BIOINFORMATICS INSTITUTE

Oral cavity metagenomics, or a friendly reminder to floss daily

Amina Ibragimova[1, 2 *] and Nikita Vaulin[1, 2, * †]

[1]Bioinformatics Institute, Saint-Petersburg, Russia
[2]Skolkovo Institute of Science and Technology, Moscow, Russia

[*]Contributed equally.
[†]vaulin@ro.ru

Abstract

Dental problems are an enduring cause of pain from ancient times to the present day. Microbial communities that were intact in dental calculus for more than 1000 years were extracted and analysed using both 16S rRNA sequencing and shotgun whole genome sequencing approaches. In this study we found that some species tend to appear more frequently in diseased dental samples and thus associated with diseases. We list such species, some of which are among the elusive three, the Red Complex, which is wanted throughout the Wild West for attacks on teeth. We also list the genetic changes that have occurred over 1000 years in one such species. Taking together, there is no hope that our results can be useful in solving the problem of dental diseases.

Key words: metagenomics, 16S rRNA, WGS

Introduction

Identifying the microbial taxon present in complex biological samples is one of the oldest challenges in microbiology. Metagenomics sequencing allows scientists to directly interrogate the community composition in an unbiased manner, enabling more rapid species detection and the discovery of novel species [1].

NGS sequencing is a method of choice to assess a microbial community composition of biological samples. It can be achieved by two alternative approaches: sequencing of conserved 16S ribosomal RNA (rRNA) gene or whole genome shotgun sequencing [2] .

The first approach is based on PCR amplification with primers specific to highly conserved regions of the 16S gene. [3] the annotation is based on putative association of the 16S rRNA gene with a taxa defined as an operational taxonomic unit (OTU). In general, OTUs are analyzed at the phyla or genera level, and can be less precise at the species level. In addition, specific genes are not directly sequenced, but rather predicted based on the OTUs. Due to horizontal gene transfer and the existence of numerous bacterial strains [4], the lack of direct gene identification potentially limits understanding of a microbiome [5].

An alternative approach is shotgun sequencing which utilizes random primers for sequencing overlapping regions of a whole genome. Though this method is more expensive and requires more extensive data analysis, the major advantage is that the taxa can be more accurately defined at the species level [5].

In the present study, we analyzed the samples of dental calculus – a cover on the dental surface – extracted from human skulls dug out on a monastic site in Dunheim, Germany in 1990. Dental calculus preserved DNA and thus provided scientists with the opportunity to explore the microbial community that was intact for more than a thousand years.

Figure 1. Ancient teeth archaeological excavations [6]

Compiled on: March 31, 2023.
Don't forget to rest and take care of yourself.

Materials and methods

Dental metagenomic data

The dental material was collected in the monastic site in Dunheim, Germany (Fig. 2 [7]). DNA samples extracted form the dental root and calculus were sequenced using V5 16S rRNA and shotgun sequencing with single-end Roche 454 technique. The raw data is accessible *via* the NCBI SRA under number SRP029257 (BioProject PRJNA216965). The data used in this study contains 9 samples from 4 individuals (B17, B61, B78 and G12, see sample metadata in Supplemental materials). For each individual samples were collected both from the root and the calculus. Individuals B17 and B78 carried periodontal disease while B61 and G12 were not.

Figure 2. Dalheim Church of St. Peter and women's monastery location, adapted from [7]

V5 16S rRNA amplicon sequencing data processing

16S rRNA amplicon sequencing data were analysed with the Qiime2 v. 2023.2.0 pipeline [8]. The *import* function from *qiime tools* were used to convert the data from fastq to qza format. For qiime qzv reports the web-tool Qiime View was used for data visualisation. Raw qza data quality control was performed with *demux summarize* and *tools validate* functions.

For the feature table construction the *denoise-single* function from the DADA2 pipeline (as a part of Qiime2 alongside with *feature-table summarize* and *feature-table tabulate-seqs* functions) was used [9]. The adapter length of 35 bp and amplicon size of 140 bp were accounted during the procedure. To get the DADA2 step analysis statistics *qiime metadata tabulate* function was used.

Taxonomic analysis

For the taxonomic analysis, firstly, the Qiime2 Naive Bayes classifiers with Greengenes2 based on Scikit-learn library were used [10, 11, 12]. For visualisation *taxa barplot* function was used.

Secondly, the R-based web-tool MicrobiomeAnalyst in *Marker data profiling* mode with default filtering parameters was used [13, 14]. Prior the analysis ASV data were converted to the BIOM format *via* the *tools export* function [15]. To compare species families abundances in different samples *Single-factor Statistical Comparisons* tool was used with Mann-Whithey/Kruskal-Wallis test at FDR level of 0.05. To evaluate α-diversity on a species level Mann-Whithey/Kruskal-Wallis test with Shannon distance method was used. To evaluate β-diversity on a species level PCA based on PERMANOVA with Bray-Curtis distance method was used. Species heatmap was built with Ward method using Euclidean distance. The correlation analysis to build the network was performed *via* SparCC method with 200 permutations at the *p-value* threshold of 0.1 and correlation threshold of 0.2.

Shotgun sequencing data analysis

For more detailed view on the dental metagenome species evolution we used a G12 individual (which was affected by the periodontitis disease). We were kindly provided with the G12 metagenome assembly and its Metaphlan profiling by Mike Raiko [16]. The Metaphlan results were visualised with the Pavian web-tool [17].

To address the question how the evolution might have gone in this case, we aligned the readings of the G12 sample to the genome of the modern *Tannerella forsythia* (NCBI reference accession $NC_0 16610.1$) using the samtools and bwa packages [18, 19]. To identify newly emergent regions we intersected the alignment in substractive mode with the known modern *Tannerella forsythia* annotation using the bedtools machinery [20].

Results

Data quality control and processing

The initial 15S rRNA 454 sequencing data contained 9 samples with from 4000 to 6000 reads per each sample (see Fig. 3a, Table 1) - which may be not so high for the metagenomic study. After the filtering and quality control procedures around 93% of sequences were retained. After the feature table construction procedure 465 features in total were assigned to the samples (Fig. 4).

Table 1. Number of reads analysed in this study.

Sample ID	Input	Filtered	Denoised	Non-chimeric	% of remained
S10-V5-Q-B61-calc	5957	5695	5575	5559	93.32
S14-V5-P-B17-calc	4491	4282	4218	4218	93.92
S15-V5-R-B78-calc	4212	4037	3950	3911	92.85
S16S17-V5-K1-G12-root	5788	5599	5427	5361	92.62
S16S17-V5-K2-G12-root	5272	5066	4884	4862	92.22
S18S19-V5-L-B17-root	4955	4790	4680	4680	94.45
S20S21-V5-M-B61-root	5516	5396	5240	5182	93.94
S22S23-V5-N-B78-root	4695	4534	4382	4261	90.76
S8-V5-O-G12-calc	5362	5199	5108	5092	94.96

Taxonomic analysis

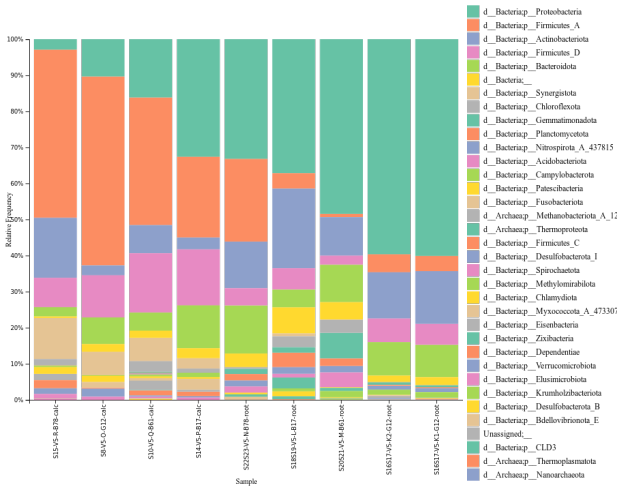
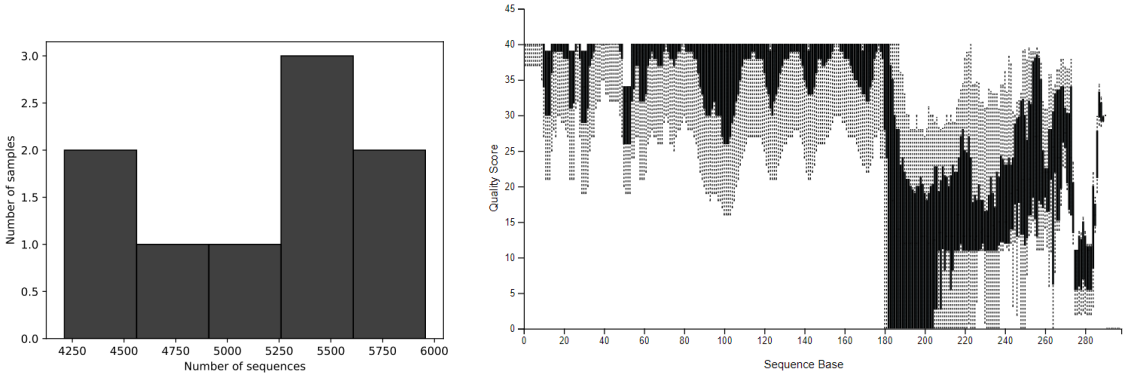


Figure 5. Species classes abundances

Most of the features found in the sample were assigned to the bacteria domain (more then 95% relative frequency). At the Fig. 10 one



(a) Initial number of reads per sample **(b)** Initial reads per base quality score

Figure 3. Raw 16S rRNA 454 reads statistics

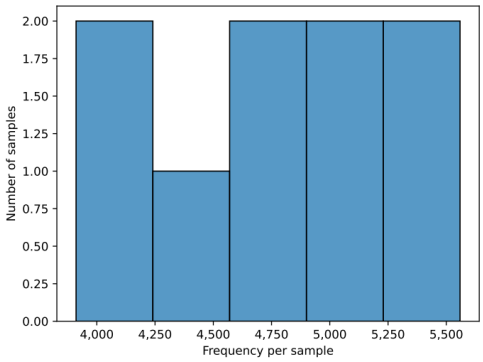


Figure 4. Number of features per sample

can observe different species abundances at a class level. Most of the features found in the root samples (the last 4 samples at Fig. 10) were found to be referred to the Proteobacteria. At the same time, for the calculus samples (the first 4 samples at Fig. 5) most of the features found to be referred to the Firmicutes A. For both types of sample also the Actinobacteriota, Firmicutes B and Bacteriodota were of the high abundance. Knowing the fact that individuals B17 and B78 carried periodontal disease while B61 and G12 were not it is still hard to identify any differences at that level. The more detailed abundances profiles (for the order, family and species level the same results can be found in the *qiime_taxonomy* folder in Supplemental materials and at Fig. 10).

For the more detailed analysis the MicrobiomeAnalyst web-tool was used. Firstly, the α- and β diversity analysis was performed for both sample type (root or calculus) and disease comparisons (Fig. 6 and Fig. 8 correspondingly). For the α-diversity we have the Shannon index values from 2.5 to 3.3 regardless of the type of sample (Fig. 6, left). For the disease status comparisons one could observe some increase of α-diversity (Fig. 6, right), but the observed differences are not such high. At the same time, for the β diversity it is impossible to differentiate between disease statuses (Fig. ??, right), but the different dental samples clearly distinct.

For a more detailed analysis of species representation in different types of samples, we constructed a heatmap with clustering by species 8a. According to the heatmap, on average, species can be found in both healthy or diseased samples. There are key differences in fact between the root or calculus samples, and there is also a strong differences can observed simply between individuals. Nevertheless, it is possible to identify some species that occur in calculus of diseased individuals but do not occur in a healthy ones or vice versa. These include: *Pseudoramibacter alactolyticus*, RUG574 and RBG_16_71_46 entries as disease-present and *Selemonas sputigena*, *Panacagrimonas perspica* as disease-absent.

We also analysed a taxonimic content with the SparCC correlation method 8b. The microbial disbyosis index for health/diseased comparison is equal to -0.83. The species mostly present in diseased sample are: *Povalibacter uvarum*, *Pseudoramibacter alactolyticus*, *Arenibacter certesii*, *Eubacterium N sphenum*.

Along with a broad analysis of the presence of species, we decided to elaborate on some of them. It is interesting to found some species that tend to occur together. For example, we found a pair of *Fretibacterium fastidiosum* and *Flexilinea sp902786265* that can be found only in healthy individuals material and their median relative abundance is 0.9888. We also investigated the well-known Red complex which found to be strongly associated with the periodontal disease: *Porphyromonas gingivalis*, *Tannerella forsythia* and *Treponema denticola*. Unfortunately, we did not find these particular species in our samples. Thus, we decided to turn our attention to other members of families Porphyromonadaceae and Treponemataceae. There was two of their representatives: *Treponema C maltophilum* and *Porphyromonas A 859423 pasteri*. They were also found to co-occur together with the median relative abundance of 0.9633. However, they can be found in both healthy and diseased samples, which gives no much reason to consider them the cause for the disease.

Metagenome assembly profiling

The G12 affected by the periodontitis disease individuals metagenome was shotgun-sequenced, assembled and profiled with Metaphlan. The results were visualised with the Pavian web-tool and the species abundancies shankey plot is presented at the Fig. 9. It is interesting, for example, to observe some species in the resulting assembly. Here we can see a *Actinomyces dentalis*, and even *Tannarella forsythia* and *Treponema denticola*. The last two are the representatives of the Red complex which we were unable to identify with the 16S rRNA sequencing.

Newly emergent regions identification

To address the question how the evolution might have gone in this case, we aligned the readings of the G12 sample to the genome of the modern *Tannerella forsythia* (NCBI reference accession $NC_016610.1$)Newly emergent regions were identified *via* the intersection of the alignment in substractive mode with the known modern *Tannerella forsythia* annotation using the bedtools machinery [20]. The full list of the identified region can be found in the *new_genes.gff3* file in *assembly* folder in Supplemental materials. The most interesting found ones are discussed in the Discussion section.

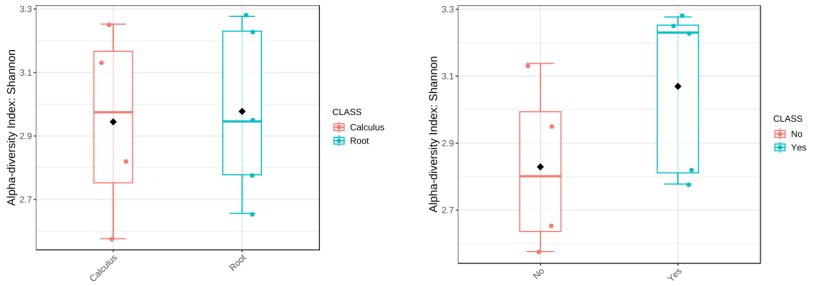


Figure 6. Species α-diversity

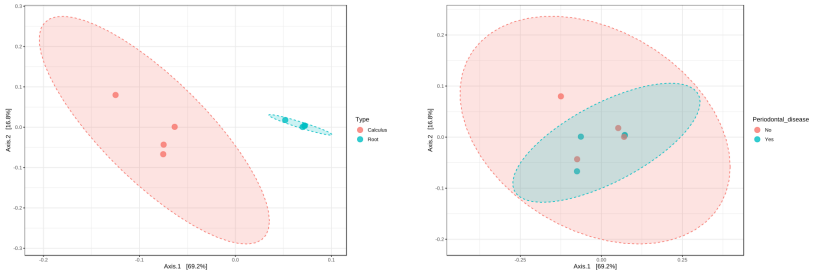


Figure 7. Species β-diversity

Discussion

Taxonomic vagueness of the periodontal disease

In this study we investigated dental metagenome sequencing samples of several men died approximately 1000 years ago. The dental material was collected in the monastic site in Dunheim, Germany (Fig. 2 [7]). The very first conclusion to be stated – the between-disease-status difference are much much less then the differences between different types of material (dental root and calculus) and even less then the differences between individuals. This statement is supported by the $\alpha-$ and $\beta-$ diversity profiles and by the actual species abundances. By the way, the interesting point to mention here - the $\alpha-$diversity for the diseased individuals are slightly higher then for the healthy ones. Taken together it might be the example of well-known **Anna Karenina principle**: «All happy families are alike; each unhappy family is unhappy in its own way» [21].

The Red complex, where are you?

We tried to find the well-known Red complex which found to be strongly associated with the periodontal disease: *Porphyromonas gingivalis*, *Tannerella forsythia* and *Treponema denticola* [22].

We found two of the three members to be with the high abundance in the G12 individual (the one that was affected by the periodontitis disease) metagenome assembly: *Tannarella forsythia* and *Treponema denticola* (Fig. 9). As a result, we can accurately conclude that 1000 years ago periodontal disease was caused by the same bacteria that we can find now in our mouth.

Unfortunately, we did not find these particular species in our 16S rRNA amplicon sequencing samples. Thus, we decided to turn our attention to other members of families Porphyromonadaceae and Treponemataceae. There was two of their representatives: *Treponema C maltophilum* and *Porphyromonas A 859423 pasteri*. They were also found to co-occur together with the median relative abundance of 0.9633. However, they can be found in both healthy and diseased samples.

These results show that despite the possible presence of other bacterial complexes, it is the Red Complex that is apparently undetectable in our data. In fact, one of the reasons may be the technical limitations of our method. During the study, we were unable to run the Silva classifier and then we decided to use the Naive Bayes Greengenes classifier. Additional research with more computer power may reveal the mystery of the Red complex. Another possible reason for this could be that the periodontitis has not yet had enough time to develop in the selected samples. This situation was met and discussed by the Marotz C. et al. [23]. This may be quite expected for the times when humans died of unnatural causes much earlier. In today's world, the causes for dental disease are growing like yeast (literally) and are more common in older people with serious diagnoses.

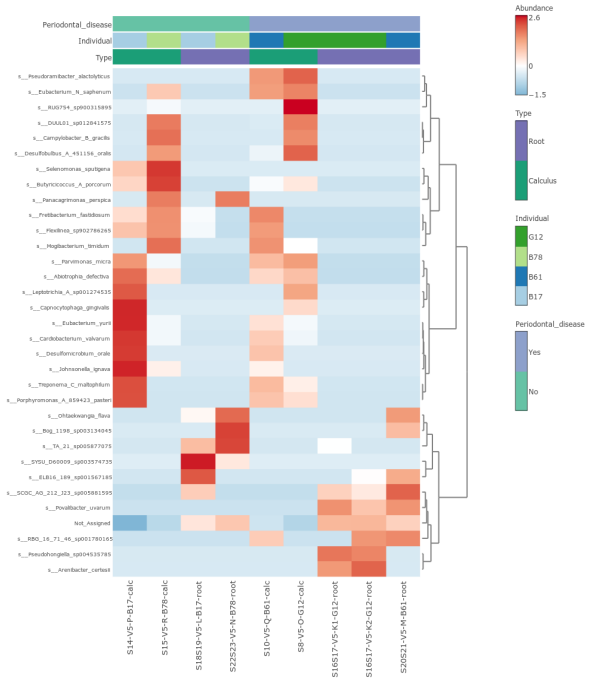
Who is guilty?

We found species that occur in calculus of diseased individuals but do not occur in the calculus of healthy ones or vice versa. The disease-present species are *Pseudoramibacter alactolyticus*, *Povalibacter uvarum*, *Arenibacter certesii* and *Eubacterium N saphenum*. The disease-absent species are *Selemonas sputigena* and *Panacagrimonas perspica*. Taking together, several promising associations might be found. For example, the highly abundant in diseased samples *Eubacterium N saphenum* is from the same genus as *Eubacterium nodatum*, that was described as a part on another one, Orange complex. Also, the *Mogibacterium timidum* that is highly co-occured with the *Eubacterium N saphenum*, was described to be strongly assotiated with the periodontitis [24, 25]. We also found the *Actinomyces dentalis* to appear in the G12 person shotgun metagenome sequence assembly with the high adundancy. This species was described to be strongly associated with the dental diseases as well [26, 27].

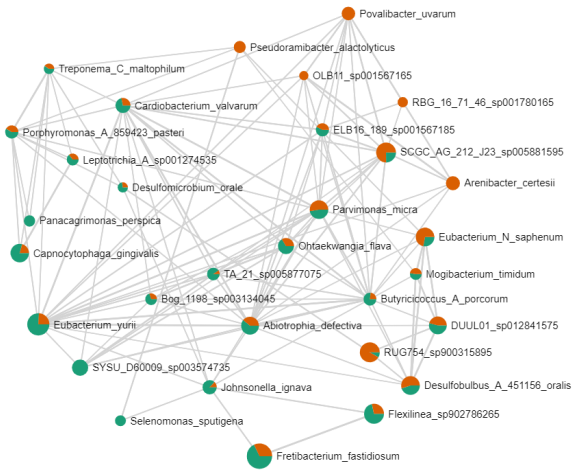
What's new?

To address the question how the evolution might have gone in this case, we extracted the regions of modern *Tannerella forsythia* annotation that are not covered by the G12 metagenome shotgun sequencing reads. As the most interesting ones we can provide:

- The several genes related to the lanthionine nonproteinogenic



(a) Heatmap of species abundancy in different types of samples



(b) SparCC correlation network (green - healthy individuals, brown - diseased individuals

Figure 8. Taxonomic analysis results from MicrobiomeAnalyst

amino acid, which is a component of lantibiotics antibiotics [28]. Here we found the class I lanthipeptide, lanthionine synthetase LanC, lantibiotic dehydratase family proteins.

- TIGR04157 family glycosyltransferase. It was described that this protein was not found in some Kuygenzhar strains while present in modern *T. forsythia* [29]. It is involved in the biosynthesis of O-glycan structures, occurring after the synthesis of glycan [30]. These genes are also absent in some ancient European strains, all ancient Mexican strains, as well as some modern strains which may indicate a recent acquisition of glycosylation genes[29].
- Several IS1, IS4 and many other family transposases (29 transposons-related entries in total). Wow! Someone dropped by *T. forsythia* genome for a visit
- Conjugative transposon protein TraJ, TraK, TraM, TraL, TraN, TraO. It feels like they're trying to put together a whole alphabet.
- Conjugal transfer protein MobA and MobB So not only phages are visiting, but also plasmids!
- Abi family proteins. Abi proteins are involved in bacteriophage resistance [31]. So our boy is not happy to have all the guests.
- Type VI secretion system tube protein TssD. The T6SS delivers multiple, diverse effector proteins directly into target cells in a way similar to to the action of contractile bacteriophage tails [32]. This is already the real bacterial high-tech of the military-industrial complex, which could have come with evolution. At the same time, there is evidence that this system can be acquired using mobile genetic elements [33].

References

1. Simon HY, Siddle KJ, Park DJ, Sabeti PC. Benchmarking metagenomics tools for taxonomic classification. Cell 2019;178(4):779–794.
2. Laudadio I, Fulci V, Palone F, Stronati L, Cucchiara S, Carissimi C. Quantitative assessment of shotgun metagenomics and 16S rDNA amplicon sequencing in the study of human gut microbiome. OMICS: A Journal of Integrative Biology 2018;22(4):248–254.
3. Sanschagrin S, Yergeau E. Next-generation sequencing of 16S ribosomal RNA gene amplicons. Journal of visualized experiments: JoVE 2014;(90).
4. Poretsky R, Rodriguez-R LM, Luo C, Tsementzi D, Konstantinidis KT. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PloS one 2014;9(4):e93827.
5. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochemical and biophysical research communications 2016;469(4):967–977.
6. Levy S. Time travellers: meet your ancestors' microbes. New Scientist 2014;222(2965):50–53.
7. Radini A, Tromp M, Beach A, Tong E, Speller C, McCormick M, et al. Medieval women's early involvement in manuscript production suggested by lapis lazuli identification in dental calculus. Science Advances 2019;5(1):eaau7126.
8. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 2019;37(8):852–857. https://doi.org/10.1038/s41587-019-0209-9.
9. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nature methods 2016;13(7):581.
10. McDonald D, Jiang Y, Balaban M, Cantrell K, Zhu Q, Gonzalez A, et al. Greengenes2 enables a shared data universe for microbiome studies. bioRxiv 2022;p. 2022–12.
11. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 2018;6(1):1–17.
12. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Jour-

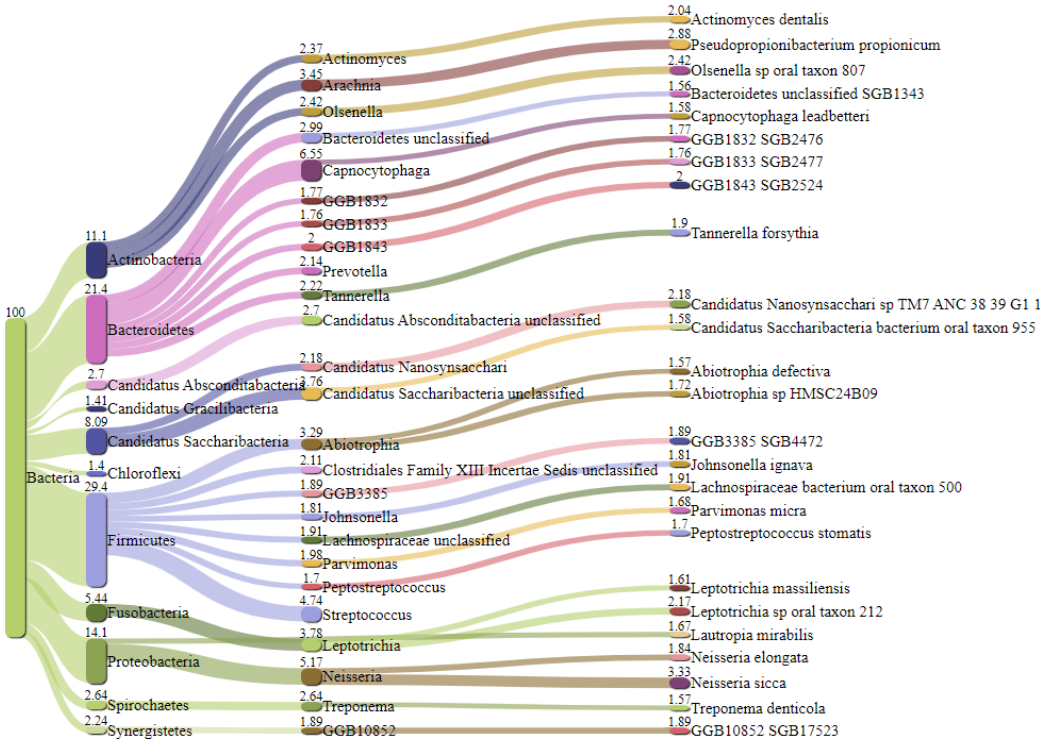


Figure 9. Shankey plot for the G12 diseased individual metagenome content

nal of machine learning research 2011;12(Oct):2825–2830.
13. Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. Nucleic acids research 2017;45(W1):W180–W188.
14. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria;, https://www.R-project.org.
15. McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, et al. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. GigaScience 2012;1(1):7.
16. Blanco-Míguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nature Biotechnology 2023;p. 1–12.
17. Breitwieser FP, Salzberg SL. Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification. Bioinformatics 2020;36(4):1303–1304.
18. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience 2021 02;10(2). https://doi.org/10.1093/gigascience/giab008, giab008.
19. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997 2013;.
20. Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. Current protocols in bioinformatics 2014;47(1):11–12.
21. Leo T. Anna Karenina. novel; 1828 - 1910.
22. Mohanty R, Asopa SJ, Joseph MD, Singh B, Rajguru JP, Saidath K, et al. Red complex: Polymicrobial conglomerate in oral flora: A review. Journal of family medicine and primary care 2019;8(11):3480.
23. Marotz C, Molinsky R, Martino C, Bohn B, Roy S, Rosenbaum M, et al. Early microbial markers of periodontal and cardiometabolic diseases in ORIGINS. npj Biofilms and Microbiomes 2022;8(1):30.
24. Casarin RCV, Saito D, Santos VR, Pimentel SP, Duarte PM, Casati MZ, et al. Detection of Mogibacterium timidum in subgingival biofilm of aggressive and non-diabetic and diabetic chronic periodontitis patients. Brazilian Journal of Microbiology 2012;43:931–937.
25. Moore W, Holdeman L, Cato E, Smibert R, Burmeister J, Palcanis K, et al. Comparative bacteriology of juvenile periodontitis. Infection and Immunity 1985;48(2):507–519.
26. Vielkind P, Jentsch H, Eschrich K, Rodloff AC, Stingu CS. Prevalence of Actinomyces spp. in patients with chronic periodontitis. International Journal of Medical Microbiology 2015;305(7):682–688.
27. Hall V. Actinomyces—gathering evidence of human colonization and infection. Anaerobe 2008;14(1):1–7.
28. Willey JM, Van Der Donk WA. Lantibiotics: peptides of diverse structure and function. Annu Rev Microbiol 2007;61:477–501.
29. White AE, de Dios T, Carrión P, Bonora GL, Llovera L, Cilli E, et al. Genomic analysis of 18th-century Kazakh individuals and their oral microbiome. Biology 2021;10(12):1324.
30. Zwickl NF, Stralis-Pavese N, Schäffer C, Dohm JC, Himmelbauer H. Comparative genome characterization of the periodontal pathogen Tannerella forsythia. BMC genomics 2020;21:1–18.
31. Chopin MC, Chopin A, Bidnenko E. Phage abortive infection in lactococci: variations on a theme. Current opinion in microbiology 2005;8(4):473–479.
32. Cianfanelli FR, Monlezun L, Coulthurst SJ. Aim, load, fire: the type VI secretion system, a bacterial nanoweapon. Trends in microbiology 2016;24(1):51–62.
33. Robinson L, Liaw J, Omole Z, Xia D, van Vliet AH, Corcionivoschi N, et al. Bioinformatic analysis of the Campylobacter jejuni type VI secretion system and effector prediction. Frontiers in Microbiology 2021;12:694824.

Supplemental materials

Working Notes: Link
GitHub repository of the project: Repo

Figure 10. Species families abundances