Computational microbial analysis of an SGB sampled from peri-implant tissues

Sara Baldinelli[1], Letizia De Pietri[1], Roan Spadazzi[1]

[1] *University of Trento, Trento 38123, Italy*

Introduction

The human microbiome varies considerably upon different anatomical regions. The oral cavity harbors the second largest and diverse microbiota after the gut; its microbiome has the role of maintaining both oral and systemic health [1]. Given the easy sampling procedure, it is one of the most studied microbial communities. Moreover, next-generation sequencing (NGS) techniques, specifically shotgun metagenomic sequencing, offer robust means to explore the whole genetic landscape of microbial communities within a sample, avoiding the possible limitations of traditional methods. In this study, we dive into the taxonomy and characterization of an SGB (Species-level Genome Bin) sampled from individuals displaying diverse conditions of the peri-implant tissue. Complementing quality control measures on 30 MAGs (Metagenome-Assembled Genomes) together with taxonomic assignment, followed by genome annotation analysis, gave us the starting ground for subsequent investigations. Through pangenomic and phylogenetic analysis we unraveled the open nature of the pangenome and relations among strains. Our ultimate aim was to elucidate the correlation, if any, within MAGs sampled from the same peri-implant condition; we hypothesize those may give rise to distinct quality measures, amount of coding sequences and phylogenetic relatedness.

1. Materials and Methods

1.1 Data

A set of 30 MAGs had been obtained by performing, after the DNA extraction from each sample and shotgun metagenomic sequencing, a *De novo* assembly approach consisting in aggregating reads into contigs, binning into MAGs, and clustering into SGBs.

Samples were extracted from peri-implant tissue collected thanks to the collaboration with dental clinics all over Italy. In particular, samples come from patients belonging to one of the following three conditions: healthy dental peri-implant tissue (3), peri-implant mucositis (11) or peri-implantitis (16) (**Figure 1**). The metadata also include the sample's smoking status (non-smoker, ex-smoker, smoker), sex, bmi, age and IDs.

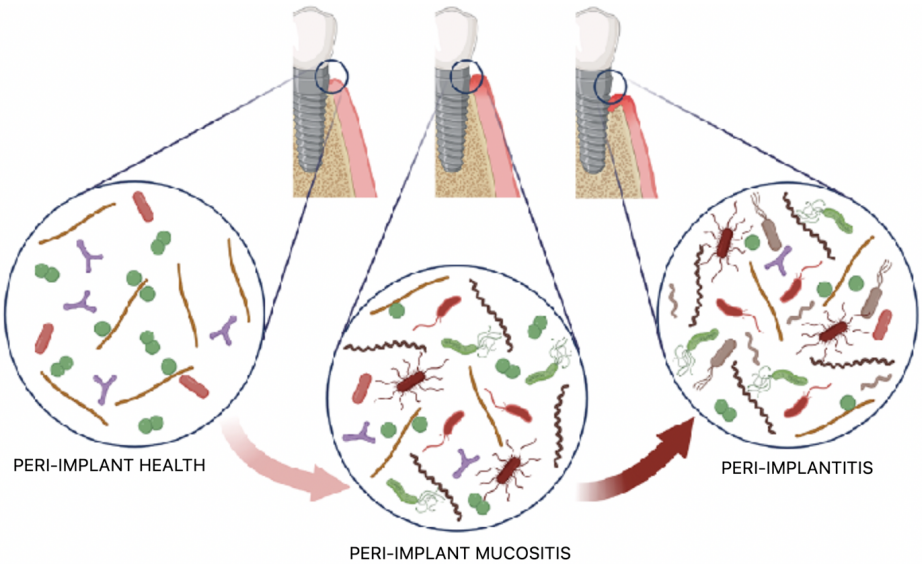


Figure 1. The three implant conditions: peri-implant health, mucositis and implantitis. These will be referred to as study groups.

1.2 Quality checking & Taxonomic assignment

To perform the quality checking, we took advantage of CheckM [2], an open source software that provides an estimate of genome completeness, contamination and heterogeneity by comparing the presence and copy number of a set of marker genes. In particular, we used the "taxonomy_wf" command by inputting a broad taxonomic assignment (Bacteria) to ensure the retrieval of the proper set of marker genes. The results were also plotted (**Figure 2**) using R [3] together with *ggplot2* [4].

```
$ checkm taxonomy_wf domain Bacteria ./mags checkm_output -t 4
```

PhyloPhlAn [5] is the pipeline we used for performing taxonomic assignment and analysis, outputting distances between our samples and the closest SGB. We discarded the other closest candidates due to them being too distant with respect to our MAGs. We then performed a second quality control using the taxonomic assignment given by PhyloPhlAn as a parameter for the CheckM command as a proof of concept in order to see whether the quality of our samples would change accordingly (**Section 2**).

1.3 Genome annotation

The genome annotation step was performed on all *.fna* files using Prokka [6]: a command line software to rapidly annotate genes and identify coding regions in prokaryotic genomes. The results, specifically the number of coding sequences and hypothetical proteins, were also visualized with a Python [7] script using the *seaborn* library [8].

1.4 Pangenome analysis

Taking in input annotated assemblies generated by Prokka as GFF files, Roary [9] is able to characterize the pangenome, identifying the core and accessory genes. As you can see from the following command, we set a percentage identity of 90 to define the core genes (-cd 90). This means that a gene is considered "core" if it appears in at least 90% of our samples, corresponding to 27 out of 30 MAGs. For what concerns the minimum percentage identity for BLASTP [10], we kept the default value of 95%.

```
$ roary prokka_output//.gff -f roary_output -p 4 -cd 90
```

1.5 Phylogenetic analysis

Following, we run a second Roary pipeline which, by using the -e and -n options, performed a core genes alignment using the MAFFT [11] algorithm instead of the default PRANK [12]. The nucleotide multiple alignment constituted the input to FastTree [13]. This tool infers phylogenetic trees by exploiting a maximum-likelihood approach; we tuned the parameters in order to achieve a reasonable computational time.

```
$ roary prokka_output/*/*.gff -f roary_output_w_aln -p 4 -cd 90 -e -n

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

Two different trees, one arising from the accessory genes presence/absence matrix resulting from the pangenome analysis Roary run and the other from the output of the second Roary run, were visualized and customized using iTOL [14]. In both cases, tree branches were colored based on the dental implant state of each sample.

2. Results and Discussion

2.1 Quality checking & Taxonomic assignment

As a first result, we observed the overall quality of our MAGs in terms of completeness, contamination and heterogeneity. Only 2 MAGs had a completeness lower than 50% (M1272513674 with 47,41% and M1364356612 with 49,14%) and only one sample (M1710390515) presented a contamination slightly higher than 5% (5.17%). Given that, these three samples have to be considered low quality ones. All the others are medium to high quality MAGs; in particular, only two (M1249152653 and M1336062757) have an overall high quality due to their completeness being higher than 90%. The GC content, which is stable at around 24% in all the samples, gives us further confidence about the good quality of our SGB.
Moreover, when searching for a correlation between both "Smoking State" and "Study Group" versus completeness and contamination, no specific patterns emerged from the plot (**Figure 2**), making our MAGs comparable in terms of quality measures.
The PhyloPhlAn taxonomic assignment revealed that, with an average distance of 0.0379, our samples are closest to the *Metamycoplasmataceae* bacterial family. Many species are part of this taxonomic group and they were found to be spread

across different organisms and locations, including the human oral cavity [15]. An in-depth literature search also highlighted the inconsistency of the nomenclature in terms of redundancy, stability and error-proneness [16]. Given these discrepancies, we performed a second quality check to verify whether the completeness and contamination improved by applying CheckM with a more specific bacterial classification (Phylum: *Tenericutes*). This hypothesis was confirmed and all our samples reached an average completeness of 80.72% (from 73.61%) and a contamination of less than 5%, officially classifying all of them as medium and high quality ones.

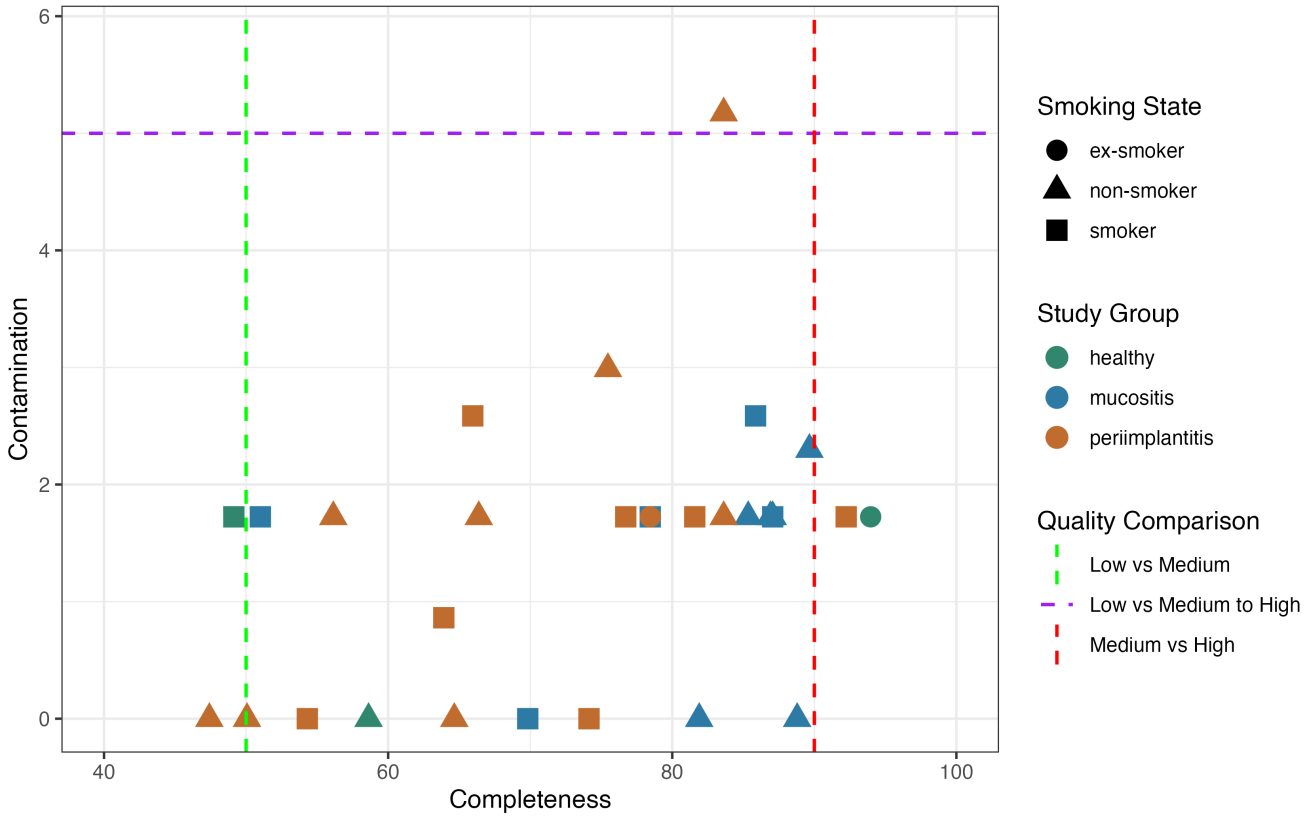


Figure 2. A scatter-plot depicting the distribution of our samples in terms of contamination and completeness. The points are colored by study group and they are shaped based on their smoking state. Furthermore, the dashed lines help the visualization of the quality of our samples: the green and red lines correspond respectively to a completeness of 50% and 90% and the purple line to a contamination of 5%.

2.2 Genome annotation

This step revealed that the amount of CDSs ranges from ∼700 to slightly above 1200 across all samples. We also visualized the proportion of hypothetical proteins (HPs) with respect to the total CDSs. From the plot we can infer that the abundance of the HPs does not change significantly between the three study groups. Note that the sum of HPs and known proteins (KPs) constitutes the total number of CDSs (**Figure 3**).
Also, it is interesting to observe that the HPs count outmatches the KPs one in all samples; this may suggest that we are dealing with an open pangenome. These proteins may be classified as ”hypothetical” due to them being found in fewer MAGs and therefore not well-characterized, possibly implying that the number of HPs could be related to the size of the accessory genome.

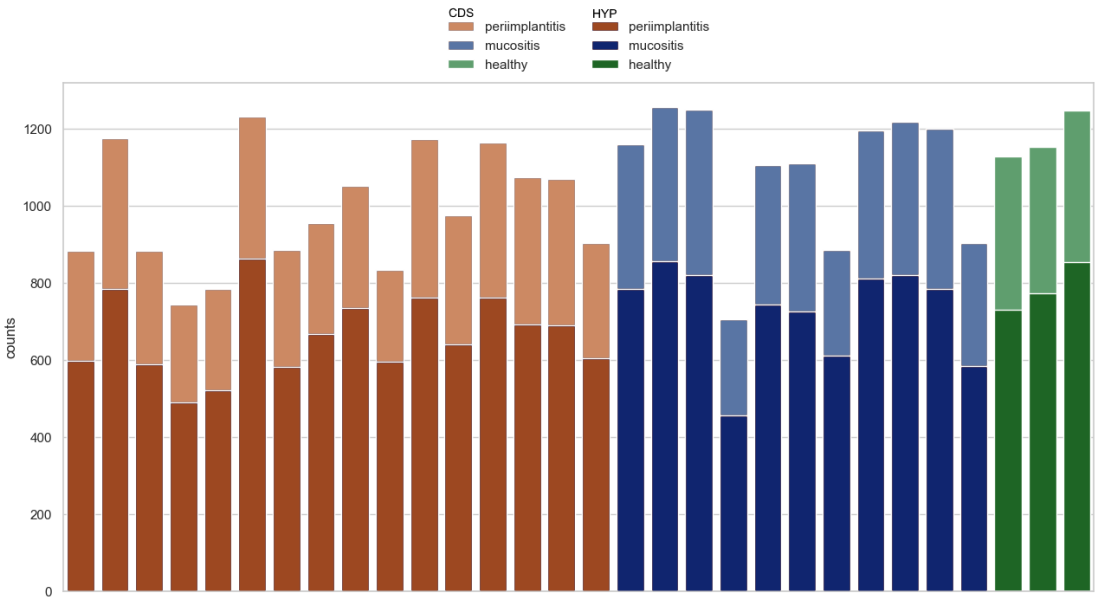
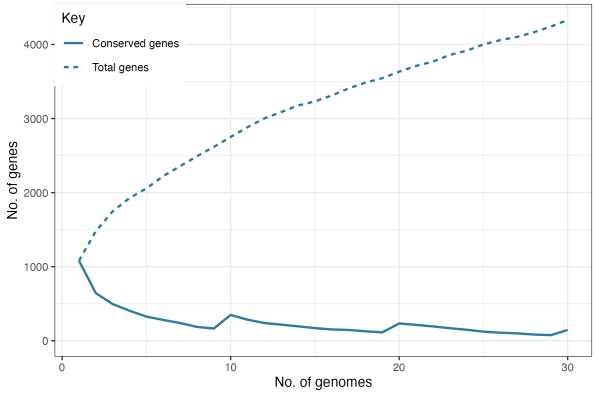


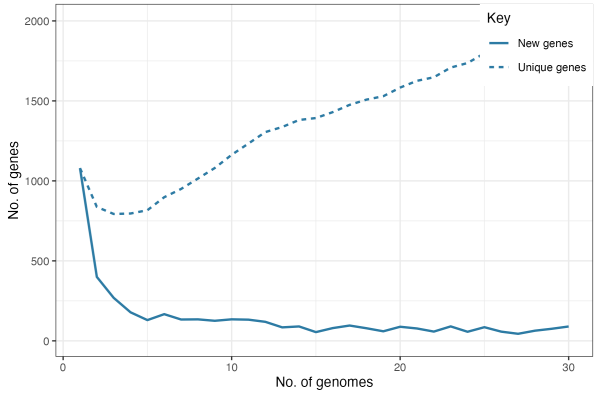
Figure 3. A barplot displaying the the amount of coding sequences and hypothetical proteins for each MAG. The darker shades are superimposed to the total amount of CDSs (represented by lighter coloring) in order to show the HPs. The colors corresponding to the study groups are also depicted.

2.3 Pangenome analysis

To explore the open genome hypothesis, we performed a pangenome analysis using Roary. **Figure 4a** shows that, since the total genes trend line continues to increase together with the number of analyzed genomes, the genome of our SGB possibly is an open one. This can be confirmed also by observing that the amount of unique genes does not reach a plateau with a growing number of analyzed genomes, while the new genes seem to stabilize around ∼100 (**Figure 4b**).
Next, we defined as core genes the ones present in at least 95% of our MAGs and as soft core the ones found in the 90-95% range. **Figure 5** shows the exact proportion of such genes, with the sum of core and soft core ones constituting the 3.371% of the total. Ultimately, we decided to include both categories in the core genome for the subsequent analyses in order to build more reliable phylogenetic trees with more genes to compare against each other.



(a) Amount of conserved and total genes against an increasing number of genomes analyzed. The stair-like behavior of the conserved genes line is caused by an approximation resulting from a low sample size.



(b) Amount of unique and new genes against an increasing number of genomes analyzed. The increasing trend of the unique genes as newer ones are found is explained by having an open pangenome.

Figure 4. Plots resulting from the pangenome analysis, displaying the trend of the number of genes found as more genomes are considered.

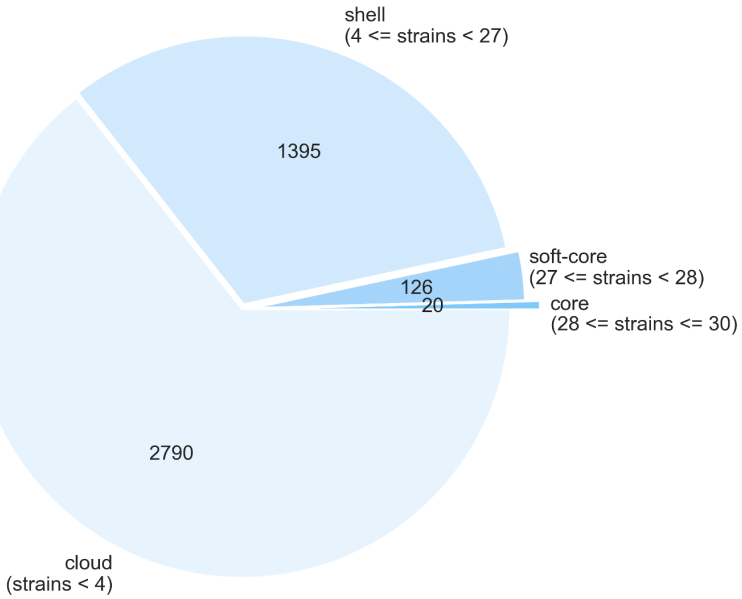
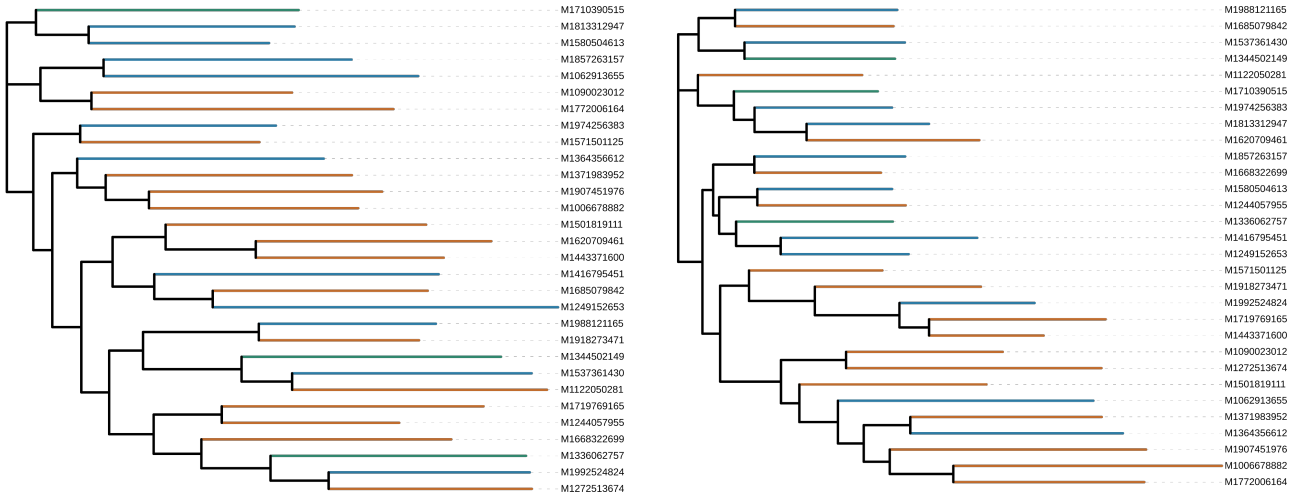


Figure 5. A piechart representing the quantitative categorization of core and accessory genes (shell and cloud). Note that the threshold for defining core and soft-core genes has been modified directly on the *roary_plots.py* Python script in order to be consistent with the previously chosen threshold of 90%.

2.4 Phylogenetic analysis

The two trees generated in the phylogenetic analysis (**Figure 6**) were respectively built using the core genes alignment and the accessory genes presence/absence matrix. Although they share some similarities in terms of branching at the highest levels, we notice a difference between the two trees given that they are created using different gene pools of different lengths as well. In both trees we can observe that the samples do not belong to few macro-clusters but they are split across many smaller clades. The tree created using only the core genes (**Figure 6a**) is more robust due to the fact that it was built by comparing the same genes across our MAGs. For this reason, it was utilized to assess whether our MAGs display a pattern based on the implant study group they belong to. Although we did not detect a clear distinction, there are some instances where samples coming from the same group are phylogenetically close: for example M1813312947 with M1580504613 and M1857263157 with M1062913655 for the peri-implant mucositis condition. Something similar happens for the peri-implantitis batch while the healthy state does not exhibit any clustering.



(a) Phylogenetic tree built using core genes.

(b) Phylogenetic tree built using accessory genes.

Figure 6. Phylogenetic trees visualized with iTOL colored by study group (green: healthy, blue: mucositis, orange: peri-implantitis). MAGs identifiers are shown on the right.

3. Conclusion

To conclude, we assessed that our SGB was assigned to the *Metamycoplasmataceae* bacterial family, also found in the oral microbiome, and that it was mainly composed of medium and high quality MAGs. In particular, as expected, no differences in terms of quality were encountered between MAGs belonging to distinct study groups. The genome annotation analysis showed a consistent proportion of HPs and CDSs across the peri-implant tissue types. On the other hand, studying the phylogenetic trees, highlighted some subclusters of the mucositis and peri-implantitis conditions. Note that the limited number of MAGs (30) could have influenced our findings in terms of statistical significance due to the small sample size. Lastly, our SGB was found to have an open pangenome as the total genes count continues to increase with the number of analyzed genomes.

Acknowledgements

We want to express our appreciation to Professor Nicola Segata for his support during the ”Computational Microbial Genomics” course. A special thank you goes to Francesco Asnicar and Vitor Heidrich for their precious help throughout the development of our project.

Abbreviations

NGS: Next-generation sequencing; **SGB:** Species-level Genome Bin; **MAG:** Metagenome-Assembled Genome; **CDS:** Coding Sequence; **HP:** hypothetical protein; **KP:** known protein.

Supplementary Material

Supplementary material and figures, together with the R and python codes utilized in our analysis can be found in the GitHub repository at the following link: `https://github.com/Sara-Baldinelli/CMG-project/`

References

[1] Priya Nimish Deo and Revati Deshmukh. Oral microbiome: Unveiling the fundamentals. *Journal of oral and maxillofacial pathology*, 23(1):122–128, 2019.

[2] 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.

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[4] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[5] Francesco Asnicar, Andrew Maltez 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.

[6] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014.

[7] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[8] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[9] Andrew J Page, Carla A Cummins, Martin Hunt, Vanessa K Wong, Sandra Reuter, Matthew TG Holden, Maria Fookes, Daniel Falush, Jacqueline A Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693, 2015.

[10] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[11] Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

[12] Ari Löytynoja. Phylogeny-aware alignment with prank. *Multiple sequence alignment methods*, pages 155–170, 2014.

[13] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26(7):1641–1650, 2009.

[14] Ivica Letunic and Peer Bork. Interactive tree of life (itol): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–128, 2007.

[15] Tina Keller-Costa, Lydia Kozma, Sandra G Silva, Rodolfo Toscan, Jorge Gonçalves, Asunción Lago-Lestón, Nikos C Kyrpides, Ulisses Nunes da Rocha, and Rodrigo Costa. Metagenomics-resolved genomics provides novel insights into chitin turnover, metabolic specialization, and niche partitioning in the octocoral microbiome. *Microbiome*, 10(1):151, 2022.

[16] Mitchell Balish, Assunta Bertaccini, Alain Blanchard, Daniel Brown, Glenn Browning, Victoria Chalker, Joachim Frey, Gail Gasparich, Ludwig Hoelzle, Tom Knight Jr, et al. Recommended rejection of the names malacoplasma gen. nov., mesomycoplasma gen. nov., metamycoplasma gen. nov., metamycoplasmataceae fam. nov., mycoplasmoidaceae fam. nov., mycoplasmoidales ord. nov., mycoplasmoides gen. nov., mycoplasmopsis gen. nov.[gupta, sawnani, adeolu, alnajar and oren 2018] and all proposed species comb. nov. placed therein. *International journal of systematic and evolutionary microbiology*, 69(11):3650–3653, 2019.