BIOINFORMATICS
INSTITUTE

Introduction to metagenomics analysis

May 31, 2024

Homework number 7

by Sapozhnikov N.A.

1 Introduction

Metagenomics has emerged as a powerful tool for unraveling the intricate microbial communities inhabiting diverse environments, ranging from soil and water ecosystems to the human microbiome. Traditional culture-based methods provide only a limited view of microbial diversity, often missing the majority of microbes that are difficult to culture in the laboratory. Metagenomics, on the other hand, circumvents these limitations by directly sequencing DNA extracted from environmental samples. This approach not only allows for the identification of known microbes but also enables the discovery of novel species and functions.

Two primary strategies are commonly employed in metagenomics analysis: 16S rRNA gene sequencing and shotgun metagenomic sequencing. 16S rRNA gene sequencing targets a specific phylogenetic marker gene, providing insights into the taxonomic composition of microbial communities. In contrast, shotgun metagenomic sequencing offers a more comprehensive view by sequencing all DNA present in a sample, allowing for the characterization of both microbial diversity and functional potential.

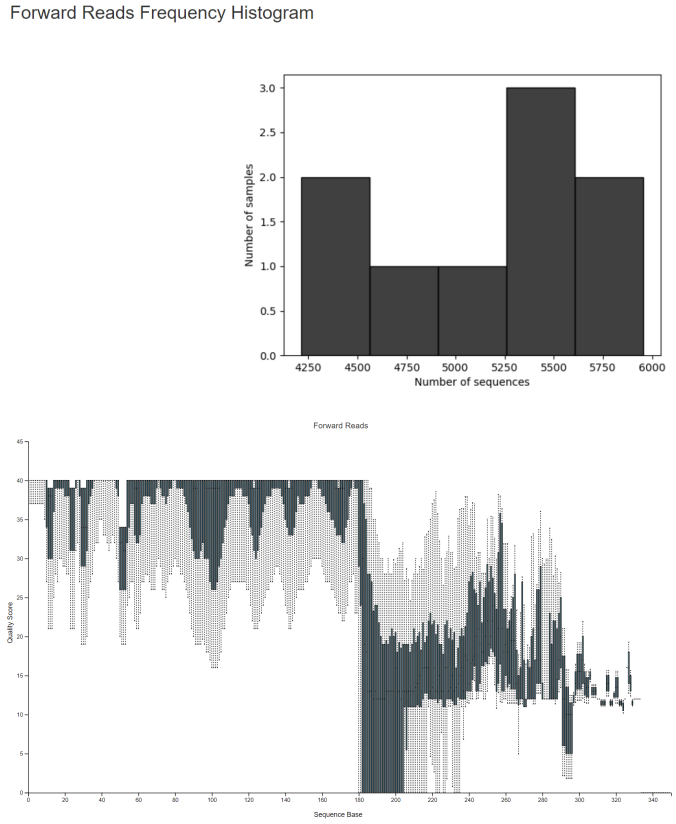
2 Methods

2.1 Raw Data

Samples were obtained from dental calculus, and will be compared with the samples from the ancient tooth roots. All data from the original research are available in the NCBI Short Read Archive (SRA) under number SRP029257 (BioProject PRJNA216965).

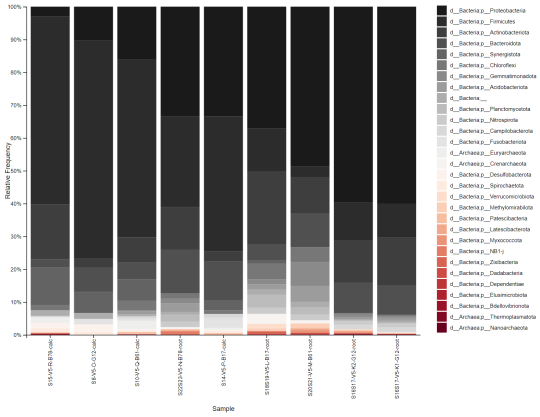
2.2 Tools and Parameters

Single-end reads in FASTQ format with Phred33 encoding were imported to qiime2. The primary QC showed how many sequences were obtained per sample (fig.1), and to get a summary of the distribution of sequence qualities (fig.2).

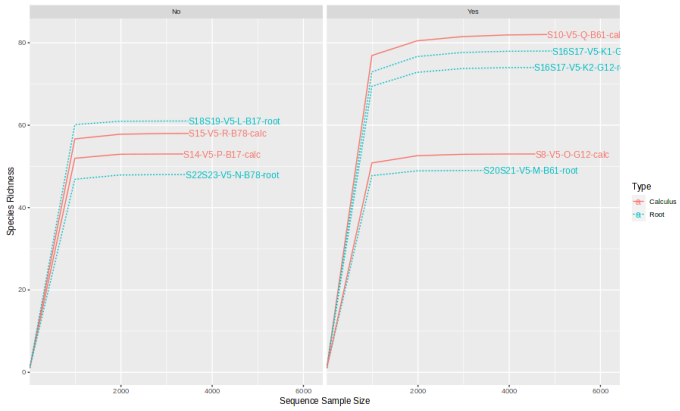


Barcodes and primer sequences were removed and chimeric sequences were filtered out. Value m for –p-trim-left as a total length of the artificial sequences (barcodes) was set as m = 35. And value n for –p-trunc-len according to Interactive Quality Plot (a total length of the sequence after the trimming), an

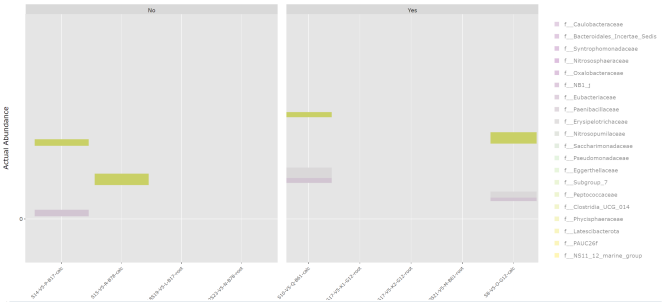
amplicon size was set to 140.
An amplicon sequence variants (ASV) - a higher-resolution analogue of the traditional Operational Taxonomical Units (OTU) were then clustered. These were used with a pre-trained Naive Bayess classifier to obtain a taxonomic composition of our samples. Visualization of a bar-plot (fig.3) was made using a MicrobiomeAnalyst online tool. A cutoff on level of families was chosen.



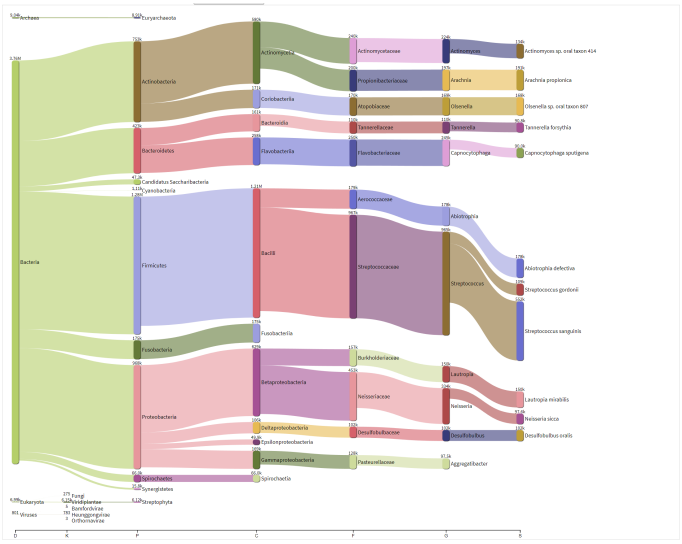
We were looking for bacteria from the "red complex": Porphyromonas gingivalis, Tannerella forsythia, Treponema denticola which are responsible for causing severe forms of periodontal disease. Rarefaction curves help to visuallize the data after filtering and normalization (fig.4).



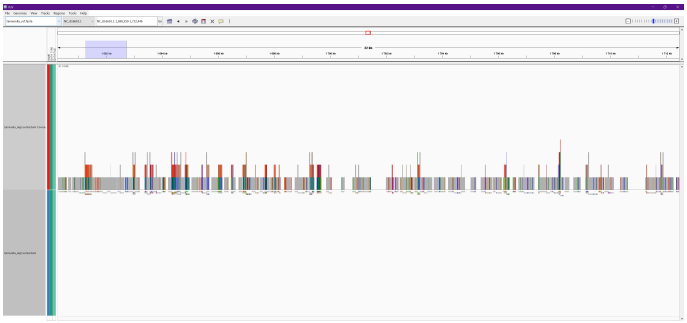
On this filtered abundance view (fig.5) samples are organized by peridontal disease and only families from the "red complex" are shown. Samples with the disease do have all 3 families of microorganisms while samples with out it have only 2. Althogh niether of these families are shown on the top-10 taxa which indicates their relatively small abundance.



We realized that 1000 years ago periodontal disease was caused by the same bacteria that we can find now in our mouth. But we know that bacteria evolve very quickly, and we have a unique opportunity to explore how it happens. To investigate it, an affected individual G12 was selected for a dental calculus whole metagenome shotgun sequencing, and reads were assembled into contigs.
Kraken2 was used to assign the reads to different taxa based on a reference database of genomes. Results are presented on a Sankey diagram (fig.6)



Our shotgun assembly was aligned on a T. forsythia strain reference genome. The coverage left much to be desired (fig.7).



Thus an assumption was made that the strain evolutionized. To check this assumption regions that are present in the reference but abscent in our assembly were crossed and briefly annotated.

3 Discussion

The annotation of crossed regions showed a fair amount of transposase coding genes. Transposase genes encode enzymes responsible for the movement of transposable elements within the genome, contributing to genome plasticity, evolution, and adaptation.

- Genomic Plasticity: Transposable elements are DNA sequences capable of changing their positions within the genome. These elements can lead to genomic rearrangements, duplications, deletions, and mutations, thereby contributing to genomic plasticity.
- Evolutionary Dynamics: Transposable elements play a significant role in driving evolutionary processes by generating genetic diversity. Their movement within the genome can result in the creation of new gene arrangements, regulatory sequences, and functional innovations, contributing to species diversification and adaptation.
- Adaptation and Genome Evolution: The presence of transposase coding genes suggests ongoing genomic changes that may facilitate adaptation to changing environments. Transposable elements can carry and disseminate adaptive genetic elements, such as genes conferring antibiotic resistance, stress response, or metabolic advantages, thus promoting rapid adaptation to new selective pressures.
- Regulation of Gene Expression: Transposable elements can influence gene expression patterns by inserting into regulatory regions, altering chromatin structure, or providing new regulatory sequences.

The presence of transposase coding genes in crossed regions may indicate potential regulatory effects on nearby genes, leading to changes in their expression profiles.

Differences in microbiome content between samples can arise from various factors, including environmental conditions, host factors, geographical location. These differences can impact the evolution of pathogens through mechanisms such as host-associated evolution, horizontal gene transfer, and ecological interactions, shaping their virulence, antibiotic resistance, and adaptation to different environments. Understanding these dynamics is crucial for elucidating the drivers of pathogen evolution and developing strategies for disease prevention and control.