Project 7. Dead Man’s Teeth. Introduction to metagenomics analysis.

Tags

PART 1B DADA2-based

Analysis of the DNA from the material underneath the dental calculus. results of sequencing portions of V5 16S ribosomal RNA obtained by an instrument Roche GS Junior (454).

```
mkdir Project7
cd Project7
mkdir raw_data
cd raw_data
```

Download the raw sequencing data of the samples from dental calculus from
https://figshare.com/articles/dataset/_Dead_man_s_teeth_dataset/12152040

Analyse the downloaded data with the code in R, which can be found here

Silva libraries can be downloaded here

As a result we obtained 4 files: asv_table, metadata, rep_seqs, tax_table

As we can see, there are two samples - B61 and G12 with periodontal disease.

#NAME	sample.id	BarcodeSequence	LinkerPrimerSequence	Type	Individual	Periodontal_disease
SRR957750.fastq	S14-V5-P-B17-calc	ACGAGTGCGT	CAGGATTAGATACCCTGGTAGTCC	Calculus	B17	No
SRR957753.fastq	S15-V5-R-B78-calc	ACGAGTGCGT	CAGGATTAGATACCCTGGTAGTCC	Calculus	B78	No
SRR957756.fastq	S18S19-V5-L-B17-root	ACGAGTGCGT	CAGGATTAGATACCCTGGTAGTCC	Root	B17	No
SRR957760.fastq	S22S23-V5-N-B78-root	ACGAGTGCGT	CAGGATTAGATACCCTGGTAGTCC	Root	B78	No
SRR986774.fastq	S10-V5-Q-B61-calc	ACGAGTGCGT	CAGGATTAGATACCCTGGTAGTCC	Calculus	B61	Yes
SRR986778.fastq	S16S17-V5-K1-G12-root	ACGAGTGCGT	CAGGATTAGATACCCTGGTAGTCC	Root	G12	Yes
SRR986779.fastq	S16S17-V5-K2-G12-root	ACGAGTGCGT	CAGGATTAGATACCCTGGTAGTCC	Root	G12	Yes
SRR986782.fastq	S20S21-V5-M-B61-root	ACGAGTGCGT	CAGGATTAGATACCCTGGTAGTCC	Root	B61	Yes
SRR986773.fastq	S8-V5-O-G12-calc	ACGAGTGCGT	CAGGATTAGATACCCTGGTAGTCC	Calculus	G12	Yes

PART 2

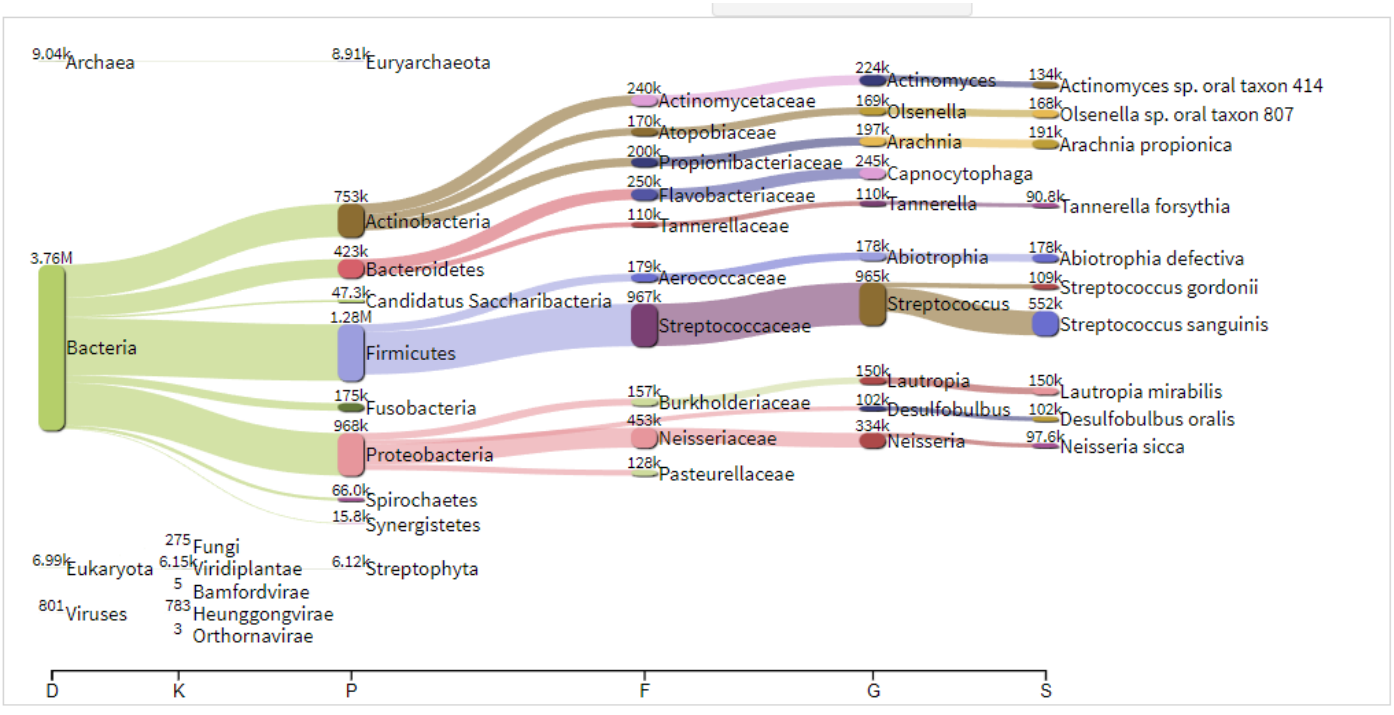
An affected individual G12 was selected for a dental calculus whole metagenome shotgun sequencing, and reads were assembled into contigs. We will skip the actual assembly process, because the raw data is too large and it will take a lot of time. Assembly results can be downloaded here

1. Shotgun sequence data profiling

The problem with working with kraken is the very large size of the databases. So we have already calculated results for this sample - report and taxonomic prediction for each sequence.

2. Visualization of the Kraken results as a Sankey diagram

We used the Pavian web application to visualise the classification results obtained with Kraken.



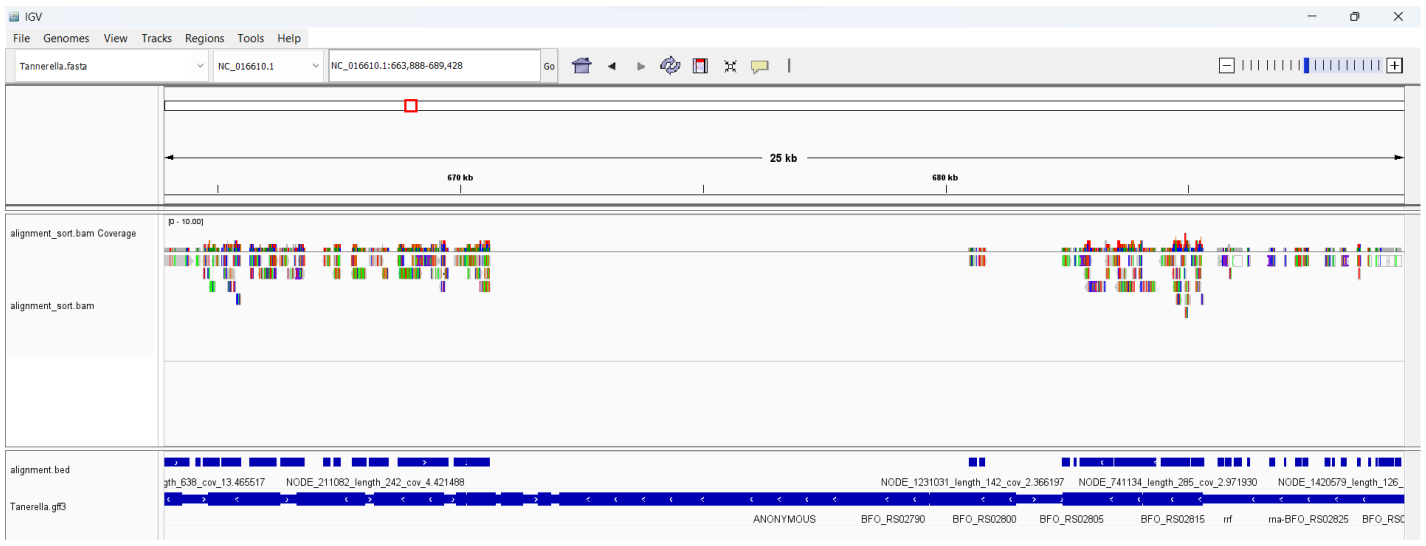
3. Comparison with ancient *Tannerella forsythia* genome

Our shotgun assembly is still pretty fragmented, so we will have to align our contigs to reference. We downloaded data for the *T. forsythia* strain (there is only one complete genome in GenBank so far) - we will need the genome itself (fasta) and annotation (GFF3).

After that we aligned contigs on the downloaded reference

```
# indexing file
bwa index Tannerella.fasta

# aligning contigs to the reference genome of Tannerella
bwa mem -t 8 ./Tannerella.fasta ../G12_assembly.fna > alignment.sam

# convert bam to sam
samtools view -b -S -h alignment.sam > alignment.bam

# sort bam file
samtools sort alignment.bam -o alignment_sort.bam

# convert sorted bam to bed file format
bedtools bamtobed -i alignment_sort.bam > alignment.bed

# find intersections automatically
bedtools intersect -a ../sequence.gff3 -b alignment.bed -v > intersections.bed
# index bam file
samtools index -b alignment_sort.bam
# select protein accessions
awk -F'\t' '$3 == "CDS" {split($NF, a, "ID=cds-"); split(a[2], b, ";"); print b[1]}' intersec
tions.bed > selected_intersections.txt
```

Alignment visualization with IGV shows what the coverage looks like. We can see that some of the regions in the modern strain have zero coverage, and probably were obtained during the strain evolution.



WP_001300563.1 MULTISPECIES: IS4-like element IS421 family transposase [Bacteria]
WP_004584212.1 MULTISPECIES: hypothetical protein [Bacteroidales]
WP_004584877.1 MULTISPECIES: Abi family protein [Bacteroidales]
WP_005944180.1 MULTISPECIES: conjugal transfer protein MobA [Bacteroidota]
WP_007366490.1 MULTISPECIES: hypothetical protein [Bacteroidota]
WP_007366491.1 MULTISPECIES: hypothetical protein [Bacteroidota]
WP_007366492.1 MULTISPECIES: PcfK-like family protein [Bacteroidota]
WP_007366494.1 MULTISPECIES: hypothetical protein [Bacteroidota]
WP_007366496.1 MULTISPECIES: hypothetical protein [Bacteroidota]
WP_007366497.1 MULTISPECIES: hypothetical protein [Bacteroidota]
WP_007366498.1 MULTISPECIES: hypothetical protein [Bacteroidota]
WP_007366499.1 MULTISPECIES: lysozyme [Bacteroidota]
WP_007366500.1 MULTISPECIES: DUF3872 domain-containing protein [Bacteroidota]
WP_007366501.1 MULTISPECIES: toprim domain-containing protein [Bacteroidota]
WP_007366502.1 MULTISPECIES: conjugal transfer protein TraO [Bacteroidota]
WP_007366503.1 MULTISPECIES: conjugative transposon protein TraN [Bacteroidota]
WP_007366504.1 MULTISPECIES: conjugative transposon protein TraM [Bacteroidota]
WP_007366505.1 MULTISPECIES: TraL conjugative transposon family protein [Bacteroidota]
WP_007366506.1 MULTISPECIES: conjugative transposon protein TraK [Bacteroidota]
WP_007366507.1 MULTISPECIES: conjugative transposon protein TraJ [Bacteroidota]
WP_007366508.1 MULTISPECIES: DUF4141 domain-containing protein [Bacteroidota]
WP_007366509.1 MULTISPECIES: DUF3876 domain-containing protein [Bacteroidota]
WP_007366512.1 MULTISPECIES: DUF4133 domain-containing protein [Bacteroidota]
WP_007366514.1 MULTISPECIES: hypothetical protein [Bacteroidota]
WP_007366516.1 MULTISPECIES: DUF3408 domain-containing protein [Bacteroidota]
WP_007366517.1 MULTISPECIES: ParA family protein [Bacteroidales]
WP_007366518.1 MULTISPECIES: conjugal transfer protein MobB [Bacteroidota]
WP_007366520.1 MULTISPECIES: hypothetical protein [Bacteroidota]
WP_007366521.1 MULTISPECIES: ATP-binding protein [Bacteroidota]
WP_007366522.1 MULTISPECIES: RteC domain-containing protein [Bacteroidota]
WP_007366523.1 MULTISPECIES: dihydrofolate reductase family protein [Bacteroidota]
WP_007366524.1 MULTISPECIES: sigma-54 dependent transcriptional regulator [Bacteroidota]
WP_007366525.1 MULTISPECIES: ATP-binding protein [Bacteroidota]
WP_007366526.1 MULTISPECIES: tetracycline resistance ribosomal protection protein [Bacteroidota]
WP_007366527.1 MULTISPECIES: GNAT family N-acetyltransferase [Bacteroidota]
WP_007366528.1 MULTISPECIES: nuclear transport factor 2 family protein [Bacteroidota]
WP_007366531.1 MULTISPECIES: DUF1896 domain-containing protein [Bacteroidota]
WP_007366534.1 MULTISPECIES: helix-turn-helix domain-containing protein [Bacteroidota]
WP_007366536.1 MULTISPECIES: helix-turn-helix domain-containing protein [Bacteroidales]
WP_007366538.1 MULTISPECIES: hypothetical protein [Bacteroidota]
WP_007366539.1 MULTISPECIES: site-specific integrase [Bacteroidota]

WP_009016663.1 MULTISPECIES: DUF4134 domain-containing protein [Bacteroidota]
WP_010956331.1 MULTISPECIES: helix-turn-helix domain-containing protein [Bacteroidales]
WP_014223573.1 helix-turn-helix domain-containing protein [Tannerella forsythia]
WP_014223582.1 TIGR04157 family glycosyltransferase [Tannerella forsythia]
WP_014223583.1 lanthionine synthetase C family protein [Tannerella forsythia]
WP_014223598.1 hypothetical protein [Tannerella forsythia]
WP_014223665.1 four helix bundle protein [Tannerella forsythia]
WP_014223716.1 hypothetical protein [Tannerella forsythia]
WP_014223804.1 hypothetical protein [Tannerella forsythia]
WP_014223806.1 NVEALA domain-containing protein [Tannerella forsythia]
WP_014223811.1 6-bladed beta-propeller [Tannerella forsythia]
WP_014223814.1 hypothetical protein [Tannerella forsythia]
WP_014223816.1 6-bladed beta-propeller [Tannerella forsythia]
WP_014224019.1 hypothetical protein [Tannerella forsythia]
WP_014224021.1 hypothetical protein [Tannerella forsythia]
WP_014224073.1 AIPR family protein [Tannerella forsythia]
WP_014224074.1 PD-(D/E)XK motif protein [Tannerella forsythia]
WP_014224075.1 MULTISPECIES: Z1 domain-containing protein [Bacteroidales]
WP_014224076.1 MULTISPECIES: ATP-binding protein [Bacteroidales]
WP_014224136.1 phosphorylase family [Tannerella forsythia]
WP_014224267.1 hypothetical protein [Tannerella forsythia]
WP_014224286.1 radical SAM peptide maturase [Tannerella forsythia]
WP_014224287.1 TIGR04150 pseudo-rSAM protein [Tannerella forsythia]
WP_014224298.1 hypothetical protein [Tannerella forsythia]
WP_014224299.1 galactosyltransferase-related protein [Tannerella forsythia]
WP_014224551.1 rhodanese-like domain-containing protein [Tannerella forsythia]
WP_014224553.1 IS1595 family transposase [Tannerella forsythia]
WP_014224586.1 hypothetical protein [Tannerella forsythia]
WP_014224629.1 MULTISPECIES: helix-turn-helix domain-containing protein [Bacteroidota]
WP_014224631.1 MULTISPECIES: helix-turn-helix domain-containing protein [Bacteroidales]
WP_014224641.1 DUF3873 domain-containing protein [Tannerella forsythia]
WP_014224642.1 MULTISPECIES: PcfJ domain-containing protein [Bacteroidota]
WP_014224645.1 MULTISPECIES: hypothetical protein [Bacteroidales]
WP_014224782.1 hypothetical protein [Tannerella forsythia]
WP_014224885.1 ISL3 family transposase [Tannerella forsythia]
WP_014225126.1 hypothetical protein [Tannerella forsythia]
WP_014225157.1 IS110 family transposase [Tannerella forsythia]
WP_014225252.1 hypothetical protein [Tannerella forsythia]
WP_014225286.1 hypothetical protein [Tannerella forsythia]
WP_014225373.1 hypothetical protein [Tannerella forsythia]
WP_014225380.1 hypothetical protein [Tannerella forsythia]
WP_014225422.1 hypothetical protein [Tannerella forsythia]
WP_014225661.1 SusD/RagB family nutrient-binding outer membrane lipoprotein [Tannerella forsythia]
WP_014225702.1 hypothetical protein [Tannerella forsythia]
WP_014225810.1 hypothetical protein [Tannerella forsythia]
WP_014225909.1 hypothetical protein [Tannerella forsythia]
WP_014226260.1 hypothetical protein [Tannerella forsythia]
WP_014226274.1 hypothetical protein [Tannerella forsythia]
WP_014226277.1 hypothetical protein [Tannerella forsythia]
WP_014226278.1 beta-ketoacyl-ACP synthase III [Tannerella forsythia]
WP_014226279.1 MMPL family transporter [Tannerella forsythia]
WP_014226280.1 hypothetical protein [Tannerella forsythia]
WP_014226283.1 MULTISPECIES: hypothetical protein [Bacteroidales]
WP_014226298.1 MULTISPECIES: hypothetical protein [Bacteroidales]
WP_014226302.1 hypothetical protein [Tannerella forsythia]
WP_014226304.1 conjugal transfer protein TraO [Tannerella forsythia]
WP_014226305.1 DUF3872 domain-containing protein [Tannerella forsythia]
WP_014226309.1 MULTISPECIES: type VI secretion system tube protein TssD [Bacteroidales]

WP_014226312.1 hypothetical protein [Tannerella forsythia]
WP_014226314.1 hypothetical protein [Tannerella forsythia]
WP_021644789.1 MULTISPECIES: outer membrane lipoprotein-sorting protein [Bacteroidales]
WP_025880900.1 MULTISPECIES: hypothetical protein [Bacteroidales]
WP_028899187.1 MULTISPECIES: DUF3408 domain-containing protein [Bacteroidota]
WP_041590503.1 class I lanthipeptide [Tannerella forsythia]
WP_041590506.1 lantibiotic dehydratase family protein [Tannerella forsythia]
WP_041590507.1 thiopeptide-type bacteriocin biosynthesis protein [Tannerella forsythia]
WP_041590516.1 hypothetical protein [Tannerella forsythia]
WP_041590537.1 IS1595-like element ISTfo1 family transposase [Tannerella forsythia]
WP_041590544.1 histidinol phosphate phosphatase [Tannerella forsythia]
WP_041590612.1 MULTISPECIES: hypothetical protein [Bacteroidales]
WP_041590680.1 hypothetical protein [Tannerella forsythia]
WP_041590710.1 IS1595-like element ISTfo1 family transposase [Tannerella forsythia]
WP_041590794.1 hypothetical protein [Tannerella forsythia]
WP_041590818.1 DUF4372 domain-containing protein [Tannerella forsythia]
WP_041590833.1 hypothetical protein [Tannerella forsythia]
WP_041590836.1 hypothetical protein [Tannerella forsythia]
WP_041590907.1 DUF4974 domain-containing protein [Tannerella forsythia]
WP_041590940.1 hypothetical protein [Tannerella forsythia]
WP_041590952.1 hypothetical protein [Tannerella forsythia]
WP_041590957.1 outer membrane beta-barrel family protein [Tannerella forsythia]
WP_041590961.1 MULTISPECIES: type VI secretion system tube protein TssD [Bacteroidales]
WP_041590964.1 hypothetical protein [Tannerella forsythia]
WP_041590984.1 IS1380 family transposase [Tannerella forsythia]
WP_041590995.1 lanthionine synthetase LanC family protein [Tannerella forsythia]
WP_041591157.1 IS110 family transposase [Tannerella forsythia]
WP_051322484.1 MULTISPECIES: TetR/AcrR family transcriptional regulator [Bacteroidales]
WP_052299227.1 hypothetical protein [Tannerella forsythia]
WP_052299228.1 hypothetical protein [Tannerella forsythia]
WP_052299234.1 ISL3 family transposase [Tannerella forsythia]
WP_052299279.1 Arm DNA-binding domain-containing protein [Tannerella forsythia]
WP_052299290.1 DUF1566 domain-containing protein [Tannerella forsythia]
WP_074453014.1 hypothetical protein [Tannerella forsythia]
WP_080561826.1 hypothetical protein [Tannerella forsythia]
WP_080561836.1 NVEALA domain-containing protein [Tannerella forsythia]
WP_099046116.1 IS1 family transposase [Tannerella forsythia]
WP_099046118.1 IS1 family transposase [Tannerella forsythia]
WP_099046121.1 IS1 family transposase [Tannerella forsythia]
WP_099046123.1 IS1 family transposase [Tannerella forsythia]
WP_099046125.1 DUF3289 family protein [Tannerella forsythia]
WP_143596664.1 hypothetical protein [Tannerella forsythia]
WP_157755267.1 hypothetical protein [Tannerella forsythia]
WP_157755299.1 hypothetical protein [Tannerella forsythia]
WP_157755307.1 hypothetical protein [Tannerella forsythia]
WP_157755324.1 hypothetical protein [Tannerella forsythia]
WP_157755326.1 hypothetical protein [Tannerella forsythia]
WP_157755327.1 hypothetical protein [Tannerella forsythia]
WP_167536494.1 hypothetical protein [Tannerella forsythia]
WP_208854691.1 hypothetical protein [Tannerella forsythia]
WP_208854693.1 hypothetical protein [Tannerella forsythia]
WP_231964057.1 transposase [Tannerella forsythia]
WP_231964473.1 hypothetical protein [Tannerella forsythia]
WP_236684403.1 hypothetical protein [Tannerella forsythia]
WP_236684414.1 transposase [Tannerella forsythia]
WP_244262912.1 transposase [Tannerella forsythia]
WP_244262913.1 hypothetical protein [Tannerella forsythia]
WP_244262916.1 hypothetical protein [Tannerella forsythia]

WP_244262917.1 hypothetical protein [Tannerella forsythia]
WP_244262918.1 hypothetical protein [Tannerella forsythia]
WP_244262919.1 hypothetical protein [Tannerella forsythia]
WP_244262921.1 DVUA0089 family protein [Tannerella forsythia]
WP_244262922.1 hypothetical protein [Tannerella forsythia]
WP_244262923.1 radical SAM protein [Tannerella forsythia]
WP_244262924.1 PH domain-containing protein [Tannerella forsythia]
WP_244262932.1 hypothetical protein [Tannerella forsythia]
WP_244262935.1 IS4 family transposase [Tannerella forsythia]
WP_244262939.1 very short patch repair endonuclease [Tannerella forsythia]
WP_244262944.1 helix-turn-helix domain-containing protein [Tannerella forsythia]
WP_244262945.1 transposase [Tannerella forsythia]
WP_244262972.1 antirestriction protein ArdA [Tannerella forsythia]
WP_244262973.1 hypothetical protein [Tannerella forsythia]
WP_262508926.1 hypothetical protein [Tannerella forsythia]