
# Computational Microbial Genomics

## Analysis and characterization of an SGB from the oral microbiome

Submitted by: Virginia Leombruni

A.Y. 2023-2024

University of Trento - Via Sommarive 9, 38123 Povo (TN), Italy

# 1. INTRODUCTION

Microbes, prevalent in nature and within the human body, vary in their impact from beneficial to disease-causing. [8-9] While the vast majority of microbial species are still unknown, the recent advance in high-throughput and computational technologies [1] represents a great opportunity to expand the understanding of the link between health and the state of the microbiome. Through metagenomics, the DNA sequencing of microbial communities, it's possible to reconstruct their genome, uncover their complexity and their role in human health.

This study focuses on investigating how the microbiome of dental plaque relates to inflammatory conditions around dental implants, such as inflammation of the mucosa (peri-implant mucositis) and the extended surrounding area, including the bone (peri-implantitis). [7] By analysing MAGs reconstructed from 30 oral samples, this work aims to uncover the role of a specific microbial species in the post-implantation context.

# 2. MATERIALS AND METHODS

Computational analysis was performed on a set of 30 MAGs obtained from oral cavity samples. The samples were collected from patients who previously received a dental implant, which in some cases resulted in inflammatory states known as mucositis or peri-implantitis. The dataset is the result of a previous binning process which assigned each reconstructed genome to the same, undetermined taxonomic cluster.

## 2.1 TAXONOMIC ASSIGNMENT

The taxonomic assignment of each genome was performed using PhyloPhlAn [1], a software specialised in large-scale phylogenetic profiling: in particular, it was used to try to assign the MAGs to species-level genome bins (SGBs) present in a chosen database.
Details about the command and the main parameters are reported as follows:

```
phylophlan_metagenomic -i <bin folder> -o <output folder> --nproc 4 -n 1 \
-d CMG2324
```

- `-nproc` defines the number of cores to use (set as 1 by default)
- `-n` specify the number of SGBs to report in the output (set at 10 by default)
- `-d` defines the chosen database containing information about the SGBs

## 2.2 QUALITY-CHECKING

Quality-checking on the MAGs was performed using the checkM software [3], which provides a set of tools to evaluate the quality of the reconstructed genomes by assessing the presence and copy number of a given set of single-copy genes.
Since all of the given MAGs belong to the same taxonomic cluster, the choice fell on a taxonomic-specific workflow: in this way, each genome was analysed in respect to the same set of markers associated with a specified taxonomic label. The command utilised to launch this analysis has the following shape:

```
checkm taxonomy_wf <rank> <taxon> <bin folder> <output folder>
```

## 2.3 GENOME ANNOTATION

Genome annotation was performed using Prokka [4], a tool specifically designed for detecting genome features of prokaryotes, viruses, and archaea. The only input Prokka requires are the genomes sequences in FASTA format. At the end of the analysis, it generates several output files: most importantly, it produces GFF3 files containing details about each contig (ID and name), along with their respective lengths. The basic syntax for this software is:

```
prokka <contigs.fna>
```

## 2.4 PANGENOME ANALYSIS

Pangenome analysis was performed using Roary [5], a computational tool designed for genome analysis and pangenome construction. The software takes as input GFF3 files (produced by Prokka) and converts all coding regions into protein sequences, in order to compare them to identify shared genes across different genomes. At the end of the analysis, Roary computes the core and accessory genome and generates several output files. To perform the analysis, the following command was executed:

```
roary <GFF3 files> -f <output folder> -i 95 -cd 90 -p 4
```

This command applies the following specifications:

- `-i` defines the minimum percentage of identity between two sequences needed to consider them as the same gene (set as 95% by default)
- `-cd` defines the percentage of strains a gene must be present in to be considered a core gene; it's set as 99% by default, but was lowered to 90% to account for the low number of genomes.
- `-p` increases the thread count from 1 to 4 to speed up the computation

## 2.5. PHYLOGENETIC ANALYSIS AND ASSOCIATION OF HOST METADATA

Phylogenetic analysis was performed using a combination of Roary and FastTreeMP [6]. The phylogenetic tree was built considering only the core genes, which were compared between different strains through a process of multiple alignment (performed by Roary). This step was conducted using the following line of code:

```
FastTreeMP -pseudo -spr 4 -mlacc 2 -slownni -fastest -no2nd -mlnni 4 -gtr -nt \
-out core_gene_phylogeny.nwk core_gene_alignment.aln
```

- `Core_gene_phylogeny.nwk and core_gene_alignment.aln` are two files produced by Roary needed to build the tree
- `-pseudo` is an option recommended for fragmented sequences with little overlap
- `-spr 4, -mlacc 2 -slownni, -fastest -no2nd and -mlnni 4` are recommended configurations to speed up the computation
- `-gtr -nt` are parameters needed to generate a phylogeny of nucleotide sequences

# 3. RESULTS AND DISCUSSION

## 3.1 TAXONOMIC ASSIGNMENT OF THE MAGS

Based on the analysis conducted using PhyloPhlan, the taxonomy identity of each sample was determined: the Mash distance [2] relative to each genome is <5% with respect to the kSGB1256, associated with *Prevotella pleuritidis*. The full taxonomic label is here reported:

- Kingdom: Bacteria
- Phylum: Bacteroidota
- Class: Bacteroidia
- Order: Bacteroidales
- Family: Prevotellaceae
- Genus: Prevotella
- Species: Prevotella pleuritidis

*Prevotella*, an anaerobic, gram-negative bacterium, is abundant in the oral environment of mammals, but it can also be found in the genitourinary tract, vaginal mucosa, skin, and digestive system. While its role in the human microbiome remains unclear, this genus is associated with various pathological conditions [10-11].

*P. pleuritidis*, first identified in 2007 from the pleural fluid of a patient affected by pleuritis, was previously linked only to oral or dental infections, but it's now frequently implicated in many inflammatory states, like lung abscesses and rheumatoid arthritis [12-13].

## 3.2 QUALITY-CHECKING

Quality-checking was performed after successfully assigning the taxonomic label to the species, which allowed the selection of a Genus-specific gene set. The main characteristics of the gene set that was used are summarised in the following table:

| **Rank** | **# Reference genomes** | **# Marker genes** | **# Marker sets** |
|---|---|---|---|
| Genus: Prevotella | 65 | 522 | 281 |

This analysis revealed that about half of the MAGs are considered high quality, meaning that when looking for a list of marker genes for the corresponding taxa the completeness was estimated to be >90%, with <5% contamination. The other half of the MAGs were considered as medium quality genomes, since their completeness oscillated between 50% and 90%. These statistics reveal that the genomes analysed in this study are the result of an overall successful assembly, which excluded most non-related sequences coming from different microbes.

The consistent CG content estimated around 45% is another strong indication that the genomes belong to the same species. The average genome size and number of predicted genes, calculated considering only high quality genomes, are respectively 2.1 Mb and over 2000 predicted genes per strain. The results of this analysis are reported in the following table:

| | |
|---|---|
| **# High quality genomes (>90% completeness and <5% contamination)** | 16/30 |
| **# Medium quality genomes (>50% <90% completeness and <5% contamination)** | 14/30 |
| **# Low quality genomes (<50% completeness or >5% contamination)** | 0/30 |
| **Average completeness** | 86.5% |
| **Average contamination** | 0.754% |
| **CG content ± std** | 0.452 ± 0.003 |
| **Average genome size** | ~ 2.1 Mb |
| **Average # predicted genes** | ~ 2059 |

## 3.3 GENOMIC ANNOTATION

Through genomic annotation, a total of 54,103 proteins were detected in the MAGs, consisting of 25,013 known proteins and 29,090 hypothetical ones. Among the known proteins, 341 are present in a single MAG.
Hypothetical proteins make up a significant portion (about 54%) of the proteome; they are termed hypothetical because they are proteins whose existence has been predicted by bioinformatic analysis but whose function has not been experimentally confirmed. With further experimental or functional bioinformatic analysis, the function and biological role of these proteins can be determined.

## 3.4 PANGENOME ANALYSIS

Pangenome analysis revealed a total of 5336 in the pangenome of *P. pleuritidis*, with 764 core genes (present in at least 90% of the strains, or 27 out of 30) and 4572 accessory genes. The relative size of the core genome with respect to the full pangenome is quite low (about 14% of the total genes detected in this analysis): considering that during quality-checking the number of predicted genes was around 2000 in each strains, this suggests that *P. pleuritidis* harbours extensive genetic flexibility, which is a feature of open pangenomes.
This is confirmed by Figure 1, representing the number of conserved and total genes found as the analysis was performed over the 30 MAGs. While the amount of core genes stabilises, the number of total genes found continues to increase without reaching a plateau. While this could be the result of a low number of samples, it's possible to assume that this trend would be maintained even as more and more strains are included in the study; thus, the total number of genes would continue to increase in size, as typical of an open pangenome.

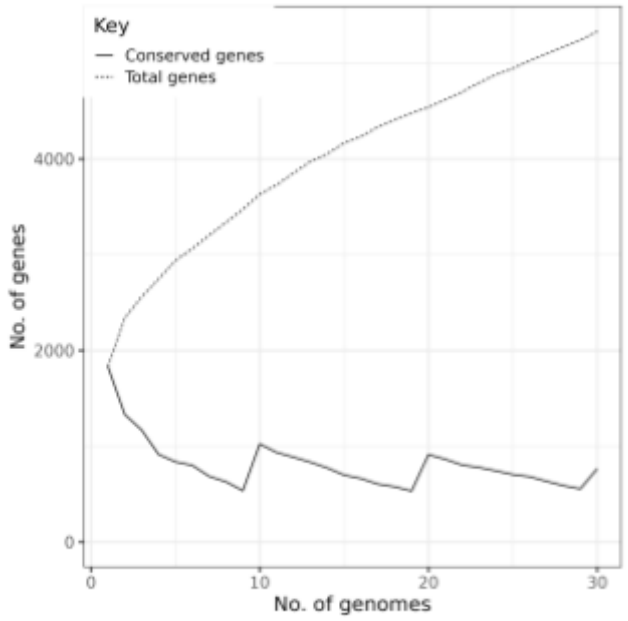


*Figure 1 - conserved vs total genes*

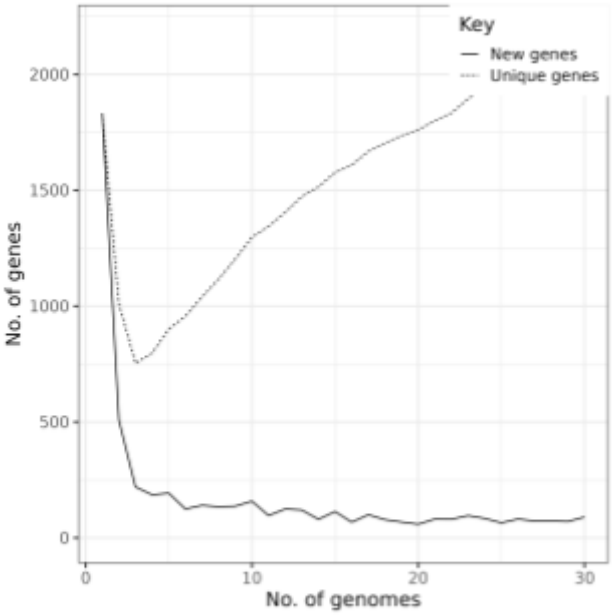


*Figure 2 - unique vs new genes*

Finally, Figure 2 represents the number of new and unique genes found while progressing through the analysis. The number of unique genes detected with each new MAG is steeply increasing, meaning that the genome has a high tolerance to variability while still maintaining the fundamental characteristics associated with *P. pleuritidis* as a species.

## 3.5 PHYLOGENETIC ANALYSIS AND ASSOCIATION WITH HOST METADATA

Figure 3 represents the phylogenetic tree built through the comparison of core genes. The tree visualisation was developed in iTOL, where the central point-root and rectangular shape settings were applied in order to ease the visualisation of the distances between MAGs.
Two main subtrees are highlighted, each further subdivided into two other main clades with additional branches. The tree is annotated with the patients' metadata to investigate the potential correlation between specific conditions and the distribution of branches: the metadata collected for each host includes sex, BMI, age, smoking status (smoker, ex-smoker, non-smoker), and study group (peri-implantitis, mucositis, healthy).
Some subgroups appear to have similar characteristics: for example, a correlation is identified in the second subgroup between four MAGs (M1873802605, M1421637969, M1260886280, M11944372573) and inflammation of the mucosa. Still, no definite correlation could be derived from this dataset.
Figure 4 shows the phylogenetic tree built through the comparison of accessory genes. Despite utilising the same visualisation settings, this tree displays a different arrangement compared to the previous one. Notably, the first sub-tree appears smaller than the second, exhibiting more extensive ramifications.
As in Figure 3, Figure 4 shows that despite the observable correlations between MAGs and metadata, the results remain statistically insignificant due to the small size of the dataset.

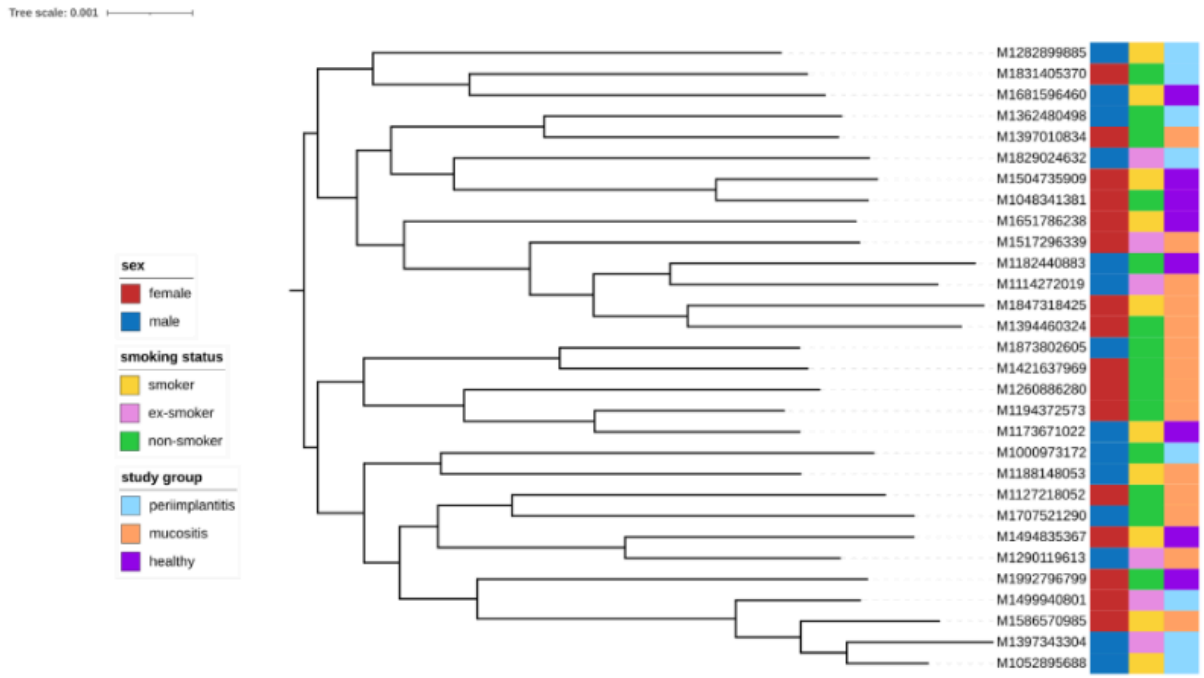


*Figure 3 - Core gene phylogeny*

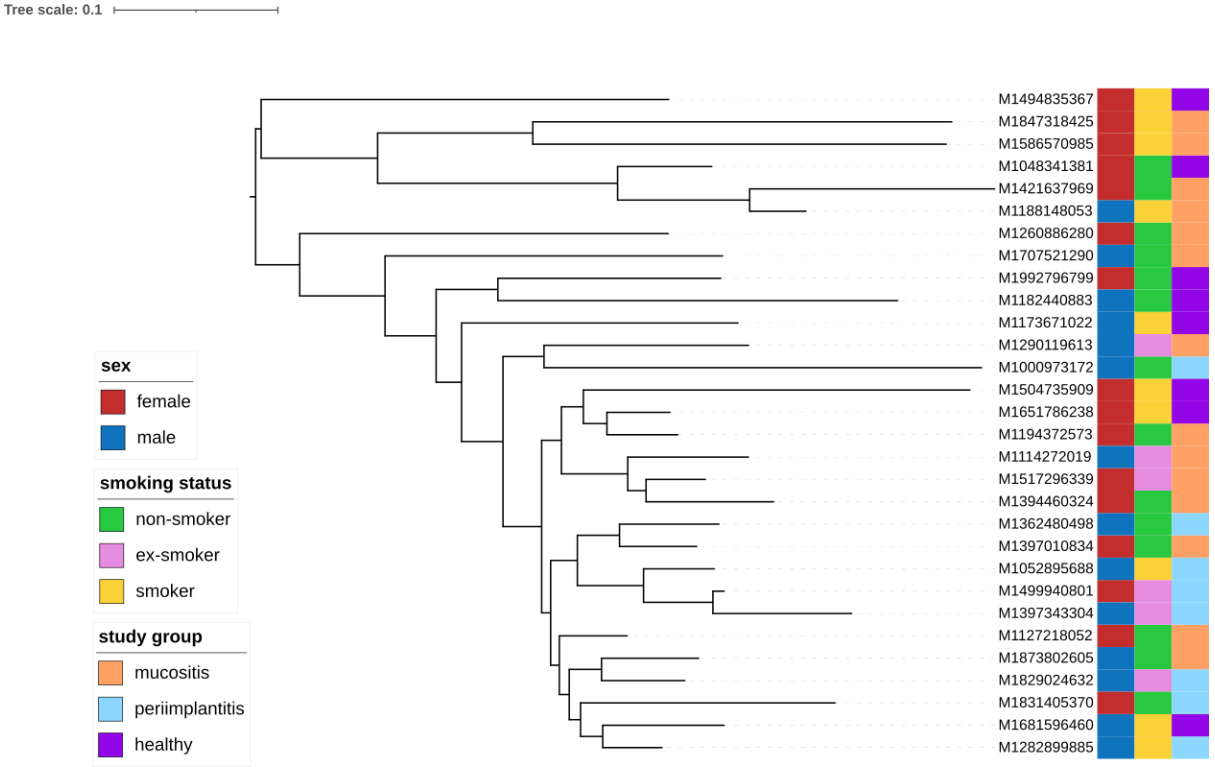


*Figure 4 - Accessory genes phylogeny*

# 4. CONCLUSIONS

In this study, 30 different MAGs belonging to the SGB associated with *Prevotella pleuritidis* were analysed: both qualitative and quantitative insights were successfully collected through the employment of computational tools. Analysis of the pangenome revealed extensive genetic plasticity, whereas genomic annotation underlined how most of this species' gene content is yet to be fully characterised. Still, the role of this species in the context of the oral microbiome, especially during inflammatory states, remains unknown. Future studies should focus on determining the functions encoded in the genome, especially in genes that are differentially expressed between strains in different inflammatory conditions.

# 5. REFERENCES


[1] Asnicar, F. *et al.* (2020) *Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using phylophlan 3.0*, *Nature News*. Available at: https://www.nature.com/articles/s41467-020-16366-7

[2] Ondov, B.D. *et al.* (2016) *MASH: Fast genome and metagenome distance estimation using Minhash - Genome Biology*, *BioMed Central*. Available at: https://doi.org/10.1186/s13059-016-0997-x

[3] Parks, D.H. *et al.* (2015) *CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes*, *Genome Research*. Available at: https://genome.cshlp.org/content/25/7/1043.short

[4] Seemann, T. (2014) *Prokka: Rapid Prokaryotic Genome Annotation*, *Bioinformatics (Oxford, England)*. Available at: https://pubmed.ncbi.nlm.nih.gov/24642063/

[5] Page, A.J. *et al.* (2015) *Roary: Rapid large-scale prokaryote Pan Genome Analysis*, *OUP Academic*. Available at: https://academic.oup.com/bioinformatics/article/31/22/3691/240757

[6] Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) *FastTree 2 – approximately maximum-likelihood trees for large alignments*, *PLOS ONE*. Available at: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0009490

[7] Schwarz, F., Derks, J., Monje, A., Wang, H.L. (2018) *Peri-implantitis*, *Journal of clinical periodontology*. Available at: https://pubmed.ncbi.nlm.nih.gov/29926484/

[8] Yamashita, Y., Takeshita, T. (2017) *The oral microbiome and human health*, *Journal of oral science*. Available at: https://pubmed.ncbi.nlm.nih.gov/28637979/

[9] Sedghi, L.,DiMassa, V., Harrington, A., Lynch, S.V., Kapila, Y.L.; (2000) *The oral microbiome: Role of key organisms and complex networks in oral health and disease*, *Periodontology 2000*. Available at: https://pubmed.ncbi.nlm.nih.gov/34463991/

[10] Tett, A. *et al.* (2021) *Prevotella diversity, niches and interactions with the human host*, *Nature News*. Available at: https://www.nature.com/articles/s41579-021-00559-y

[11] Sakamoto, M. *et al.* (2007) *Prevotella pleuritidis sp. nov., isolated from pleural fluid*, *microbiologyresearch.org*. Available at: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.64885-0#tab2

[12] Asif, A.A., Roy, M. and Ahmad, S. (2020) *Rare case of* prevotella pleuritidis *lung abscess*, *BMJ case reports*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7478012/

[13] Esberg, A. *et al.* (2021) *Oral microbiota identifies patients in early onset rheumatoid arthritis*, *Microorganisms*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8400434/