UNIVERSITÀ DI TRENTO

MAG analysis reveals *Tannerella serpentiformis* relationship with dental implant inflammation

Alex Callegaro, Alberto Catalano, Alessandro Fiume

1 Introduction

The **oral cavity** represents the second most abundant and diverse microbiota after the gut [1]. This complex ecosystem is characterized by microbial colonization of both dental hard tissues and oral mucosal soft tissues.
Dental implants are a common and essential treatment modality in modern dentistry [2]. However, implant failures can occur due to infection, resulting in peri-implant diseases. These are categorized as peri-implant **mucositis**, a reversible inflammatory condition initiated by biofilm accumulation that disrupts host-microbe homeostasis at the implant-mucosa interface [3], and **peri-implantitis**, a progressive and irreversible disease involving inflammation and destruction of surrounding hard and soft tissues [4].

Advances in molecular techniques, such as shotgun sequencing and metagenomic analysis of Metagenome-Assembled Genomes (**MAGs**), have significantly advanced our understanding of the microbiome's role in human health [5]. The application of genomic tools, including *CheckM* [6, 7], *PhyloPhlAn* [8, 9], *Prokka* [10, 11], *Roary* [12, 13], and *FastTree* [14], now facilitates more precise metagenomic analysis of oral surfaces, offering deeper insights into the microbial composition linked to peri-implantitis and peri-implant mucositis.

In this study, a metagenomic analysis was conducted on samples from individuals with varying implant conditions. The objectives were to accurately determine the **taxonomic classification** of bacteria associated with these conditions and to analyze their **phylogenetic profiles** through comparisons between samples from affected and healthy individuals. Furthermore, **genome annotation** and **pangenome analysis** were performed to characterize the bacterial strain diversity associated with these conditions.

2 Methods

The initial SGB2048 dataset comprised 30 MAGs, each accompanied by patient metadata including sex, BMI, age, **smoking status**, and **implant condition** (healthy, mucositis, or peri-implantitis).

2.1 Quality check

MAG quality assessment, focusing on **completeness** and **contamination**, was performed using *CheckM v1.2.3*.
A taxonomic workflow, employing lineage-specific marker gene sets, was implemented for genome quality evaluation.
MAGs were classified based on the following quality thresholds:

- **High quality**: completeness $\geq$ 90% AND contamination $<$ 5%;
- **Medium quality**: 50% $\leq$ completeness $<$ 90% AND contamination $<$ 5%;
- **Low quality**: completeness $<$ 50% OR contamination $>$ 5%

2.2 Taxonomic assignment

Taxonomic assignment of all MAGs was conducted using *PhyloPhlAn 3.0.67*. MAGs were mapped to the nearest species genome bin (SGB) based on genomic similarity, using a **5%** Mash distance threshold. MAGs exhibiting Mash distances exceeding this threshold were not assigned to an SGB, potentially indicating novel species.

2.3 Genome annotation

Genomic feature identification and annotation were performed using *Prokka v1.14.6*. The generated .gff file was of particular interest due to its detailed annotations of genomic regions within each contig, including **CDS**, tRNA, and rRNA features.

2.4 Pangenome analysis

Pangenome analysis was performed using *Roary v3.13.0*, which categorizes the gene into **core**, soft-core, and **accessory** genes (further divided into shell and cloud genes). The BLASTP protein similarity threshold was set at the default 95%. Core genes were defined by a **custom 90%** presence threshold across genomes, while soft-core, shell and cloud were set respectively to 89%-90%, 15%-89% and 0%-15%.

2.5 Phylogenetic analysis

Maximum-likelihood phylogenetic analysis was conducted using *FastTreeMP (v2.1.11)*. Gene alignments generated by *Roary* served as input, enabling the construction of a comprehensive **phylogenetic tree**, annotated using iTOL. *Tannerella forsythia* reference genome *(refSeq:GCF_000238215.1)*, the sole other known species within the *Tannerella* genus, was used as the **outgroup** *(NCBI:txid28112)*. Additionally, an **accessory gene tree** was constructed from the binary gene presence/absence file produced by *Roary*. **Patristic branch distances** for both trees were calculated using a tailor-made developed Python script.

3 Discussion and Results

3.1 Quality assessment

Initial quality assessment of MAGs was performed using *CheckM*, with the Bacteria domain as the reference taxonomic group. This preliminary analysis revealed one low-quality MAG due to high contamination. Despite it is generally advised to exclude low-quality MAGs, this data was **retained** due to the limited dataset size, which could compromise statistical validity. Most MAGs were classified as medium quality, and two as high quality. Notably, one high-quality MAG originated from the peri-implantitis group, which, due to its smaller size, is particularly valuable as its inclusion increased the statistical power of the analysis. Subsequent to the initial analysis, a second *CheckM* evaluation was conducted after assigning the MAGs to a single SGB under the taxon *Tannerella serpentiformis*. This modification led to a significant **quality reduction**; no MAGs were classified as high quality, and the number of low-quality MAGs increased to four. This quality decline was primarily attributed to a **marked increase in contamination**. Examination of completeness and contamination metrics revealed a significant deterioration in both (*Figure* 1; *Supplementary table* S1 and S2)

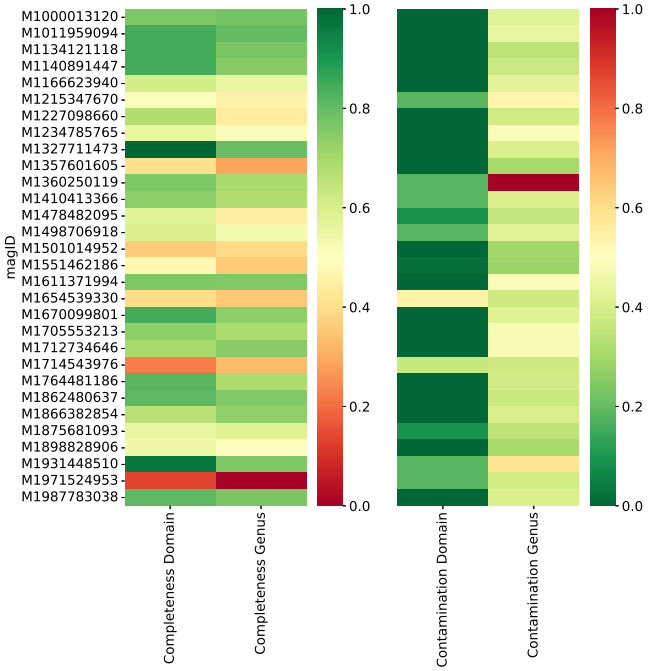


Figure 1: *Heatmap showing the overall decrease in quality achieved by using the CheckM tool [6, 7] with and without appropriate taxonomic metadata.*

Average completeness decreased from 78.71% to 75.51% (approx. -4.07% relative change), while average contamination increased from 0.69% to 4.01% (+481% relative change).
As observed, the primary quality decline is attributed to the substantial contamination increase, mainly due to the more specific taxonomic assignment of our MAGs. The refined taxonomic assignment enabled *CheckM* to identify **more reliable** marker genes, leading to **more accurate** contamination assessments and the detection of higher levels of foreign DNA.

The average genome length of the MAGs was determined to be 2,201,476 base pairs, compared to 2,973,544 base pairs for the *T. serpentiformis* reference genome *(RefSeq:GCF_001717525.2)*. The ratio of these values indicates that, on average, the MAGs covered 74.40% of the reference genome. This coverage percentage is consistent with the completeness values independently estimated by *CheckM*, **demonstrating the reliability and consistency** of this tool in assessing genome completeness.

The GC content of the MAGs remained **consistent**, averaging 57.8% with a standard deviation of 0.021, which closely aligns with the reference genome GC content of 56.5%. This consistency suggests a stable genomic composition across the MAGs and supports the reliability of the assemblies.

A brief review of the scientific literature indicates that the average N50 score of 33.108 Mbp calculated for the 30 MAGs falls within the extreme upper tail of N50 value distributions typically reported for metagenome-assembled genomes. [15]
To investigate the potential correlation between assembly quality and *CheckM* quality for the generated MAGs, an analysis was initiated. Assembly quality was assessed using the number of contigs and the **N50** score as metrics. Given that the majority of MAGs were initially categorized as medium quality by *CheckM*, a re-grouping of MAGs into four custom quality groups was performed prior to graphical representation (Supplementary figure 2). The plot presented in Figure 2 revealed several noteworthy insights. Firstly, an inverse proportionality between contig number and N50 score was observed. As anticipated, an increase in the number of contigs necessary to span the approximate 2.2 Mb length of each MAG, corresponded to a decrease in the N50, which is indeed a contiguity metric. Furthermore, and most significantly, a clear correlation between the regrouped CheckM quality and the assembly quality was evident. These findings suggest that the integrity of the source genetic material, coupled with the reliability of the sequencing technology and assembly software, are crucial determinants in obtaining biologically relevant MAGs.

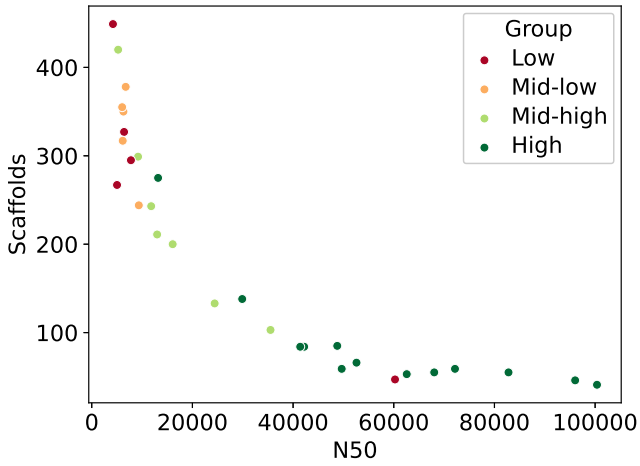


Figure 2: *Assembly quality VS. CheckM quality scatter plot; regrouping criterion shown in Figure S1.*

3.2 Taxonomic assignment

Taxonomic classification of the metagenome-assembled genomes (MAGs) was conducted using *PhyloPhlAn*.

It was determined that all MAGs were affiliated with a **single** species-level genome bin (SGB), which was subsequently identified as ***Tannerella serpentiformis*** (*NCBI:txid712710*). Based on average phylogenetic distance, *T. serpentiformis* is by far the closest match, in terms of average distance, compared to the other top three poteintial taxonomic classifications: *Porphyromonas denticola*, *Porphyromonas koreensis*, and an unidentified *Selenomonas* species (*Supplementary table* S3).

Tannerella serpentiformis is recognized as a **common** component of the human oral cavity microbiome. [16, 17] Although a planktonic lifestyle is possible [18], it is predominantly associated with subgingival plaque formations (biofilms) [18]. A key distinction from *Tannerella forsythia* is the inability of *T. serpentiformis* to form single-species biofilms. [18] Consequently, within oral plaque, it is primarily found in association with other bacterial species. Despite this limited capacity for independent environmental colonization, *T. serpentiformis* is reported to exhibit **high prevalence** in human populations, although typically at low abundance levels. [17] In contrast to *T. forsythia*, which is recognized for its pathogenic potential, *T. serpentiformis* is generally regarded as a **health-associated member** of the oral microbiota, largely attributed to the lack of proteins implicated in host tissue degradation [17, 19, 20] and cytotoxicity [21], as opposed to *T. forsythia*, while also beeing less masked from the host's inmune system, due to its unique S-layer composition [18, 22]. *T. serpentiformis* is also known to compete with and inhibit the growth of *T. forsythia*, which also contributes to overall peridontal health. [18]

3.3 Genome annotation

Genome annotation of the 30 MAGs, performed using *Prokka*, revealed a total of **54,011** coding sequences (**CDS**). Within these CDS, 29,788 were identified as encoding **hypothetical** proteins, while 23,681 CDS were associated with proteins of **known** function. Further analysis indicated that a smaller proportion of CDS corresponded to tRNAs, putative proteins, and other genetic elements. Initial plans to conduct a virulence analysis of the known proteins, with the aim of investigating potential horizontal acquisition of virulence genes from *Tannerella forsythia*, were considered. However, due to time constraints and the absence of *Tannerella species* in the NCBI Virulence Factor Database (*VFDB*), this analysis was not pursued.

To **assess** the annotations **consistency** we analyzed the number of different annotation objects relative to MAG length. The ratios demonstrated a striking consistency across all MAGs (*Figure* 3). For instance, the ratio of hypothetical proteins to MAG length, as well as the ratios for known proteins and tRNAs, remained relatively constant. This uniformity not only indicates the robustness of the *Prokka* annotations but also highlights the inherent consistency of the MAG dataset itself.

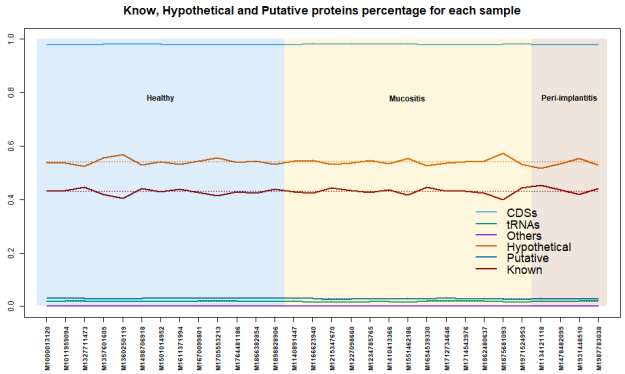


Figure 3: *Known, hypothetical and putative percentage for each sample.*

3.4 Pangenome analysis

Pangenome analysis of *Tannerella serpentiformis* using *Roary* revealed insights into its genetic diversity. The pangenome was categorized into core, soft-core, shell, and cloud genomes. As expected, the core genome was small (**9%** of the total pangenome), and the cloud genome was large (70%). Unexpectedly, soft-core genes were absent.

Initial bifurcation analysis (Figure 4) showed that total gene number **increases** with genome number, while conserved gene number plateau after circa five genomes, giving a clue about an **open pangenome**. Conserved genes averaged 629.24 (std. 115.34) on the long run. An unexpected periodic increase in conserved genes every ten genomes was observed, attributed to *Roary*'s dynamic recalculation of this number and to the 90% custom core genome threshold.

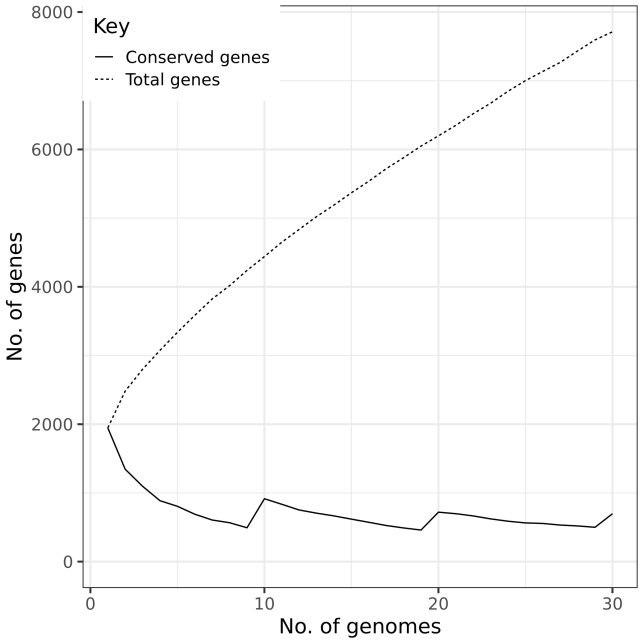


Figure 4: *Classic biforcation analysis: Conserved VS. Total genes.*

The second bifurcation analysis (Figure 5) examined the relationship between new and unique genes. After the first four genomes are plotted, new genes contributed by each additional genome averaged at 181.92 (std. 39.23) and decrease

slowly. This observations lead us to the implementation of a linear model (red line in Figure 5) including data from genome 4 onwards. The model was successful with $r^2 = 0.77$, $p < 0.001$ with a slope of -4.35 suggesting a decreasing rate of contribution from each new genome to pangenome diversity, indicating a negative acceleration in pangenome growth, challenging the initial open pangenome hypothesis.

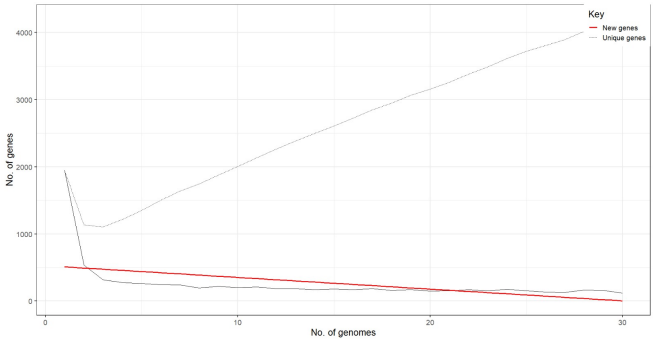


Figure 5: *Unique vs new genes: Linear model fitting curve suggesting pangenome closure.*

Further analysis performed by fitting a **Heap's law model** to the new genes plot [23] lead to the obtainment of the model curve: $y = 423.4566 \cdot x^{-0.3205}$, ($p < 0.01$). The negative power of the x coordinate is a commonly cosidered **indicative of asymptotic behavior**. [24] Evaluation of the curve's value at high genome counts predicted minimal new genes discovery (22 new genes at 10,000 genomes, ˜0.3 at 10^{11} genomes), supporting pangenome closure. While we acknowledge the potential limits of the model, we are confident that these results, obtained using a standard pangenome analysis method, suggest that the *T. serpentiformis* pangenome is **approaching closure**.

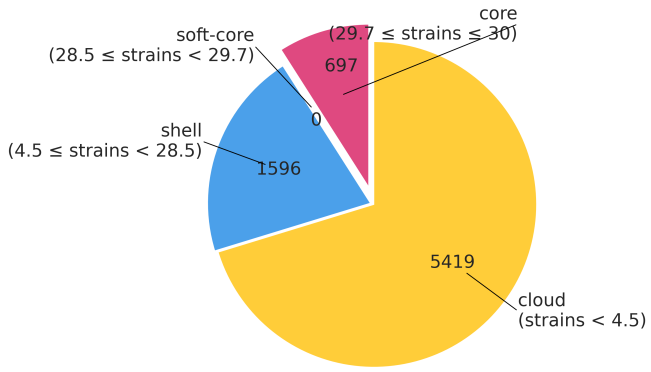


Figure 6: *Pie chart showing percentages of pangenome composition.*

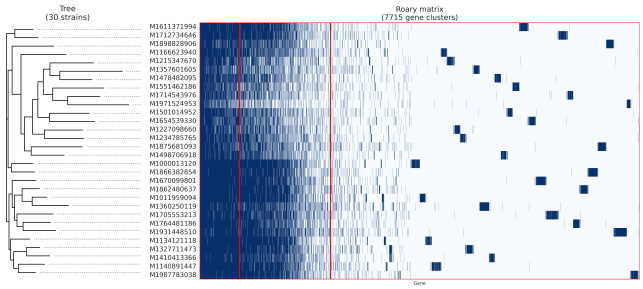


Figure 7: *Heatmap and dendrogram that intuitively and visually show the regions corresponding to the components of the pangenome.*

3.5 Phylogenetic analysis

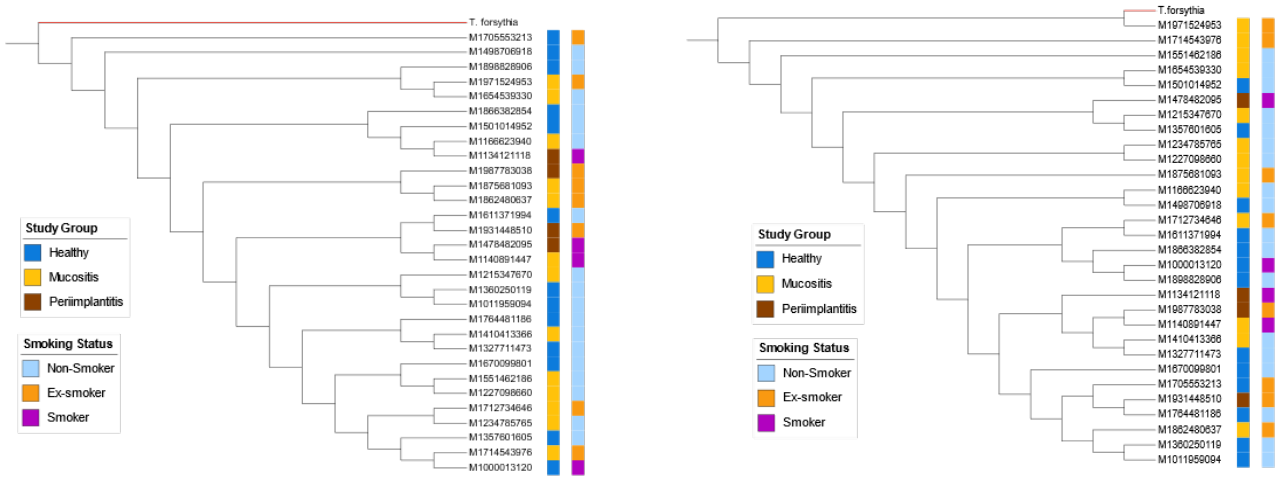
To assess the phylogenetic **relationships** among the strains, two phylogenetic trees were constructed 8. To enhance the robustness and reliability of the phylogenetic inference [25], an **outgroup** reference *(T.forsythia)* was incorporated. Tree annotations were performed, incorporating metadata related to the health and and smoking status. Body Mass Index (BMI) was not considered as potentially useful metadatum due to the lack of underweight or obese individuals. No noticeable correlation between the health status was observed in the **core genome phylogeny**. The average ingroup (excluding the outgroup) branch distance (average patristic distance) was computed to be 0.0355, with a standard deviation of 0.0038. These values indicate a high degree of similarity among taxa within the ingroup at the core genome level. The average distance from the outgroup was also calculated, resulting in a value of 0.3373 with a standard deviation of 0.0042. This value reflects the expected clear separation between the ingroup and the outgroup.
In contrast, an **accessory genome** phylogenetic tree was also constructed. The average ingroup patristic in the accessory genome tree, excluding the outgroup, was 0.4543, with a standard deviation of 0.1345. This relatively high average indicates a **greater** degree of divergence among the ingroup taxa, suggesting substantial variability in the accessory genes. The average patristic distance from the outgroup, was even higher at 0.9827, with a standard deviation of 0.1334. Notably, despite the average distance between the ingroups and the outgroup being larger in the accessory genomes tree, nuanced clustering of diseased case is noticeable near the outgroup.

Overall the phylogenetic analysis results further confirm a greater separation between the ingroup and the outgroup compared to the core genome tree. The comparatively higher standard deviation observed in the accessory genome tree suggests a potential role for the accessory genome in adaptation and its susceptibility to environmental pressures or horizontal gene transfer events.

4 Conclusion

Starting with 30 MAGs, this study performed a comprehensive analysis. Quality assessment using CheckM showed average quality MAGs *decreased* with correct taxonomic assignment, and revealed a **positive correlation** between assembly quality and CheckM quality. Taxonomic assignment using PhyloPhlAn confirmed all MAGs belong to ***Tannerella serpentiformis*** bin. Genome annotation showed **consis-**

tent CDS percentages across MAGs. Pangenome analysis revealed a typical structure with a **large** accessory genome and **small** core genome, while also suggesting a **likely closed pangenome – a key finding.** Phylogenetic analysis, using both core and accessory genomes based trees, confirmed the phylogenetic distance of *T. serpentiformis* from *T. forsythia*.



(a) *Core tree.*

(b) *Accessory tree.*

Figure 8: *Phylogenetic trees*

Supplementary informations

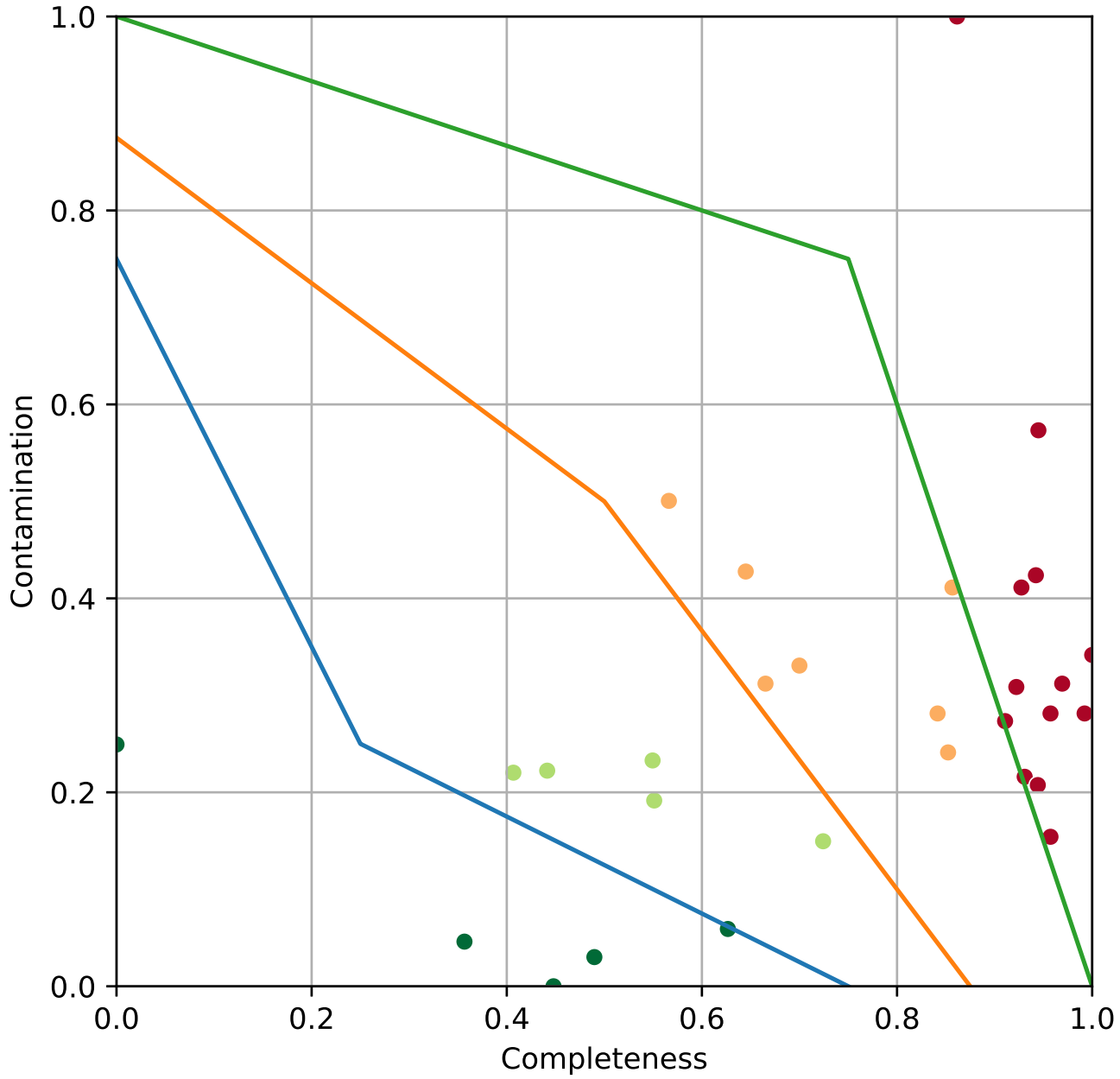


Figure S1: *Quality-based MAGs re-grouping criterion used for the plot in Figure 2.*

MagID	Marker lineage	Completeness	Contamination	Strain heterogeneity	Genome size	GC	N50
M1000013120	Bacteria	84.4828	0.0	0.0	2,512,087	0.5746	42,200
M1011959094	Bacteria	88.5057	0.0	0.0	2,545,038	0.5746	49,657
M1134121118	Bacteria	88.5057	0.0	0.0	2,474,344	0.5743	41,414
M1140891447	Bacteria	88.5057	0.0	0.0	2,575,540	0.5726	62,576
M1166623940	Bacteria	76.7241	0.0	0.0	2,176,665	0.5798	11,808
M1215347670	Bacteria	71.4734	1.7241	100.0	1,860,509	0.5849	5,247
M1227098660	Bacteria	79.8851	0.0	0.0	1,852,772	0.5824	9,351
M1234785765	Bacteria	74.1379	0.0	0.0	1,979,940	0.5846	12,991
M1327711473	Bacteria	96.4211	0.0	0.0	2,579,225	0.5702	96,034
M1357601605	Bacteria	66.3009	0.0	0.0	1,725,121	0.5835	6,425
M1360250119	Bacteria	84.1954	1.7241	100.0	2,477,783	0.5772	13,181
M1410413366	Bacteria	83.3333	1.7241	100.0	2,349,304	0.5730	24,416
M1478482095	Bacteria	75.0	0.8621	0.0	1,864,740	0.5833	6,296
M1498706918	Bacteria	75.7837	1.7241	100.0	2,063,662	0.5824	9,203

MagID	Marker lineage	Completeness	Contamination	Strain heterogeneity	Genome size	GC	N50
M1501014952	Bacteria	63.7931	0.0	0.0	1,880,534	0.5827	7,789
M1551462186	Bacteria	69.7492	0.1567	0.0	1,713,132	0.5844	4,216
M1611371994	Bacteria	84.4828	0.0	0.0	2,436,439	0.5749	52,595
M1654539330	Bacteria	65.7524	5.1724	66.67	1,825,243	0.5830	6045
M1670099801	Bacteria	88.5057	0.0	0.0	2,519,518	0.5734	48,772
M1705553213	Bacteria	83.3333	0.0	0.0	2,507,524	0.5719	35,515
M1712734646	Bacteria	81.0345	0.0	0.0	2,464,654	0.5748	82,795
M1714543976	Bacteria	57.0533	3.4482	100.0	1,664,457	0.5845	6,155
M1764481186	Bacteria	87.0690	0.0	0.0	2,366,678	0.5768	16,079
M1862480637	Bacteria	86.7816	0.0	0.0	2,562,504	0.5730	100,375
M1866382854	Bacteria	79.3103	0.0	0.0	2,498,285	0.5737	68,038
M1875681093	Bacteria	74.0596	0.8621	100.0	2,081,394	0.5807	6,744
M1898828906	Bacteria	73.4326	0.0	0.0	2,119,768	0.5757	60,249
M1931448510	Bacteria	94.6970	1.7241	100.0	2,646,394	0.5719	29,892
M1971524953	Bacteria	52.3510	1.7241	100.0	1,215,563	0.5935	5,024
M1987783038	Bacteria	86.7816	0.0	0.0	2,505,455	0.5734	72,163

Table S1: *Detailed summary of the qualities produced by the CheckM tool [6, 7] at domain level, considering it as Bacteria.*

MagID	Marker lineage	Completeness	Contamination	Strain heterogeneity	Genome size	GC	N50
M1000013120	*Tannerella*	85.1361	3.9966	10.34	2,512,087	0.5746	42,200
M1011959094	*Tannerella*	86.3870	4.1478	0.0	2,545,038	0.5746	49,657
M1134121118	*Tannerella*	84.6485	3.2785	0.0	2,474,344	0.5743	41,414
M1140891447	*Tannerella*	83.5714	3.5431	4.35	2,575,540	0.5726	62,576
M1166623940	*Tannerella*	74.1772	4.0911	19.23	2,176,665	0.5798	11,808
M1215347670	*Tannerella*	68.7338	5.0624	42.86	1,860,509	0.5849	5,247
M1227098660	*Tannerella*	68.0547	3.6187	20.0	1,852,772	0.5824	9,351
M1234785765	*Tannerella*	71.9396	4.6202	31.25	1,979,940	0.5846	12,991
M1327711473	*Tannerella*	86.0708	3.8454	0.0	2,579,225	0.5702	96,034
M1357601605	*Tannerella*	60.2087	2.8628	21.05	1,725,121	0.5835	6,425
M1360250119	*Tannerella*	80.7515	9.4712	41.43	2,477,783	0.5772	13,181
M1410413366	*Tannerella*	79.9395	3.8454	11.11	2,349,304	0.5730	24,416
M1478482095	*Tannerella*	68.1217	3.4360	22.73	1,864,740	0.5833	6,296
M1498706918	*Tannerella*	72.7632	3.9966	34.38	2,063,662	0.5824	9,203
M1501014952	*Tannerella*	65.6266	2.8061	33.33	1,880,534	0.5827	7,789
M1551462186	*Tannerella*	63.9218	2.7022	10.53	1,713,132	0.5844	4,216
M1611371994	*Tannerella*	84.0401	4.5981	10.34	2,436,439	0.5749	52,595
M1654539330	*Tannerella*	63.6590	3.5713	55.56	1,825,243	0.5830	6,045
M1670099801	*Tannerella*	83.2237	3.9794	30.0	2,519,518	0.5734	48,772
M1705553213	*Tannerella*	80.5593	4.5257	6.9	2,507,524	0.5719	35,515
M1712734646	*Tannerella*	83.4354	4.5257	2.44	2,464,654	0.5748	82,795
M1714543976	*Tannerella*	62.2492	3.5620	30.77	1,664,457	0.5842	6,155
M1764481186	*Tannerella*	80.3779	3.6565	7.69	2,366,678	0.5768	16,079
M1862480637	*Tannerella*	84.1194	3.5053	0.0	2,562,504	0.5730	100,375
M1866382854	*Tannerella*	82.7551	3.8076	4.0	2,498,285	0.5737	68,038
M1875681093	*Tannerella*	75.1657	3.2596	10.0	2,081,394	0.5807	6,744
M1898828906	*Tannerella*	71.1936	2.9100	0.0	2,119,768	0.5757	60,249
M1931448510	*Tannerella*	84.1446	5.5461	31.58	2,646,394	0.5719	29,892
M1971524953	*Tannerella*	45.6975	3.6943	61.54	1,215,563	0.5935	5,024
M1987783038	*Tannerella*	84.6485	3.8454	0.0	2,505,455	0.5734	72,163

Table S2: *Detailed summary of the qualities produced by the CheckM tool at genus level, considering it as Tannerella.*

MagID	Genome Bin	Genus	Species	Strain	Average distance
M1000013120	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03506
M1011959094	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03450
M1134121118	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03495
M1140891447	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03646
M1166623940	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03708
M1215347670	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03779
M1227098660	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03660
M1234785765	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03721
M1327711473	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03557
M1357601605	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.04138
M1360250119	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03683
M1410413366	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03604
M1478482095	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03726
M1498706918	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03708
M1501014952	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03812
M1551462186	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03963
M1611371994	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03588
M1654539330	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03922
M1670099801	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03547
M1705553213	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03811
M1712734646	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03583
M1714543976	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03969

MagID	Genome Bin	Genus	Species	Strain	Average distance
M1764481186	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03468
M1862480637	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03526
M1866382854	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03543
M1875681093	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03829
M1898828906	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03877
M1931448510	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03754
M1971524953	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.04803
M1987783038	kSGB2048	*Tannerella*	*Tannerella serpentiformis*	SGB2048	0.03481

Table S3: *Detailed summary of the most probable taxonomy assigned by PhyloPhlAn [8, 9] on the set of MAGs considered in this study (described in Section 1).*

References

[1] P. Deo and R. Deshmukh. "Oral microbiome: Unveiling the fundamentals". In: *Journal of Oral and Maxillofacial Pathology* 23.1 (2017), pp. 122–122. DOI: 10.4103/jomfp.jomfp_304_18.

[2] Ranjan Gupta, Neha Gupta, and DDS Kurt K. Weber. *Dental Implants*. PMID: 29262027, Bookshelf ID: NBK470448. StatPearls Publishing LLC, 2025.

[3] L. J. A. Heitz-Mayfield and G. E. Salvi. "Peri-implant mucositis". In: *Journal of Clinical Periodontology* 45.Suppl 20 (2018), S237–S245. DOI: 10.1111/jcpe.12953.

[4] R. Smeets et al. "Definition, etiology, prevention and treatment of peri-implantitis–a review". In: *Head and Face Medicine* 10 (2014), p. 34. DOI: 10.1186/1746-160x-10-34.

[5] S Yen and J. S. Johnson. "Metagenomics: a path to understanding the gut microbiome". In: *Mammalian genome* 32.4 (2021), pp. 282–296. DOI: 10.1007/s00335-021-09889-x.

[6] D. H. Parks et al. "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes". In: *Genome Research* 25.7 (2015), pp. 1043–1055. DOI: 10.1101/gr.186072.114.

[7] *CheckM*. URL: https://github.com/Ecogenomics/CheckM. Version 1.0.7, Parks et al., 2015.

[8] N. Segata et al. "PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes". In: *Nature Communications* 4 (2013), p. 2304. DOI: 10.1038/ncomms3304.

[9] *PhyloPhlAn*. DOI: 10.1038/s41467-020-16366-7. URL: https://github.com/biobakery/phylophlan. Version 3.0, Asnicar et al., 2020.

[10] Torsten Seemann. "Prokka: rapid prokaryotic genome annotation". In: *Bioinformatics* 30.14 (2014), pp. 2068–2069. DOI: 10.1093/bioinformatics/btu153.

[11] *Prokka*. DOI: 10.1093/bioinformatics/btu153. URL: https://github.com/tseemann/prokka. Torsten Seemann, 2014.

[12] Andrew J. Page et al. "Roary: rapid large-scale prokaryote pan genome analysis". In: *Bioinformatics* 31.22 (2015), pp. 3691–3693. DOI: 10.1093/bioinformatics/btv421.

[13] *Roary*. DOI: 10.1093/bioinformatics/btv421. URL: https://github.com/sanger-pathogens/Roary. Andrew J. Page et al., 2015.

[14] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments". In: *PLoS ONE* 5.3 (2010), e9490. DOI: 10.1371/journal.pone.0009490.

[15] Aditya Bandla et al. "910 metagenome-assembled genomes from the phytobiomes of three urban-farmed leafy Asian greens". In: *Scientific Data* 7.1 (Aug. 2020), p. 278. DOI: 10.1038/s41597-020-00617-9.

[16] K. Ansbro, W. G. Wade, and G. P. Stafford. "Tannerella serpentiformis sp. nov., isolated from the human mouth". In: *International Journal of Systematic and Evolutionary Microbiology* 70.6 (2020), pp. 3749–3754. DOI: 10.1099/ijsem.0.004229.

[17] J. Züger, H. Lüthi-Schaller, and R. Gmür. "Uncultivated Tannerella BU045 and BU063 are slim segmented filamentous rods of high prevalence but low abundance in inflammatory disease-associated dental plaques". In: *Microbiology* 153 (Nov. 2027), pp. 3809–3816. DOI: 10.1099/mic.0.2007/010926-0.

[18] Fabian L. Kendlbacher et al. "Multispecies biofilm behavior and host interaction support the association of Tannerella serpentiformis with periodontal health". In: *Molecular Oral Microbiology* 38.2 (Aug. 2022), pp. 115–133. DOI: 10.1111/omi.12385.

[19] Abdulkarim Y. Karim et al. "A novel matrix metalloprotease-like enzyme (karilysin) of the periodontal pathogen Tannerella forsythia ATCC 43037". In: *Biological Chemistry* 391.1 (Jan. 2010), pp. 65–75. DOI: 10.1515/BC.2010.009.

[20] Verena Friedrich et al. "Outer membrane vesicles of Tannerella forsythia: biogenesis, composition, and virulence". In: *Molecular Oral Microbiology* 30.2 (May 2015), pp. 105–120. DOI: 10.1111/omi.12104.

[21] Anne C. R. Tanner and Jacques Izard. "Tannerella forsythia, a periodontal pathogen entering the genomic era". In: *Periodontology 2000* 42 (Jan. 2006), pp. 88–113. DOI: 10.1111/j.1600-0757.2006.00184.x.

[22] G. Sekot et al. "Potential of the Tannerella forsythia S-layer to delay the immune response." In: *Journal of dental research* 90.1 (2011), pp. 109–114. DOI: 10.1177/0022034510384622.

[23] George Vernikos et al. "Ten years of pan-genome analyses". In: *Current Opinion in Microbiology* 23 (Jan. 2015), pp. 148–154. DOI: 10.1016/j.mib.2014.11.016.

[24] Hervé Tettelin et al. "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial pan-genome". In: *Proceedings of the National Academy of Sciences* 102.39 (Sept. 2005), pp. 13950–13955. DOI: 10.1073/pnas.0506758102.

[25] T. Kinene et al. "Rooting Trees, Medots for," in: *Encyclopedia of Evolutionary Biology* (2016), pp. 489–493. DOI: 10.1016/B978-0-12-800049-6.00215-8.