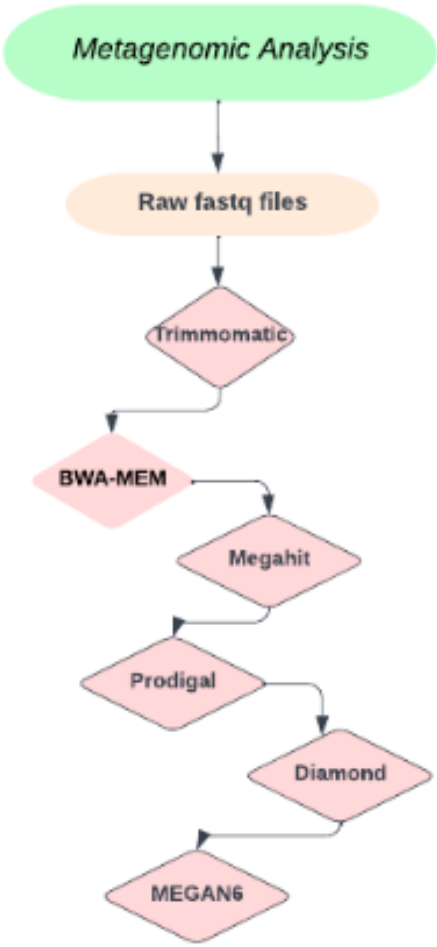
Metagenomics Project

The dataset I used for this Metagenomics Analysis are SRR3586068, SRR3586069, SRR3586070 which are oral cancer and one normal cancer sample from Human oral microbiome produced by Illumina sequencing technology. I performed the workflow with different tools. All the analysis part was done on Google cloud server; PATH : /home/vaidehipatel46/final_exam

The proposed workflow for Metagenomic Analyis was created by using LUCIDchart



All Tools were installed into google cloud server using conda;
Installing Conda is very useful to install all the bioinformatics tool required for this study, https://docs.anaconda.com/anaconda/install/linux/. Conda has all the tool that can be installed by just one or two command line and it also has all the dependencies

- conda install -c bioconda trimmomatic
- https://anaconda.org/bioconda/bwa

- conda install -c bioconda megahit
- conda install -c bioconda prodigal
- `conda install -c bioconda diamond`
- https://software-ab.cs.uni-tuebingen.de/download/megan6/welcome.html-
- `MEGAN6 was installed as Desktop Version`

Trimmomatic – A flexible read trimming tools for NGS tool

vaidehipatel46@class-allhomework:~/final_exam$ trimmomatic SE -phred33 SRR3586068.fastq SRR3586068_trim.fastq LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

TrimmomaticSE: Started with arguments:

-phred33 SRR3586068.fastq SRR3586068_trim.fastq LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

Automatically using 2 threads

Input Reads: 323954 Surviving: 303367 (93.65%) Dropped: 20587 (6.35%)

TrimmomaticSE: Completed successfully

vaidehipatel46@class-allhomework:~/final_exam$ trimmomatic SE -phred33 SRR3586069.fastq SRR3586069_trim.fastq LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

TrimmomaticSE: Started with arguments:

-phred33 SRR3586069.fastq SRR3586069_trim.fastq LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

Automatically using 2 threads

Input Reads: 818738 Surviving: 745635 (91.07%) Dropped: 73103 (8.93%)

vaidehipatel46@class-allhomework:~/final_exam$ trimmomatic SE -phred33 SRR3586070.fastq SRR3586070_trim.fastq LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

TrimmomaticSE: Started with arguments:

-phred33 SRR3586070.fastq SRR3586070_trim.fastq LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36Automatically using 2 threads

Input Reads: 455850 Surviving: 428095 (93.91%) Dropped: 27755 (6.09%)

Getting the Human Reference genome(Latest version of human reference genome)

vaidehipatel46@class-allhomework:~$ wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.40_GRCh38.p14/GCF_000001405.40_GRCh38.p14_genomic.fna.gz
--2022-12-10 18:50:03--
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.40_GRCh38.p14/GCF_000001405.40_GRCh38.p14_genomic.fna.gz
Resolving ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)... 130.14.250.13, 130.14.250.11, 2607:f220:41f:250::228,
...
Connecting to ftp.ncbi.nlm.nih.gov (ftp.ncbi.nlm.nih.gov)|130.14.250.13|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 972898531 (928M) [application/x-gzip]
Saving to: ‘GCF_000001405.40_GRCh38.p14_genomic.fna.gz’

GCF_000001405.40_GRCh38.p14 100%[===>] 927.83M 94.4MB/s in 12s

2022-12-10 18:50:15 (75.5 MB/s) - ‘GCF_000001405.40_GRCh38.p14_genomic.fna.gz’ saved [972898531/972898531]

vaidehipatel46@class-allhomework:~$ unzip GCF_000001405.40_GRCh38.p14_genomic.fna.gz
Archive: GCF_000001405.40_GRCh38.p14_genomic.fna.gz
End-of-central-directory signature not found. Either this file is not
a zipfile, or it constitutes one disk of a multi-part archive. In the
latter case the central directory and zipfile comment will be found on
the last disk(s) of this archive.
unzip: cannot find zipfile directory in one of GCF_000001405.40_GRCh38.p14_genomic.fna.gz or
GCF_000001405.40_GRCh38.p14_genomic.fna.gz.zip, and cannot find
GCF_000001405.40_GRCh38.p14_genomic.fna.gz.ZIP, period.

(base) vaidehipatel46@class-allhomework:~$ mv GCF_000001405.40_GRCh38.p14_genomic.fna final_exam

BWA- Software for mapping sequences against the reference genome.

1. Index the Genome file is required for BWA Mapper
2. BWA-MEM is fast to map the genome

https://bio-bwa.sourceforge.net/bwa.shtml

BWA Index (indexing the genome file)

```
(base) vaidehipatel46@class-allhomework:~/final_exam$ bwa index
GCF_000001405.40_GRCh38.p14_genomic.fna
[bwa_index] Pack FASTA... 33.01 sec
[bwa_index] Construct BWT for the packed sequence...
[BWTIncCreate] textLength=6596861272, availableWord=476179232
[BWTIncConstructFromPacked] 10 iterations done. 99999992 characters processed.
[BWTIncConstructFromPacked] 20 iterations done. 199999992 characters processed.
[BWTIncConstructFromPacked] 30 iterations done. 299999992 characters processed.
[BWTIncConstructFromPacked] 40 iterations done. 399999992 characters processed.
[BWTIncConstructFromPacked] 50 iterations done. 499999992 characters processed.
[BWTIncConstructFromPacked] 60 iterations done. 599999992 characters processed.
[BWTIncConstructFromPacked] 70 iterations done. 699999992 characters processed.
[BWTIncConstructFromPacked] 80 iterations done. 799999992 characters processed.
[BWTIncConstructFromPacked] 90 iterations done. 899999992 characters processed.
[BWTIncConstructFromPacked] 100 iterations done. 999999992 characters
processed.
[BWTIncConstructFromPacked] 110 iterations done. 1099999992 characters
processed.
[BWTIncConstructFromPacked] 120 iterations done. 1199999992 characters
processed.
[BWTIncConstructFromPacked] 130 iterations done. 1299999992 characters
processed.
[BWTIncConstructFromPacked] 140 iterations done. 1399999992 characters
processed.
[BWTIncConstructFromPacked] 150 iterations done. 1499999992 characters
processed.
^[[BWTIncConstructFromPacked] 160 iterations done. 1599999992 characters
processed.
[BWTIncConstructFromPacked] 170 iterations done. 1699999992 characters
processed.
[BWTIncConstructFromPacked] 180 iterations done. 1799999992 characters
processed.
[BWTIncConstructFromPacked] 190 iterations done. 1899999992 characters
processed.
```

```
[BWTIncConstructFromPacked] 200 iterations done. 1999999992 characters
processed.
[BWTIncConstructFromPacked] 210 iterations done. 2099999992 characters
processed.
[BWTIncConstructFromPacked] 220 iterations done. 2199999992 characters
processed.
[BWTIncConstructFromPacked] 230 iterations done. 2299999992 characters
processed.
[BWTIncConstructFromPacked] 240 iterations done. 2399999992 characters
processed.
[bwt_gen] Finished constructing BWT in 728 iterations.
[bwa_index] 3875.70 seconds elapse.
[bwa_index] Update BWT... 25.11 sec
[bwa_index] Pack forward-only FASTA... 19.73 sec
[bwa_index] Construct SA from BWT and Occ...
2179.09 sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa index GCF_000001405.40_GRCh38.p14_genomic.fna
[main] Real time: 6239.456 sec; CPU: 6132.635 sec
```

The output of this was all these index file of Human reference genome

GCF_000001405.40_GRCh38.p14_genomic.fna
GCF_000013425.1_ASM1342v1_genomic.fna.bwt
GCF_000001405.40_GRCh38.p14_genomic.fna.amb
GCF_000013425.1_ASM1342v1_genomic.fna.pac
GCF_000001405.40_GRCh38.p14_genomic.fna.ann
GCF_000013425.1_ASM1342v1_genomic.fna.sa

BWA-MEM Mapper

```
(base) vaidehipatel46@class-allhomework:~/final_exam$ bwa mem
GCF_000001405.40_GRCh38.p14_genomic.fna SRR3586068_trim.fastq >
SRR3586068_trim_bwa_ref.sam
[M::bwa_idx_load_from_disk] read 0 ALT contigs
[M::process] read 100360 sequences (10000173 bp)...
[M::process] read 100124 sequences (10000010 bp)...
[M::mem_process_seqs] Processed 100360 reads in 51.173 CPU sec, 51.105 real sec
[M::process] read 99966 sequences (10000027 bp)...
[M::mem_process_seqs] Processed 100124 reads in 50.403 CPU sec, 50.126 real sec
[M::process] read 2917 sequences (291695 bp)...
[M::mem_process_seqs] Processed 99966 reads in 50.386 CPU sec, 50.266 real sec
[M::mem_process_seqs] Processed 2917 reads in 1.599 CPU sec, 1.458 real sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa mem GCF_000001405.40_GRCh38.p14_genomic.fna
SRR3586068_trim.fastq
[main] Real time: 195.752 sec; CPU: 164.448 sec
```

```
(base) vaidehipatel46@class-allhomework:~/final_exam$ bwa mem
GCF_000001405.40_GRCh38.p14_genomic.fna SRR3586069_trim.fastq >
SRR3586069_trim_bwa_ref.sam
[M::bwa_idx_load_from_disk] read 0 ALT contigs
[M::process] read 99286 sequences (10000169 bp)...
[M::process] read 99542 sequences (10000101 bp)...
[M::mem_process_seqs] Processed 99286 reads in 42.374 CPU sec, 42.251 real sec
[M::process] read 99472 sequences (10000041 bp)...
[M::mem_process_seqs] Processed 99542 reads in 42.048 CPU sec, 41.776 real sec
[M::process] read 99602 sequences (10000076 bp)...
[M::mem_process_seqs] Processed 99472 reads in 43.882 CPU sec, 43.611 real sec
[M::process] read 99764 sequences (10000035 bp)...
[M::mem_process_seqs] Processed 99602 reads in 42.853 CPU sec, 42.598 real sec
[M::process] read 99730 sequences (10000012 bp)...
[M::mem_process_seqs] Processed 99764 reads in 41.816 CPU sec, 41.547 real sec
[M::process] read 99606 sequences (10000092 bp)...
[M::mem_process_seqs] Processed 99730 reads in 43.591 CPU sec, 43.336 real sec
[M::process] read 48633 sequences (4870963 bp)...
[M::mem_process_seqs] Processed 99606 reads in 43.702 CPU sec, 43.518 real sec
[M::mem_process_seqs] Processed 48633 reads in 21.004 CPU sec, 20.881 real sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa mem GCF_000001405.40_GRCh38.p14_genomic.fna SRR3586069_trim.fastq
[main] Real time: 358.833 sec; CPU: 329.855 sec
```

```
(base) vaidehipatel46@class-allhomework:~/final_exam$ bwa mem
GCF_000001405.40_GRCh38.p14_genomic.fna SRR3586070_trim.fastq >
SRR3586070_trim_bwa_ref.sam
[M::bwa_idx_load_from_disk] read 0 ALT contigs
[M::process] read 100472 sequences (10000020 bp)...
[M::process] read 100086 sequences (10000006 bp)...
[M::mem_process_seqs] Processed 100472 reads in 48.892 CPU sec, 48.777 real sec
[M::process] read 100186 sequences (10000161 bp)...
[M::mem_process_seqs] Processed 100086 reads in 48.300 CPU sec, 48.022 real sec
[M::process] read 99818 sequences (10000030 bp)...
[M::mem_process_seqs] Processed 100186 reads in 45.459 CPU sec, 45.179 real sec
[M::process] read 27533 sequences (2756567 bp)...
[M::mem_process_seqs] Processed 99818 reads in 44.276 CPU sec, 44.127 real sec
[M::mem_process_seqs] Processed 27533 reads in 11.882 CPU sec, 11.760 real sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa mem GCF_000001405.40_GRCh38.p14_genomic.fna
SRR3586070_trim.fastq
[main] Real time: 237.165 sec; CPU: 205.716 sec
```

Megahit- It is NGS assembler.

https://github.com/voutcn/megahit

```
(base) vaidehipatel46@class-allhomework:~/final_exam$ sudo docker run -v ~/final_exam/:/data -
i -t megahit /bin/bash
root@dee3ad80ebb3:/# apt install python-is-python3
```

```
root@dee3ad80ebb3:/data# megahit -r /data/SRR3586068_trim.fastq -o SRR3586068_out
7.768Gb memory in total.
Using: 6.991Gb.
MEGAHIT v1.0.6
--- [Sat Dec 10 21:45:45 2022] Start assembly. Number of CPU threads 2 ---
--- [Sat Dec 10 21:45:45 2022] Available memory: 8340824064, used: 7506741657
--- [Sat Dec 10 21:45:45 2022] k list: 21,41,61,81,99 ---
--- [Sat Dec 10 21:45:45 2022] Converting reads to binaries ---
b'   [read_lib_functions-inl.h  : 209]     Lib 0 (/data/SRR3586068_trim.fastq): se, 303367 reads,
101 max length'
b'   [utils.h                 : 126]     Real: 1.0497\tuser: 0.3222\tsys: 0.1304\tmaxrss: 16756'
--- [Sat Dec 10 21:45:46 2022] Extracting solid (k+1)-mers for k = 21 ---
--- [Sat Dec 10 21:45:55 2022] Building graph for k = 21 ---
--- [Sat Dec 10 21:46:02 2022] Assembling contigs from SdBG for k = 21 ---
--- [Sat Dec 10 21:46:17 2022] Local assembling k = 21 ---
--- [Sat Dec 10 21:46:19 2022] Extracting iterative edges from k = 21 to 41 ---
--- [Sat Dec 10 21:46:25 2022] Building graph for k = 41 ---
--- [Sat Dec 10 21:46:28 2022] Assembling contigs from SdBG for k = 41 ---
--- [Sat Dec 10 21:46:37 2022] Local assembling k = 41 ---
--- [Sat Dec 10 21:46:39 2022] Extracting iterative edges from k = 41 to 61 ---
--- [Sat Dec 10 21:46:42 2022] Building graph for k = 61 ---
--- [Sat Dec 10 21:46:43 2022] Assembling contigs from SdBG for k = 61 ---
--- [Sat Dec 10 21:46:48 2022] Local assembling k = 61 ---
--- [Sat Dec 10 21:46:51 2022] Extracting iterative edges from k = 61 to 81 ---
--- [Sat Dec 10 21:46:53 2022] Building graph for k = 81 ---
--- [Sat Dec 10 21:46:53 2022] Assembling contigs from SdBG for k = 81 ---
--- [Sat Dec 10 21:46:57 2022] Local assembling k = 81 ---
--- [Sat Dec 10 21:46:58 2022] Extracting iterative edges from k = 81 to 99 ---
--- [Sat Dec 10 21:46:59 2022] Building graph for k = 99 ---
--- [Sat Dec 10 21:47:00 2022] Assembling contigs from SdBG for k = 99 ---
--- [Sat Dec 10 21:47:02 2022] Merging to output final contigs ---
--- [STAT] 1002 contigs, total 483675 bp, min 200 bp, max 2362 bp, avg 483 bp, N50 484 bp
--- [Sat Dec 10 21:47:02 2022] ALL DONE. Time elapsed: 77.540911 seconds ---
```

```
root@9e5924e999cc:/data# megahit -r /data/SRR3586069_trim.fastq -o SRR3586069.out
7.768Gb memory in total.
Using: 6.991Gb.
MEGAHIT v1.0.6
--- [Sun Dec 11 00:01:54 2022] Start assembly. Number of CPU threads 2 ---
--- [Sun Dec 11 00:01:54 2022] Available memory: 8340824064, used: 7506741657
--- [Sun Dec 11 00:01:54 2022] k list: 21,41,61,81,99 ---
--- [Sun Dec 11 00:01:54 2022] Converting reads to binaries ---
b'   [read_lib_functions-inl.h  : 209]     Lib 0 (/data/SRR3586069_trim.fastq): se, 745635 reads, 101 max
length'
b'   [utils.h                  : 126]     Real: 3.7030\tuser: 1.4192\tsys: 0.7676\tmaxrss: 49288'
--- [Sun Dec 11 00:01:58 2022] Extracting solid (k+1)-mers for k = 21 ---
--- [Sun Dec 11 00:02:23 2022] Building graph for k = 21 ---
--- [Sun Dec 11 00:02:43 2022] Assembling contigs from SdBG for k = 21 ---
--- [Sun Dec 11 00:03:29 2022] Local assembling k = 21 ---
--- [Sun Dec 11 00:03:35 2022] Extracting iterative edges from k = 21 to 41 ---
--- [Sun Dec 11 00:03:55 2022] Building graph for k = 41 ---
```

```
root@9e5924e999cc:/data# megahit -r /data/SRR3586070_trim.fastq -o SRR3586070.out
7.768Gb memory in total.
Using: 6.991Gb.
MEGAHIT v1.0.6
--- [Sun Dec 11 00:08:14 2022] Start assembly. Number of CPU threads 2 ---
```

```
--- [Sun Dec 11 00:08:14 2022] Available memory: 8340824064, used: 7506741657
--- [Sun Dec 11 00:08:14 2022] k list: 21,41,61,81,99 ---
--- [Sun Dec 11 00:08:14 2022] Converting reads to binaries ---
b'   [read_lib_functions-inl.h  : 209]    Lib 0 (/data/SRR3586070_trim.fastq): se, 428095 reads, 101 max length'
b'   [utils.h                 : 126]    Real: 1.1559\tuser: 0.4875\tsys: 0.1023\tmaxrss: 26184'
--- [Sun Dec 11 00:08:15 2022] Extracting solid (k+1)-mers for k = 21 ---
--- [Sun Dec 11 00:08:28 2022] Building graph for k = 21 ---
--- [Sun Dec 11 00:08:40 2022] Assembling contigs from SdBG for k = 21 ---
--- [Sun Dec 11 00:09:04 2022] Local assembling k = 21 ---
--- [Sun Dec 11 00:09:08 2022] Extracting iterative edges from k = 21 to 41 ---
--- [Sun Dec 11 00:09:17 2022] Building graph for k = 41 ---
--- [Sun Dec 11 00:09:22 2022] Assembling contigs from SdBG for k = 41 ---
--- [Sun Dec 11 00:09:40 2022] Local assembling k = 41 ---
--- [Sun Dec 11 00:09:45 2022] Extracting iterative edges from k = 41 to 61 ---
--- [Sun Dec 11 00:09:49 2022] Building graph for k = 61 ---
--- [Sun Dec 11 00:09:52 2022] Assembling contigs from SdBG for k = 61 ---
--- [Sun Dec 11 00:10:04 2022] Local assembling k = 61 ---
--- [Sun Dec 11 00:10:10 2022] Extracting iterative edges from k = 61 to 81 ---
--- [Sun Dec 11 00:10:12 2022] Building graph for k = 81 ---
--- [Sun Dec 11 00:10:14 2022] Assembling contigs from SdBG for k = 81 ---
--- [Sun Dec 11 00:10:25 2022] Local assembling k = 81 ---
--- [Sun Dec 11 00:10:31 2022] Extracting iterative edges from k = 81 to 99 ---
--- [Sun Dec 11 00:10:32 2022] Building graph for k = 99 ---
--- [Sun Dec 11 00:10:33 2022] Assembling contigs from SdBG for k = 99 ---
--- [Sun Dec 11 00:10:42 2022] Merging to output final contigs ---
--- [STAT] 4348 contigs, total 2364849 bp, min 200 bp, max 11035 bp, avg 544 bp, N50 547 bp
--- [Sun Dec 11 00:10:42 2022] ALL DONE. Time elapsed: 148.121994 seconds ---
```

Prodigal software- It is gene prediction software that predicts the protein coding genes. It gave two output files such as .gbk and .faa files.
https://github.com/hyattpd/Prodigal

```
base) vaidehipatel46@class-allhomework:~/final_exam$ prodigal -i
~/final_exam/SRR3586068_out/final.contigs.fa -o SRR2586068_ORFs.gbk -a
SRR2586068_pro.faa -p meta
------------------------------------
PRODIGAL v2.6.3 [February, 2016]
Univ of Tenn / Oak Ridge National Lab
Doug Hyatt, Loren Hauser, et al.
------------------------------------
Request:  Metagenomic, Phase:  Training
Initializing training files...done!
------------------------------------
Request:  Metagenomic, Phase:  Gene Finding
Finding genes in sequence #1 (300 bp)...done!
Finding genes in sequence #2 (353 bp)...done!
Finding genes in sequence #3 (327 bp)...done!
Finding genes in sequence #4 (329 bp)...done!
Finding genes in sequence #5 (624 bp)...done!
Finding genes in sequence #6 (464 bp)...done!
Finding genes in sequence #7 (479 bp)...done!
Finding genes in sequence #8 (1119 bp)...done!
Finding genes in sequence #9 (1282 bp)...done!
Finding genes in sequence #10 (381 bp)...done!
Finding genes in sequence #11 (428 bp)...done!
Finding genes in sequence #12 (299 bp)...done!
Finding genes in sequence #13 (385 bp)...done!
Finding genes in sequence #14 (301 bp)...done!
Finding genes in sequence #15 (469 bp)...done!
Finding genes in sequence #16 (471 bp)...done!
Finding genes in sequence #17 (316 bp)...done!
Finding genes in sequence #18 (302 bp)...done!
Finding genes in sequence #19 (308 bp)...done!
Finding genes in sequence #20 (407 bp)...done!
Finding genes in sequence #21 (912 bp)...done!
Finding genes in sequence #22 (321 bp)...done!
Finding genes in sequence #23 (308 bp)...done!
Finding genes in sequence #24 (515 bp)...done!
Finding genes in sequence #25 (690 bp)...done!
Finding genes in sequence #26 (304 bp)...done!
Finding genes in sequence #27 (319 bp)...done!
Finding genes in sequence #28 (351 bp)...done!
Finding genes in sequence #29 (336 bp)...done!
Finding genes in sequence #30 (405 bp)...done!
Finding genes in sequence #31 (614 bp)...done!
Finding genes in sequence #32 (590 bp)...done!
Finding genes in sequence #33 (463 bp)...done!
```

Finding genes in sequence #34 (815 bp)...done!

```
base) vaidehipatel46@class-allhomework:~/final_exam$ prodigal -i
/home/vaidehipatel46/final_exam/SRR3586069.out/final.contigs.fa -o SRR3586069_ORFs.gbk -a
SRR3586069_pro.faa -p meta
-------------------------------------
PRODIGAL v2.6.3 [February, 2016]
Univ of Tenn / Oak Ridge National Lab
Doug Hyatt, Loren Hauser, et al.
-------------------------------------
Request:  Metagenomic, Phase:  Training
Initializing training files...done!
-------------------------------------
Request:  Metagenomic, Phase:  Gene Finding
Finding genes in sequence #1 (298 bp)...done!
Finding genes in sequence #2 (301 bp)...done!
Finding genes in sequence #3 (372 bp)...done!
Finding genes in sequence #4 (1048 bp)...done!
Finding genes in sequence #5 (345 bp)...done!
Finding genes in sequence #6 (334 bp)...done!
Finding genes in sequence #7 (311 bp)...done!
Finding genes in sequence #8 (400 bp)...done!
Finding genes in sequence #9 (519 bp)...done!
Finding genes in sequence #10 (311 bp)...done!
Finding genes in sequence #11 (407 bp)...done!
Finding genes in sequence #12 (1600 bp)...done!
Finding genes in sequence #13 (299 bp)...done!
Finding genes in sequence #14 (320 bp)...done!
Finding genes in sequence #15 (326 bp)...done!
Finding genes in sequence #16 (344 bp)...done!
Finding genes in sequence #17 (298 bp)...done!
Finding genes in sequence #18 (297 bp)...done!
Finding genes in sequence #19 (353 bp)...done!
Finding genes in sequence #20 (649 bp)...done!
Finding genes in sequence #21 (455 bp)...done!
Finding genes in sequence #22 (948 bp)...done!
Finding genes in sequence #23 (545 bp)...done!
Finding genes in sequence #24 (467 bp)...done!
Finding genes in sequence #25 (852 bp)...done!
Finding genes in sequence #26 (372 bp)...done!
Finding genes in sequence #27 (346 bp)...done!
Finding genes in sequence #28 (661 bp)...done!
Finding genes in sequence #29 (354 bp)...done!
```

Finding genes in sequence #30 (664 bp)...done!
Finding genes in sequence #31 (569 bp)...done!
Finding genes in sequence #32 (331 bp)...done!
Finding genes in sequence #33 (312 bp)...done!
Finding genes in sequence #34 (847 bp)...done!

```
base) vaidehipatel46@class-allhomework:~/final_exam$ prodigal -i
/home/vaidehipatel46/final_exam/SRR3586070.out/final.contigs.fa -o SRR3586070_ORFs.gbk -a
SRR3586070_pro.faa -p meta
------------------------------------
PRODIGAL v2.6.3 [February, 2016]
Univ of Tenn / Oak Ridge National Lab
Doug Hyatt, Loren Hauser, et al.
------------------------------------
Request:  Metagenomic, Phase:  Training
Initializing training files...done!
------------------------------------
Request:  Metagenomic, Phase:  Gene Finding
Finding genes in sequence #1 (357 bp)...done!
Finding genes in sequence #2 (439 bp)...done!
Finding genes in sequence #3 (484 bp)...done!
Finding genes in sequence #4 (456 bp)...done!
Finding genes in sequence #5 (311 bp)...done!
Finding genes in sequence #6 (328 bp)...done!
Finding genes in sequence #7 (446 bp)...done!
Finding genes in sequence #8 (568 bp)...done!
Finding genes in sequence #9 (1158 bp)...done!
Finding genes in sequence #10 (383 bp)...done!
Finding genes in sequence #11 (445 bp)...done!
Finding genes in sequence #12 (501 bp)...done!
Finding genes in sequence #13 (381 bp)...done!
Finding genes in sequence #14 (712 bp)...done!
Finding genes in sequence #15 (571 bp)...done!
Finding genes in sequence #16 (399 bp)...done!
Finding genes in sequence #17 (416 bp)...done!
Finding genes in sequence #18 (410 bp)...done!
Finding genes in sequence #19 (521 bp)...done!
Finding genes in sequence #20 (362 bp)...done!
Finding genes in sequence #21 (468 bp)...done!
Finding genes in sequence #22 (438 bp)...done!
Finding genes in sequence #23 (316 bp)...done!
Finding genes in sequence #24 (431 bp)...done!
Finding genes in sequence #25 (331 bp)...done!
Finding genes in sequence #26 (543 bp)...done!
Finding genes in sequence #27 (444 bp)...done!
Finding genes in sequence #28 (328 bp)...done!
Finding genes in sequence #29 (393 bp)...done!
```

Diamond- software that helps for finding homologs of protein in a reference database.

- I used Swissprot database and NCBI NR database was around 198 GB so I opt out to use swissprot database instead
- The Diamond manual is useful in deciding what options and filter I should add it to the command line for Diamond software
- The output files were daa format and tab format

https://www.softwareradius.com/best-gene-and-genome-annotation-tools-and-software/

https://metagenomics-workshop.readthedocs.io/en/latest/annotation/taxonomic_annotation.html#megan

https://usermanual.wiki/Pdf/diamondmanual.1718530976/view

SRR3586068(Cancer)

(base) vaidehipatel46@class-allhomework:~/final_exam$ diamond blastp --threads 8 --query SRR3586068_pro.faa --db swissprot.dmnd --taxonnodes nodes.dmp --evalue .00001 --salltitles --daa SRR3586068.blastp_result

```
 --daa SRR3586068.blastp_result
diamond v0.9.14.115 | by Benjamin Buchfink <buchfink@gmail.com>
Licensed under the GNU AGPL <https://www.gnu.org/licenses/agpl.txt>
Check http://github.com/bbuchfink/diamond for updates.

#CPU threads: 8
Scoring parameters: (Matrix=BLOSUM62 Lambda=0.267 K=0.041 Penalties=11/1)
#Target sequences to report alignments for: 25
Temporary directory:
Opening the database...  [0.071856s]
Loading taxonomy nodes...  [1.48316s]
Opening the input file...  [0.015935s]
Opening the output file...  [0.000116s]
Loading query sequences...  [0.022215s]
Masking queries...  [0.016754s]
Building query seed set...  [0.002812s]
Algorithm: Query-indexed
Building query histograms...  [0.000453s]
Allocating buffers...  [4.6e-05s]
Loading reference sequences...  [2.42029s]
Building reference histograms...  [0.594902s]
Allocating buffers...  [6.9e-05s]
Initializing temporary storage...  [0.002384s]
Processing query chunk 0, reference chunk 0, shape 0, index chunk 0.
Building reference index...  [0.764365s]
Building query index...  [0.00322s]
Building seed filter...  [0.008467s]
Searching alignments...  [0.337201s]
Deallocating buffers...  [0.001874s]
Computing alignments...  [0.079422s]
Deallocating reference...  [0.012035s]
Loading reference sequences...  [4.1e-05s]
Deallocating buffers...  [2.7e-05s]
Deallocating queries...  [2.9e-05s]
Loading query sequences...  [3.1e-05s]
Closing the input file...  [3.1e-05s]
Closing the output file...  [0.000425s]
Closing the database file...  [3.2e-05s]
Total time = 5.84171s
Reported 668 pairwise alignments, 668 HSPs.
59 queries aligned.
(base) vaidehipatel46@class-allhomework:~/final_exam$
```

(base) vaidehipatel46@class-allhomework:~/final_exam$ diamond view -a SRR3586068.blastp_result.daa > SRR3586068.search_result.tab

SRR3586069(Normal)

(base) vaidehipatel46@class-allhomework:~/final_exam$ diamond blastp --threads 8 --query SRR3586069_pro.faa --db swissprot.dmnd --taxonnodes nodes.dmp --evalue .00001 --salltitles --daa SRR3586069.blastp_result

```
(base) vaidehipatel46@class-allhomework:~/final_exam$ diamond blastp  --threads 8 --query SRR3586069_pro.faa --db swissprot.dmnd --taxonnodes nodes.dmp --evalue .00001 --salltitles
 --daa SRR3586069.blastp_result
diamond v0.9.14.115 | by Benjamin Buchfink <buchfink@gmail.com>
Licensed under the GNU AGPL <https://www.gnu.org/licenses/agpl.txt>
Check http://github.com/bbuchfink/diamond for updates.

#CPU threads: 8
Scoring parameters: (Matrix=BLOSUM62 Lambda=0.267 K=0.041 Penalties=11/1)
#Target sequences to report alignments for: 25
Temporary directory:
Opening the database...  [3.6e-05s]
Loading taxonomy nodes...  [0.806606s]
Opening the input file...  [0.012962s]
Opening the output file...  [0.000105s]
Loading query sequences...  [0.024384s]
Masking queries...  [0.063455s]
Building query seed set...  [0.004129s]
Algorithm: Query-indexed
Building query histograms...  [0.000868s]
Allocating buffers...  [5.1e-05s]
Loading reference sequences...  [0.541509s]
Building reference histograms...  [0.68855s]
Allocating buffers...  [9.2e-05s]
Initializing temporary storage...  [0.002585s]
Processing query chunk 0, reference chunk 0, shape 0, index chunk 0.
Building reference index...  [1.65188s]
Building query index...  [0.012825s]
Building seed filter...  [0.037982s]
Searching alignments...  [1.76265s]
Deallocating buffers...  [0.009484s]
Computing alignments...  [1.16202s]
Deallocating reference...  [0.012768s]
Loading reference sequences...  [1.8e-05s]
Deallocating buffers...  [1.1e-05s]
Deallocating queries...  [1.3e-05s]
Loading query sequences...  [1.9e-05s]
Closing the input file...  [1.8e-05s]
Closing the output file...  [0.005371s]
Closing the database file...  [8e-06s]
Total time = 6.80298s
Reported 12287 pairwise alignments, 12287 HSPs.
855 queries aligned.
```

base) vaidehipatel46@class-allhomework:~/final_exam$ diamond view -a SRR3586069.blastp_result.daa > SRR3586069.search_result.tab

SRR3586070(Cancer sample)

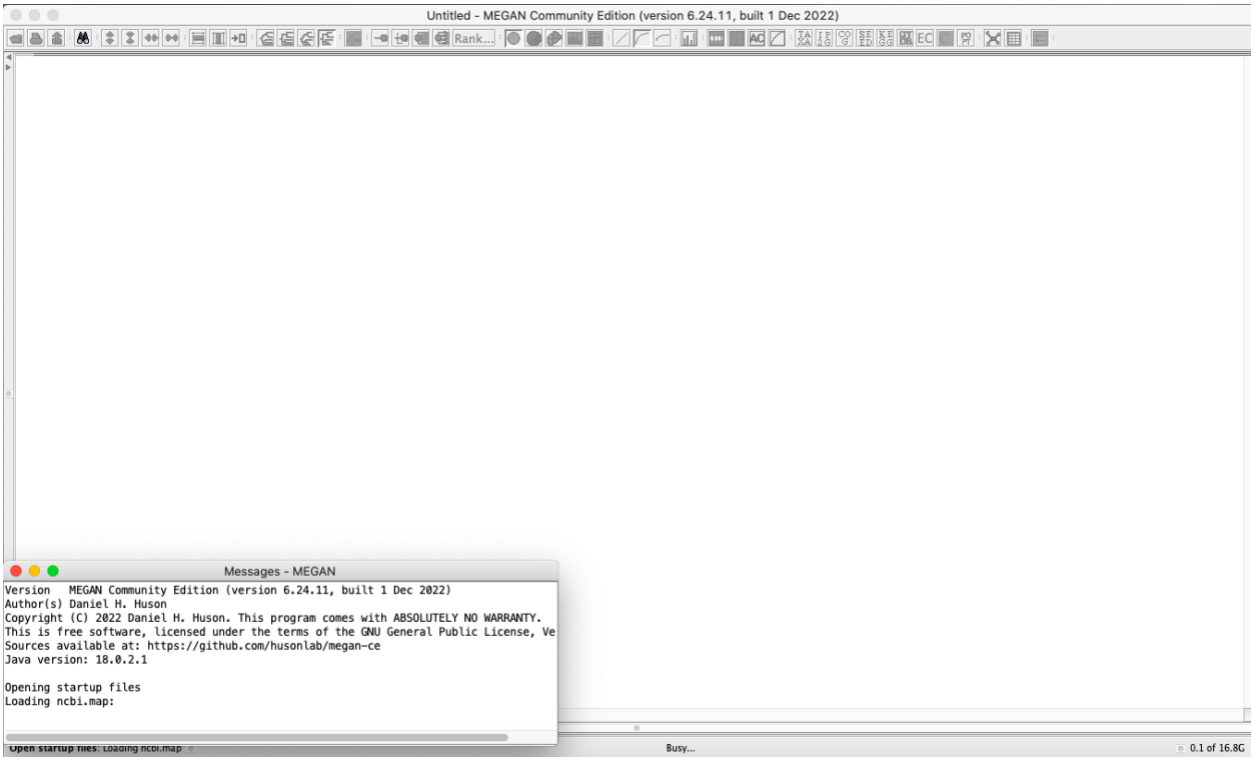
(base) vaidehipatel46@class-allhomework:~/final_exam$ diamond blastp --threads 8 --query SRR3586070_pro.faa --db swissprot.dmnd --taxonnodes nodes.dmp --evalue .00001 --salltitles --daa SRR3586070.blastp_result

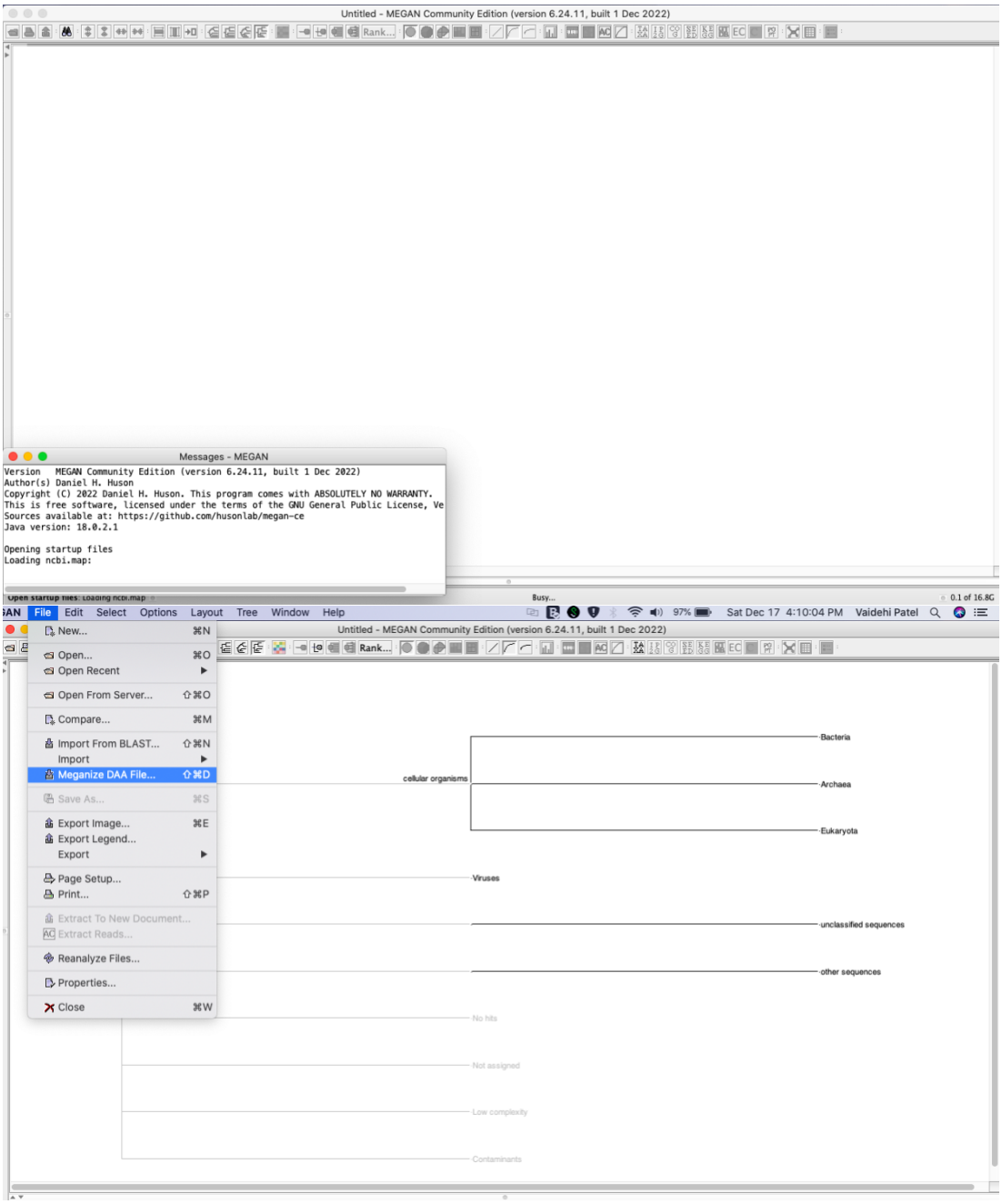
```
(base) vaidehipatel46@class-allhomework:~/final_exam$ diamond blastp  --threads 8 --query SRR3586070_pro.faa --db swissprot.dmnd --taxonnodes nodes.dmp --evalue .00001 --sall
 --daa SRR3586070.blastp_result
diamond v0.9.14.115 | by Benjamin Buchfink <buchfink@gmail.com>
Licensed under the GNU AGPL <https://www.gnu.org/licenses/agpl.txt>
Check http://github.com/bbuchfink/diamond for updates.

#CPU threads: 8
Scoring parameters: (Matrix=BLOSUM62 Lambda=0.267 K=0.041 Penalties=11/1)
#Target sequences to report alignments for: 25
Temporary directory:
Opening the database...  [2.3e-05s]
Loading taxonomy nodes...  [0.789353s]
Opening the input file...  [0.000406s]
Opening the output file...  [6.9e-05s]
Loading query sequences...  [0.052431s]
Masking queries...  [0.125655s]
Building query seed set...  [0.006515s]
Algorithm: Query-indexed
Building query histograms...  [0.001371s]
Allocating buffers...  [3.3e-05s]
Loading reference sequences...  [0.531466s]
Building reference histograms...  [0.818432s]
Allocating buffers...  [4.8e-05s]
Initializing temporary storage...  [0.002558s]
Processing query chunk 0, reference chunk 0, shape 0, index chunk 0.
Building reference index...  [2.73417s]
Building query index...  [0.028826s]
Building seed filter...  [0.069435s]
Searching alignments...  [3.81122s]
Deallocating buffers...  [0.018669s]
Computing alignments...  [3.53299s]
Deallocating reference...  [0.013422s]
Loading reference sequences...  [2e-05s]
Deallocating buffers...  [1.4e-05s]
Deallocating queries...  [7e-06s]
Loading query sequences...  [1.7e-05s]
Closing the input file...  [1.8e-05s]
Closing the output file...  [0.010899s]
Closing the database file...  [9e-06s]
Total time = 12.5494s
Reported 30061 pairwise alignments, 30062 HSPs.
1995 queries aligned.
```

base) vaidehipatel46@class-allhomework:~/final_exam$ diamond view -a SRR3586070.blastp_result.daa > SRR3586070.search_result.tab

MEGAN6- It is software can be downloaded in linux and laptop. I downloaded it on laptop so it is easier me to save all files. It provides taxnonomic information for metagenomic dataset. It is great for taxonomic analysis of sequences.

https://metagenomics-workshop.readthedocs.io/en/latest/annotation/taxonomic_annotation.html#megan

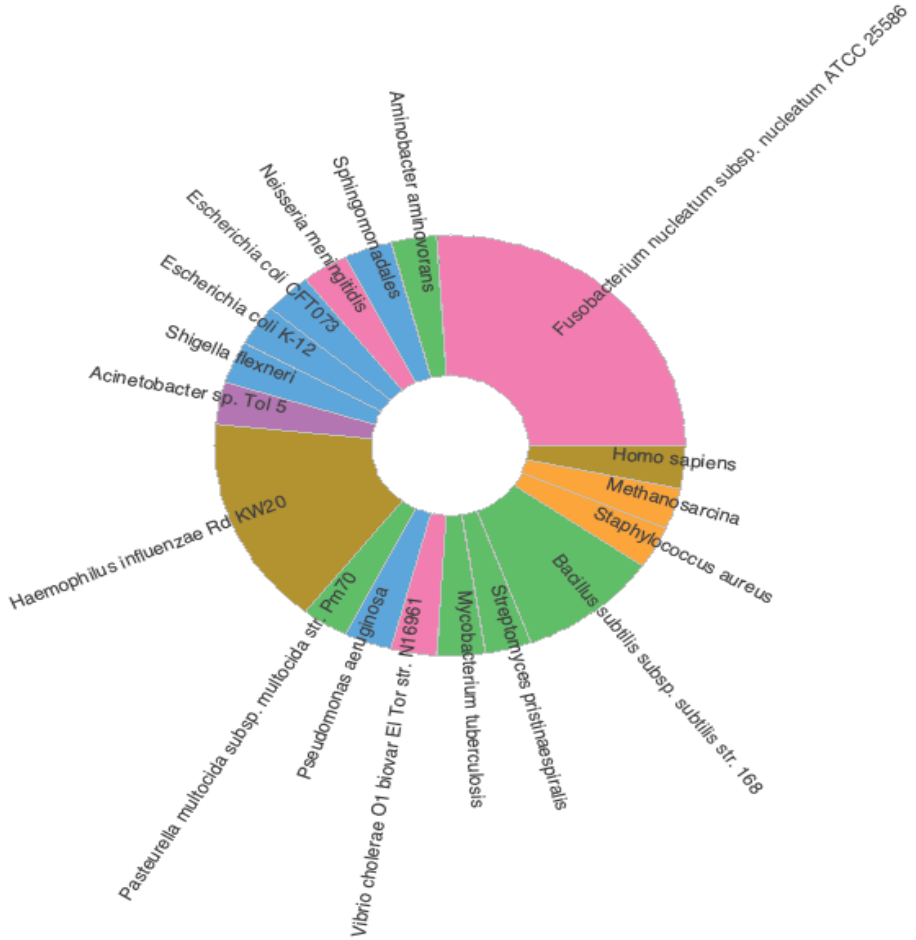


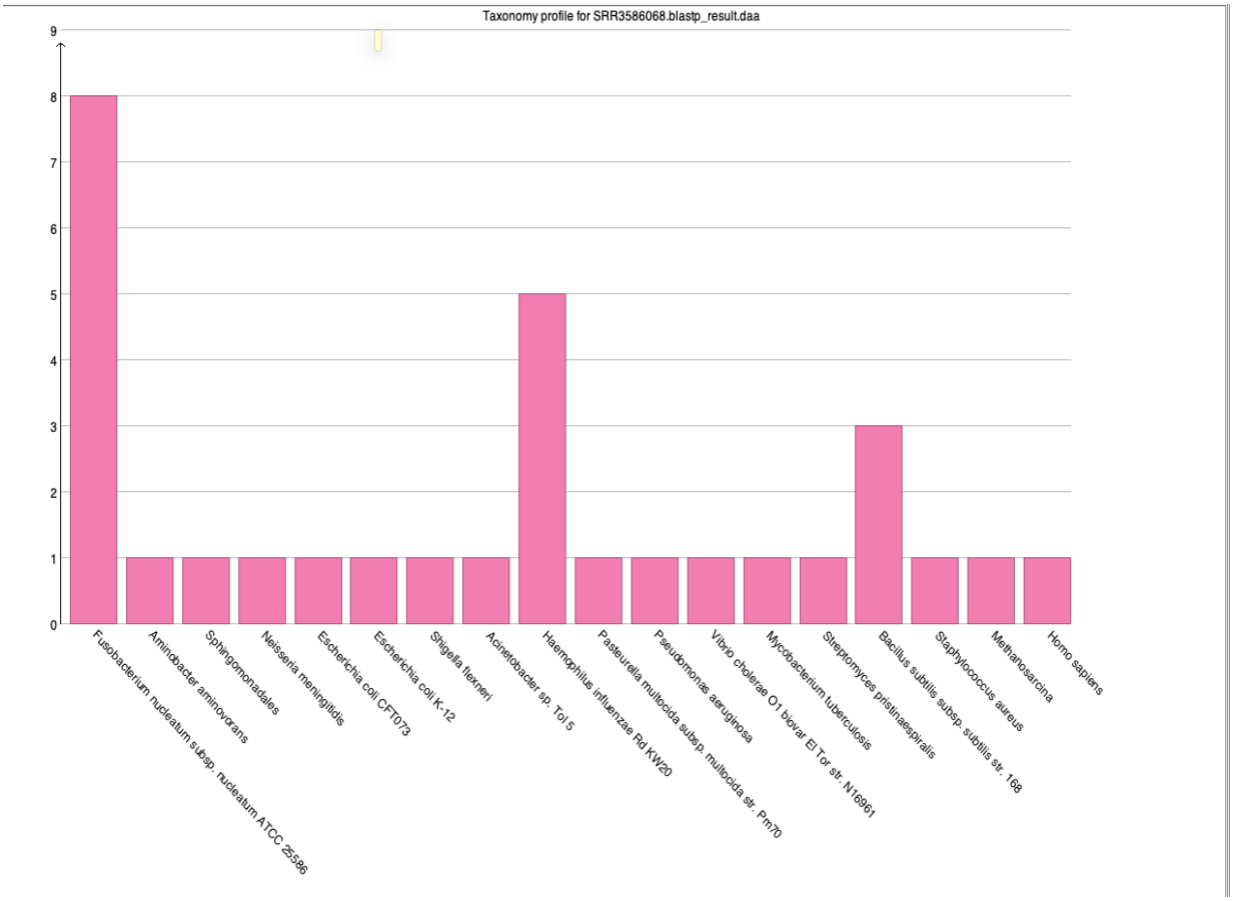


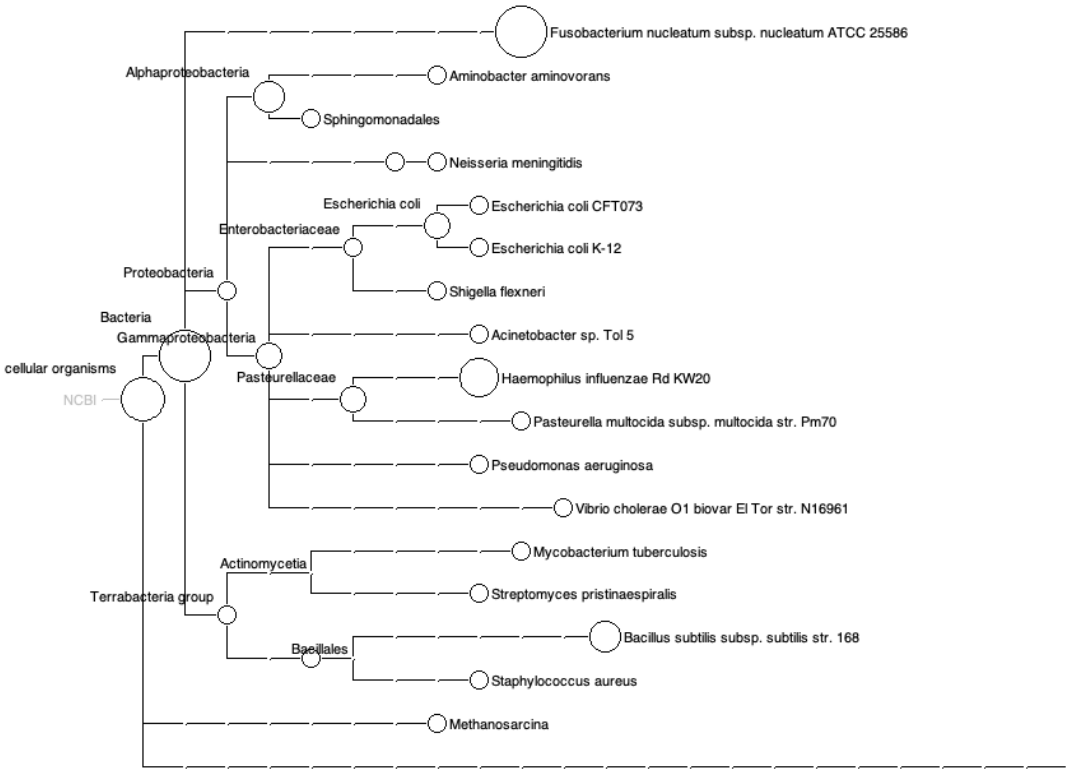
- First step is to meganize the DAA file output from DIAMOND software then it will open a new file with megan format with all taxonomy information

SRR3586068

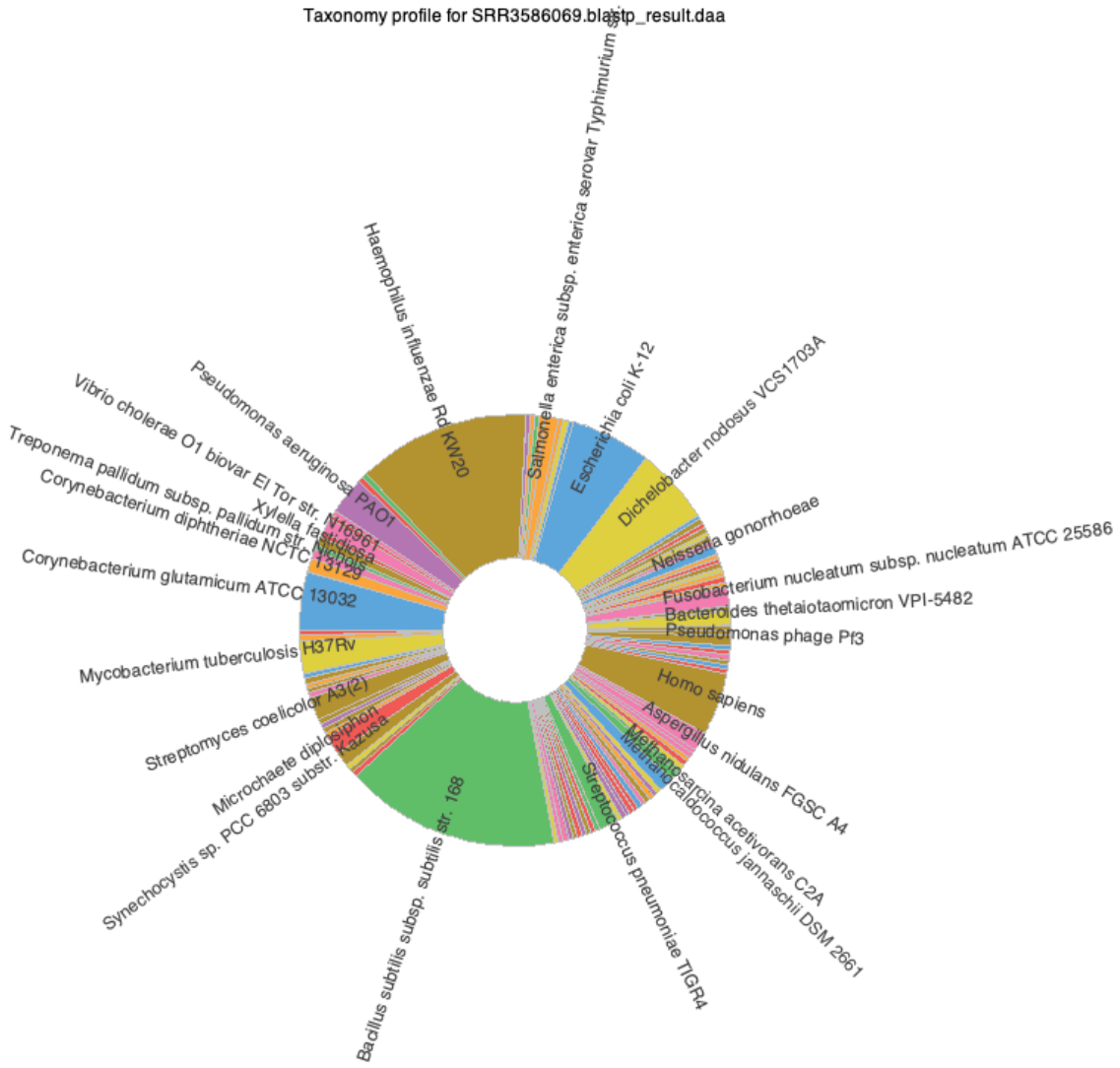
Taxonomy profile for SRR3586068.blastp_result.daa

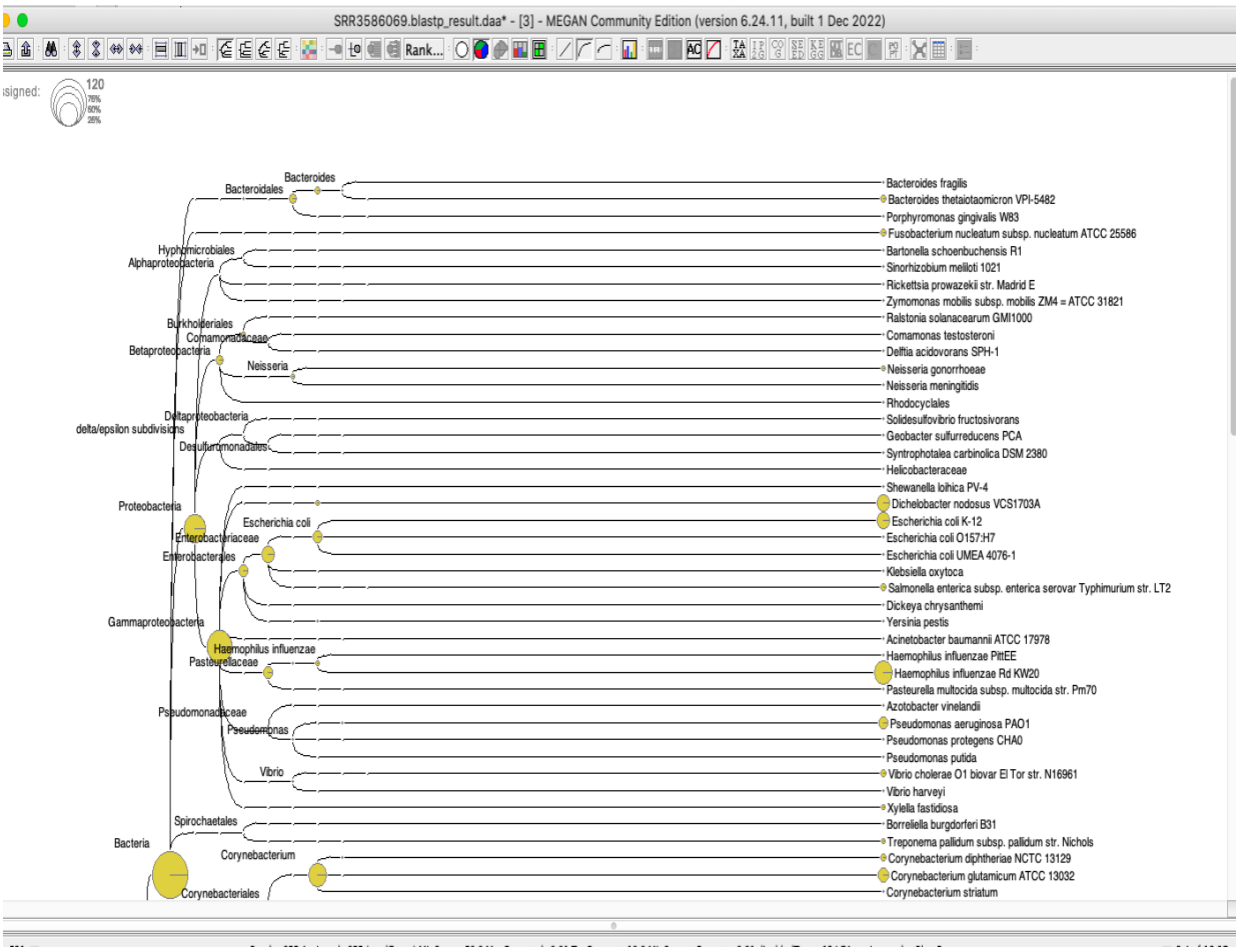



Taxonomy profile for SRR3586068.blastp_result.daa
9
8
7
6
5
4
3
2
1
0
Fusobacterium nucleatum subsp. nucleatum ATCC 25586
Amindobacter aminovorans
Sphingomonadales
Neisseria meningitidis
Escherichia coli CFT073
Escherichia coli K-12
Shigella flexneri
Acinetobacter sp. Tol 5
Haemophilus influenzae Rd KW20
Pasteurella multocida subsp. multocida str. Pm70
Pseudomonas aeruginosa
Vibrio cholerae O1 biovar El Tor str. N16961
Mycobacterium tuberculosis
Streptomyces pristinaespiralis
Bacillus subtilis subsp. subtilis str. 168
Staphylococcus aureus
Methanosarcina
Homo sapiens

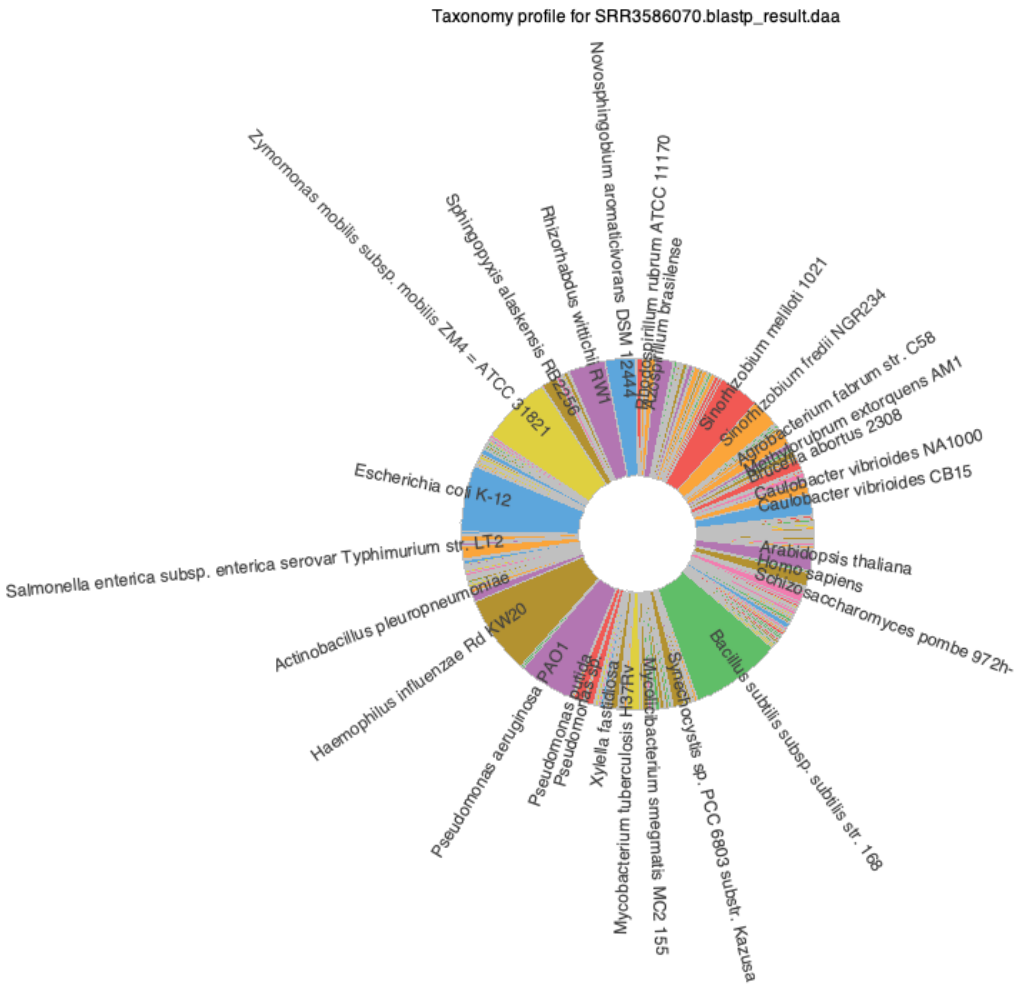

Fusobacterium nucleatum subsp. nucleatum ATCC 25586
Alphaproteobacteria
Aminobacter aminovorans
Sphingomonadales
Neisseria meningitidis
Escherichia coli
Escherichia coli CFT073
Enterobacteriaceae
Escherichia coli K-12
Proteobacteria
Shigella flexneri
Bacteria
Acinetobacter sp. Tol 5
Gammaproteobacteria
cellular organisms
Pasteurellaceae
Haemophilus influenzae Rd KW20
NCBI
Pasteurella multocida subsp. multocida str. Pm70
Pseudomonas aeruginosa
Vibrio cholerae O1 biovar El Tor str. N16961
Mycobacterium tuberculosis
Actinomycetia
Streptomyces pristinaespiralis
Terrabacteria group
Bacillus subtilis subsp. subtilis str. 168
Bacillales
Staphylococcus aureus
Methanosarcina

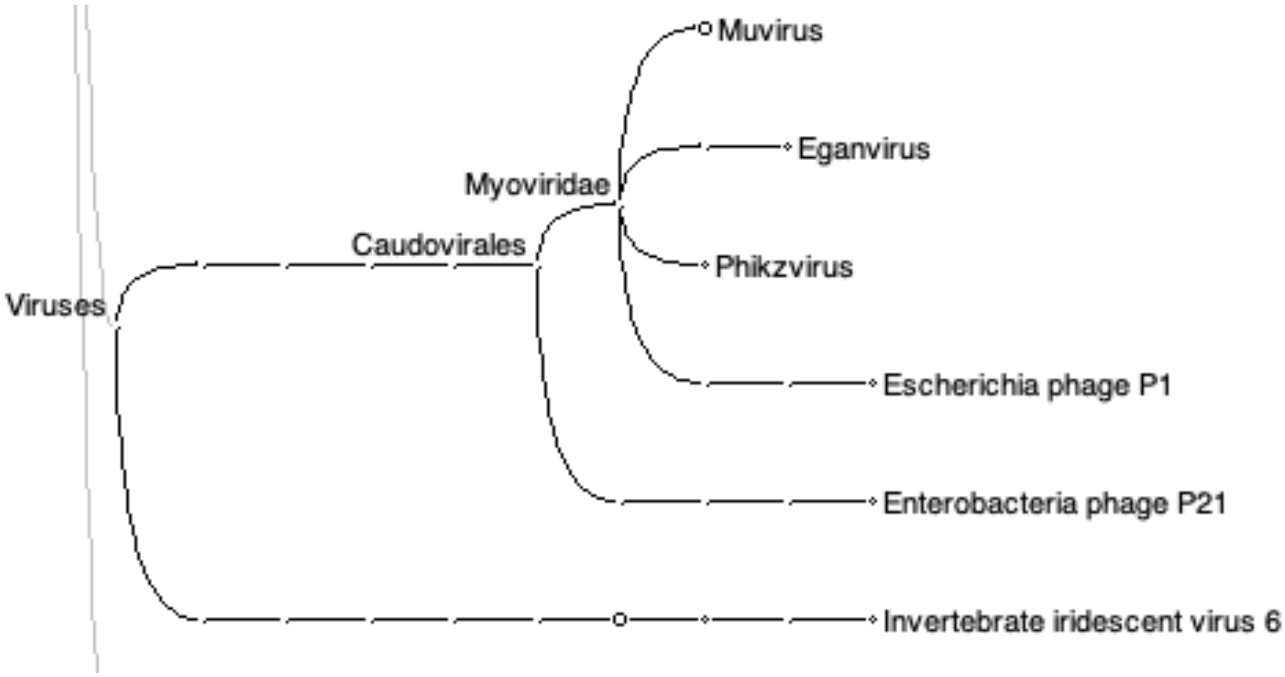
SRR3586069


Taxonomy profile for SRR3586069.blastp_result.daa
Haemophilus influenzae Rd KW20
Salmonella enterica subsp. enterica serovar Typhimurium str.
Escherichia coli K-12
Dichelobacter nodosus VCS1703A
Neisseria gonorrhoeae
Fusobacterium nucleatum subsp. nucleatum ATCC 25586
Bacteroides thetaiotaomicron VPI-5482
Pseudomonas phage Pf3
Homo sapiens
Aspergillus nidulans FGSC A4
Methanosarcina acetivorans C2A
Methanocaldococcus jannaschii DSM 2661
Streptococcus pneumoniae TIGR4
Bacillus subtilis subsp. subtilis str. 168
Synechocystis sp. PCC 6803 substr.
Microchaete diplosiphon
Streptomyces coelicolor A3(2)
Mycobacterium tuberculosis H37Rv
Corynebacterium glutamicum ATCC 13032
Corynebacterium diphtheriae NCTC 13129
Treponema pallidum subsp. pallidum str. Nichols
Vibrio cholerae O1 biovar El Tor str. N16961
Pseudomonas aeruginosa PAO1


SRR3586069.blastp_result.daa* - [3] - MEGAN Community Edition (version 6.24.11, built 1 Dec 2022)
Rank...
Bacteroides
Bacteroidales
Bacteroides fragilis
Bacteroides thetaiotaomicron VPI-5482
Porphyromonas gingivalis W83
Fusobacterium nucleatum subsp. nucleatum ATCC 25586
Hyphomicrobiales
Alphaproteobacteria
Bartonella schoenbuchensis R1
Sinorhizobium meliloti 1021
Rickettsia prowazekii str. Madrid E
Zymomonas mobilis subsp. mobilis ZM4 = ATCC 31821
Burkholderiales
Comamonadaceae
Ralstonia solanacearum GMI1000
Comamonas testosteroni
Betaproteobacteria
Delftia acidovorans SPH-1
Neisseria
Neisseria gonorrhoeae
Neisseria meningitidis
Rhodocyclales
Deltaproteobacteria
delta/epsilon subdivisions
Solidesulfovibrio fructosivorans
Geobacter sulfurreducens PCA
Desulfuromonadales
Syntrophotalea carbinolica DSM 2380
Helicobacteraceae
Shewanella lohica PV-4
Proteobacteria
Dichelobacter nodosus VCS1703A
Escherichia coli
Escherichia coli K-12
Enterobacteriaceae
Escherichia coli O157:H7
Enterobacterales
Escherichia coli UMEA 4076-1
Klebsiella oxytoca
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2
Dickeya chrysanthemi
Gammaproteobacteria
Yersinia pestis
Acinetobacter baumannii ATCC 17978
Haemophilus influenzae
Haemophilus influenzae PittEE
Pasteurellaceae
Haemophilus influenzae Rd KW20
Pasteurella multocida subsp. multocida str. Pm70
Azotobacter vinelandii
Pseudomonadaceae
Pseudomonas
Pseudomonas aeruginosa PAO1
Pseudomonas protegens CHA0
Pseudomonas putida
Vibrio
Vibrio cholerae O1 biovar El Tor str. N16961
Vibrio harveyi
Xylella fastidiosa
Spirochaetales
Borreliella burgdorferi B31
Bacteria
Treponema pallidum subsp. pallidum str. Nichols
Corynebacterium
Corynebacterium diphtheriae NCTC 13129
Corynebacterium glutamicum ATCC 13032
Corynebacteriales
Corynebacterium striatum

SRR3586070


Taxonomy profile for SRR3586070.blastp_result.daa
Novosphingobium aromaticivorans DSM 12444
Azospirillum rubrum ATCC 11170
Azospirillum brasilense
Sinorhizobium meliloti 1021
Sinorhizobium fredii NGR234
Agrobacterium fabrum str. C58
Methylorubrum extorquens AM1
Brucella abortus 2308
Caulobacter vibrioides NA1000
Caulobacter vibrioides CB15
Arabidopsis thaliana
Homo sapiens
Schizosaccharomyces pombe 972h-
Bacillus subtilis subsp. subtilis str. 168
Synechocystis sp. PCC 6803 substr. Kazusa
Mycobacterium smegmatis MC2 155
Mycobacterium tuberculosis H37Rv
Xylella fastidiosa
Pseudomonas putida
Pseudomonas sp.
Pseudomonas aeruginosa PAO1
Haemophilus influenzae Rd KW20
Actinobacillus pleuropneumoniae
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2
Escherichia coli K-12
Zymomonas mobilis subsp. mobilis ZM4 = ATCC 31821
Sphingopyxis alaskensis RB2256
Rhizorhabdus wittichii RW1



References

https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR3586065&display=download

https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40#/st

https://www.ncbi.nlm.nih.gov/assembly/GCF_000013425.1

https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/

https://slequime.github.io/HTS-tutorial/de-novo-assembly.html

https://www.gungorbudak.com/blog/2018/05/16/how-to-download-hg38-grch38-fasta-human-reference-genome/

https://github.com/usadellab/Trimmomatic

https://bedtools.readthedocs.io/en/latest/content/tools/genomecov.html

http://www.bioinformatics.org/cd-hit/cd-hit-user-guide.pdf

https://bio.tools/diamond

https://bio.tools/megan

https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/megan6/

https://software-ab.cs.uni-tuebingen.de/download/megan6/welcome.html
https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/megan6/

https://www.youtube.com/watch?v=r6EVAS3DA40

https://drive5.com/usearch/manual/install.html