

Bean Beetle Microbiome Identification of Picked Colony 16S rRNA DNA Sequence Student Handout

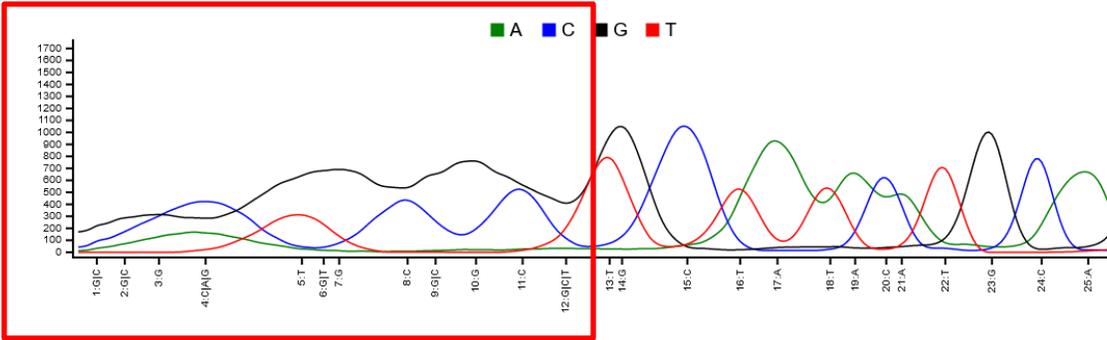
Sanger sequencing is a biochemical method of determining the complete nucleotide sequence of a piece of DNA (see the following video for a detailed description of this process: https://www.youtube.com/watch?v=-QIMkQ4E_wE). The DNA that each of you amplified using PCR with 16S rRNA gene specific primers was sent to a commercial laboratory where they conducted Sanger sequencing on each sample. In 2024, the Sanger sequencing was conducted with a 24-hour turn-around at a cost of \$9.00 per sample. Most but not all the samples were successfully sequenced.

For each successfully sequenced sample, we retrieved two files from the sequencing center's website – an .ab1 file and a .seq file. The .ab1 file contains the sequence and a chromatogram that convey information about the quality of the sequencing for any particular base pair. These files can only be opened with special software. They can be viewed using the Teal Trace Analysis Viewer (<https://www.gear-genomics.com/teal/>). The .seq file is a data file called a "FASTA" file (pronounced "fast-A"). The FASTA file contains a string of letters corresponding to the nitrogenous bases in the 16S rRNA gene (DNA) of the bacterium you picked from a culture plate. The .seq files can be opened with any text editor. You might also be given access to a multi-FASTA file (.fasta), which is a file that has the sequence data for all of the sequences. This file is just all of the .seq files combined. Again, this file can be opened with any text editor.

Conducting a Nucleotide BLAST (BLASTn) (basic local alignment search tool) will permit you to compare your bacterium 16S rRNA gene with sequences in a library of DNA sequences maintained by the National Center for Biotechnology Information, in the National Library of Medicine at the US National Institutes of Health.

Inspecting sequence quality and trimming sequences (optional)

1. Navigate to the Teal Trace Analysis Viewer (<https://www.gear-genomics.com/teal/>).
2. Use the browse button and select an .ab1 file.
3. Click "Launch Analysis".
4. View the chromatogram. Typically, at the beginning and end of the sequence, the quality is low, as indicated by broad peaks in the chromatogram. Also, multiple different nucleotides are shown at a particular location.
5. Make note of the first location at the beginning of the sequence where the peaks are sharp and only a single nucleotide is listed. Do the same for last location at the end of the sequence. These values can be entered into the BLAST search to trim the sequence.



In the example chromatogram above, we have put a red box around the poor-quality sequence. In this case, we would record “13” as the location at the beginning of the sequence where high-quality sequence starts.

BLASTing Sanger Sequences to Determine Identity

1. Navigate to <https://blast.ncbi.nlm.nih.gov/Blast.cgi> and select the Nucleotide BLAST (BLASTn)
2. In a text editor, open the .seq file of the sample that you wish to evaluate. Copy the sequence. The sequence must include the identifier “>sample ID code”. Paste sequence into text box for query sequence. Alternatively, click the “Browse” button and select the .seq file.
3. If you inspected the quality of the sequence, enter the locations of the high-quality base pairs from the beginning and end of the sequence in the “From” and “To” boxes in the “Query Sub-range”.
4. Under database, select 16S ribosomal RNA database.
5. Hit BLAST. Below is part of an example results page.

Descriptions		Graphic Summary	Alignments	Taxonomy				
Sequences producing significant alignments								
<input checked="" type="checkbox"/> select all 10 sequences selected		Download	Select columns	Show 10				
		GenBank	Graphics	Distance tree of results				
		MSA Viewer						
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> Enterobacter hormaechei subsp. xiangfangensis strain 10-17 16S ribosomal RNA, partial sequence	Enterobacter hormaechei subsp. ...	1716	1716	95%	0.0	97.85%	1429	NR_126208.1
<input checked="" type="checkbox"/> Enterobacter cloacae strain ATCC 13047 16S ribosomal RNA, partial sequence	Enterobacter cloacae	1716	1716	95%	0.0	97.85%	1369	NR_118568.1
<input checked="" type="checkbox"/> Enterobacter asburiae strain JCM6051 16S ribosomal RNA, partial sequence	Enterobacter asburiae	1683	1683	95%	0.0	97.34%	1422	NR_024640.1
<input checked="" type="checkbox"/> Enterobacter bugandensis strain 247BMC 16S ribosomal RNA, partial sequence	Enterobacter bugandensis	1674	1674	95%	0.0	97.13%	1444	NR_148649.1
<input checked="" type="checkbox"/> Enterobacter cancerogenus strain LMG 2693 16S ribosomal RNA, partial sequence	Enterobacter cancerogenus	1672	1672	95%	0.0	97.03%	1495	NR_044977.1
<input checked="" type="checkbox"/> Enterobacter hormaechei strain 0992-77 16S ribosomal RNA, partial sequence	Enterobacter hormaechei	1661	1661	94%	0.0	96.92%	1365	NR_042154.1
<input checked="" type="checkbox"/> Pseudoscherichia vulneris strain NBRC 102420 16S ribosomal RNA, partial sequence	Pseudoscherichia vulneris	1655	1655	95%	0.0	96.83%	1465	NR_114080.1
<input checked="" type="checkbox"/> Enterobacter cancerogenus strain LMG 2693 16S ribosomal RNA, partial sequence	Enterobacter cancerogenus	1653	1653	92%	0.0	97.48%	1387	NR_116756.1
<input checked="" type="checkbox"/> Klebsiella grimontii strain SB73 16S ribosomal RNA, partial sequence	Klebsiella grimontii	1650	1650	95%	0.0	96.62%	1454	NR_159317.1
<input checked="" type="checkbox"/> Pseudoscherichia vulneris strain ATCC 33821 16S ribosomal RNA, partial sequence	Pseudoscherichia vulneris	1648	1648	93%	0.0	97.09%	1468	NR_041927.1

6. Look carefully at the top 5-10 hits. You may select “Show 10” in the “Sequences producing significant alignments” window.

In the example output above, the BLAST alignments are listed in order of the E Values and alignment scores, which are influenced by the length of your sequence (query) and the quality of the alignment. We are interested in the columns labelled: **Description** (the taxonomic names of the bacterium matches [alignments], generally we only match to Genus), **Query Coverage** (the fraction of your sequence that is covered by each known reference sequence, greater values are better), **E Value** (the probability that the match between your sequence [query] and the known reference sequence was by chance alone, smaller values are better, 0.0 is the best), and **Percentage Identity** (the fraction of your sequence that exactly matches that of the known reference sequence, this value should be 90% or greater). A good match is one in which the top 5-10 matches identify the same genus and the E Values are small. BLAST results that yield Query Coverage of less than 50% but have a small E Value, a large Percentage Identity value (>90%), and the same genus identified in the top 5 alignments is acceptable.

- a. What is the % identity of your sequence and the database sequences?
- b. What % of your sequence aligns with the database sequences?
- c. What is the taxonomic identification of the top 5-10 hits?
- d. Can you assign your sample confidentially to family? To genus? To species?

Generally, an identification to genus is reliable using the 16S gene that each of you amplified by PCR. The primers we used are specific for the 16S gene of the rRNA DNA of bacteria.

If you were provided with a multi-fasta file, you can use the same procedure as described above by pasting more than one sequence in the Enter Query Sequence window or browsing and selecting the multi-fasta file. On the left side of results page for BLAST, there is a dropdown next to “Results for”. Here you can select the individual sequence results you want to view.