**2)** [**https://github.com/bcgsc/abyss**](https://github.com/bcgsc/abyss)**. This is the link to the documentation for ABySS. In your own words, please summarize the function of each of the commands (e.g., abyss-pe, k, B, etc) that you included in your code.**

abyss-pe name=assembly k=96 B=2G in='SRR32657023\_1.fastq.gz SRR32657023\_2.fastq.gz'

abyss-pe  
This is what program is going to run in the command line

name=assembly  
Specifies that the output files’ names will all start with “assembly

k=96  
Specifies that the word size is 96. The program will process data in 96-bit long “chunks.”

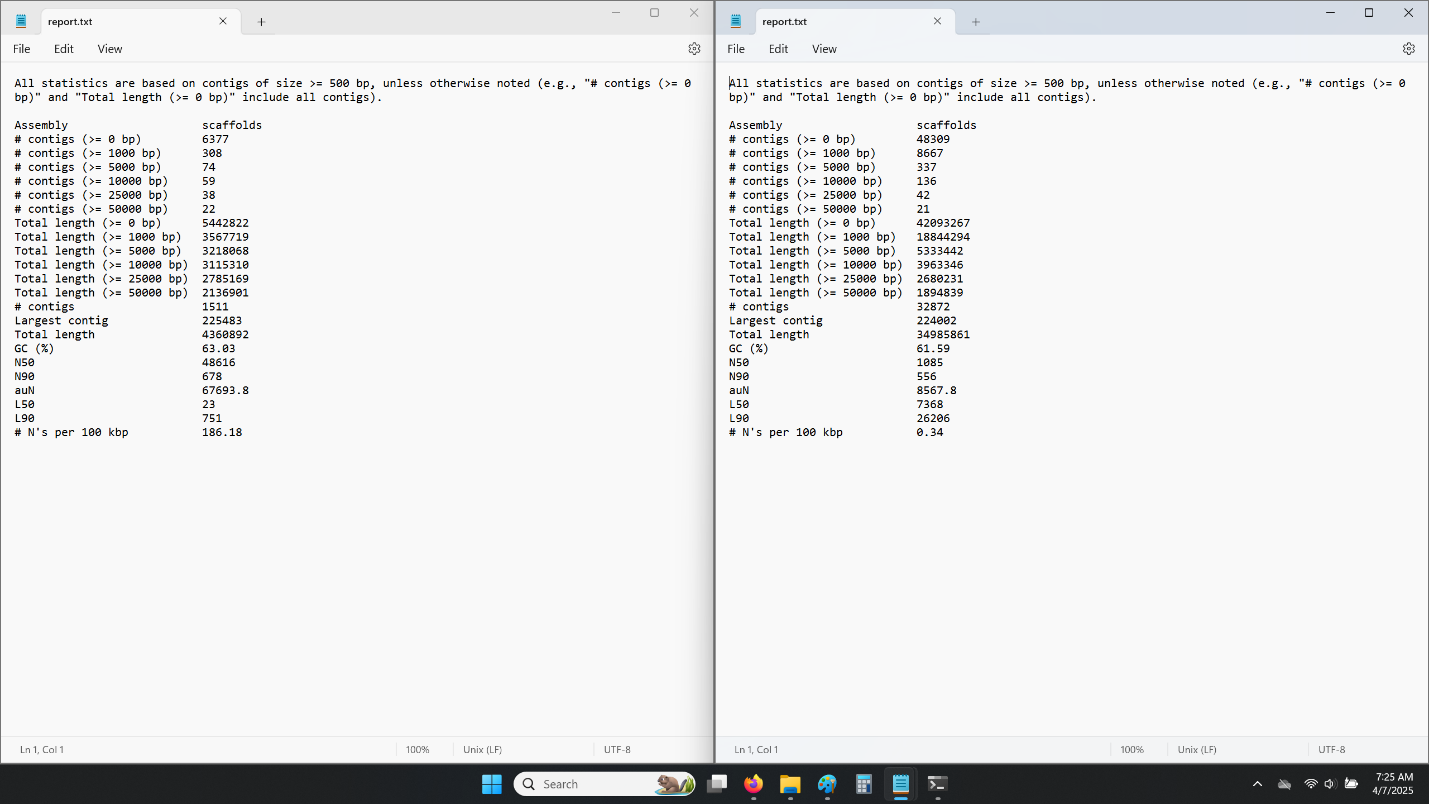
B=2G  
This specifies a maximum amount of RAM that the program is allowed to use.

in='SRR32657023\_1.fastq.gz SRR32657023\_2.fastq.gz'  
Specifies the input files

**3)** [**https://ablab.github.io/spades/index.html.**](https://ablab.github.io/spades/index.html) **This is the documentation for SPAdes. Based on this manual, can you identify how you could modify the code you used to do a hybrid assembly with nanopore reads? Please explain what a hybrid assembly is and why someone might want to do that.**

To include a file with Oxford Nanopore reads in a hybrid assembly, you would add  
--nanopore <file\_name>  
to the line of code

**4) Include a screenshot of the QUAST assembly statistics for the ABySS and SPAdes assembly. This is a demo from the sample files we worked on.**



The left half of the image contains the reports.txt file from quastabyssgalaxy

The right half of the image contains the reports.txt file from quastspadesgalaxy

**5) Based on the statistics from your genome, which assembly do you think is best? Why? This is the assembly you can use going forward.**

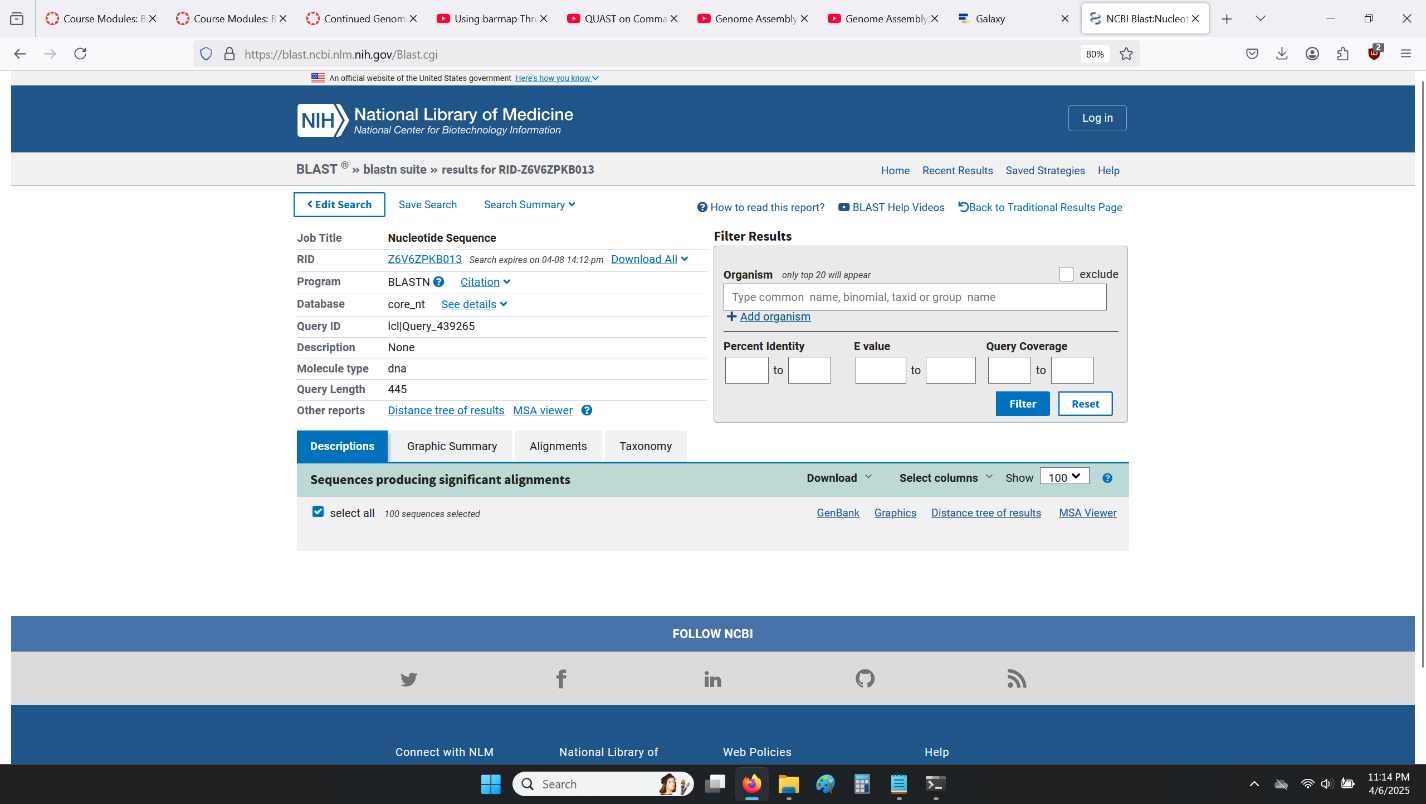
These two reports are wildly different. They had the same input data but seem like they’re of different genomes. I used the spades data for the rest of the assignment. Attempting to use the Abyss data returned more errors.

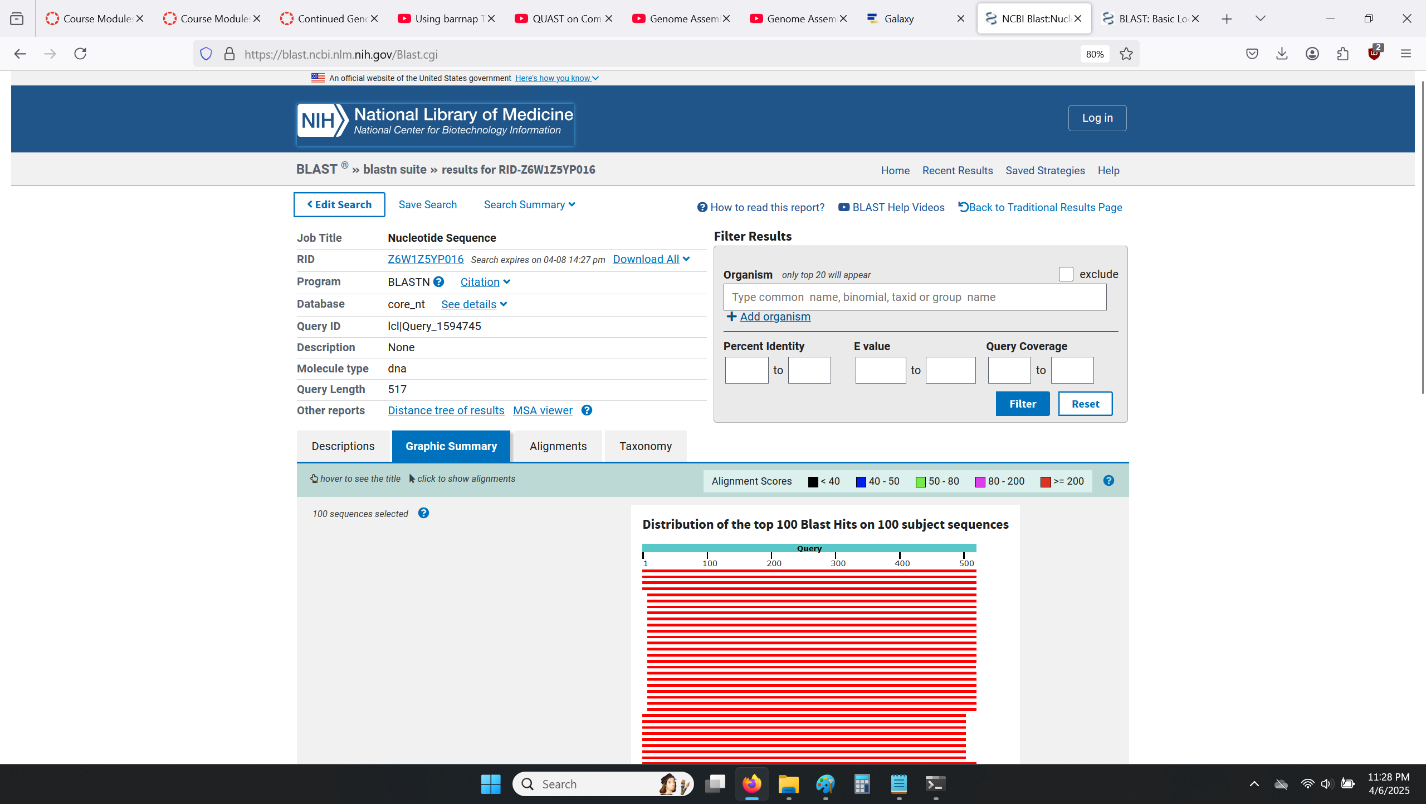
**6) How can we use barrnap to figure out what species we have? Why is using the 16S rRNA sequence a good, but imperfect, tool for identifying species identity?**

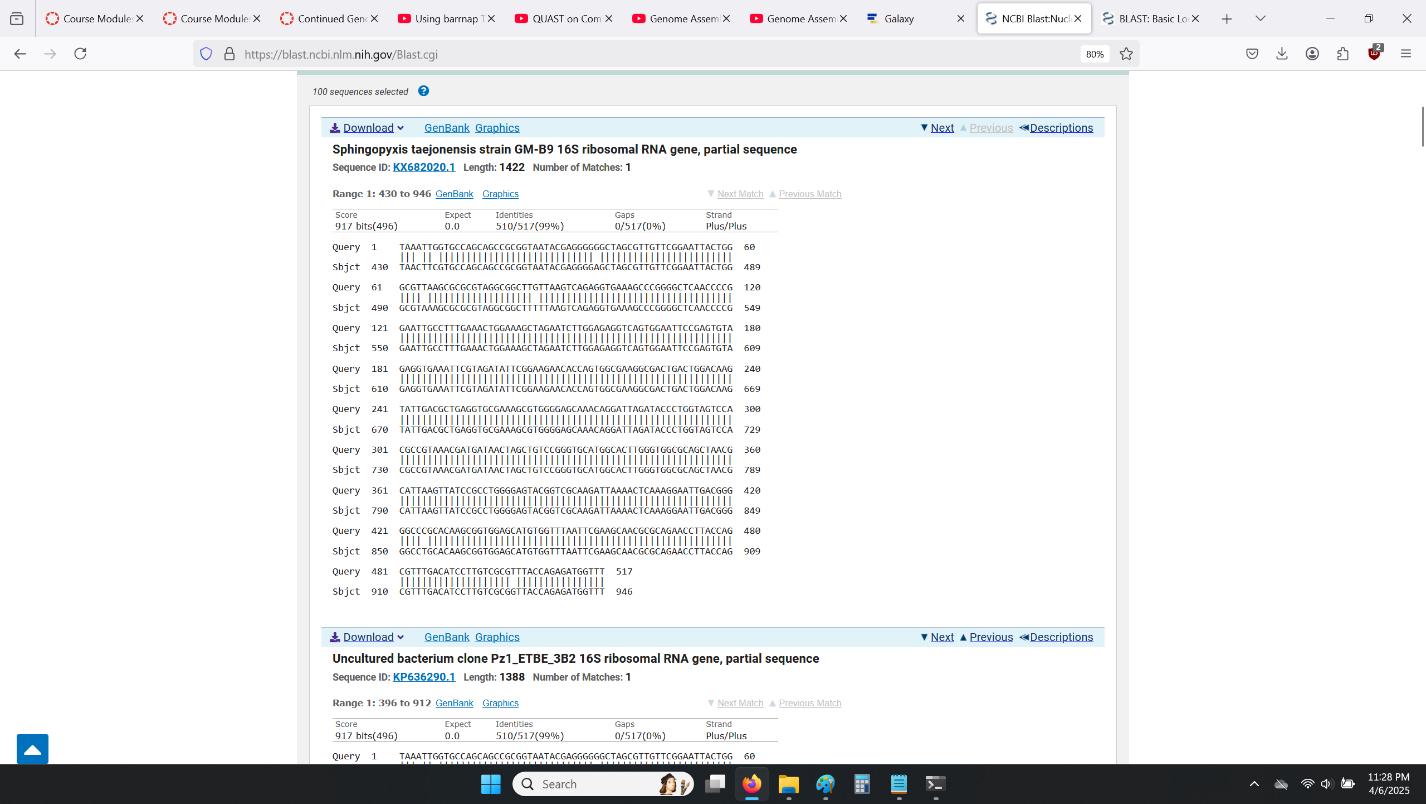
Barrnap finds out where the 16S rRNA sequence is in the genome, while bedtools pulls those sequences  
16S rRNA sequences are good for identification because every species of bacteria has a 16S rRNA sequence, but they’re all unique to each species.

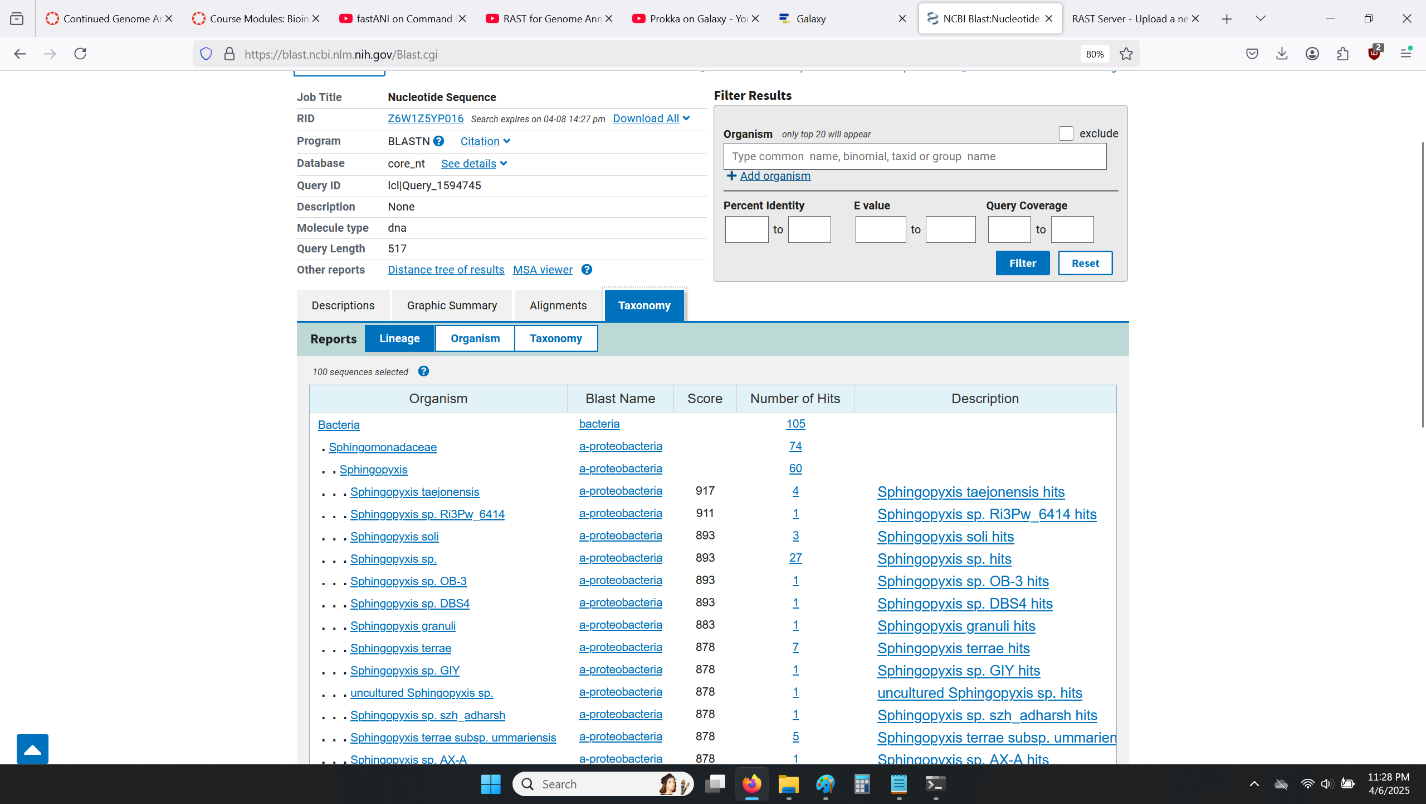
**7) What species do you have? Include a screenshot of your top NCBI results.**

My rRNAsequences.fasta file had several different nodes labeled as “16S\_rRNA.”  
Each of these returned different results when put into BLASTn.  
BLASTn would not display results on the “Descriptions” tab, but the other tabs worked as expected.

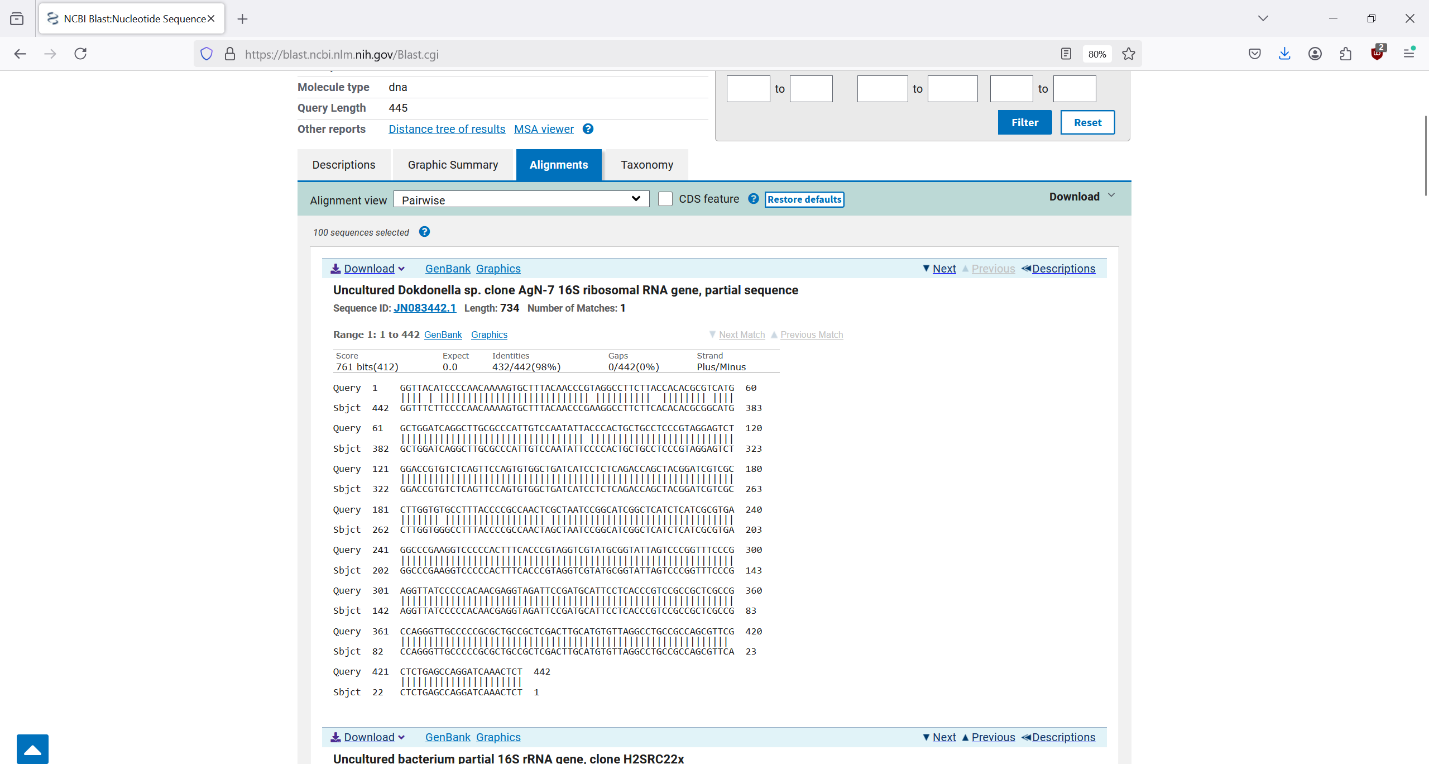




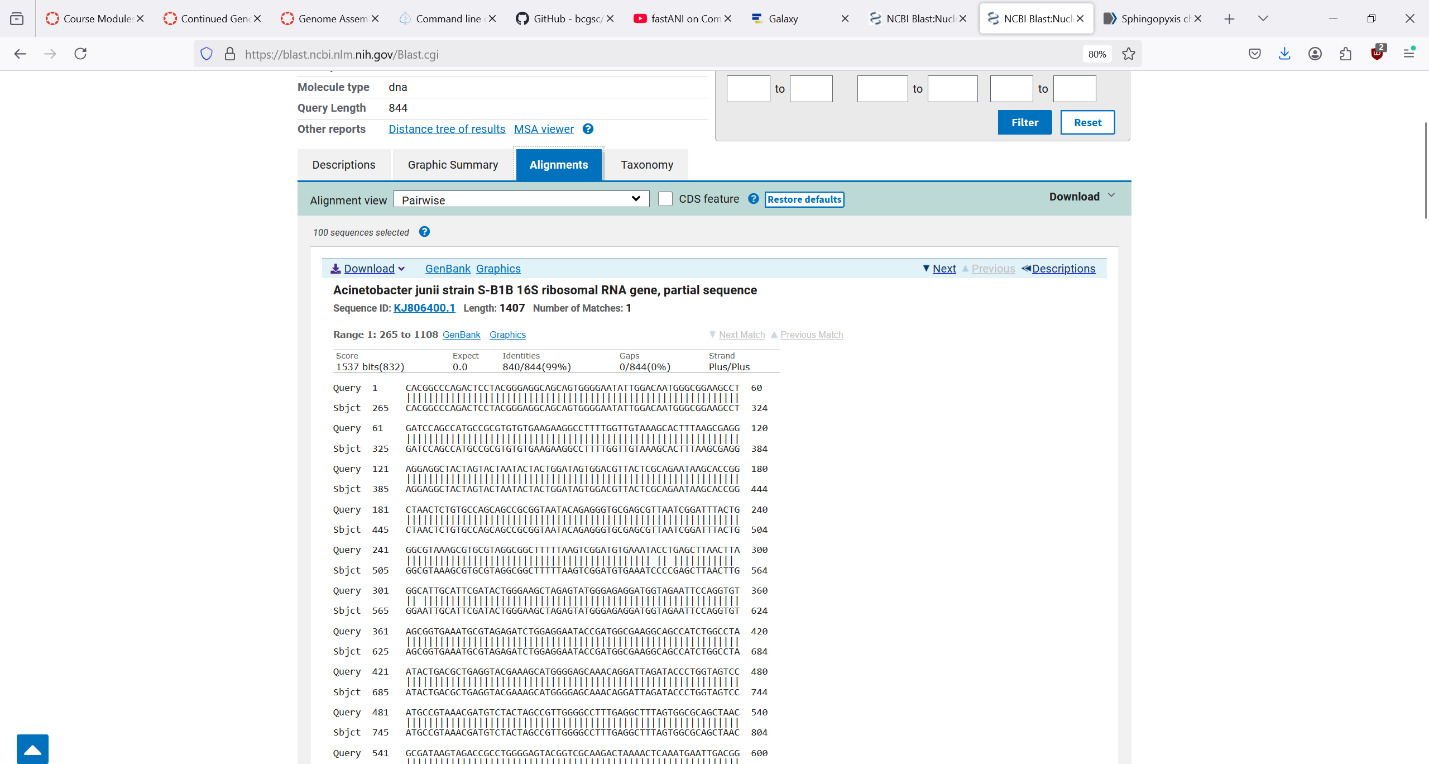




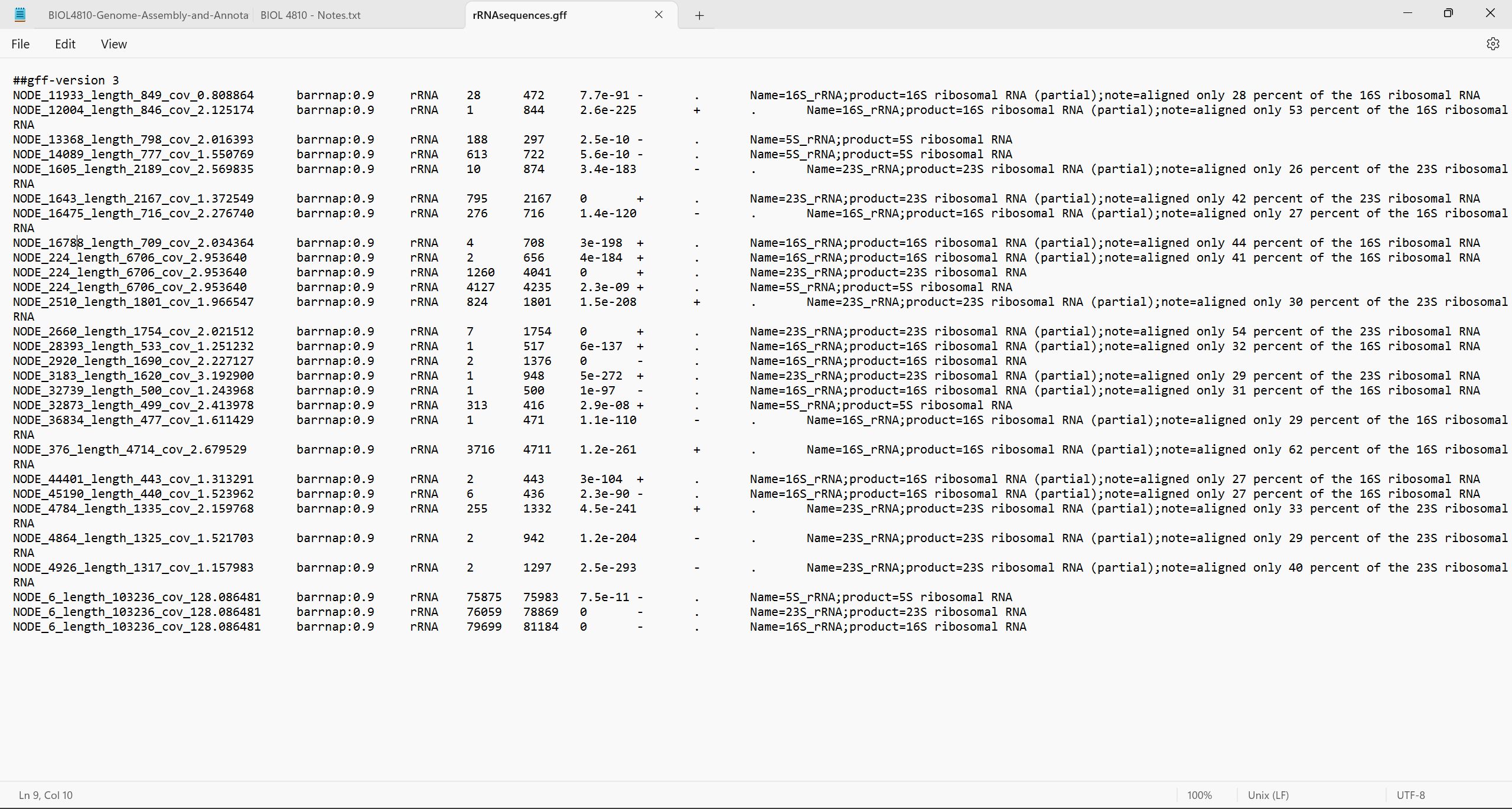
These four screenshots are of the results when NODE\_ 28393 is used as input.



This is a screenshot of the results when NODE\_11933 is used as input



These are the results when NODE\_12004 is used.

  
This shows the multiple rRNA sequences.

**8) What is genome annotation? Why is it important to do that?**

Genome annotation is a process of identifying which sections of a genome are genes and what those genes do. This is important because it allows us to determine what the organism is capable of. Even if it’s a novel organism, by comparing sections of its genome with similar sections from known organisms, we can glean some of the functions of the novel organism.

**9) Perform a genome annotation using two different programs. Find 3 of the 5 genes/features in your results file and create a table of those results: recA, gyrA, 16S rRNA, rpsB, dnaA. What is the location of the genes you chose? What does each program tell you about the gene? How are the outputs different between the two programs.**

Attempting to use RAST online resulted in the following error. The same error was reached using files from spadesoutgalaxy or abyssoutgalaxy.

A screenshot of a computer

AI-generated content may be incorrect.  
Attempting to use SPAdes or ABySS in the command line resulted in errors and there was no scaffolds file in the output.

**9) Perform a genome annotation using two different programs. Find 3 of the 5 genes/features in your results file and create a table of those results: recA, gyrA, 16S rRNA, rpsB, dnaA. What is the location of the genes you chose? What does each program tell you about the gene? How are the outputs different between the two programs.**

|  |  |  |
| --- | --- | --- |
| **Gene** | **DFAST Location** | **Prokka Location** |
| recA | 10946..12034 | 10946..12034 |
| gyrA | complement(79268..82009) | complement(79268..82009)  ( showed up as gene="gyrA\_1" ) |
| rpsB | complement(370..1278) | complement(144945..145943) ( showed up as gene="rpsB\_1" ) |
| dnaA | complement(6330..7664) | 39971..40963  ( showed up as gene="dnaA\_1" ) |

The programs give the start and end loci of the genes, the identification percentage, and a description of what the gene is/does.  
DFAST and Prokka gave the same locations for some of the genes, but displayed different locations for others. Prokka had multiple hits for most of the genes (gyrA\_1, gyrA\_2, etc.) while DFAST returned only one result for each gene.

**10) Create a table for your ANI results. How do you interpret these results? What do each of the columns represent? Hint: You can refer to the original paper to find that information.**

Assuming the organism was Sphingopyxis taejonesis, I downloaded the genomes for Sphingopyxis alaskensis and Sphingopyxis chilensis. When comparing these as neighbors, the output files for fastANI were blank (0 bytes), regardless of if fastANI was run in the command line or in Galaxy.

Redoing this step using the longest nucleotide sequence I could find at a glance and assuming the organism was Nitrobacter wingradskyi, I downloaded the genomes for Nitrobacter hamburgensis and Nitrobacter vulgaris.

Rerunning fastANI from the command line using these neighbors, I got the following results:

spadesoutgalaxy/scaffolds.fasta neighbors/vulgaris.fasta 86.6124 818 2059

spadesoutgalaxy/scaffolds.fasta neighbors/hamburgensis.fasta 84.6228 789 2059

My genome was 86.6124% similar to N. vulgaris and 84.6228% similar to N. hamburgensis.

Running fastANI through galaxy returned the same results, using SPAdes Scaffolds as the query sequence and vulgaris.fasta and hamburgensis.fasta as the reference sequences.