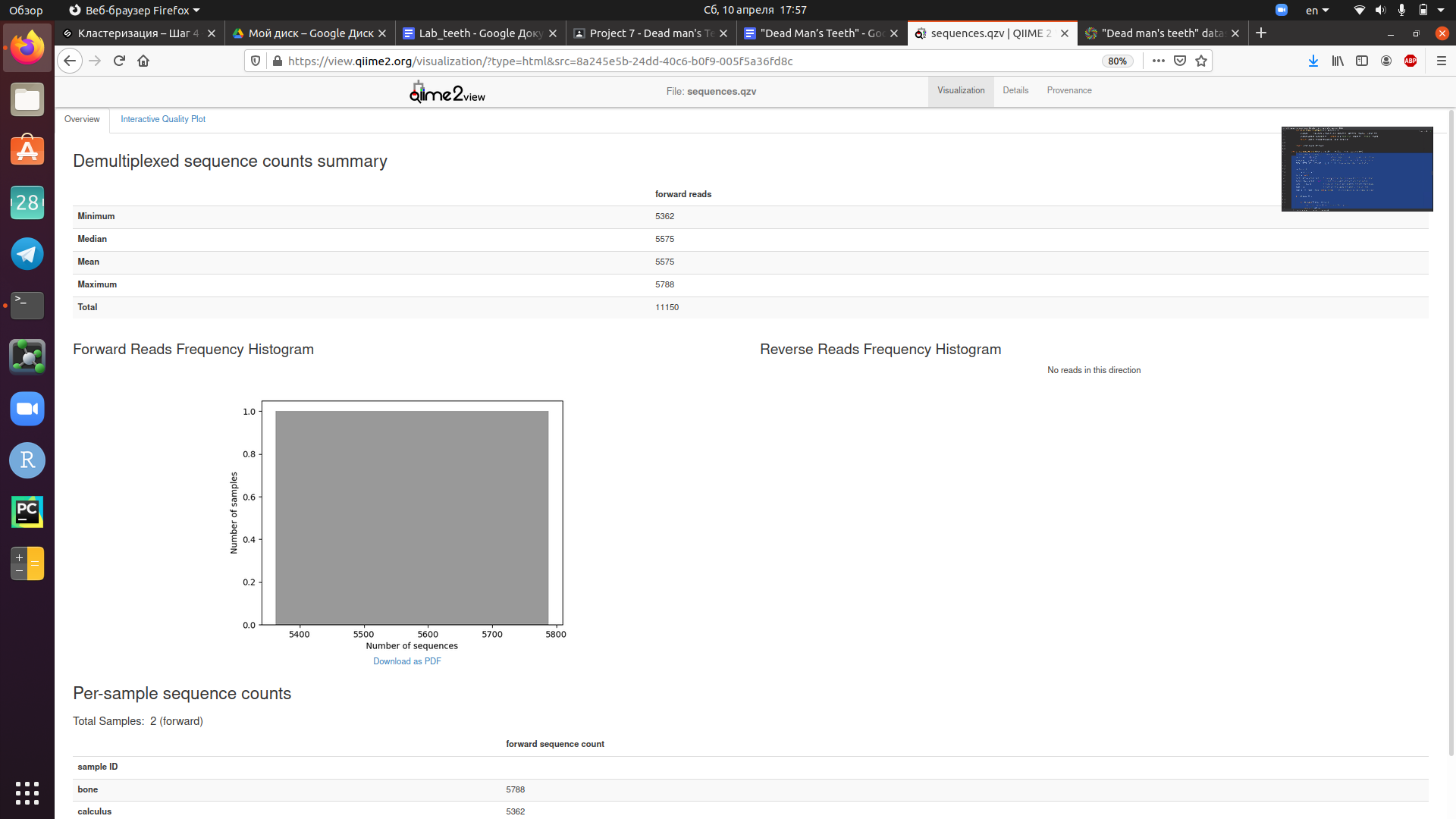
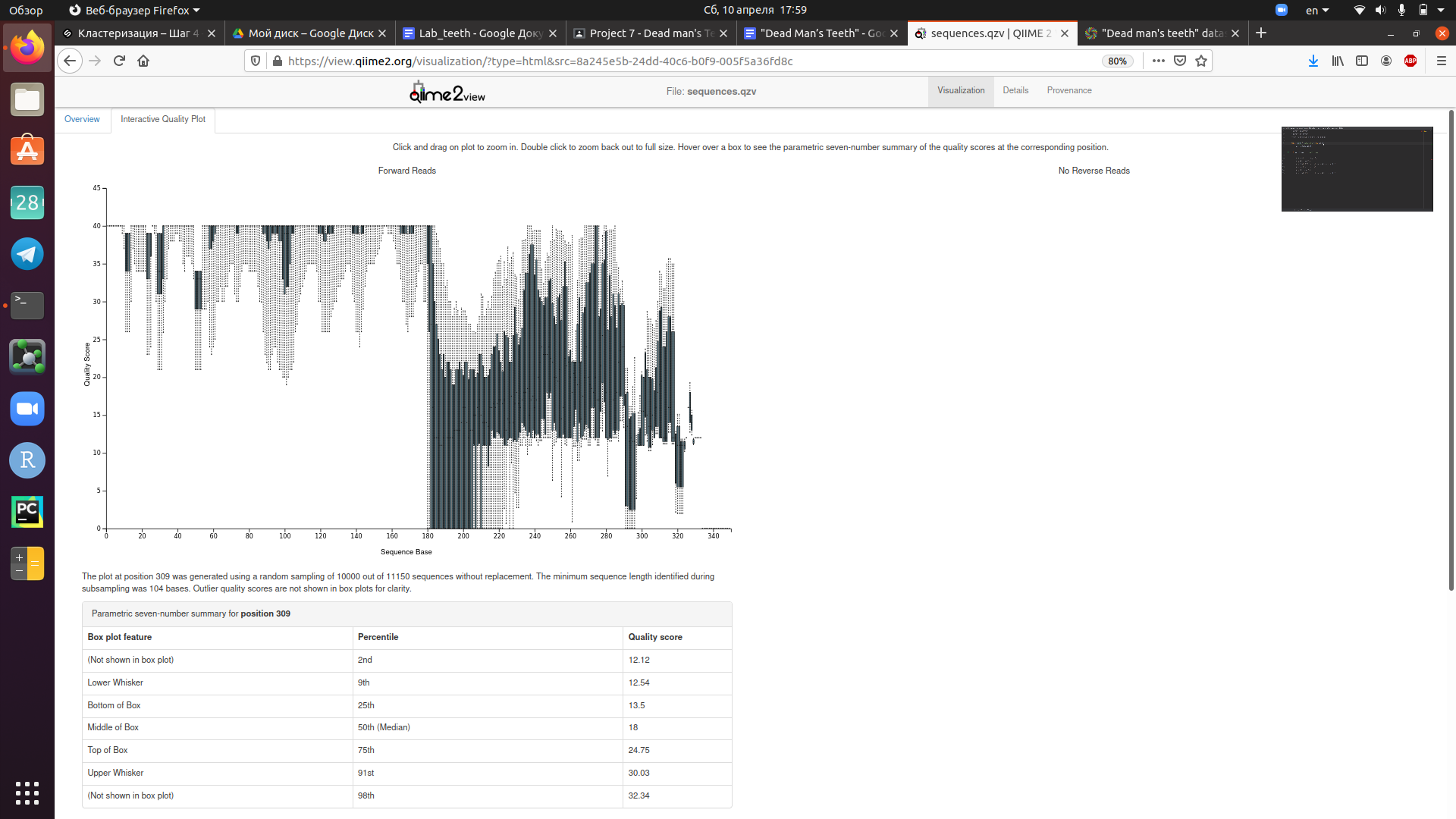
**Supplementary materials  
The history of human oral diseases through metagenomics analysis**

**PART 1. AMPLICON SEQUENCING**

1. **Installation of the Miniconda and QIIME2.**./Miniconda3-latest-Linux-x86\_64.sh # install conda  
   conda update conda  
   conda install wget  
   wget <https://data.qiime2.org/distro/core/qiime2-2021.2-py36-linux-conda.yml>  
   conda env create -n qiime2-2021.2 --file qiime2-2021.2-py36-linux-conda.yml  
   conda activate qiime2-2021.2 # to activate qiime, deactivate: conda deactivate
2. **Importing the data.**Changing of the absolute paths for fastq files in manifest.tsv.  
     
   # import  
   qiime tools import --type 'SampleData[SequencesWithQuality]' --input-path manifest.tsv --output-path sequences.qza --input-format SingleEndFastqManifestPhred33V2  
   # out  
   Imported manifest.tsv as SingleEndFastqManifestPhred33V2 to sequences.qza.# check correctness  
   qiime tools validate sequences.qza  
   # out  
   Result sequences.qza appears to be valid at level=max.
3. **Demultiplexing and QC.**# how many sequences were obtained per sample  
   qiime demux summarize --i-data sequences.qza --o-visualization sequences.qzv  
     
   # visualization of the distribution of sequence qualities  
     
   
4. **Feature table construction (and more QC).**# m = len(barcode) + len(primer) = 10 + 22 = 32  
   # n = len(sequence before quality fall) = 180 - 32 = 148  
   qiime dada2 denoise-single --i-demultiplexed-seqs sequences.qza --p-trim-left 32 --p-trunc-len 148 --o-representative-sequences rep-seqs.qza --o-table table.qza --o-denoising-stats stats.qza  
   # out  
   Saved FeatureTable[Frequency] to: table.qza  
   Saved FeatureData[Sequence] to: rep-seqs.qza  
   Saved SampleData[DADA2Stats] to: stats.qza  
     
   # how many reads passed filtration  
   qiime metadata tabulate --m-input-file stats.qza --o-visualization stats.qzv  
   # out  
   Saved Visualization to: stats.qzv  
   Table 1 - Summary of the filtration.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sample ID | Input | Filtered | % of input passed | Denoised | Non-chimeric | % of input non-chimeric |
| bone | 5788 | 5589 | 96.56 | 5377 | 5377 | 92.9 |
| calculus | 5362 | 5183 | 96.66 | 5059 | 4817 | 89.84 |

1. **FeatureTable and FeatureData summaries.**# visual summary  
   qiime feature-table summarize --i-table table.qza --o-visualization table.qzv --m-sample-metadata-file sample-metadata.tsv  
   # out  
   Saved Visualization to: table.qzv  
     
   Table 2 - Summary of the features.

|  |  |
| --- | --- |
| Number of samples | 2 |
| Number of features | 163 |
| Total frequency | 10,194 |

Table 3 - Frequency per sample.

|  |  |
| --- | --- |
| Minimum frequency | 4,817.0 |
| 1st quartile | 4,957.0 |
| Median frequency | 5,097.0 |
| 3rd quartile | 5,237.0 |
| Maximum frequency | 5,377.0 |
| Mean frequency | 5,097.0 |

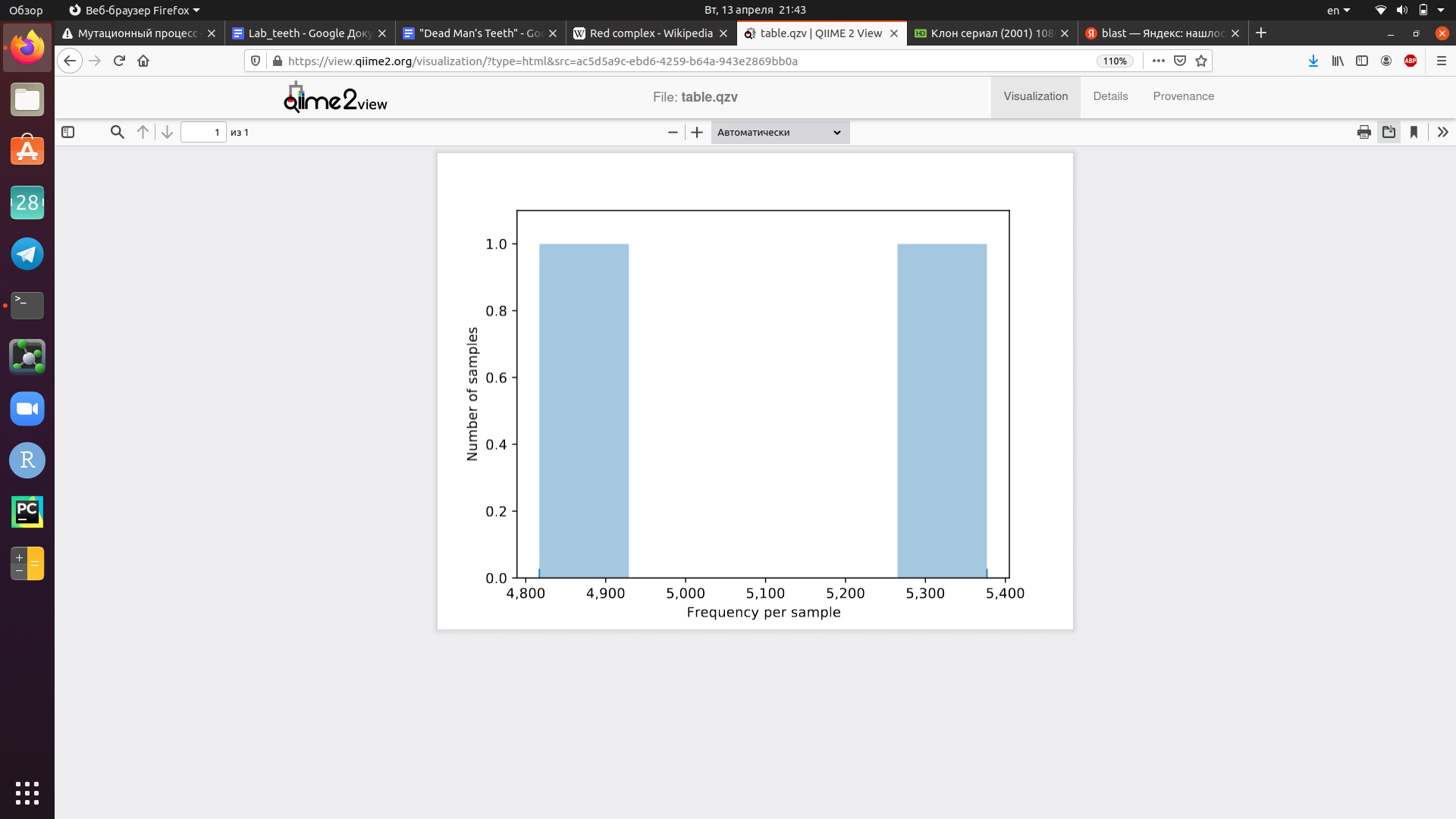
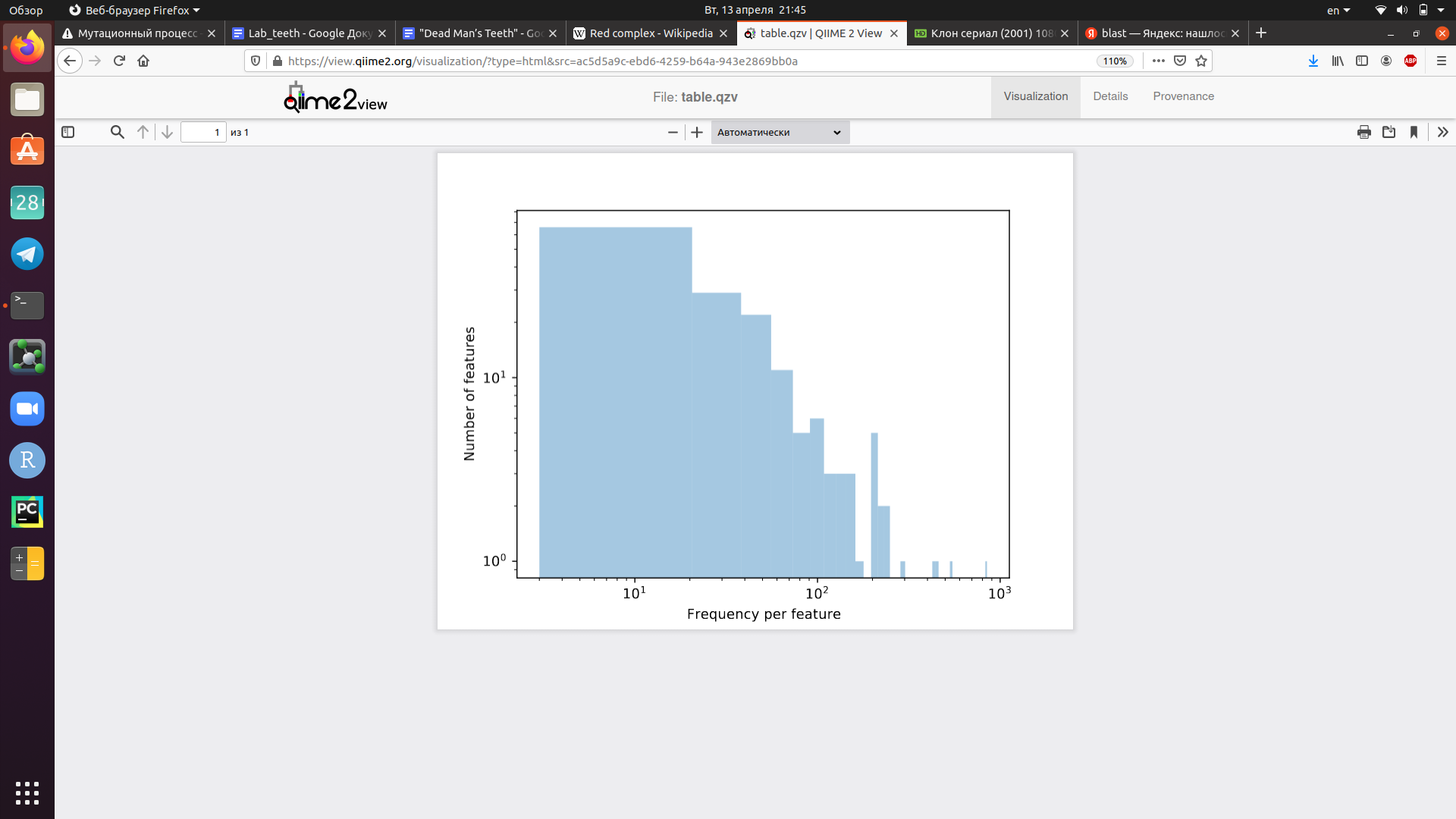
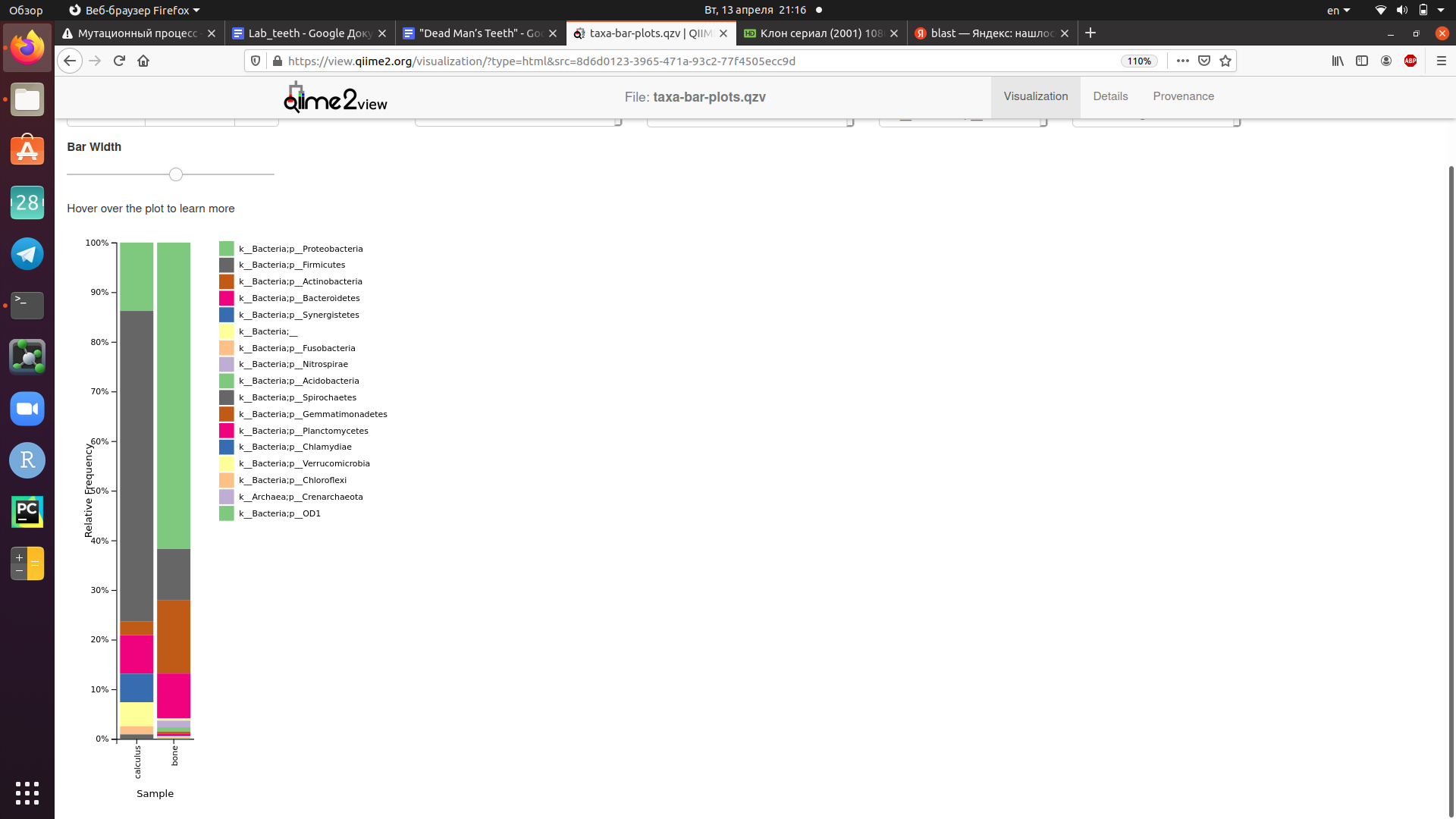


Table 4 - Frequency per feature.

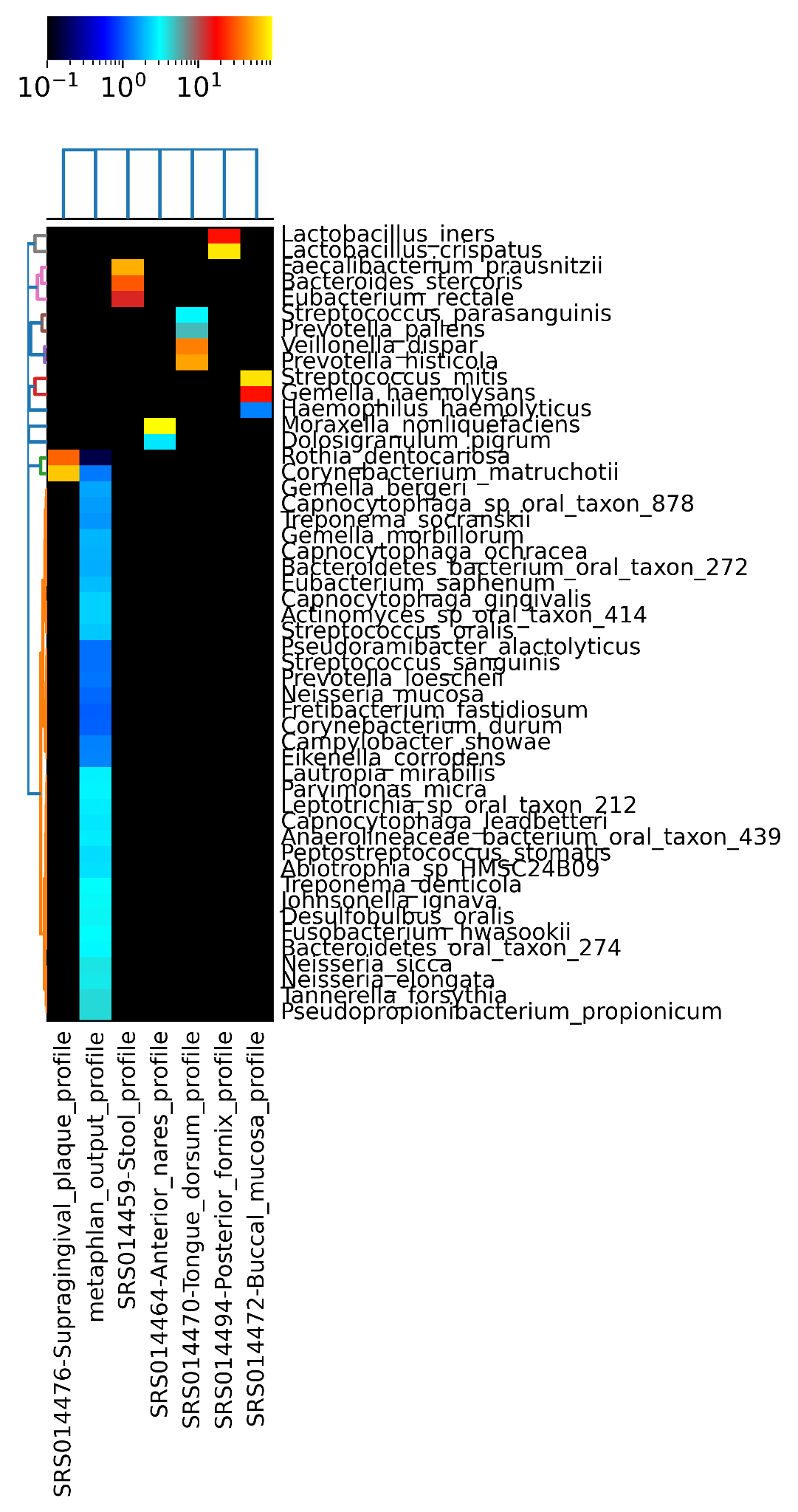
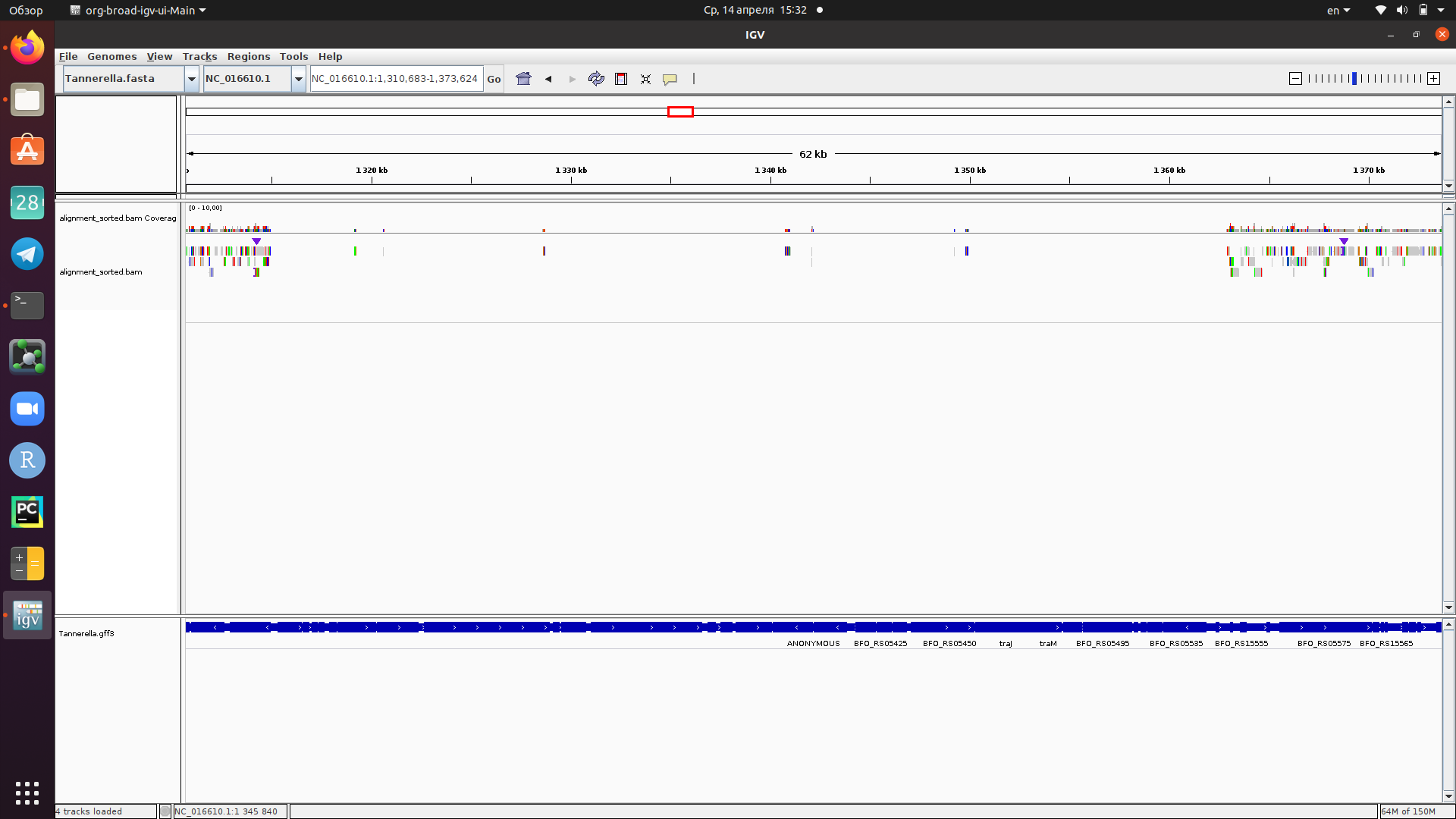
|  |  |
| --- | --- |
| Minimum frequency | 3.0 |
| 1st quartile | 13.0 |
| Median frequency | 29.0 |
| 3rd quartile | 61.5 |
| Maximum frequency | 847.0 |
| Mean frequency | 62.54 |

  
  
# mapping feature IDs to sequences  
qiime feature-table tabulate-seqs --i-data rep-seqs.qza --o-visualization rep-seqs.qzv  
# out  
Saved Visualization to: rep-seqs.qzv  
Representative sequences are saved in file “sequences.fasta” from qiime visualization site.

1. **Taxonomic analysis.**  
     
   # training feature classifier  
   qiime feature-classifier classify-sklearn --i-classifier gg-13-8-99-nb-classifier.qza --i-reads rep-seqs.qza --o-classification taxonomy.qza  
   # out  
   Plugin error from feature-classifier:  
   The scikit-learn version (0.22.1) used to generate this artifact does not match the current version of scikit-learn installed (0.23.1). Please retrain your classifier for your current deployment to prevent data-corruption errors.  
   # installation of the scikit-learn 0.22.1  
   pip uninstall scikit-learn  
   pip install scikit-learn==0.22.1  
   # training feature classifier  
   qiime feature-classifier classify-sklearn --i-classifier gg-13-8-99-nb-classifier.qza --i-reads rep-seqs.qza --o-classification taxonomy.qza  
   # out  
   Saved FeatureData[Taxonomy] to: taxonomy.qza  
     
   # visualization  
   qiime metadata tabulate --m-input-file taxonomy.qza --o-visualization taxonomy.qzv  
   # out  
   Saved Visualization to: taxonomy.qzv  
   # barplots  
   qiime taxa barplot \  
    --i-table table.qza \  
   --i-taxonomy taxonomy.qza \  
   --m-metadata-file sample-metadata.tsv \  
   --o-visualization taxa-bar-plots.qzv  
   # out  
   Saved Visualization to: taxa-bar-plots.qz  
   
2. **Bacterial Teamwork.**calculus | k\_\_Bacteria;p\_\_Bacteroidetes;c\_\_Bacteroidia;o\_\_Bacteroidales;f\_\_Porphyromonadaceae;g\_\_Tannerella;s\_\_ | 0.519%  
   calculus | k\_\_Bacteria;p\_\_Bacteroidetes;c\_\_Bacteroidia;o\_\_Bacteroidales;f\_\_Porphyromonadaceae;g\_\_Porphyromonas;s\_\_ | 0.291%  
   calculus | k\_\_Bacteria;p\_\_Spirochaetes;c\_\_Spirochaetes;o\_\_Spirochaetales;f\_\_Spirochaetaceae;g\_\_Treponema;s\_\_socranskii | 0.561%  
   calculus | k\_\_Bacteria;p\_\_Spirochaetes;c\_\_Spirochaetes;o\_\_Spirochaetales;f\_\_Spirochaetaceae;g\_\_Treponema;s\_\_ | 0.353%

**PART 2. SHOTGUN SEQUENCING**

1. **Shotgun sequence data profiling.**# downloading genome assembly  
   <https://www.dropbox.com/s/f5j52tliumt6etm/G12_assembly.fna.gz?dl=0>  
     
   # installing MetaPhlAn  
   pip install metaphlan  
   # alignment of the data to the microbiota database  
   metaphlan G12\_assembly.fna --input\_type fasta --nproc 2 > metaphlan\_output\_profile.txt
2. **Comparison with samples from HMP.**# downloading samples from HMP  
   <https://www.dropbox.com/s/aeq9fsoax68h9qk/SRS014459-Stool.fasta?dl=0>  
   <https://www.dropbox.com/s/r6du0tazu7ocyop/SRS014464-Anterior_nares.fasta?dl=0>  
   <https://www.dropbox.com/s/xyoy6oo12r9mjry/SRS014470-Tongue_dorsum.fasta?dl=0>  
   <https://www.dropbox.com/s/mwxyzjumb6gc9sz/SRS014472-Buccal_mucosa.fasta?dl=0>  
   <https://www.dropbox.com/s/ajt786ac2ijnra8/SRS014476-Supragingival_plaque.fasta?dl=0>  
   <https://www.dropbox.com/s/6oiowqeg447kq1j/SRS014494-Posterior_fornix.fasta?dl=0>  
     
   # processing samples from HMP  
   for f in \*.fasta; do metaphlan $f --input\_type fasta --nproc 2 > ${f%.fasta}\_profile.txt; done  
   # out  
   WARNING: The metagenome profile contains clades that represent multiple species merged into a single representant.   
   An additional column listing the merged species is added to the MetaPhlAn output.

1. **Visualization of the metaphlan results with a heat map.**# preparing of the file  
   merge\_metaphlan\_tables.py \*\_profile.txt > merged\_abundance\_table.txt  
     
   grep -E "s\_\_|clade" merged\_abundance\_table.txt | sed 's/^.\*s\_\_//g' | cut -f1,3-9 | sed -e 's/clade\_name/body\_site/g' > merged\_abundance\_table\_sp.txt  
     
   # installing tool for drawing heatmap  
   pip install hclust2  
     
   # heatmap  
   hclust2.py -i merged\_abundance\_table\_sp.txt -o abundance\_heatmap\_sp.png --f\_dist\_f braycurtis --s\_dist\_f braycurtis --cell\_aspect\_ratio 0.5 -l --flabel\_size 10 --slabel\_size 10 --max\_flabel\_len 100 --max\_slabel\_len 100 --minv 0.1 --dpi 600 --ftop 50  
   
2. **Comparison with ancient Tannerella forsythia genome.**# downloading Tannerella genome  
   from <https://www.ncbi.nlm.nih.gov/nuccore/NC_016610.1>  
   Saved in fasta and GFF3.  
     
   # indexing the genome  
   bwa index Tannerella.fasta  
   bwa index Tannerella.gff3  
     
   # alignment  
   bwa mem Tannerella.fasta G12\_assembly.fna > alignment.sam  
     
   # sam to bam, statistics, sort and indexing  
   samtools view -S -b alignment.sam > alignment.bam  
   samtools flagstat alignment.bam  
   samtools sort alignment.bam -o alignment\_sorted.bam  
   samtools index alignment\_sorted.bam  
     
   # visualization in IGVbrowser  
   Added files Tannerella.fasta, Tannerella.gff3, alignment\_sorted.bam  
     
     
   # bam to bed  
   bedtools bamtobed -i alignment\_sorted.bam > alignment\_sorted.bed  
     
   # intersection  
    bedtools intersect -v -a Tannerella.gff3 -b alignment\_sorted.bed > intersected.bed  
     
   # extraction of the gene products  
   grep "CDS" intersect.bed | cut -f 9 | grep -o 'product=[^;]\*' | awk -F'=' '{print $2}' | sort | uniq  
     
   Table 5 - New features of the modern T. forsythia genome.

|  |
| --- |
| **Product of a gene** |
| 6-bladed beta-propeller |
| Abi family protein |
| AIPR family protein |
| antirestriction protein ArdA |
| ATP-binding protein |
| beta-ketoacyl-ACP synthase III |
| cell filamentation protein Fic |
| class I lanthipeptide |
| conjugal transfer protein TraO |
| conjugative transposon protein TraJ |
| conjugative transposon protein TraK |
| conjugative transposon protein TraM |
| conjugative transposon protein TraN |
| DDE-type integrase/transposase/recombinase |
| dihydrofolate reductase family protein |
| DNA mismatch endonuclease Vsr |
| DUF1566 domain-containing protein |
| DUF1896 domain-containing protein |
| DUF3289 family protein |
| DUF3408 domain-containing protein |
| DUF3872 domain-containing protein |
| DUF3873 domain-containing protein |
| DUF3876 domain-containing protein |
| DUF3989 domain-containing protein |
| DUF4133 domain-containing protein |
| DUF4134 domain-containing protein |
| DUF4141 domain-containing protein |
| DUF418 domain-containing protein |
| DUF4372 domain-containing protein |
| DVUA0089 family protein |
| four helix bundle protein |
| GIY-YIG nuclease family protein |
| GNAT family N-acetyltransferase |
| helicase |
| helix-turn-helix domain containing protein |
| helix-turn-helix domain-containing protein |
| histidinol phosphate phosphatase |
| hypothetical protein |
| IS110 family transposase |
| IS1380 family transposase |
| IS1595 family transposase |
| IS1595-like element ISTfo1 family transposase |
| IS1 family transposase |
| IS4-like element IS421 family transposase |
| IS5 family transposase |
| ISL3 family transposase |
| lanthionine synthetase C family protein |
| lantibiotic dehydratase family protein |
| lysozyme |
| nuclear transport factor 2 family protein |
| NVEALA domain-containing protein |
| outer membrane lipoprotein-sorting protein |
| ParA family protein |
| PcfJ family protein |
| PcfK-like family protein |
| PD-(D/E)XK motif protein |
| PH domain-containing protein |
| phosphorylase family |
| putative DNA binding domain-containing protein |
| radical SAM peptide maturase |
| relaxase/mobilization nuclease domain-containing protein |
| response regulator |
| rhodanese-like domain-containing protein |
| RND family transporter |
| RteC domain-containing protein |
| sigma-54-dependent Fis family transcriptional regulator |
| site-specific integrase |
| SPASM domain-containing protein |
| SusD/RagB family nutrient-binding outer membrane lipoprotein |
| T9SS type A sorting domain-containing protein |
| TetR/AcrR family transcriptional regulator |
| tetracycline resistance ribosomal protection protein |
| thiopeptide-type bacteriocin biosynthesis protein |
| TIGR04149 family rSAM-modified RiPP |
| TIGR04150 pseudo-rSAM protein |
| TIGR04157 family glycosyltransferase |
| TonB-dependent receptor |
| toprim domain-containing protein |
| transposase |
| Z1 domain-containing protein |