Elastic-BLAST exercises

For these exercises, you search sequences from the chicken gut metagenome against a well curated 16S database for bacteria and archaea. You produce the BLAST results in a tabular format that allows you to calculate some specific values for each query-subject match. From this information, you can infer some properties of the query sequences such as a putative taxonomic classification or whether the query is a partial sequence.

The 16S database consists of 16S ribosomal RNA sequences that correspond to bacteria and archaea type materials. This database is part of the RefSeq Targeted Loci Project. See https://www.ncbi.nlm.nih.gov/refseq/targetedloci/ for more information.

The chicken gut metagenome is from the WGS project KCET01 (https://www.ncbi.nlm.nih.gov/Traces/wgs/KCET01?display=contigs&page=1). This project contains nearly 4 million sequences, but here we look at only the first 299,000 sequences.

Definition of a "good match":

A query has a good match to the 16S_ribosomal_RNA database if these statements are true:

1.) The match consists of one alignment
2.) The match shows an 80% identity between the query and database sequence.
3.) In the match, the database sequence covers (aligns to) at least 90% of the query. The qcovhsp value in the tabular output specifies the coverage.

Tasks:

1.) Run Elastic-BLAST using the config file blastn.16S.ini. Notes:

- You will need to edit the file to add your results bucket. Use the bucket you created earlier and add "/results" to the end.

- Elastic-BLAST will construct a default cluster name for you (elasticblast-${USER}). Other values in the cluster section of the configuration file have been optimized for the 16S database you will be searching.
- The query file is in a cloud bucket and can be accessed directly by Elastic-BLAST.
- The tabular format is specified as '7 std qcovhsp staxid sscinname'. Slide 15 of the presentation lists the fields that "std" specifies.
- The Elastic-BLAST search should take about 37 minutes.

2.) Identify the good matches in your results and produce a file listing, for each match, query ID, subject ID, genus. From this:

- Find the genus that occurs most often per query sequence, or
- Find the genera of the top five matches per query sequence, or
- Use some other method of your choosing to identify a genus for the query.

3.) Produce a list of the query sequences **without** good matches.

4.) Search a few of the query sequences from 3.) against nt using the BLAST webpage to see if you can identify them.