Computational Microbial Genomics

07/04/2024

Introduction

The world of microbes is characterized by organisms with size ranging from nanometers to micrometers, but yet are powerful enough to influence our planet.

Nowadays, they are known for multiple contributions, noticeably in breaking down waste products of living organisms, adding fertility to soil, taking up CO_2 converting it into organic products [7]. With increasing studies and better sequencing technologies, a deeper understanding of microbial evolution and interaction with the environment will be the outcome.

Recently, high-throughput sequencing technologies and a suite of computational pipelines have been combined into the shotgun metagenomics method that transformed microbiology, making it easier to catalogue members and to understand how communities of microbes function.

This method starts with the the collection, processing and sequencing of the samples. The products of this step are lots of short reads that are consequently broken into overlapping k-mers and assembled in contigs. At this point metagenomic assemblies are comprising thousands of contigs from different species. The last step is contig 'binning', that groups the contigs trying to divide the species, creating a MAG (Metagenomic-assembled genome). [13]

In this paper, we propose an analysis of 30 MAGs, obtained from three groups of patients:

- healthy patients;
- patients suffering from peri-implantitis;
- patients suffering from mucositis.

In the dataset are collected health condition, sex, BMI, age, and smoking status of the patients. Our goal is to find more information about the bacteria present in the implant site of these patients, and investigate whether differences in health conditions are associated with different strains. To do so, we employed computational tools such as CheckM for checking the quality of MAGs, PhyloPlan for taxonomic assignment, Prokka for genome annotation, Roary for prokaryote pan genome analysis, and FastTree for phylogenetic analysis.

Methods

Quality checking

CheckM is a tool used to assess the quality of microbial genomes, mainly via the two parameters of completeness and contamination [12]. It relies on a set of single copy genes that are core genes for the taxonomic level of interest (in the taxonomy workflow option), and are hence used as reference markers that are expected to be found in the metagenome-assembled genomes.

The parameter of completeness measures the percentage of MAG that presents the expected marker genes, and contamination measures the unexpected finding of multiple copies of the theoretically single copy gene markers inside the same MAG. High contamination signals that the binning may have grouped different species into one MAG.

To measure whether this wrongly grouped species are at least closely related, i. e. from how far phylogenetically the contamination is coming from, strain heterogeneity is calculated too.

The taxonomy workflow was selected to run under the domain of Bacteria, using 104 marker genes (divided into 58 marker sets) but more specific runs, inspecting some genus and families of taxonomy, were run in further analyses.

The alternative lineage workflow provided by checkm in the command options is mostly used to discriminate genomes in phylogenetic groups.

Taxonomic assignment

PhyloPhlAn is a method for large-scale microbial genome characterization and phylogenetic analysis at multiple levels of resolution [17, 3]. This tool can assign genomes to species-level genome bins (SGBs) and reconstruct strain-level phylogenies using clade-specific maximally informative phylogenetic markers [2].

In particular, species labels are assigned to consistently clustered genomes by majority voting. An input genome is assigned to an SGB (and its associated taxonomy, if any) if the Mash average distance to the genomes in the bin is below 5%, as this threshold has been suggested to be optimal for species definition[3]. If a genome results far from every SGB, we are discovering new diversity.

The SGBs we employed were 11, and just some of them were known (at least one isolate). We chose to display just the 3 closest SGBs identified by setting the parameter -n in the phylophlan command to 3.

Genome annotation

With the aim of annotating some relevant bacterial genomic features we employed Prokka, a tool that relies on external feature prediction tools to put together the information needed to identify features of interest in a set of genomic DNA sequences [16]. In particular, Prokka allows to retrieve information regarding coding sequences, ribosomal RNA genes, transfer RNA genes, signal leader peptides, and other non-coding sequences. The expected input consists just of the preassembled genomic DNA sequences in FASTA format, and the output is a series of files containing the annotations.

Pangenome analysis

The goal of pangenome analysis methods is to characterize the distribution of genes across different genomes, identifying the cardinality of the union of the set of genes of each genome (*pangenome*) and of their intersection (*core genome*). We set the threshold to identify the core and *accessory genes* (pangenome - core) to 90% of presence, meaning that a gene is considered as core gene if present in at least 27 genomes out of 30. Accessory genes are further discriminated in *cloud genes*, which are unique for one strain (here 3.33% of presence), and the *shell genes* which make up the rest of the accessory.

We input genomic feature file previously generated by Prokka to Roary [11], with the parameter -i to set the percentage of identity to 95.

Phylogenetic analysis

We dive into how strains are related in term of evolutionary by doing multiple sequence alignment on core genes and contruct a phylogenetic tree. We take the .gff file from prokka, telling the program that the genes should present in 90% of the strains to be core genes, then run multiple sequence alignment with MAFFT and build the tree using FastTreeMP.

Host metadata association

To associate with host metadata, we decided to find bacterial genes associated with certain groups of patients to reason on the correlation.

Results

Quality checking

The MAGs showed an acceptable quality when checked using a taxonomy workflow. In particular, the high quality thresholds of more than 90% completeness (*near completeness*) and less than 5% contamination (*low*) are satisfied by 17 of the 30 genomes. All the remaining are more than 50% complete (*moderate completeness*) and at most 8% contaminated (*medium*), falling in the category of medium quality. Therefore, we decide to keep them all for further analysis.

The GC content is around 40% percent with a low variance, and consistent among different MAGs, this could be explained by the binning process which considers similar GC content, differential coverage and k-mer frequency to assign the contigs to MAGs. Since similar genome size is expected among the strains, variations in size possibly reflect a loss of information during the sequencing. Lower completeness is indeed observed in 3 out of the 4 strains with the lowest genome size, validating this hypothesis. The remaining low quality MAGs are associated with short contigs size or a low N50 that signals a poor continuity of the assembled genomes. We notice that one of the 4 genomes associated to healthy patients has a low quality, lowering the significance of this already small control group, anyway all these 4 have a null strain heterogeneity.

Taxonomic assignment

All the metagenome-assembled genomes are classified into the same unknown SGB, from which they result less than 5% distant, obtaining a taxonomic classification with precision up to the phylum level of Bacteroidota. The second less distant SGBs assign 3 genomes to the Filifactor alocis species (family Peptostreptococcaceae) and 4 genomes to the Prevotella pleuritidis one (family Prevotellaceae), but the distances ranging from 20% to 50% make these assignments almost random. The remaining genomes are assigned to the Bacteroidota phylum of another uSGB.

Genome annotation

The output of the genome annotation is 12 files containing information about coding or non-coding sequences, known genes and their corresponding proteins and information whether the proteins are known or only hypothetical. In the following table are summarized these information, extracted from the .txt and .tsv files.

MAG	# CDS	# hypothetical proteins	# known proteins
M1008713361	1247	560	2012
M1009573584	1469	710	2308
M1056017853	1253	704	1850
M1081599018	1406	677	2210
M1129475058	1390	691	2166
M1136986178	1423	674	2248
M1155483984	1270	716	1863
M1156291224	1392	668	2197
M1202011224	1513	812	2297
M1222374074	1592	784	2483
M1227789332	1515	732	2369
M1237908033	1521	732	2400
M1251207731	1520	726	2399
M1251959113	1333	623	2124
M1351607188	1190	619	1813
M1374897024	1253	577	2013
M1403268331	696	305	1122
M1560202640	1483	691	2355
M1604650000	1316	609	2089
M1690387012	1024	499	1612
M1690599824	1560	768	2440
M1696505562	1720	901	2624
M1718303558	1516	781	2336
M1778821642	1451	673	2318
M1789695349	1779	956	2687
M1807791478	1466	677	2338
M1814436263	1052	565	1582
M1937366822	1586	816	2438
M1940574871	1467	710	2307
M1980221194	1522	707	2420

Pangenome analysis

As shown in Fig. 1, output of the roary command, the size of the pangenome is 6213 genes, composed by: 547 genes considered as core genes, 1443 shell genes, and 4223 cloud genes, where cloud and shell are a partition of accessory genes. The pangenome of these 30 MAGs can be considered as an open pangenome, as only the 8.8% of the genes is part of the core genome. In Fig. 2, the number of genes keep increasing when more genomes are added and does not seem to reach an asymptotic behaviour. In Fig. 3, 3 main branches suggest 8 strain groups sharing similar presence or absence of different accessory genes.

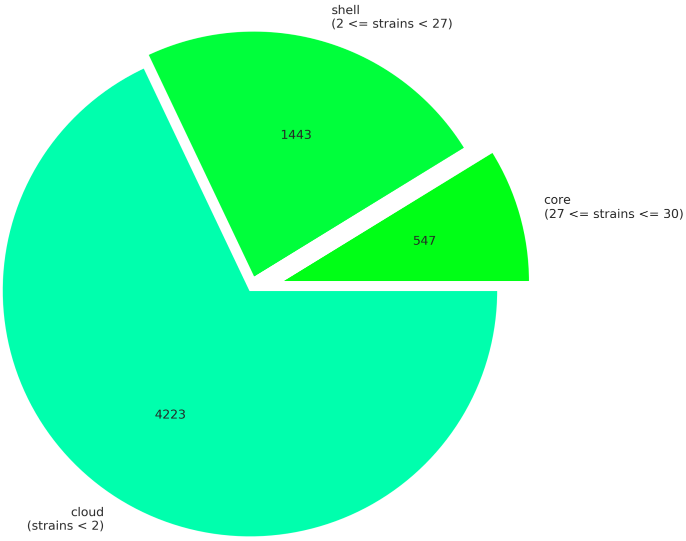


Figure 1: Pie chart of the pangenome

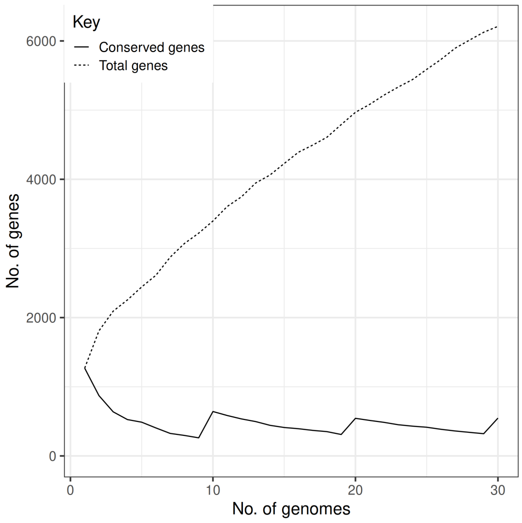


Figure 2: Pangenome vs Core genes

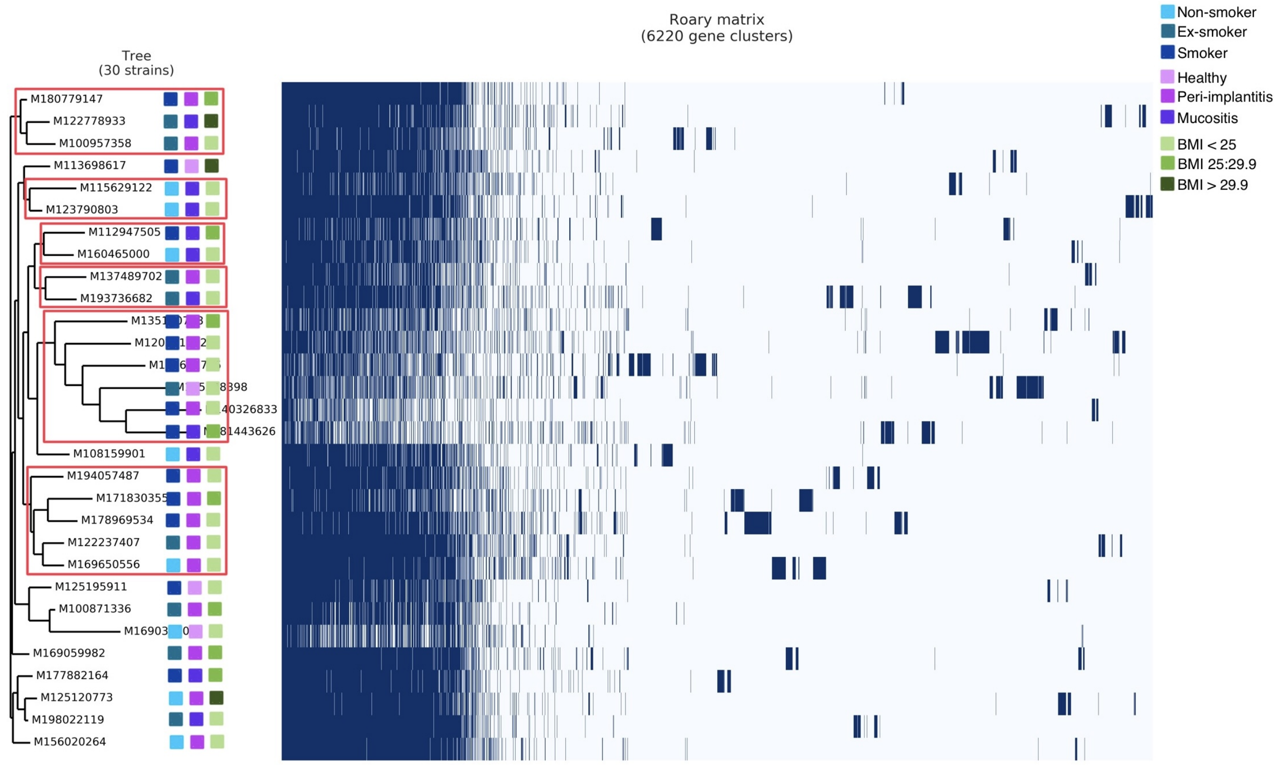


Figure 3: Accessory genes tree with presence/absence matrix

Phylogenetic analysis

The constructed phylogenetic tree are shown in Fig.4. In comparison with the tree constructed from accessory genes, the different order of clades suggests there may not be a strong connection between core and accessory genes.

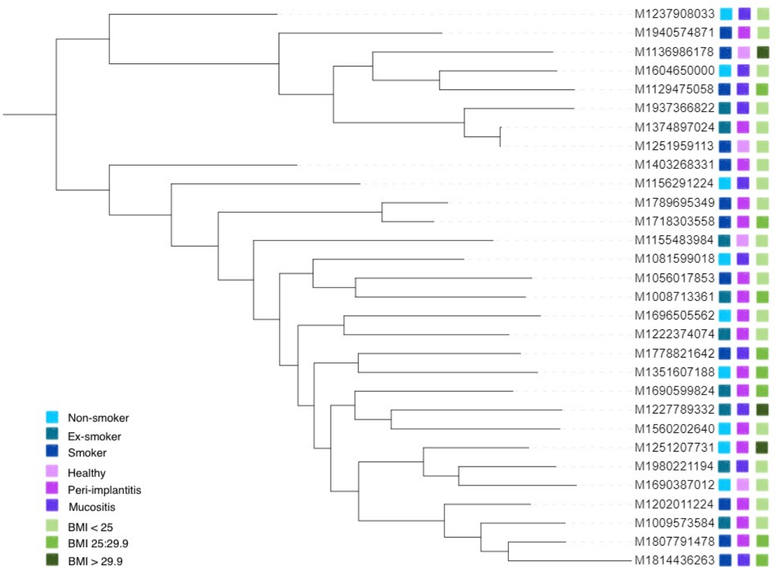


Figure 4: Tree of Core genes

Association with host metadata

In this analysis we found genes that are only present in mucositis and peri-implantitis but not in healthy patients. Similarly, some genes are present in ex-smokers and current smokers but not in non smokers.

For the disease groups, 2141 genes were found from healthy strains, 3327 in mucositis, and 4522 in peri-implantitis. Among them, 1677 bacterial genes were in mucositis but not in healthy samples, with 425 of them known and 1252 hypothetical proteins.
We also found 2809 bacterial genes in peri-implantitis samples not present in healthy ones, with 584 known genes and 2225 hypothetical.
In both cases, many genes are linked with ribosomal protein, CRISPR-associated, and transposate. Additionally, peri-implantitis samples count more genes regulating transcription, cell division, DNA replication and repair are found. But also genes related to the amino sugar, nucleotide sugar metabolism pathway, antibiotic resistance and cell detoxification.

Regarding smoking state, 4078 genes were found in strains from smoking patients, 3188 from ex-smoking, and 2899 from non-smoking. Among them, 2212 bacterial genes present in strains from smoking patients but not found in non-smoking, of which 520 known and 1692 are hypothetical.
13344 bacterial genes found in strains from ex-smoking patients and not in non-smoking ones, 297 are known and 1047 are hypothetical.
In general, both cases share similarity in having genes for CRISPR-associated, transpotase, and tyrosine recombinase. However, in the case of smoking, 2 genes for DNA mismatches repair were found in smoking, and more genes related to antibiotic resistance and detoxification were found in smoking (8 genes) in comparison with ex-smoking (6 genes). We hypothesize that since keeping more genes is energy and resource costly, these genes may have special functions of interest in our associations.

Discussion

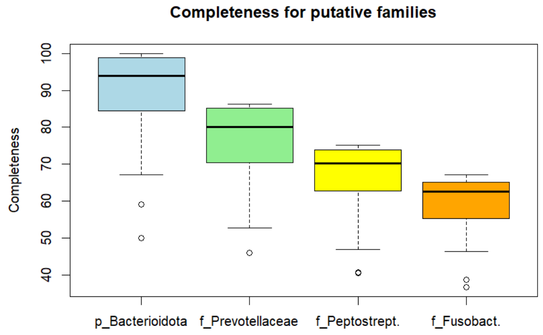
Quality checking and taxonomic assingnment

We have reason to trust our classification of all the MAGs in the Bacteroidota phylum: this is the closest assignation, even below a distance of 3% for most of the genomes; the same phylum is found in 27 of the assignations to the second closest SGB: only the 3 MAGs assigned to Filifactor are classified under the different phylum of Firmicutes; and we know that our data was selected running a QC with the lineage workflow in order not to obtain too divergent genomes. The third closest SGBs are useless since they report a distance of 90%.

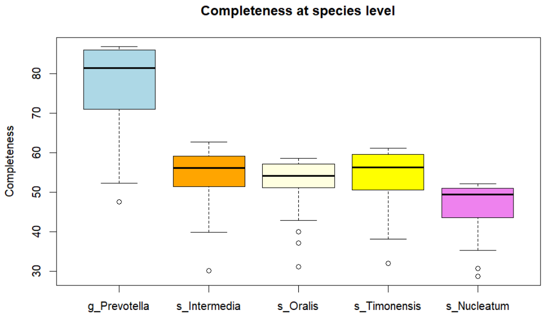
We note that most of the genomes resulting in a bigger strain heterogeneity score in the QC, fall in the category that is assigned to an unknown SGB even when looking at the second closest SGB. This must be considered when taking into account the possibility of having found an unknown species.

The identification of Prevotella pleuritidis and Filifactor alocis must be interpreted as a possible similarity of our genomes to these families, taking into account a relevant margin of diversity shown by the >20% distances. These two species have in common an anaerobic metabolism, suggesting this could be a feature shared by our strain too. To check this hypothesis the origin of our sample should be considered, since anaerobic bacteria are indeed common in the subgingival microbiome (oxygen availability - an environmental factor - may be a major driver of this community composition [18]). It is possible that our MAGs were obtained from subgingival samples, since the stability of that microbiome make it a good indicator of periodontal diseases [19], and this aspect should be further evaluated.

Having two putative families of species similar to ours, we were motivated to perform other quality checking in the same workflow option, with more specific levels of taxonomy. At first we ran checkm specifying three families: Prevotellaceae (for P. pleuritidis), Peptostreptococcaceae (for F. alocis) and Fusobacteriaceae. The results for Prevotellaceae showed worse but still acceptable medium quality for more than 90% of the genomes, and were better than that for the other two families, which served as a control (Fusobacteriaceae are common in same conditions [18] [19]). This is reasonable since many Prevotella species are common residents on various surfaces of the mouth. Exploring in this direction we tried to find some similarities with different Prevotellaceae strains either known to be pathogenic and found in oral samples (P. intermedia, P. oralis [9]) or to be similar to P. pleuritidis (P. timonensis, P. oralis [14]). These results show an expected low completeness and still a medium contamination for most of the MAGs.



(a) Comparison of different completeness results for quality checking regarding Bacteroidota phylum (reference) and Prevotellaceae, Peptostreptococcaceae and Fusobacteriaceae putative families.



(b) Comparison of different completeness results for quality checking regarding Prevotella genus (reference) and P. intermedia, P. oralis, P. timonensis and F. nucleatum putative species.

Genome annotation

The total number of proteins seems to double the number of coding sequences. This may be explained by overlapping genes, different genes located in similar regions but slightly different in transcription starting sites.

Pangenome analysis

In Fig. 3, some clusters presenting some pattern are highlighted in red. For example, the first cluster shows 3 people who smoke or have smoked and have the disease, or the second cluster consists of 2 identical patient situation. In the 5th cluster, the bigger one, they have some evident similarity in the presence/absence matrix, differing a lot from all the others. Moreover, 83% of them smoke and the 66% has the peri-implantitis. In the last highlighted cluster, all the patients taken in account have peri-implantitis, showing a strong difference between this disease and mucositis.
The bacteria have an open pangenome, so they have a high rate of horizontal gene transfer that makes them gain a lot of accessory genes. As a result, they are more complex and can quickly adapt to multiple environments [4]. This can be further supported by comparing the accessory genes tree versus core genes tree. The accessory tree is branching a lot at the very starting point. The first split of this tree gives 3 branches, the second split 6 and the third 10, while the core tree first splits in 2 branches, then in 4 branches and the third time in 6 branches.

Phylogenetic analysis

In the first branch the percentage of smokers or ex-smokers is 75%, this can be correlated to the fact that also the 50% of the mucositis are in this small group of patients. In the bigger branch, that contains the 73% of patients, 63% of them have peri-implantitis. The 87.5% of overweight people are in this second group, showing a probable connection between this two features. This two group of core genes can be considered as different core genomes based on SNPs and SNVs. In conclusion, smoking and weight can be related to some variation of some bp in the core genes. Anyway we note that this tree, built via multiple alignment, should be built on a relevant number of core genes to be trusted. Our open pangenome may be a limitation in this sense.

Association with hostmetadata

We did not expect the number of genes associated to the different groups to differ much, because the bacteria all belong to the same species. One way to explain the difference is the open pangenome and an imbalance in the number of sample in each group. Open pangenome may also explain CRISPR-associated protein and transposate found, since these facilitate the bacteria flexiblity to accumulate new genes from the environment.

From the clusters in accessory tree, we found some association between the disease and smoking. These groups also share similar genetic characteristics regarding DNA repairing, antibiotic resistance and cell detoxification genes. DNA mismatch repair protein MutL and MutS were found from both peri-implantitis and smoking cases. These two proteins work together to recognize DNA base mismatches in replication that are missed by polymerase proofreading and fix it [6, 5]. Additionally, multiple genes related to the class of multidrug efflux/export ATP-binding protein and MepA were found in peri-implantitis, smoking, and ex-smoking. These are transport proteins present in the cell membrane to expel harmful substances [8, 10]. MepA belongs to the multidrug and toxin extrusion (MATE) family and is found in Gram-positive bacteria [15]. The multidrug efflux/export ATP-binding protein belongs to bacterial ATP-binding cassette (ABC) transporters, and they obtained energy from ATP binding and hydrolysis to function [1]. The presence of DNA repair and toxic removal genes indicate an adaptive mechanism of bacterial strains under host conditions.

Future perspective

Our analysis encountered some difficulties due to limits in the number of samples. This have prevented us to narrow down the scope of our search for bacterial genes associated with particular host conditions. Therefore, we suggest more work from sampling and comparison of bacterial genes among different patient groups for future perspective. This could help us to gain more understanding on the evolution of the strains and their interactions with host environment.

Conclusion

In conclusion, we found unknown bacterial species with a pangenome size of 6213 and which may be related to the Prevotellaceae family, with no further taxonomic resolution. The open pangenome makes the species individuals likely to be flexible in adding and removing accessory genes from their genome in coping with different host conditions. In more hostile host environment such as smoking, they develop adaptive mechanism in increasing DNA repair and cell detoxification genes. As a result, some of them may acquire pathogenicity and be associated with the host peri-implant disease. We suggest a deeper study with more samples and comparisons based on patient groups and clusters to confirm the link between host conditions and pathogenic strains.

Bibliography

[1] Irshad Ahmad, Nighat Nawaz, Fatemeh Karimi Dermani, Alisa Khodadadi Kohlan, Massoud Saidijam, and Simon G Patching. Bacterial multidrug efflux proteins: A major mechanism of antimicrobial resistance. *Curr. Drug Targets*, 19:1–13, 2018.

[2] et al. Asnicar F. Phylophlan 3.0. https://segatalab.github.io/tools/phylophlan/.

[3] et al. Asnicar F. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using phylophlan 3.0. *Nature Communications*, 2020.

[4] Michael A Brockhurst, Ellie Harrison, James PJ Hall, Thomas Richards, Alan McNally, and Craig MacLean. The ecology and evolution of pangenomes. *Current Biology*, 29(20):R1094–R1103, 2019.

[5] Peter Friedhoff, Pingping Li, and Julia Gotthardt. Protein-protein interactions in dna mismatch repair. *DNA repair*, 38:50–57, 2016.

[6] Christopher M Furman, Ryan Elbashir, and Eric Alani. Expanded roles for the mutl family of dna mismatch repair proteins. *Yeast*, 38(1):39–53, 2021.

[7] Ankit Gupta, Rasna Gupta, and Ram Lakhan Singh. Microbes and environment. *Principles and applications of environmental biotechnology for a sustainable future*, pages 43–84, 2017.

[8] Lulu Huang, Cuirong Wu, Haijiao Gao, Chao Xu, Menghong Dai, Lingli Huang, Haihong Hao, Xu Wang, and Guyue Cheng. Bacterial multidrug efflux pumps at the frontline of antimicrobial resistance: An overview. *Antibiotics*, 11(4):520, 2022.

[9] et al. Könönen E, Fteita D. Prevotella species as oral residents and infectious agents with potential impact on systemic conditions. *Journal of oral microbiology*, 2022.

[10] Kunihiko Nishino, Seiji Yamasaki, Ryosuke Nakashima, Martijn Zwama, and Mitsuko Hayashi-Nishino. Function and inhibitory mechanisms of multidrug efflux pumps. *Frontiers in Microbiology*, 12:737288, 2021.

[11] 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.

[12] 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.

[13] Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata. Shotgun metagenomics, from sampling to analysis. *Nature biotechnology*, 35(9):833–844, 2017.

[14] Mitsuo Sakamoto, Hidefumi Kumada, Nobushiro Hamada, Yusuke Takahashi, Masaaki Okamoto, Mohammad Bakir, and Yoshimi Benno. Prevotella falsenii sp. nov., a prevotella intermedia-like organism isolated from monkey dental plaque. *International journal of systematic and evolutionary microbiology*, 59, 2009.

[15] Bryan D Schindler and Glenn W Kaatz. Multidrug efflux pumps of gram-positive bacteria. *Drug Resistance Updates*, 27:1–13, 2016.

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

[17] N Segata, D Börnigen, XC Morgan, and C Huttenhower. Phylophlan is a new method for improved phylogenetic and taxonomic placement of microbes, 2013.

[18] et al. Segata N. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome biology*, 2012.

[19] et al. Tamashiro, Strange. Stability of healthy subgingival microbiome across space and time. *Scientific Reports*, 2021.