1. *From the ABySS output, create a table for the unitigs, contigs, and scaffolds with the number of each, N50 for each, and predicted genome length.*
2. [*https://github.com/bcgsc/abyssLinks to an external site.*](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=Unknown k=96 B=2G in='Reads1.fastq.gz Reads2.fastq.gz'

Abyss-pe: means that we are using abyss with paired end reads

Name=unkown: means the name of the files will start with “unkown”

K=96: this is the kmer which is a a nucleotide sequence of a specific length, in this case the length is 96

B=2G: means that the computer will only use 2G of memory for this project

In='Reads1.fastq.gz Reads2.fastq.gz': this is telling the computer where to find the information to accomplish this task

1. *Using either output, perform a BLAST search to identify your species. Write your species name here:*

*Klebsiella pneumoniae*

*4) Perform quality assessment using QUAST. You need find a reference genome and reference annotation to upload to QUAST for the best quality check. Which assembler gave you the higher quality output? How do you know?*

Spades gave the higher quality output for several reasons. The N50 on the Spades output (194,930) was higher than the Abyss output (71,324). This means the half the pieces in the Spades output are 194,930 or longer. This indicates higher quality because it has fewer small pieces than the Abyss output. The L50 of the Spades output (9) is lower than the L50 of the Abyss output (25). This means that it only takes 9 contigs/scaffolds to have 50% of the genome in the Spades assembly, in comparison to 25 in the Abyss output. This supports the previous conclusion, that the Spades output generated fewer, longer segments, indicating a higher quality assembly.

*5) Describe what BUSCO is used for. What were the BUSCO values for your assembly?*

BUSCO is used to measure the completeness of the assembled genome. For SPADES, the BUSCO value was 97.97%. For Abyss, the BUSCO value was 97.97%.

*6) Perform a genome annotation using Prokka. Find 3 of the 5 genes/features in your results file and create a table of those results: recA, gyrA, 16S rRNA, rpsB, dnaA.*



*7)*[*https://github.com/tseemann/prokkaLinks to an external site.*](https://github.com/tseemann/prokka)*Here is the documentation for prokka. In your own words, what is the function of each of the commands in your line of code?*

Prokka --outdir prokkaGAA--prefix unknown AbyssOutputGAA/unknown-8.fa

Prokka: tells the computer to run the program called prokka

--outdir: creates an output directory called prokkaGAA

--prefix unknown: tells the computer to create files with the prefix “unknown”

AbyssOutputGAA/unknown-8.fa: tells prokka to look in the folder AbyssOutputGAA for the file unknown-8.fa and use that for the analysis

*8) What is the function of the genes/features you chose?*

recA functions in DNA repair

gyrA functions in winding the DNA tightly so that it will fit inside the cell

rpsB produces a protein that is a part of the 30S ribosomal subunit, which helps translate mRNA

*9) Find those same genes/features in your RAST annotation. What information did you learn about them from RAST?*

recA is part of a system that uses MutS and a hypothetical protein called recX that regulates recA.

The protein produced by gyrA is a subunit of a larger system of proteins called DNA gyrase.

The proteins composing the small subunit of ribosomes are extremely similar in all strains of bacteria.

*10) Upload the folder of this information to your GitHub in your Bioinformatics Repository. Please share the link to your repository.*

https://github.com/NathanielDavies/Bioinformatics