Dead Man's Teeth. Introduction to metagenomics analysis.

Authors: Anna Rybina, Daria Nikanorova

Abstract

In this study we analyzed DNA samples extracted from dental calculus of mediaeval people. We first performed taxonomic classification of 16S ribosomal RNA reads with QIIME2 pipeline. We found seven dominant phyla in the calculus sample: Firmicutes (relative frequency of 65.53%), Proteobacteria (13.784%), Bacteroidetes (7.755%), Synergistetes (5.79%), Actinobacteria (2.8%), Fusobacteria (1.62%), and Spirochaetes (0.915%). We also confirmed the presence of "red-complex" pathogens in ancient oral microbiomes. Next, we analyzed metagenome assembly derived from shotgun sequencing for taxonomy composition and relative organismal abundance. Final profile was compared with profiles obtained for subsampled shotgun read data from the Human Microbiome Project. We found out that the ancient sample is similar to the modern sample of supralingual plaque. Both samples are abundant with *Streptococcus oralis*, *Tanerella forsythia* and *Capnocytophaga gingivalis*. Interestingly, *Rothia dentocariosa* was found mostly in the modern sample. Finally, we concentrated on possible ways of evolution of detected bacteria. Particularly, we pinpointed 191 coding sequences that possibly were obtained by *Tannerella forsythia* during evolution. We distinguished the following groups of acquired genes: mobile genetic elements, involved in the horizontal gene transfer of plasmids, genes related to DNA repair, post-translational modification, antibiotic biosynthesis, drug resistance, metabolism, and restriction modification system.

Introduction

The field of metagenomics is a promising area to unravel the content of microbial communities sampled from different places and even time periods. Analyzing DNA from natural samples, metagenomics produces a profile of microbial diversity, without the necessity of obtaining pure cultures of bacteria or viruses.

Metagenomic analysis may be carried out via sequencing of specific genes (often the 16S rRNA gene) or shotgun sequencing of the aggregated metagenomic DNA (mgDNA) of the entire microbial community (Breitwieser et al. 2019). In the first approach, the environmental DNA is extracted and the 16S rRNA gene is amplified using "universal" (actually, not so universal according to Fuks et al. 2018) primers. Next, these amplified genes are sequenced using NGS, which results in the thousands of rRNA reads that are aligned with reference databases in search of similarities. While the amplicon-based approach is limited by incompleteness of reference databases, shotgun technologies can cope with analyzing novel sequences belonging to previously unidentified taxonomic lineages (Lapidus and Korobeynikov 2021).

In shotgun metagenomics, the total genomic DNA is extracted from a sample and sequenced. It is a powerful technique, because apart from detecting new species it can also unravel the functional relationship in a microbial community. However, shotgun metagenomics tends to bias towards the dominant species and only sparsely covers the genomic content of the low abundance members of the community. Astonishingly, these approaches allow us to study not only modern, but also ancient microbial communities, preserved in different substances. Recently, ancient metagenomics made it possible to analyze microbial communities of dental calculus and reconstruct oral microbiomes and lifestyles of humans from the past. Calcified dental plaque (dental calculus) preserves for thousands of years and hosts biomolecules from oral cavity bacteria.

In this study we will analyze 16S ribosomal RNA reads as well as shotgun sequencing results in order to investigate ancient oral microbiomes.

Methods

The 16S rRNA library targeting V5 region was generated by an instrument Roche GS Junior (454) from archaeological human dental calculus dated to c. 950-1200 CE. For the resulting experimental dental calculus sample SRX351237, the tooth root (sample SRX351242) was used as a proxy for environmental control. Using QIIME2 v2020.2 (Bolyen et al. 2019) with DADA2 pipeline (Callahan et al. 2016), amplicon sequencing data was subjected to the following analysis: quality control, trimming, denoising, dereplication, filtering chimeric sequences, clustering into amplicon sequence variants (ASVs). Taxonomy composition of samples was identified by Naive Bayes algorithm implemented in QIIME2 using GreenGenes database (https://data.qiime2.org/2020.2/common/gg-13-8-99-nb-classifier.qza). Visualization of QIIME results was done on the web-server https://view.qiime2.org/ .

Dental calculus microbiome from the individual G12 skeleton (ca. AD 950-1200) with periodontal disease was subjected to shotgun sequencing and resulting reads were assembled into contigs. Metagenome assembly was analyzed for taxonomy composition and relative organismal abundance using MetaPhlAn v3.0.4 (Segata et al. 2012). Final profile was compared with profiles obtained for subsampled shotgun read data from the Human Microbiome Project (SRS014459 (stool), SRS014464 (anterior nares), SRS014470 (tongue dorsum), SRS014472 (buccal mucosa), SRS014476 (supragingival plaque), SRS014494(posterior fornix)). Species only heatmap of organism abundances was generated using hclust2. Contigs of metagenome assembly were aligned to the *Tannerella forsythia* 92A2 genome (assembly accession: ASM23821v1) using BWA v0.7.17 (bwa mem) (Li and Durbin 2009). Resulting alignment was converted to binary format, sorted, and indexed using SAMtools v1.12-8 (Li et al. 2009), and converted to BED format by BEDtools v2.20.0 (bedtools bamtobed) (Li et al. 2009; Quinlan and Hall 2010). New regions in the modern strain absent in the ancient strain were subtracted from GFF3 using bedtools intersect (option -v). Resulting annotation file was manually inspected using bash.

Results

Amplicon sequencing

Initially, calculus sample SRX351237 and bone sample SRX351242 contained 5362 and 5788 sequences, respectively, with the median length of 179 nucleotides (nt). According to quality assessment results, quality score across sequences began to drop from 180 base position. As the total length of the artificial sequences (barcode and primer) was 32 nt, sequences were truncated at the position of 150 nt. After denoising, dereplicating, and filtering chimeras, 92.95 and 89.71 % of bone and calculus sequences, respectively, have retained (**Table 1**).

Table 1. Statistics summary on sequences passed filtration

sample-id	input	filtered	percentage of input passed filter	denoised	non-chimeric	percentage of input non-chimeric
bone	5788	5589	96.56	5377	5377	92.9
calculus	5362	5183	96.66	5068	4837	90.21

Clustering samples into ASVs revealed 163 feature sequences of the length 118 nt that were distributed among bone and calculus samples with counts 5380 and 4810, respectively, with a mean frequency per feature of about 62.
According to taxonomy classification results (**Fig. 1a**), we found seven dominant phyla in the calculus sample: Firmicutes (relative frequency of 65.53%), Proteobacteria (13.784%), Bacteroidetes (7.755%), Synergistetes (5.79%), Actinobacteria (2.8%), Fusobacteria (1.62%), and Spirochaetes (0.915%). The bone sample also contained bacteria from Firmicutes, Proteobacteria, Bacteroidetes, and Actinobacteria but they constituted different proportions (10.46 %, 61.71 %, 8.996%, and 14.86%, respectively). Interestingly, Spirochaetes and Fusobacteria were not present in the bone sample. Moreover, calculus sample, opposed to the bone sample, contained several bacteria associated with periodontal disease (members of so-called “the red complex”): Treponema socranskii (relative frequency of 0.561%), Treponema (0.353%), Tannerella (0.52 %), and Porphyromonas (0.291%) (**Fig. 1b**).

a) b)

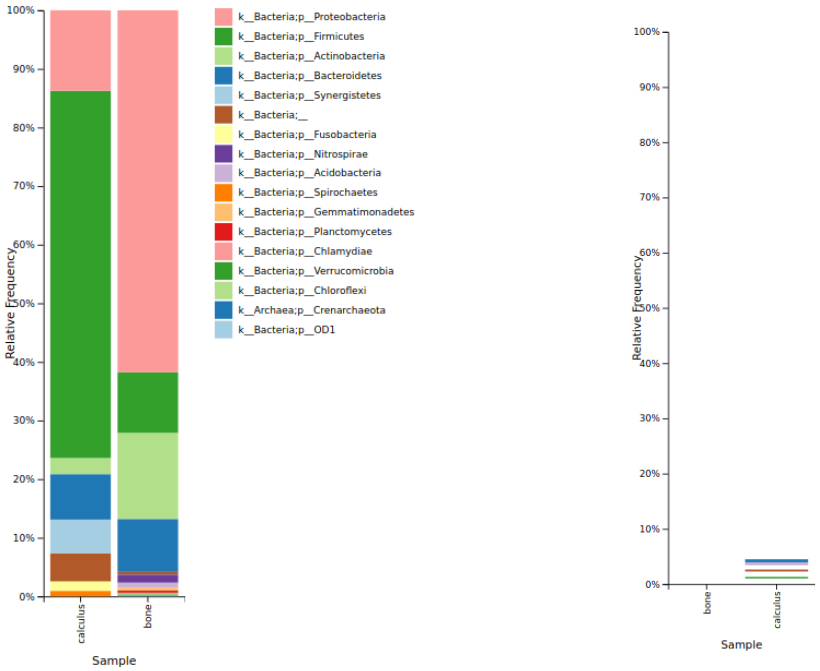


Figure 1. Taxonomy composition barplot: a) phylum-level; b) members of "the red complex": Treponema socranskii (relative frequency of 0.561%), Treponema (0.353%), Tannerella (0.52 %), and Porphyromonas (0.291%) are present in "calculus" sample.

Shotgun sequencing

According to analysis of profile taxonomy abunduces (**Fig. 2a**), calculus microbiome from the G12 individual is quite diverse and composed of different bacteria species with much higher abundance compared to other samples (e.g., bacteria from Tannerella, Treponema, and Porphyromonas genera). According to comparison with subsampled read data from HMP (**Fig. 2a**), we may identify several bacteria mostly specific to the calculus microbiome: *Streptococcus gordonii*, *Streptococcus sanguinis*, *Corynebacterium matruchotii*, *Campylobacter showae*, *Treponema denticola*, *Capnocytophaga gingivalis*, *Capnocytophaga ochracea*, and *Fusobacterium nucleatum*. We found *Corynebacterium matruchotii* in both calculus and supragingival plaque, while *Rothia dentocariosa* was more abundant in the latter sample. After recalculation with raw read data (**Fig. 2b**), we obtained a little different picture when much more species from the calculus sample were shared with supragingival plaque and with buccal mucosa but to a lesser extent. Based on heatmap, we may conclude that calculus metagenome is more similar to supragingival plaque microbiome and a little less

resemble buccal mucosa than any other microbiomes used in analysis. Similarity between calculus and supragingival plaque metagenomes might be observed on Sankey plots (**Fig. 3**).

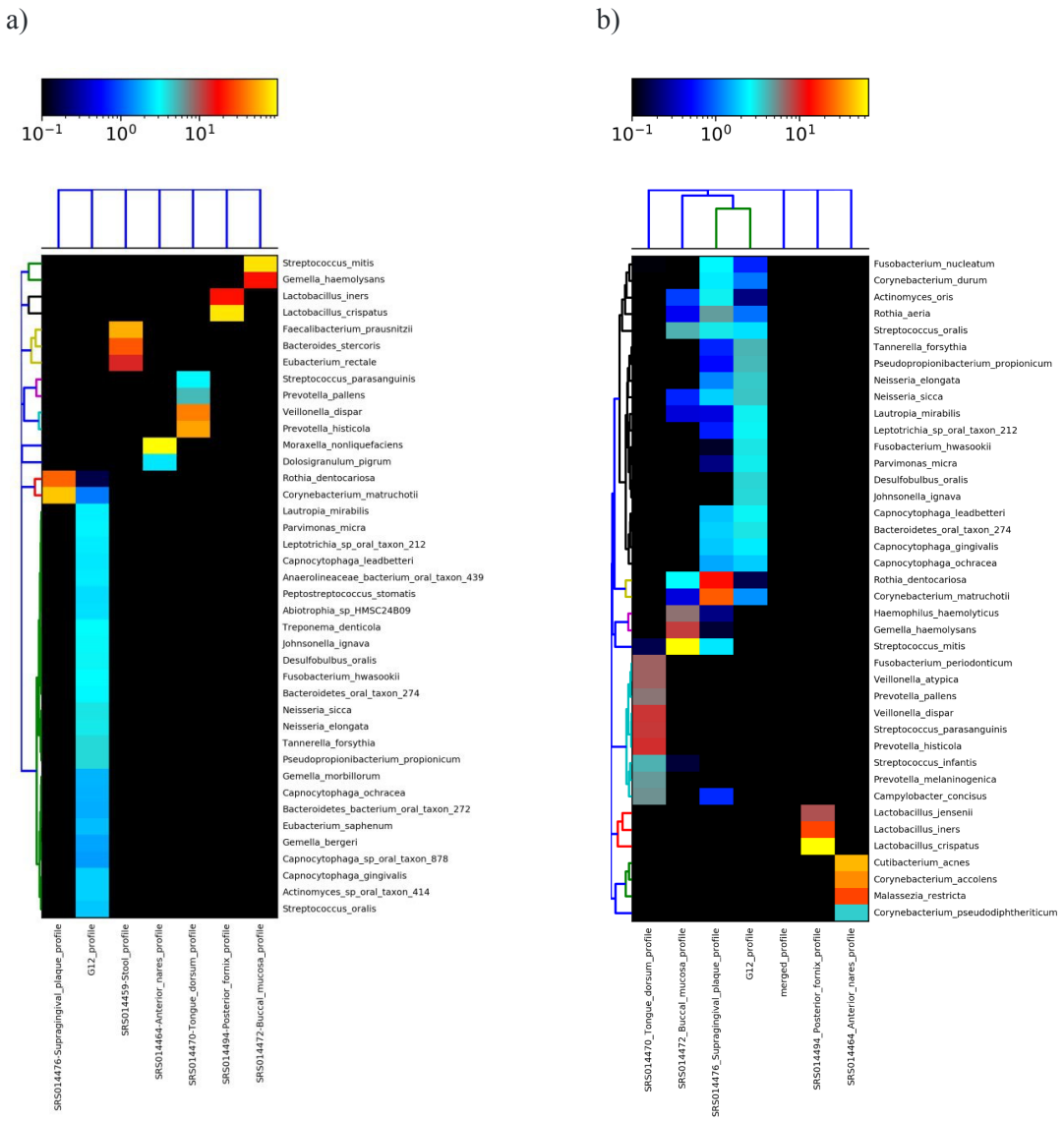
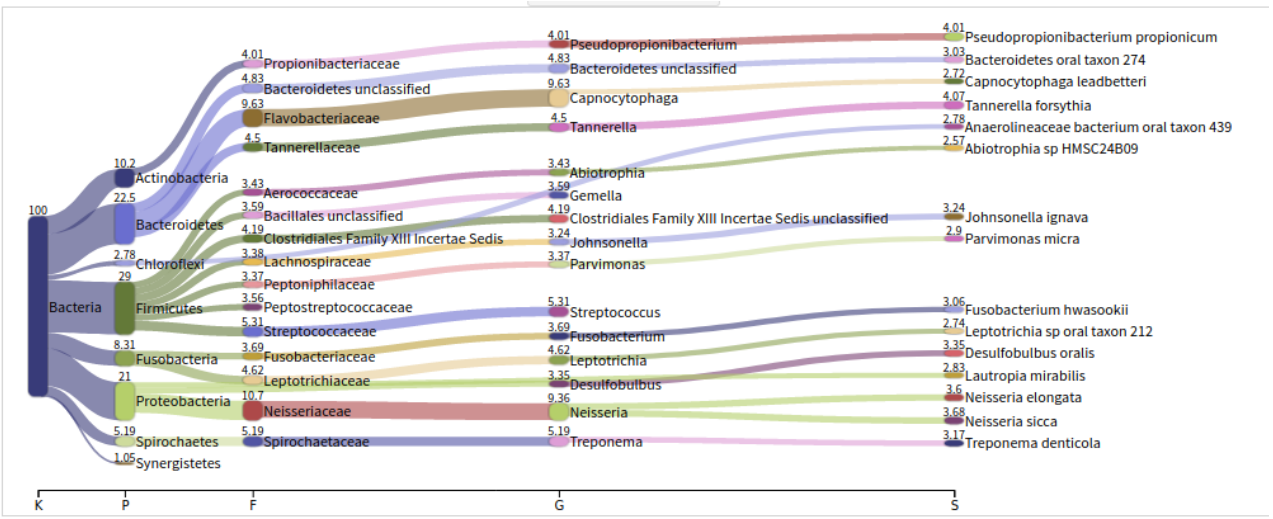


Figure 2. The species-level abundance heatmap of the G12 sample and several microbiome samples from the Human Microbiome Project (HMP): a) subsampled read data from HMP b) raw data from HMP. Bray-Curtis was used as the distance measure both between samples and between features. Only the top 40 species are present. Heatmap colors were assigned in a log scale. Profile data was obtained using MetaPhlan. Heatmap was calculated via hclust2.

a)



b)

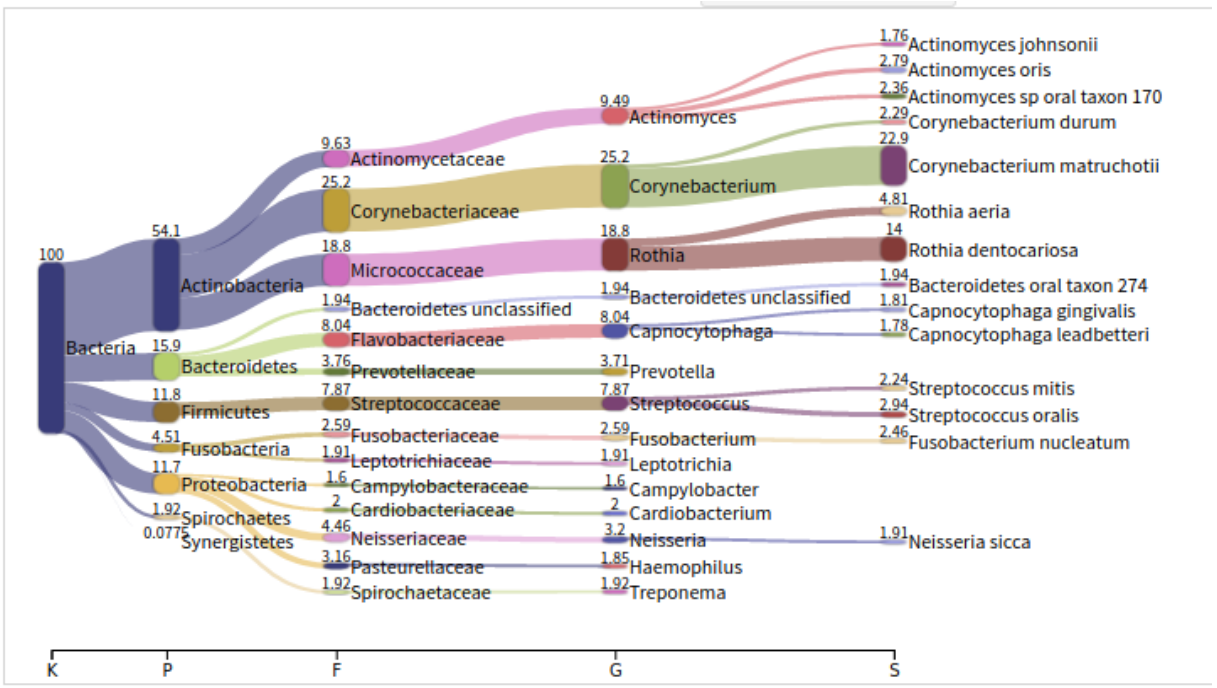


Figure 3. The species-level abundance Sankey plot of the a) G12 sample and b) supragingival plaque sample from the Human Microbiome Project (B). Profile data was obtained using MetaPhlan. The Sankey plot was visualized via Pavian.

Only 1.83% of calculus metagenomic contigs were aligned to the *Tannerella forsythia* 92A2 genome. Along with regions mapped with contigs, we could reveal several regions of zero coverage in the modern strain that were probably gained during evolution. Respective regions contained 191 coding sequences. Inspecting their annotation, we distinguished the following groups of acquired genes (**Table 2**): mobile genetic elements, involved in the

horizontal gene transfer of plasmids, genes related to DNA repair, post-translational modification, antibiotic biosynthesis, drug resistance, metabolism, and restriction modification system.

Table 2. Summary of several genes acquired by the *Tannerella forsythia* 92A2.

product	protein_id/locus_tag	related to
AIPR family protein	WP_014224073.1	restriction modification system
DDE-type integrase/transposase/recombinase	WP_014223573.1	mobile genetic element
IS1 family transposase	WP_099046116.1	mobile genetic element
IS110 family transposase	WP_014225157.1	mobile genetic element
IS1380 family transposase	WP_041590984.1	mobile genetic element
IS1595 family transposase	WP_014224553.1	mobile genetic element
IS1595-like element ISTfo1 family transposase	WP_041590537.1	mobile genetic element
IS4-like element IS421 family transposase	WP_001300563.1	mobile genetic element
IS5 family transposase	BFO_RS14050	mobile genetic element
ISL3 family transposase	WP_014224885.1	mobile genetic element
site-specific integrase	WP_007366539.1	mobile genetic element
transposase	WP_157755281.1	mobile genetic element
DNA mismatch endonuclease Vsr	WP_041590610.1	DNA repair
GIY-YIG nuclease family protein	WP_041591057.1	DNA repair
toprim domain-containing protein	WP_007366501.1	DNA repair
PD-(D/E)XK motif protein	WP_014224074.1	DNA repair
GNAT family N-acetyltransferase	WP_007366527.1	post-translational modification
SPASM domain-containing protein	WP_157755359.1	post-translational modification
TIGR04149 family rSAM-modified RiPP	WP_208854694.1	post-translational modification
TIGR04150 pseudo-rSAM protein	WP_014224287.1	post-translational modification
TIGR04157 family glycosyltransferase	WP_014223582.1	post-translational modification
radical SAM peptide maturase	WP_014224286.1	post-translational modification
class I lanthipeptide	WP_041590503.1	antibiotic biosynthesis
lanthionine synthetase C family protein	WP_014223583.1	antibiotic biosynthesis
lantibiotic dehydratase family protein	WP_041590506.1	antibiotic biosynthesis
thiopeptide-type bacteriocin biosynthesis protein	WP_041590507.1	antibiotic biosynthesis
ParA family protein	WP_007366517.1	HGT of plasmids
conjugal transfer protein TraO	WP_007366502.1	HGT of plasmids
conjugative transposon protein TraJ	WP_007366507.1	HGT of plasmids
conjugative transposon protein TraK	WP_007366506.1	HGT of plasmids
conjugative transposon protein TraM	WP_007366504.1	HGT of plasmids
conjugative transposon protein TraN	WP_007366503.1	HGT of plasmids
relaxase/mobilization nuclease domain-containing protein	WP_007366518.1	HGT of plasmids
antirestriction protein ArdA	BFO_RS15645	anti-restriction modification
tetracycline resistance ribosomal protection protein	WP_007366526.1	antibiotic resistance
RteC domain-containing protein	WP_007366522.1	antibiotic resistance

TetR/AcrR family transcriptional regulator	WP_051322484.1	antibiotic resistance
dihydrofolate reductase family protein	WP_007366523.1	metabolic enzyme
beta-ketoacyl-ACP synthase III	WP_014226278.1	metabolic enzyme

Discussion

Calculus microbiome composition

Although we found several organisms common for calculus, the bone and plaque samples (**Fig. 1a, Fig. 2**), we may distinguish several bacteria mainly specific for the calculus sample. Remarkably, some of these organisms are pathogens associated with periodontal disease. For instance, (Etoh et al. 1993) reported *Campylobacter showae* to be linked with periodontal disease. *Fusobacterium nucleatum* is considered to be one of the crucial gram-negative pathogens involved in periodontitis (Signat et al. 2011) as well as bacteria of Capnocytophaga species (e.g., *Capnocytophaga gingivalis, Capnocytophaga ochracea*) associated with chronic character of the disease (Idate et al. 2020). *Streptococcus gordonii* is a pioneer colonizer of the oral environment, an opportunistic pathogen able to trigger infective endocarditis (Kolenbrander 2011). Various pathogens connected with periodontal disease were detected in the calculus sample, both in amplicon data and shotgun sequencing data, confirming an idea that archaeological individuals suffered from periodontitis.

According to taxonomy composition results, we suggest that the microbiome community that is the most similar to calculus metagenome, might be of supragingival plaque (**Fig. 2**). Corynebacterium matruchotii was detected in both samples which corresponds to the fact that this bacteria is associated with subgingival plaque (Paster et al. 2001) and dental plaque (Barrett et al. 2001)

Moreover, ancient dental calculus samples were abundant with another periodontal pathogens, so-called "red-complex" bacteria (Socransky and Haffajee 2005), namely *Tannerella forsythia, Porphyromonas gingivalis*, and *Treponema denticola.* This finding points towards the presence of periodontal disease during the medieval period as nowadays albeit significant alterations in lifestyle, diet and oral hygiene.

Tannerella forsythia: acquired genes

Comparing with the ancient calculus metagenome, we identified several genes that modern *Tannerella forsythia* 92A2 strain has acquired (some of them listed in **Table 2**). Respective gene products might improve bacterial survival, increase its virulence and competitiveness among other pathogens.

It is known that horizontal gene transfer plays a crucial role in the evolution, transmissing and maintaining virulence (Keen 2012). We revealed several genes (*tra* genes) that facilitate HGT of plasmids being part of F-like transfer systems (Keen 2012; Zatyka and Thomas 1998). Moreover, about 40 coding sequences obtained by modern strain are related to mobile genetic elements, including transposons and plasmids.

We identified several antibiotic resistance genes which products act against tetracycline (WP_007366526.1, WP_007366522.1) and TetR/AcrR family transcriptional regulator (WP_051322484.1) that is important for efflux-mediated antimicrobial resistance in pathogens (Colclough, Scadden, and Blair 2019). For Bacteroides, tetracycline resistance genes were reported to be acquired by large conjugative transposons (Stevens et al. 1993) and, probably, *Tannerella forsythia* 92A2 obtained drug resistance genes via HGT with some mobile genetic elements as well.

There might be a balance between acquisition of foreign DNA and prevention to this process which is manifested on genetic level. For instance, on the one hand, modern strain has gained gene coding for AIPR family protein (WP_014224073.1) which is part of a restriction-modification (RM) system and provides a defense against forein DNA and infectious agents such as a bacteriophage (Iyer, Abhiman, and Aravind 2008). On the other hand, we found anti-RM protein encoded by *ardA* (BFO_RS15645) that can enhance dissemination of mobile genetic elements by mimicking the DNA structure bound by Type I RM enzymes (Chen et al. 2014).

Among acquired genes enhancing pathogen viability are those which products are involved in DNA repair, antibiotic biosynthesis, and post-translational modification. For example, toprim domain-containing protein (WP_007366501.1) is responsible for DNA strand breakage and rejoining (Aravind, Leipe, and Koonin 1998) and DNA mismatch endonuclease Vsr (WP_041590610.1) is required for mismatch repair (Bhagwat and Lieb 2002). Several acquired genes are linked to antibiotic biosynthesis, including lanthionine dehydratase (WP_041590506.1) and lanthionine synthetase (WP_014223583.1) (Marsh et al. 2010). We discovered at least three genes encoding radical SAM peptide maturases (WP_208854694.1, WP_014224286.1, WP_014224287.1) that form a peptide thioether crosslink during post-translational modification (Marsh et al. 2010; Bruender et al. 2016). Resulting peptide-derived natural products provide a pathogen with a selective advantage over other organisms in their biological niche.

Supplementary materials

Detailed pipeline with code is provided here.

References

Aravind, L., D. D. Leipe, and E. V. Koonin. 1998. “Toprim--a Conserved Catalytic Domain in Type IA and II Topoisomerases, DnaG-Type Primases, OLD Family Nucleases and RecR Proteins.” *Nucleic Acids Research* 26 (18): 4205–13.

Barrett, S. L., B. T. Cookson, L. C. Carlson, K. A. Bernard, and M. B. Coyle. 2001. “Diversity within Reference Strains of Corynebacterium Matruchotii Includes Corynebacterium Durum and a Novel Organism.” *Journal of Clinical Microbiology* 39 (3): 943–48.

Bhagwat, Ashok S., and Margaret Lieb. 2002. “Cooperation and Competition in Mismatch Repair: Very Short-Patch Repair and Methyl-Directed Mismatch Repair in Escherichia Coli.” *Molecular Microbiology* 44 (6): 1421–28.

Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. 2019. “Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2.” *Nature Biotechnology* 37 (8): 852–57.

Bruender, Nathan A., Jarett Wilcoxen, R. David Britt, and Vahe Bandarian. 2016. “Biochemical and Spectroscopic Characterization of a Radical S-Adenosyl-L-Methionine Enzyme Involved in the Formation of a Peptide Thioether Cross-Link.” *Biochemistry* 55 (14): 2122–34.

Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. 2016. “DADA2: High-Resolution Sample Inference from Illumina Amplicon Data.” *Nature Methods* 13 (7): 581–83.

Chen, Kai, Marcel Reuter, Bansi Sanghvi, Gareth A. Roberts, Laurie P. Cooper, Matthew Tilling, Garry W. Blakely, and David T. F. Dryden. 2014. “ArdA Proteins from Different Mobile Genetic Elements Can Bind to the EcoKI Type I DNA Methyltransferase of E. Coli K12.” *Biochimica et Biophysica Acta* 1844 (3): 505–11.

Colclough, A. L., J. Scadden, and J. M. A. Blair. 2019. “TetR-Family Transcription Factors in Gram-Negative Bacteria: Conservation, Variation and Implications for Efflux-Mediated Antimicrobial Resistance.” *BMC Genomics* 20 (1): 731.

Etoh, Y., F. E. Dewhirst, B. J. Paster, A. Yamamoto, and N. Goto. 1993. “Campylobacter Showae Sp. Nov., Isolated from the Human Oral Cavity.” *International Journal of Systematic Bacteriology* 43 (4): 631–39.

Idate, Ulka, Kishore Bhat, Vijayalakshmi Kotrashetti, Manohar Kugaji, and Vijay Kumbar. 2020. “Molecular Identification of Species from the Oral Cavity of Patients with Chronic Periodontitis and Healthy Individuals.” *Journal of Oral and Maxillofacial Pathology: JOMFP* 24 (2): 397.

Iyer, Lakshminarayan M., Saraswathi Abhiman, and L. Aravind. 2008. “MutL Homologs in Restriction-Modification Systems and the Origin of Eukaryotic MORC ATPases.” *Biology Direct* 3 (March): 8.

Keen, Eric C. 2012. “Paradigms of Pathogenesis: Targeting the Mobile Genetic Elements of Disease.” *Frontiers in Cellular and Infection Microbiology* 2 (December): 161.

Kolenbrander, Paul E. 2011. *Oral Microbial Communities: Genomic Inquiry and Interspecies Communication*. Edited by Paul E. Kolenbrander. Washington, D.C., DC: American Society for Microbiology.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.”

Bioinformatics 25 (16): 2078–79.
Marsh, Alan J., Orla O'Sullivan, R. Paul Ross, Paul D. Cotter, and Colin Hill. 2010. "In Silico Analysis Highlights the Frequency and Diversity of Type 1 Lantibiotic Gene Clusters in Genome Sequenced Bacteria." *BMC Genomics* 11 (November): 679.
Paster, B. J., S. K. Boches, J. L. Galvin, R. E. Ericson, C. N. Lau, V. A. Levanos, A. Sahasrabudhe, and F. E. Dewhirst. 2001. "Bacterial Diversity in Human Subgingival Plaque." *Journal of Bacteriology* 183 (12): 3770–83.
Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42.
Segata, Nicola, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. 2012. "Metagenomic Microbial Community Profiling Using Unique Clade-Specific Marker Genes." *Nature Methods* 9 (8): 811–14.
Signat, Benoit, Christine Roques, Pierre Poulet, and Danielle Duffaut. 2011. "Fusobacterium Nucleatum in Periodontal Health and Disease." *Current Issues in Molecular Biology* 13 (2): 25–36.
Socransky, Sigmund S., and Anne D. Haffajee. 2005. "Periodontal Microbial Ecology." *Periodontology 2000* 38: 135–87.
Stevens, A. M., N. B. Shoemaker, L. Y. Li, and A. A. Salyers. 1993. "Tetracycline Regulation of Genes on Bacteroides Conjugative Transposons." *Journal of Bacteriology* 175 (19): 6134–41.
Zatyka, M., and C. M. Thomas. 1998. "Control of Genes for Conjugative Transfer of Plasmids and Other Mobile Elements." *FEMS Microbiology Reviews* 21: 291–319.