Andrew Lutsky
Sep 10, 2022
MCB432

Individual Major Assignment

*Disclaimer: I have read the instructions and followed the rules of this major assignment.

Introduction/Background:

16sRNA has been "a mainstay of sequence-based bacterial analysis for decades."(Johnson, Jethro S et al.) According to CD genomics, a genomics services company, 16sRNA is a highly conserved region with 9 specific regions of high variability and is present in "almost all bacteria."(Janda et al.) This characteristic makes it ideal to help identify the bacteria that it belongs to using PCR techniques. The general outline of this procedure involves amplifying 16sRNA in a "shotgun manner," using those reads of the RNA to form a consensus sequence, comparing the consensus to known bacteria sequences, and generating a phylogenetic tree. A phylogenetic tree is generated to identify the bacteria and relate it to known bacterial species.

Procedure:

1. The dataset assigned to me was downloaded from the course website(dataset#11).
2. The file was unzipped and moved into a directory in my home folder.
3. Through the shell terminal, the new directory was set as the working directory.
    a. $ cd ~/dataset#11
4. Then I activated the conda environment that enables me to work with the cap3(doi:10.1101/gr.9.9.868).
    a. $ conda activate myenv_x86
    b. *Cap3 or contig assembly program is a program that is used to assemble "contigs" or short reads that can form a continuous sequence. The input for the program takes in a fasta file of various sequences and then forms a continuous sequence.
    c. See Fig. 1-1 and 1-2 below.
5. Then we need to find the names of each fasta file in the dataset.
    a. $ ls *_*.fasta | awk -F. '{print $1}' | awk '{seq=seq $1" "}END{print seq}'
6. Using cap3 a bash script was used to perform cap3 on each of the fasta files given which compiles a consensus sequence for each, moves the cap files into a new directory, and writes a new file with a consensus sequence.
    a. $ for p in 2016Mar_A12 2016Mar_F01 2016Nov_A12 2016Nov_E05 2017Apr_A12 2017Apr_E03 2017Nov_B02 2017Nov_E12; do mkdir $p;cap3 $p.fasta | awk -v n=$p 'BEGIN{print ">" n "_Cap3"}/consensus/{seq=seq$2}END{print seq}'>$p"_Cap3".fasta; mv *.cap.* ./$p; done

b. *Note one file in dataset #11 did not generate a contig file with Cap3(2016Nov_A12)

7. After generating the consensus sequences each Cap_3 file or consensus sequence generated by Cap3 was entered into NCBI blastn using default parameters and searching the nucleotide collection database.
8. Then each top scoring sequence fasta file was downloaded.
9. Then Jalview was used to input each reference and BLAST sequence. Each Blast sequence name was modified to include species name.
    a. See Fig. 3 below.
10. Then a Neighbor Joining Tree was used to generate a phylogenetic tree and downloaded in Newick File Format. Then ETEToolkit was used to visualize the Newick file and generate a visualization of the tree.
    a. See Fig. 4 below.
11. All reference consensus sequences and BLAST sequences are then assembled into one fasta file.
    a. See Fig. 5 below.

Data:

Cap3 is a program that takes several reads of sequences and then assembles them into a consensus sequence. This is particularly useful for the purpose of assembling 16s RNA reads into one continuous sequence as it is easy to install and use and is well suited towards the goal of finding a consensus sequence for short reads of RNA. The following image is an example of an input for the Cap3 program.

```
>2017Apr_A12-1492R trim rev
TTTTTTTANGGNAAAGAAACCNANGTTCCGAAGAGGTTAACTTGGNTTGGTACCCTTTGANCGGTTACCTTANCCCAGAAAAGCCCCCNGGCTAAANTACGGTGCCCAGC
AGCCCGGCGGTAATACGTTAGGGTGGCAAAGCGTTNTNCCGGAATTATTGGGCGTAAAGCGCGCGCAGGNCGGTTTNNTTAAGTTTGATGTGAAAGCCCACGGCTCAACC
GTGGAGGGTCATTGGGAAACTGGGGAACTTGAGTGCAGAAGAGAAAAGCGGAATTCCACGTGTAGCGGTGAAATGCGTAGAGATGTGGAGGAACACCAGTGGCGAAGGCG
GCTTTTTTGGTCTGTAACTGACGCTGAGGCGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGTGCTAAGTGTTAGAGGGTT
TCCGCCCTTTAGTGCTGCAGCTAACGCATTAAGCACTCCGCCTGGGGAGTACGGTCGCAAGACTGAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATG
TGGTTTAATTCGAAGCAACGCGAAGAACCTTACCAGGTCTTGACATCCTCTGACAACTCTAGAGATAGAGCGTTCCCCTTCGGGGGACAGAGTGACAGGTGGTGCATGGT
TGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTGATCTTAGTTGCCAGCATTTAGTTGGGCACTCTAAGGTGACTGCCGGTGACAA
ACCGGAGGAAGGTGGGGATGACGTCAAATCATCATGCCCCTTATGACCTGGGCTACACACGTGCTACAATGGATGGTACAAAGGGCTGCAAGACCGCGAGGTCAAGCCAA
TCCCATAAAACCATTCTCAGTTCGGATTGTAGGCTGCAACTCGCCTACATGAAGCTGGAATCGCTAGTAATCGCGGATCAGCATGCCGCGGTGAATACGTTCCCGGGCCT
TGTACACCCCGCNCGTCACACCACGAGAGTTTGTAACCACCCGAAGTCGTGGAGTANCCGAAGAGCNGCCGNNNAGGGGAC

>2017Apr_A12-907R trim rev
TTTAGAGTTTGGATCNTGGCTCAGGATGAACCGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGAACTGATTAGAAGCTTGCTTTCTATGACGTTAGCGGCGGACGGG
TGAGTAACACGTGGGCAACCTGCCTGTAAGACTGGGATAACTTCGGGAAACCGAAGCTAATACCGGATAGGATCTTCTCCTTCATGGGAGATGATTGAAAGATGGTTTCG
GCTATCACTTACAGATGGGCCCGCGGTGCATTAGCTAGTTGGTGAGGTAACGGCTCACCAAGGCAACGATGCATAGCCGACCTGAGAGGGTGATCGGCCACACTGGGACT
GAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTAGGGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCAACGCCGCGTGAGTGATGAAGGCTTTCGGGTCGTAAAA
CTCTGTTGTTAGGGAAGAACAAGTACGAGAGTAACTGCTCGTACCTTGACGGTACCTAACCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTG
GCAAGCGTTATCCGGAATTATTGGGCGTAAAGCGCGCGCAGGCGGTTTCTTAAGTCTGATGTGAAAGCCCACGGCTCAACCGTGGAGGGTCATTGGAAACTGGGGAACTT
GAGTGCAGAAGAGAAAAGCGGAATTCCACGTGTAGCGGTGAAATGCGTAGAGATGTGGAGGAACACCAGTGGCGAAGGCGGCTTTTTGGTCTGTAAACTGACGCTGAGGC
GCGAAAGCGTGGGGAGCAAAACAGGATTAGATACCCTGGTAGTTCCACGCCGTAAATCGATNGAGGTGCTAAGTGTTAGAGGGTTTCCGCCCTTTAGTGCTGCAGCTAAN
CGCCTAAGCCTTCCCCNGNGNTNA

>2017Apr_A12-515F trim
TACGTTTCGGAATTATTGGGCGTAAGCGCGCGCAGGCGGTTTCTTAAGTCTGATGTGAAAGCCCACGGCTCNNCCGTGGAGGGTCATTGGAAACTGGGGAACTTGAGTGC
AGAAGAGAAAAGCGGAATTCCACGTGTAGCGGTGAAATGCGTAGAGATGTGGAGGAACACCAGTGGCGAAGGCGGCTTTTTGGTCTGTAACTGACGCTGAGGCGCGAAAG
CGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGTGCTAAGTGTTAGAGGGTTTCCGCCCTTTAGTGCTGCAGCTAACGCATTAAGCACT
CCGCCTGGGGAGTACGGTCGCAAGACTGAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCGAAGAACCTTACCAGG
TCTTGACATCCTCTGACAACTCTAGAGATAGAGCGTTCCCCTTCGGGGGACAGAGTGACAGGTGGTGCATGGTTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTC
CCGCAACGAGCGCAACCCTTGATCTTAGTTGCCAGCATTTAGTTGGGCACTCTAAGGTGACTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGTCAAATCATCATGC
CCCTTATGACCTGGGCTACACACGTGCTACAATGGATGGTACAAAGGGCTGCAAGACCGCGAGGTCAAGCCAATCCCATAAAACCATTCTCAGTTCGGATTGTAGGCTGC
AACTCGCCTACATGAAGCTGGAATCGCTAGTAATCGCGGATCAGCATGCCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCACGAGAGTTTGTTAA
CACCCGAAGTCGGTGGGATAACCGTAAGGAGCTAGCCGCCTAAGGTGGGANANATGATTGGGGTGAANTCNTANNAAAGGGTTAACCAAA
```

Figure 1-1. Cap3 Input Fasta File

Cap3 then generates several output files. The following image is an example of a consensus sequence that Cap3 outputs.

```
>2016Mar_A12_Cap3
TTTAGAGTTTGGATCNTGGCTCAGGATGAACCGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGAACTGATTAGAAGCTTGCTTTCTATGACGTTAGCGGCGGACGGG
TGAGTAACACGTGGGCAACCTGCCTGTAAGACTGGGATAACTTCGGGAAACCGAAGCTAATACCGGATAGGATCTTCTCCTTCATGGGAGATGATTGAAAGATGGTTTCG
GCTATCACTTACAGATGGGCCCGCGGTGCATTAGCTAGTTGGTGAGGTAACGGCTCACCAAGGCAACGATGCATAGCCGACCTGAGAGGGTGATCGGCCACACTGGGACT
GAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTAGGGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCAACGCCGCGTGAGTGATGAAGGCTTTCGGGTCGTAAAA
CTCTGTTGTTAGGGAAGAACAAGTACGAGAGTAACTGCTCGTACCTTGACGGTACCTAACCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTG
GCAAGCGTTATCCGGAATTATTGGGCGTAAAGCGCGCGCAGGCGGTTTCTTAAGTCTGATGTGAAAGCCCACGGCTCAACCGTGGAGGGTCATTGGAAACTGGGGAACTT
GAGTGCAGAAGAGAAAAGCGGAATTCCACGTGTAGCGGTGAAATGCGTAGAGATGTGGAGGAACACCAGTGGCGAAGGCGGCTTTTTGGTCTGTAAACTGACGCTGAGGC
GCGAAAGCGTGGGGAGCAAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGTGCTAAGTGTTAGAGGGTTTCCGCCCTTTAGTGCTGCAGCTAACGCAT
TAAGCACTCCGCCTGGGGAGTACGGTCGCAAGACTGAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCGAAGAACC
TTACCAGGTCTTGACATCCTCTGACAACTCTAGAGATAGAGCGTTCCCCTTCGGGGGACAGAGTGACAGGTGGTGCATGGTTGTCGTCAGCTCGTGTCGTGAGATGTTGG
GTTAAGTCCCGCAACGAGCGCAACCCTTGATCTTAGTTGCCAGCATTTAGTTGGGCACTCTAAGGTGACTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGTCAAAT
CATCATGCCCCTTATGACCTGGGCTACACACGTGCTACAATGGATGGTACAAAGGGCTGCAAGACCGCGAGGTCAAGCCAATCCCATAAAACCATTCTCAGTTCGGATTG
TAGGCTGCAACTCGCCTACATGAAGCTGGAATCGCTAGTAATCGCGGATCAGCATGCCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCACGAGAG
TTTGTTAACACCCGAAGTCGGTGGGATAACCGTAAGGAGCTAGCCGCCTAAGGTGGGANANATGATTGGGGTGAANTCNTANNAAAGGGTTAACCAAA
```

Figure 1-2. Cap3 Output Fasta File

Blastn was used to determine closest alignments to the contig sequences generated by Cap3. An example can be found below, where Cap3 generated a contig file and that file was inputted into BLAST to find the closest alignment to the given contig. This data will allow us to eventually generate phylogenetic information and distances.

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Bacillus aryabhattai strain IHB B 7058 16S ribosomal RNA gene, partial sequence	Priestia aryabhattai	2728	2728	99%	0.0	98.95%	1520	KJ721212.1
☑	Bacillus aryabhattai strain IGND-13 16S ribosomal RNA gene, partial sequence	Priestia aryabhattai	2726	2726	99%	0.0	99.01%	1522	MN133922.1
☑	Bacillus aryabhattai partial 16S rRNA gene, strain B39	Priestia aryabhattai	2721	2721	98%	0.0	99.14%	1679	LN890215.1
☑	Bacillus aryabhattai partial 16S rRNA gene, strain L33	Priestia aryabhattai	2721	2721	98%	0.0	99.14%	1696	LN890029.1
☑	Priestia aryabhattai strain PCA7 16S ribosomal RNA gene, partial sequence	Priestia aryabhattai	2721	2721	98%	0.0	99.14%	1533	OK083712.1
☑	Bacillus aryabhattai strain JN33 16S ribosomal RNA gene, partial sequence	Priestia aryabhattai	2721	2721	98%	0.0	99.14%	1538	KF150346.1

Figure 2. BLAST Results for 2016Mar_A12_Cap3.fasta

Then each fasta file for the top score was downloaded into the dataset folder and inputted into the program JalView. An image of the Jalview Viewer can be seen below. This program allows for generation of a Tree using an algorithm of your choice between Neighbor Joining methods and Average Distance Tree Generation. The generated tree below was created using Neighbor Joining Methods. The Phylogenetic Tree file was downloaded as a Newick File Format and then viewed via the ETEToolkit(https://doi.org/10.1093/molbev/msw046).

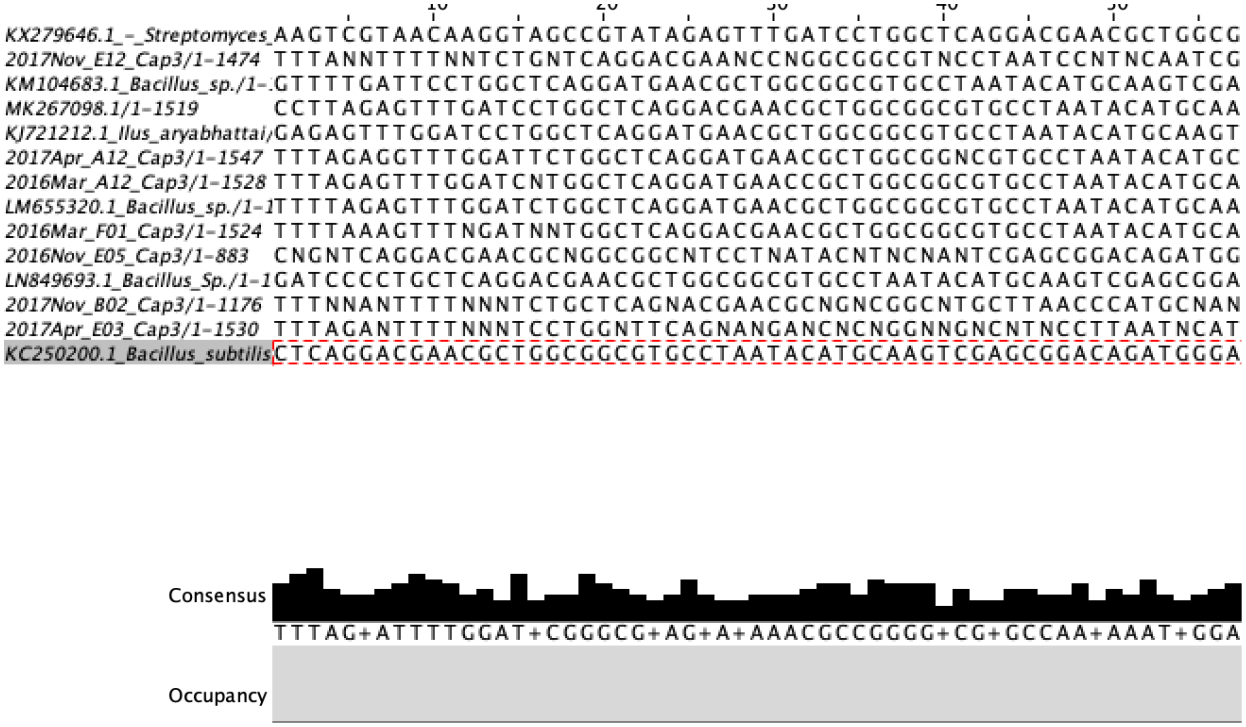


Figure 3. Jalview View of Reference and BLAST sequences

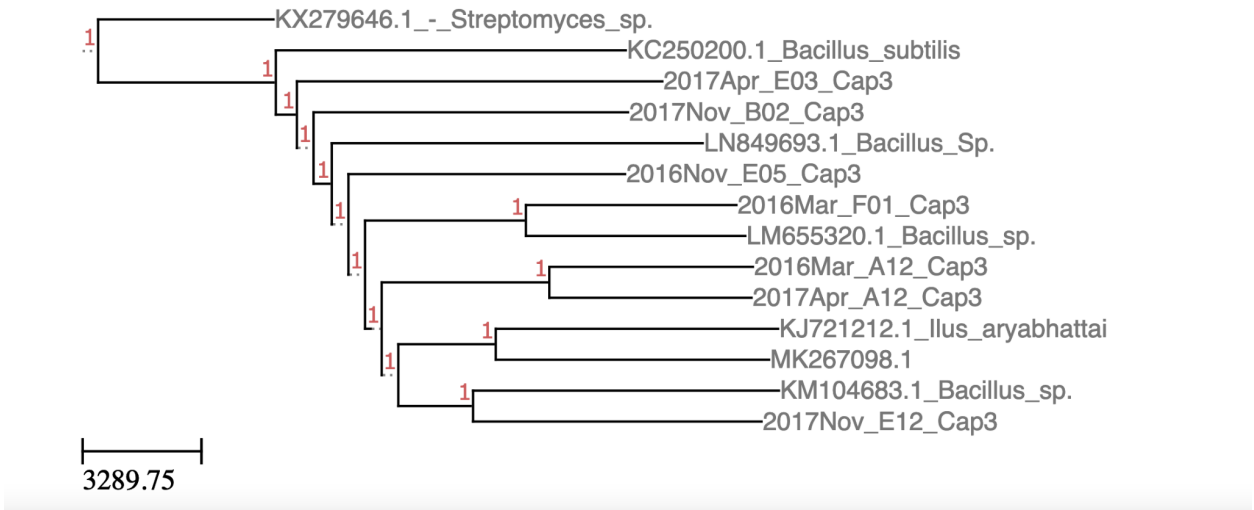
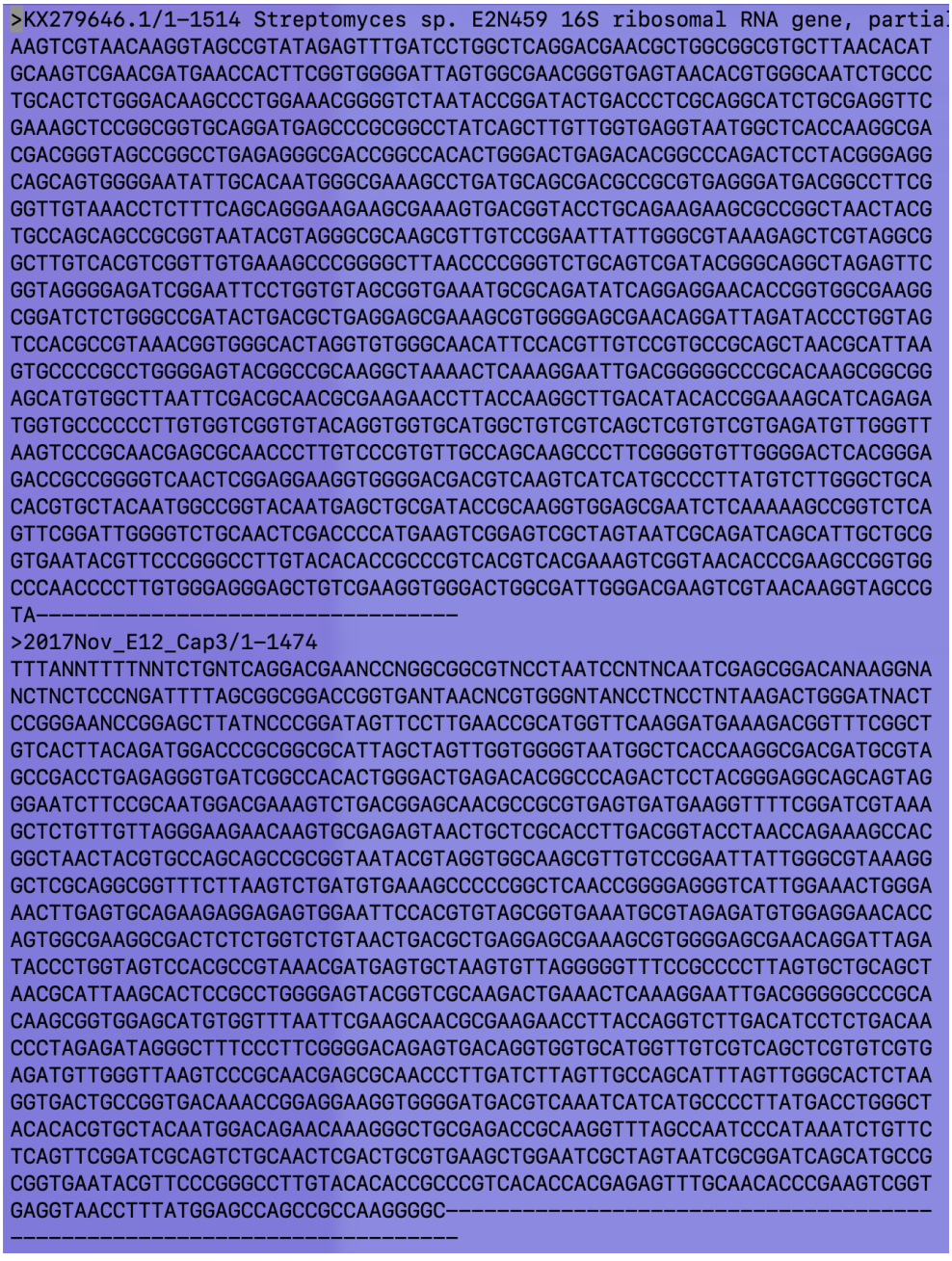


Figure 4. Phylogenetic Tree generated from Reference and Blast Sequences using Jalview and ETEToolkit

After the phylogenetic tree was generated in Jalview and using ETEToolkit a fasta file was generated that contained all the reference and Blast Sequences. A screenshot of this can be seen below.



```
>KX279646.1/1-1514 Streptomyces sp. E2N459 16S ribosomal RNA gene, partial
AAGTCGTAACAAGGTAGCCGTATAGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCTTAACACAT
GCAAGTCGAACGATGAACCACTTCGGTGGGGATTAGTGGCGAACGGGTGAGTAACACGTGGGCAATCTGCCC
TGCACTCTGGGACAAGCCCTGGAAACGGGGTCTAATACCGGATACTGACCCTCGCAGGCATCTGCGAGGTTC
GAAAGCTCCGGCGGTGCAGGATGAGCCCGCGGCCTATCAGCTTGTTGGTGAGGTAATGGCTCACCAAGGCGA
CGACGGGTAGCCGGCCTGAGAGGGCGACCGGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGG
CAGCAGTGGGGAATATTGCACAATGGGCGAAAGCCTGATGCAGCGACGCCGCGTGAGGGATGACGGCCTTCG
GGTTGTAAACCTCTTTCAGCAGGGAAGAAGCGAAAGTGACGGTACCTGCAGAAGAAGCGCCGGCTAACTACG
TGCCAGCAGCCGCGGTAATACGTAGGGCGCAAGCGTTGTCCGGAATTATTGGGCGTAAAGAGCTCGTAGGCG
GCTTGTCACGTCGGTTGTGAAAGCCCGGGGCTTAACCCCGGGTCTGCAGTCGATACGGGCAGGCTAGAGTTC
GGTAGGGGAGATCGGAATTCCTGGTGTAGCGGTGAAATGCGCAGATATCAGGAGGAACACCGGTGGCGAAGG
CGGATCTCTGGGCCGATACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGAACAGGATTAGATACCCTGGTAG
TCCACGCCGTAAACGGTGGGCACTAGGTGTGGGCAACATTCCACGTTGTCCGTGCCGCAGCTAACGCATTAA
GTGCCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGCGG
AGCATGTGGCTTAATTCGACGCAACGCGAAGAACCTTACCAAGGCTTGACATACACCGGAAAGCATCAGAGA
TGGTGCCCCCCTTGTGGTCGGTGTACAGGTGGTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTT
AAGTCCCGCAACGAGCGCAACCCTTGTCCCGTGTTGCCAGCAAGCCCTTCGGGGTGTTGGGGACTCACGGGA
GACCGCCGGGGTCAACTCGGAGGAAGGTGGGGACGACGTCAAGTCATCATGCCCCTTATGTCTTGGGCTGCA
CACGTGCTACAATGGCCGGTACAATGAGCTGCGATACCGCAAGGTGGAGCGAATCTCAAAAAGCCGGTCTCA
GTTCGGATTGGGGTCTGCAACTCGACCCCATGAAGTCGGAGTCGCTAGTAATCGCAGATCAGCATTGCTGCG
GTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACGTCACGAAAGTCGGTAACACCCGAAGCCGGTGG
CCCAACCCCTTGTGGGAGGGAGCTGTCGAAGGTGGGACTGGCGATTGGGACGAAGTCGTAACAAGGTAGCCG
TA----------
>2017Nov_E12_Cap3/1-1474
TTTANNTTTTNNTCTGNTCAGGACGAANCCNGGCGGCGTNCCTAATCCNTNCAATCGAGCGGACANAAGGNA
NCTNCTCCCNGATTTTAGCGGCGGACCGGTGANTAACNCGTGGGNTANCCTNCCTNTAAGACTGGGATNACT
CCGGGAANCCGGAGCTTATNCCCGGATAGTTCCTTGAACCGCATGGTTCAAGGATGAAAGACGGTTTCGGCT
GTCACTTACAGATGGACCCGCGGCGCATTAGCTAGTTGGTGGGGTAATGGCTCACCAAGGCGACGATGCGTA
GCCGACCTGAGAGGGTGATCGGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTAG
GGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCAACGCCGCGTGAGTGATGAAGGTTTTCGGATCGTAAA
GCTCTGTTGTTAGGGAAGAACAAGTGCGAGAGTAACTGCTCGCACCTTGACGGTACCTAACCAGAAAGCCAC
GGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGGAATTATTGGGCGTAAAGG
GCTCGCAGGCGGTTTCTTAAGTCTGATGTGAAAGCCCCCGGCTCAACCGGGGAGGGTCATTGGAAACTGGGA
AACTTGAGTGCAGAAGAGGAGAGTGGAATTCCACGTGTAGCGGTGAAATGCGTAGAGATGTGGAGGAACACC
AGTGGCGAAGGCGACTCTCTGGTCTGTAACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGAACAGGATTAGA
TACCCTGGTAGTCCACGCCGTAAACGATGAGTGCTAAGTGTTAGGGGGTTTCCGCCCCTTAGTGCTGCAGCT
AACGCATTAAGCACTCCGCCTGGGGAGTACGGTCGCAAGACTGAAACTCAAAGGAATTGACGGGGGCCCGCA
CAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCGAAGAACCTTACCAGGTCTTGACATCCTCTGACAA
CCCTAGAGATAGGGCTTTCCCTTCGGGGACAGAGTGACAGGTGGTGCATGGTTGTCGTCAGCTCGTGTCGTG
AGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTGATCTTAGTTGCCAGCATTTAGTTGGGCACTCTAA
GGTGACTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGTCAAATCATCATGCCCCTTATGACCTGGGCT
ACACACGTGCTACAATGGACAGAACAAAGGGCTGCGAGACCGCAAGGTTTAGCCAATCCCATAAATCTGTTC
TCAGTTCGGATCGCAGTCTGCAACTCGACTGCGTGAAGCTGGAATCGCTAGTAATCGCGGATCAGCATGCCG
CGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCACGAGAGTTTGCAACACCCGAAGTCGGT
GAGGTAACCTTTATGGAGCCAGCCGCCAAGGGGC----------
----------
```

Figure 5. Compiled Sequence Fasta File

Discussion/Conclusions:

In summation for the project, 16sRNA reads were used to generate phylogenetic information. The processing of the initial data was done using cap3, Jalview, and ETEToolkit. Cap3 allows for generation of contig sequences, Jalview allows for generation of trees, and the ETEToolkit allows for visualization of that tree data. The generated phylogenetic tree using Neighbor Joining methods allows for identification of protein using 16sRNA reads and comparing it to known 16sRNA information of other bacteria species.

Cap3, as previously mentioned, takes short reads and compiles them into consensus sequences. This can be done by hand or using supplementary tools like BLAST to find alignments and clip parts of the sequences off, but to a lower degree of accuracy. It is significantly more accurate and fast to use a program for data, particularly if examining more than one sequence. Cap3 in this particular case takes our several short reads and compiles them into consensus sequences, however it failed to use one of the datasets so this dataset was ignored.

Jalview is a versatile viewing tool. It allows for figure making, tree generation, and multiple sequences. alignment viewing, and much more. Jalview is well known in the industry as it is intuitive and free software for visualization. In this case, it was used to generate Tree information using Neighbor Joining methods. There are several methods for tree generation including but not limited to, Neighbor Joining, Average Distance, Maximum Parsimony, and Maximum Likelihood methods. Neighbor Joining was used in this particular case. Jalview allows for easy alignment and import/export of files. The tree generated from the dataset was exported as a Newick File Format for better visualization.

The tree Newick file was input into a program known as ETEToolkit. ETEToolkit has a free web service that allows for easy visualization of phylogenetic trees. If Fig. 4 is examined, we can see for example, that KM104683.1_Bacillus_sp. and 2017Nov_E12 are similar in terms of “distance” on the tree to each other. This tree visualization method allows for easy identification of bacteria.

Works Cited

Altschul, S F et al. “Basic local alignment search tool.” Journal of molecular biology vol. 215,3 (1990): 403-10. doi:10.1016/S0022-2836(05)80360-2

Jaime Huerta-Cepas, François Serra, Peer Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data, *Molecular Biology and Evolution*, Volume 33, Issue 6, June 2016, Pages 1635–1638, https://doi.org/10.1093/molbev/msw046

Janda, J Michael, and Sharon L Abbott. “16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls.” Journal of clinical microbiology vol. 45,9 (2007): 2761-4. doi:10.1128/JCM.01228-07

Johnson, Jethro S et al. “Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis.” *Nature communications* vol. 10,1 5029. 6 Nov. 2019, doi:10.1038/s41467-019-13036-1

Huang, X, and A Madan. “CAP3: A DNA sequence assembly program.” *Genome research* vol. 9,9 (1999): 868-77. doi:10.1101/gr.9.9.868