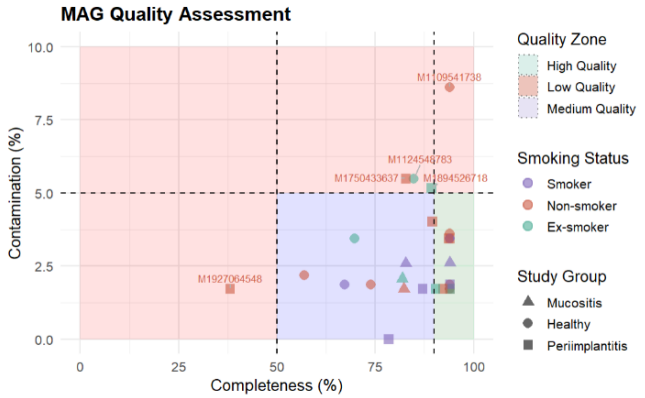


Fig. 2: Scatter-plot of MAGs completness and contamination. The points are colored by smoking status and shaped by study group.Furthermore, the dashed lines partition the space into three distinct quality regions: high, medium, and low.

The PhyloPhlAn taxonomic assignment revealed that, with an average distance of 0.0249, our samples are closest to an unknown species belonging to the *Treponemataceae* bacterial family, *Treponema* Genus. This taxonomic group includes both pathogenic and nonpathogenic species, with nonpathogenic species that can be part of the normal flora or linked to oral diseases (11). In particular, the *Treponema* spp. are described as one of the first pathogens in the gingival tissue (12).
It was also detected that all the MAGs analyzed had an average distance below 0.05. Consequently, we were able to assign a taxonomic label to all our MAGs.
Given these results, we decided to perform a second quality check to verify if the overall quality of our MAGs could be improved by applying checkM with a more specific taxonomic label. At first we tried with Genus: *Treponema*, however, by putting this taxon, checkM would search for marker genes for the *Spirochaetaceae* family, instead of *Treponemataceae*. An in-depth literature search highlighted both the ambiguity of this family name and how the *Spirochaetaceae* taxonomic group is shown to be a rather heterogeneous and weakly supported group, with many studies supporting the reinstating of the valid family name *Treponemataceae* for *Treponema* (13).
Subsequently, we applied checkM with Order Spirochaetales as the rank parameter.
Unfortunately, not all the MAGs in our dataset reached the requirements to be classified of medium and high quality, since 4 out of the 30 samples were still labeled as low-quality (M1428478814, M1485435981, M1927064548, M1983179296). In Figure 3 are listed the MAGs that underwent a change in quality

when the parameter rank of CheckM was changed from *Domain Bacteria* to *Order Spirochaetales*.

MAG	Quality Rank: Domain	Quality Rank: Order
M1109541738	Low	High
M1124548783	Low	Medium
M1428478814	High	Low
M1485435981	Medium	Low
M1750433637	Low	Medium
M1791556706	Medium	High
M1869481370	Medium	High
M1894526718	Low	High
M1983179296	High	Low

Fig. 3: Quality rank differences in MAG classification between Domain- and Order-level taxonomic resolution.

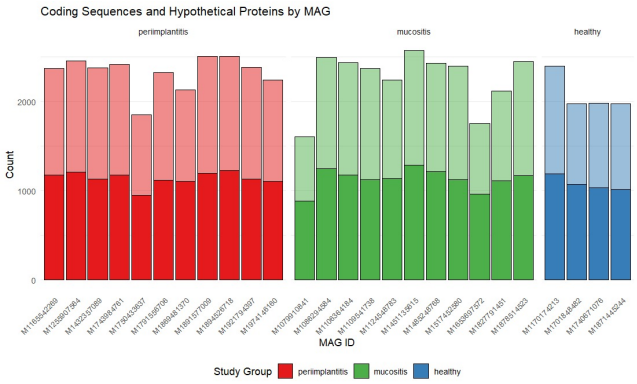
We, therefore, proceeded by excluding the four low-quality MAGs identified in the second CheckM analysis. This decision was based on the greater taxonomic specificity provided by Order-level classification compared to Domain-level assignment.
Genome annotation facilitates comparative analysis of functional, taxonomic and genomic features to identify microbial markers. The proportion of hypothetical versus known proteins within coding sequences (CDSs) serves as a critical parameter for interpreting these metagenomic datasets. Our results demonstrate a consistent equilibrium between these protein categories, with approximately equal proportions (between 46.82805% and 54.97512%) of known and hypothetical proteins maintained across all examined genomes and conditions (healthy, periimplanted and mucositis), as illustrated in Figure 4.

The high number of coding sequences (CDS) identified in the MAG suggests a substantial gene pool, potentially contributing to the metabolism, growth and functionality of the organisms in the sample. This observation also serves as an indicator of genome completeness, reinforcing the reliability of the assembled metagenome.

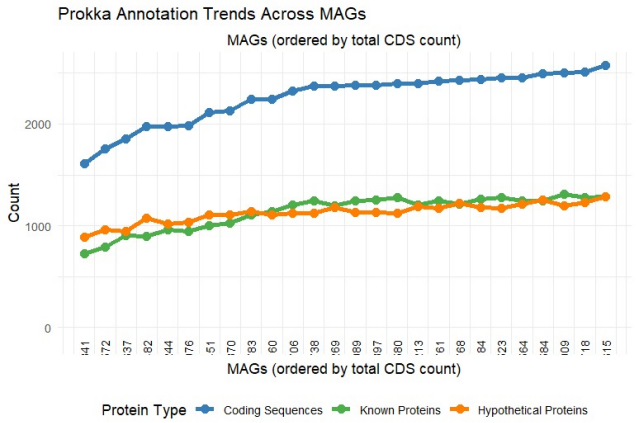
A crucial observation is regarding the hypothetical proteins. PROKKA relies on public databases therefore, previously uncharacterized proteins will not match known entries, highlighting potential novel gene functions that could be the starting point for further investigation. These results align with recent studies demonstrating that approximately 50% of open reading frames (ORFs) in uncultivated oral taxa lack functional annotation, likely due to lineage-specific proteins involved in host-microbe interactions.

Importantly, potential confounding factors such as genome fragmentation (due to low-quality MAGs) or contamination can be ruled out, as stringent quality control measures were applied during preprocessing. Thus, the abundance of hypothetical proteins is more likely to be attributed to biological novelty rather than technical artifacts.

The pangenome analysis of our dataset identified a total of 7,007 gene clusters (Figure 5). 948 genes (13.5%) were present in at least 24 MAGs (≥90%) and so classified as core genes. No genes were detected in the soft core category (89–90% prevalence). The accessory genes accounted for a substantial proportion of the total



(a)



(b)

Fig. 4: a) The plot represents the distribution of CDS in each MAGs (divided for conditions), the known proteins are the light part of the bar, the hypothetical ones are represented by the dark part. b) More clear and immediate representation of the equilibrium and abundance of known and hypothetical proteins.

gene pool, comprising 1,970 shell genes (28.1%), present in 4–23 MAGs, and 4,089 cloud genes (58.4%), found in fewer than four MAGs. A histogram of gene distribution across genomes revealed

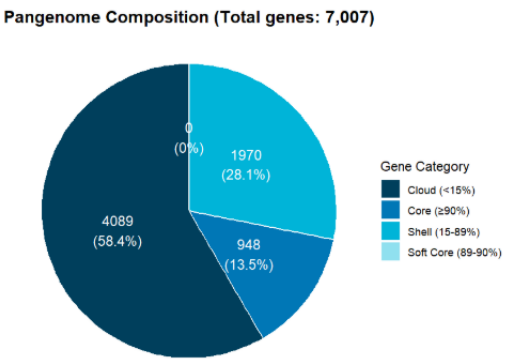


Fig. 5: Piechart representing the proportions of cloud, core, shell and soft core genes across the pangenome.

a right-skewed pattern, with the highest peak corresponding to genes shared by the fewest strains (Figure 6). Highlighting the predominance of accessory genes in the pangenome.

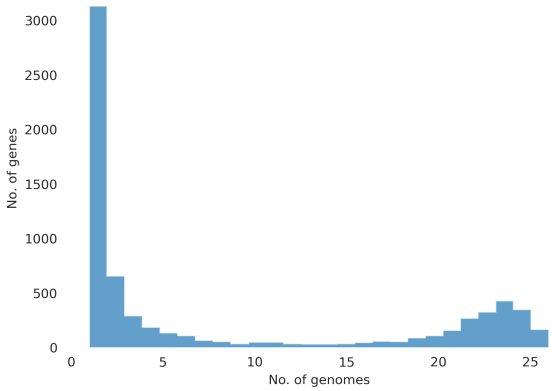


Fig. 6: Histogram representing the number of shared genes in increasing number of genomes.

These findings strongly suggest an open pangenome structure. To further validate the hypothesis of an open pangenome structure, we analyzed the trends of conserved and total gene counts. As the number of genomes increases, the total gene count continues to rise, while the fraction of conserved genes decreases, signaling the accumulation of novel genes, Figure 7. The stair-step pattern seen in conserved gene numbers can be attributed a mathematical approximation due to the small sample size.

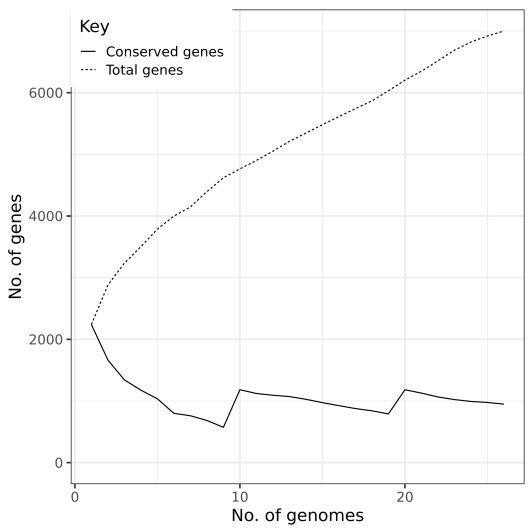


Fig. 7: Trend of conserved genes and total genes with respect to the number of genomes.

Additionally, the consistent rise in the number of unique genes compared to new genes as more genomes are incorporated further reinforces the dynamic expansion of the gene pool, providing further evidence for an open pangenome structure (Figure 8). An

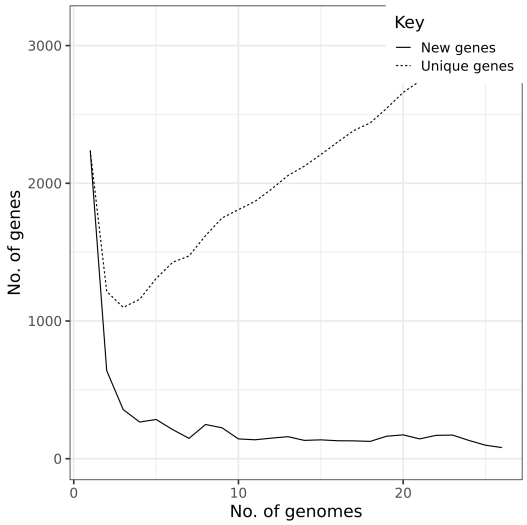


Fig. 8: Trend of unique genes and new genes with respect to the number of genomes.

important output of the Roary pipeline is also the gene presence absence matrix (Figure 9) in which core genes are not shown. On the left is represented a phylogenetic tree of the MAGs that cluster for similar accessory genes presence/absence. In the heatmap, the more dense region on the right represents the shell genes while, the more scattered part are the cloud genes.

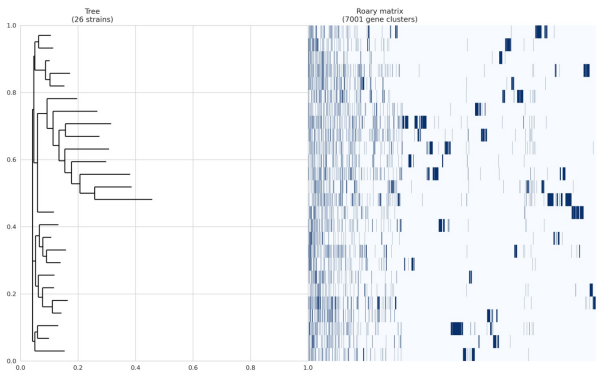


Fig. 9: The two plots represent the first 20 entries of the most frequent annotations in accessory genes and in cloud accessory genes.

It can be seen that, closely clustered MAGs, tend to have similar gene presence-absence patterns, while more distantly related strains exhibit greater differences in their accessory genome content.

To better characterize the functional roles of accessory genes, we conducted an additional analysis, through an R code, using the presence_absence_gene.csv file. We wanted to identify the most

prevalent functional annotations in the accessory genes and in the cloud accessory genes (Figure 10).

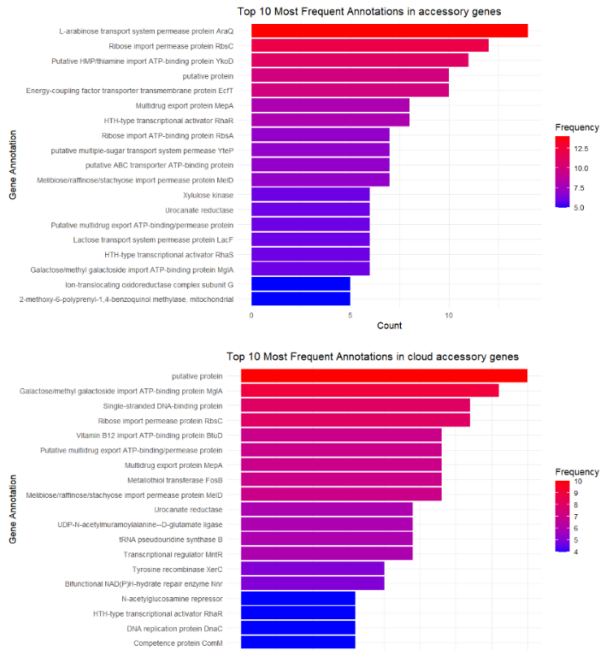


Fig. 10: The two plots represent the first 20 entries of the most frequent annotations in accessory genes and in cloud accessory genes.

We focus our investigation on the accessory genes, to elucidate non-core proteins that may contribute to specific adaptations. We noticed several interesting genes exhibiting notable functional relevance; such as: L-arabinose transport system permease protein AraQ, a transmembrane helix protein, member of the binding-protein-dependent transport system permease family, related to sugar transport in the cell (14); Multidrug export protein MepA, another transmembrane protein that contributes to resistance to the glycylcycline antibiotic tigecycline in Staphylococcus aureus (strain N315) (15); Urocanate reductase, specific of Lactiplantibacillus plantarum, that catalyzes the production of dihydrourocanate, present at higher concentrations in subjects with type 2 diabetes, and directly impairs glucose tolerance and insulin signaling at the level of insulin receptor substrate (16). These findings highlight the functional diversity of accessory genes and their potential contributions to microbial ecology, host-microbe interactions, and clinically relevant phenotypes such as antibiotic resistance and metabolic dysregulation.

Phylogenetic trees were constructed using the core gene alignment (core_gene_phylogeny.nwk, generated by FastTree) and the accessory gene presence/absence (accessory_binary_genes.fa.newick, derived from Roary). These trees were visualized and annotated in Interactive Tree of Life (iTOL), incorporating metadata related to the study groups (Figure 11). Comparative analysis revealed no clear phylogenetic clustering of strains based on their clinical condition. This pattern was consistent for both core and accessory gene phylogenies. However, in the core gene tree, two metagenome-assembled genomes (MAGs) from peri-implantitis patients (M1974146160 and M1894526718) exhibited close phylogenetic proximity.

To assess potential associations between genomic variation and

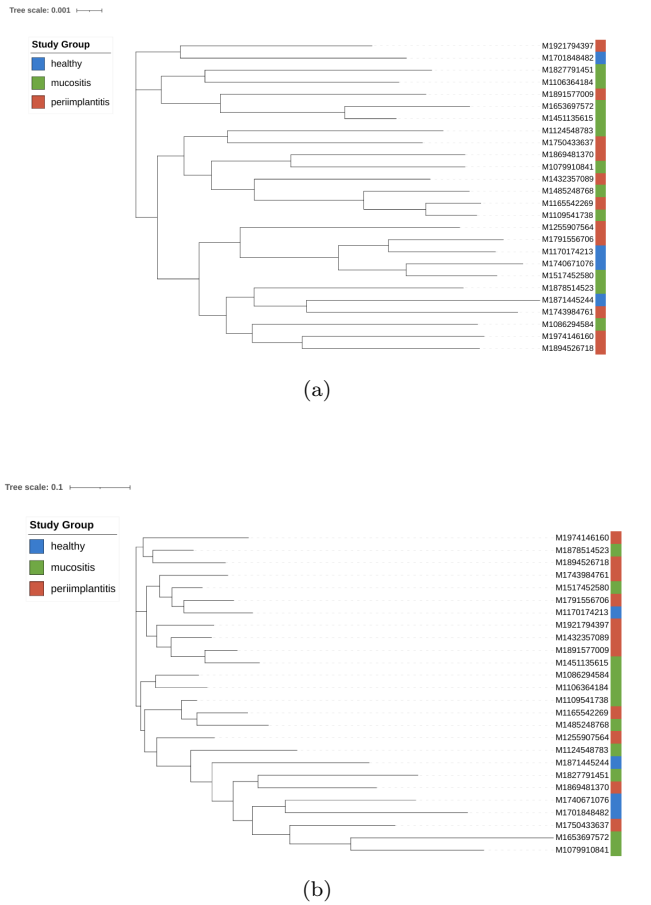


Fig. 11: a) Core genes phylogenetic tree annotated by study group. b) Accessory genes phylogenetic tree annotated by study group.

host factors, we also included metadata about study group, sex, BMI and smoking status. While phylogenetic clustering did not reveal any clear associations of the clinical status with sex, or smoking status, we identified three distinct clusters of closely related MAG pairs sharing the same clinical origin (Figure 12) . Notably, Cluster 2 comprised two MAGs from mucositis patients, both of which were associated with low BMI (BMI < 22). No other consistent phenotypic or demographic trends were observed across the remaining clusters. Including the SEX, two MAG clusters sharing identical clinical groupings (mucositis patients in Cluster 1 and Cluster 2) showed complete sex discordance by being females in cluster 2 and males in cluster 1 (Figure 13).

Discussion and Conclusion

Analysis of the MAGs composing our dataset enabled the taxonomic assignment of our Species-level Genome Bins (SGBs) to the Treponemataceae bacterial family, which was confirmed to be found in the oral microbiome. Due to the ambiguities found when assigning a taxonomic label to our MAGs, we had to discard 4 MAGs due to their low quality. Despite this, no differences in terms of quality were encountered between the different study groups.

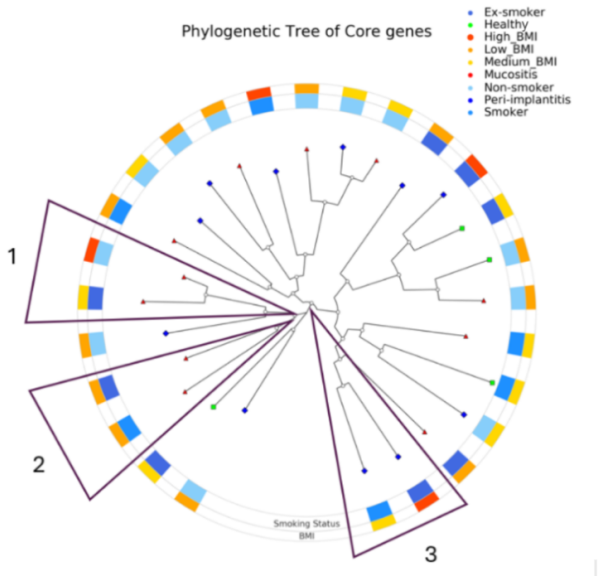


Fig. 12: Core genes phylogenetic tree annotated by smoking status (inner ring) and BMI (outer ring).

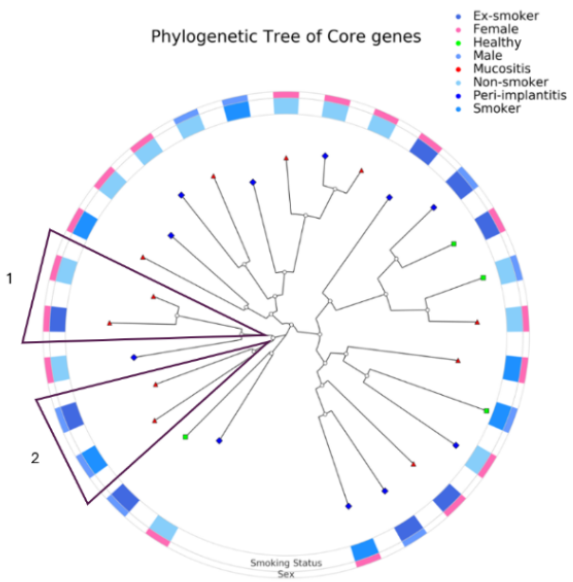


Fig. 13: Core genes phylogenetic tree annotated by smoking status (inner ring) and sex (outer ring).

The functional annotation showed a near-equal split between hypothetical (46.8–55.0%) and known proteins. Pangenome analysis revealed an open structure of our SGBs' pangenome, with only 13.5% core genes and a predominance of accessory genes linked to specific functions like antibiotic resistance (e.g. MepA) and metabolic regulation (e.g. urocanate reductase).

Phylogenetic assessment revealed no strong clustering by clinical status (healthy/mucositis/peri-implantitis), smoking history, sex or BMI. It is important to note that the limited number of given samples, which was further reduced after quality filtering, may have constrained the statistical power of these observations.

Future studies with more expanded sample sizes could elucidate potential links between these genomic features and disease progression, further validating the findings of this study.

References

1. Elisia L Cohen, Charlene A Caburnay, Shelly Rodgers, Timothy J Poor, and Matthew W Kreuter. Academic language barriers and the international divide in science. *PLoS Biology*, 17(5):e3000222, 2019. doi:10.1371/journal.pbio.3000222.
2. Natalia de Campos Kajimoto, Yvonne de Paiva Buischi, Mansour Mohamadzadeh, and Peter Loomer. The oral microbiome of peri-implant health and disease: A narrative review. *Dentistry Journal*, 12(10), 2024. URL: https://www.mdpi.com/2304-6767/12/10/299, doi:10.3390/dj12100299.
3. Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7):1043–1055, 2015. doi:10.1101/gr.186072.114.
4. Francesco Asnicar, Andrew M Thomas, Francesco Beghini, Claudia Mengoni, Serena Manara, Paolo Manghi, Qiyun Zhu, Mattia Bolzan, Fabio Cumbo, Uyen May, et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using phylophlan 3.0. *Nature Communications*, 11(1):2500, 2020. doi:10.1038/s41467-020-16366-7.
5. Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, Jul 2014. URL: https://www.ncbi.nlm.nih.gov/pubmed/24642063, doi:10.1093/bioinformatics/btu153.
6. Andrew J. Page, Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693, 07 2015. arXiv:https://academic.oup.com/bioinformatics/article-pdf/31/22/3691/49036081/bioinformatics_31_22_3691.pdf, doi:10.1093/bioinformatics/btv421.
7. Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 2002. doi:10.1093/nar/gkf436.
8. Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttreemp, 2010. Parallelized version of FastTree 2 using OpenMP. URL: http://www.microbesonline.org/fasttree/#OpenMP.
9. Ivica Letunic and Peer Bork. Interactive tree of life (itol) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, 49(W1):W293–W296, 2021. doi:10.1093/nar/gkab301.
10. Francesco Asnicar, George Weingart, Timothy L Tickle, Curtis Huttenhower, and Nicola Segata. Graphlan: a tool for visualization and analysis of microbial genomes and metagenomic data. *Bioinformatics*, 31(20):3341–3344, 2015. doi:10.1093/bioinformatics/btv404.
11. *Medical Microbiology. 4th Edition.* University of Texas Medical Branch at Galveston., 1996. URL: https://books.google.it/books?id=ykETtAEACAAJ.

12. Zhibin Li, Guanjun Lu, Zihao Li, Bo Wu, Enjie Luo, Xiaoyu Wang, Jiasong Guo, Chunyu Wang, Xiaowei Song, and Feng Wang. Gut microbiota differences between healthy older adults and individuals with parkinson's disease: a systematic review. *Translational Medicine*, 20(1):510, 2022. URL: https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-022-03636-9, doi:10.1186/s12967-022-03636-9.
13. Anton Hördt, Marina García López, Jan P. Meier-Kolthoff, Marcel Schleuning, Lisa-Maria Weinhold, Brian J. Tindall, Sabine Gronow, Nikos C. Kyrpides, Tanja Woyke, and Markus Göker. Analysis of 1,000+ type-strain genomes substantially improves taxonomic classification of alphaproteobacteria. *Frontiers in Microbiology*, 11, 2020. URL: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2020.00468, doi:10.3389/fmicb.2020.00468.
14. Mário J. Ferreira and Isabel de Sá-Nogueira. A multitask ATPase serving different ABC-type sugar importers in Bacillus subtilis. *Journal of Bacteriology*, 192(20):5312–5318, Oct 2010. arXiv:2010Aug6, doi:10.1128/JB.00832-10.
15. Fiona McAleese, Peter Petersen, Alexey Ruzin, Paul M. Dunman, Ellen Murphy, Steven J. Projan, and Patricia A. Bradford. A novel MATE family efflux pump contributes to the reduced susceptibility of laboratory-derived Staphylococcus aureus mutants to tigecycline. *Antimicrobial Agents and Chemotherapy*, 49(5):1865–1871, May 2005. doi:10.1128/AAC.49.5.1865-1871.2005.
16. Ara Koh, Antonio Molinaro, Marcus Ståhlman, Muhammad Tanweer Khan, Caroline Schmidt, Louise Mannerås-Holm, Hao Wu, Alba Carreras, Heeyoon Jeong, Louise E. Olofsson, Per-Olof Bergh, Victor Gerdes, Annick Hartstra, Maurits de Brauw, Rosie Perkins, Max Nieuwdorp, Göran Bergström, and Fredrik Bäckhed. Microbially produced imidazole propionate impairs insulin signaling through mtorc1. *Cell*, 175(4):947–961.e17, Nov 2018. arXiv:2018Oct25, doi:10.1016/j.cell.2018.09.055.