Erika Maldonado-Rosado

17 February 2019

Prokka Lab

BINF 6203

**Introduction**

For this lab we need to understand the benefits of Prokka. We are trying to learn how it can produce more accurate annotations and faster and more efficient. We will be recreating annotation for the NC files that were given to us in lab 2 by Dr. Gibas. In this lab we need to figure out how many annotated regions and how many of those gene annotations are protein coding genes are in each file. Then view it with Artemis or IGV to see if prokka gave more information per each gene.

**Method**

For this lab we need to download Prokka. For the purpose of this lab I used Ubuntu and Linux based system. I then found the Prokka manual which explained how to get prokka on my virtual machine. Once I completed this I got the files from Lab 2 in the drop box. I used NC\_007898.fasta and NC\_000913.fna. I compared the two and used IGV to visualize.

**Results**

For the ERR0008 file I viewed the .ffn file and it showed that there are 4301 of line of total annotated regions, 18168 of words, and 204958 of characters. The .ffa files showed that there are 4232 of gene annotations that are in protein coding genes, 18028 of words, and 202873 of characters. There are more annotated regions than gene annotations that code for protein coding genes.

For the ERR002 file I viewed the .ffn files and it showed that there are 4303 of line of total annotated regions, 18172 of words, and 205005 of characters. The .ffa files showed that there are 4235 of gene annotations that are in protein coding genes, 18034 of words, and 202950 of characters. There are more annotated regions than gene annotations that code for protein coding genes. For the file of E. coli of ERR002 there are more lines of annotated regions and of gene annotations that are in protein coding genes.

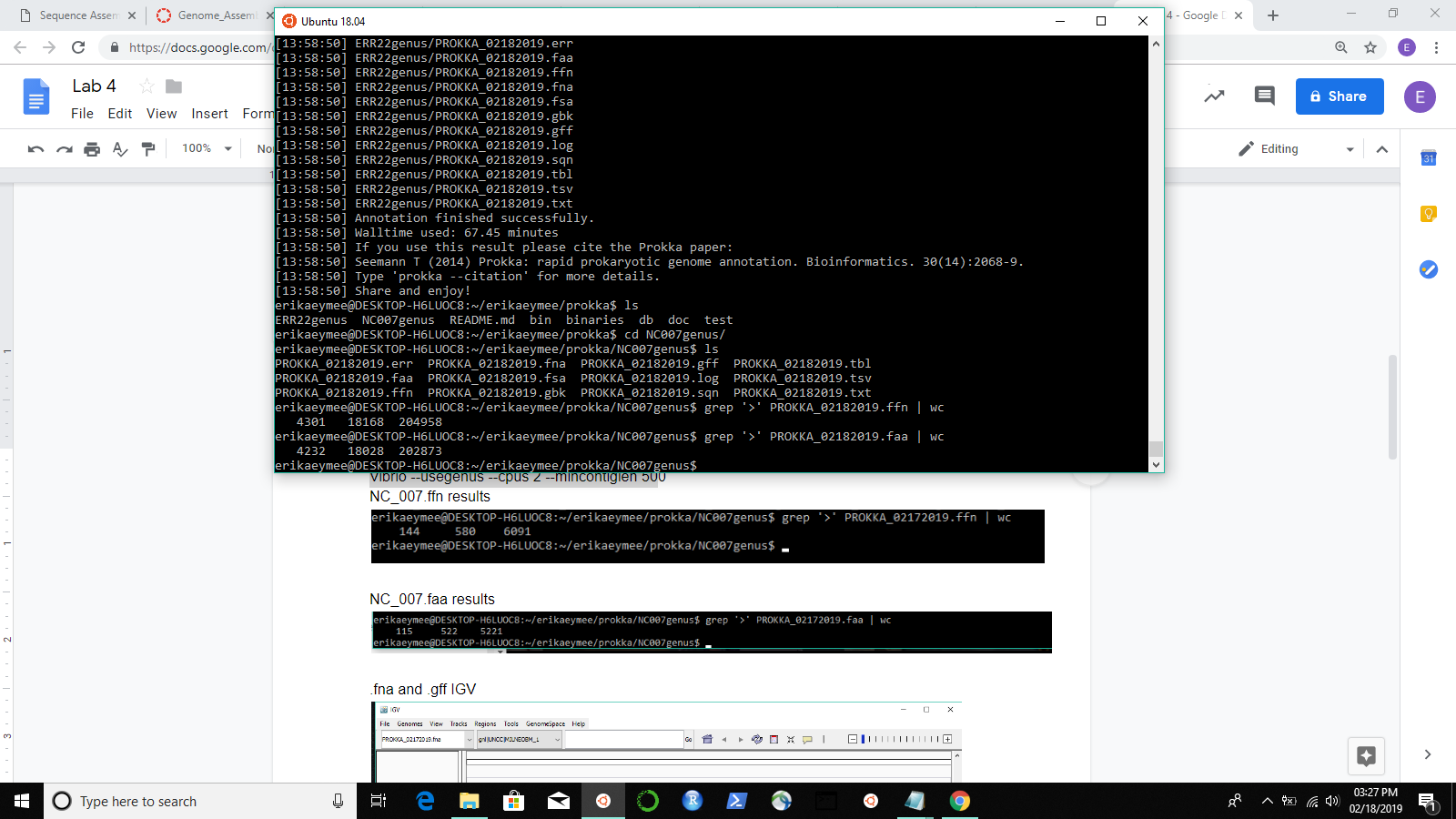
For the NC\_007898.fasta file I viewed the .ffn file and it showed that there are 144 lines of total annotated regions, 580 words, and 6091 characters. The .ffa files showed that there are 115 genes annotation that are protein coding genes, 522 total words, and 5221 characters. I believe it was less characters due to less genes being present or one gene coding for multiple proteins.

For the NC\_0009813.fna file I viewed the .ffn files and it showed that there are 4404 lines fo annotated regions, 18580 words, and 209065 characters. While with the .ffa it showed 4315 lines of genes annotation that are protein coding genes, 18400 words, and 206380 are the number of characters.

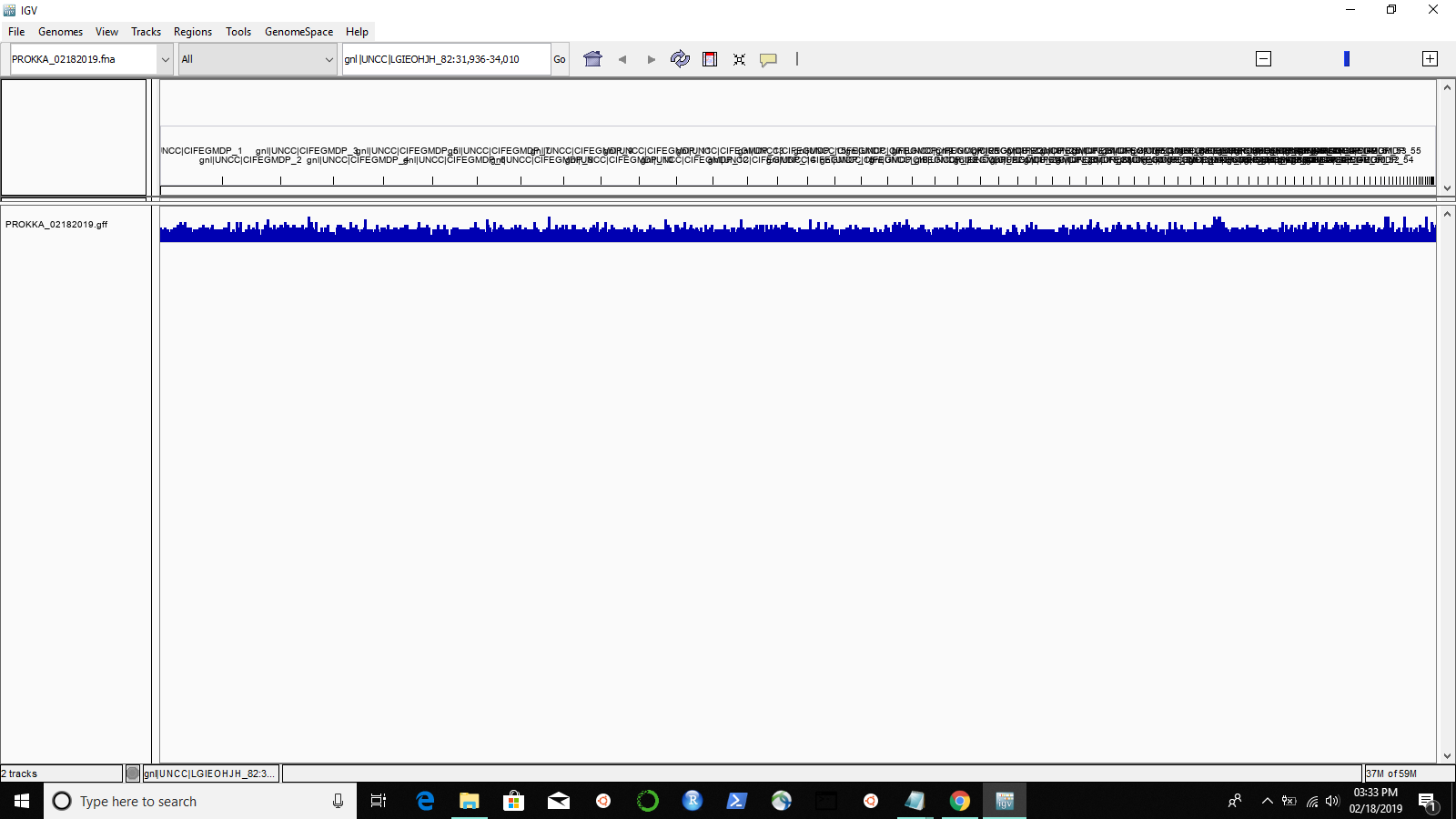
I did put both of the files in IGV. I first put the genome where I brought in the .fna files then i brought in the .gff as a file. I was able to view each contigs and view more information per contigs that prokka was able to give me. Below I attached an image of what the IGV showed in general.

bin/prokka ~/ERRcontigs.fasta --outdir NC007genus --force --compliant --centre UNCC --genus Vibrio --usegenus --cpus 2 --mincontiglen 500

ERR0008.ffn and ERR0008.ffa results

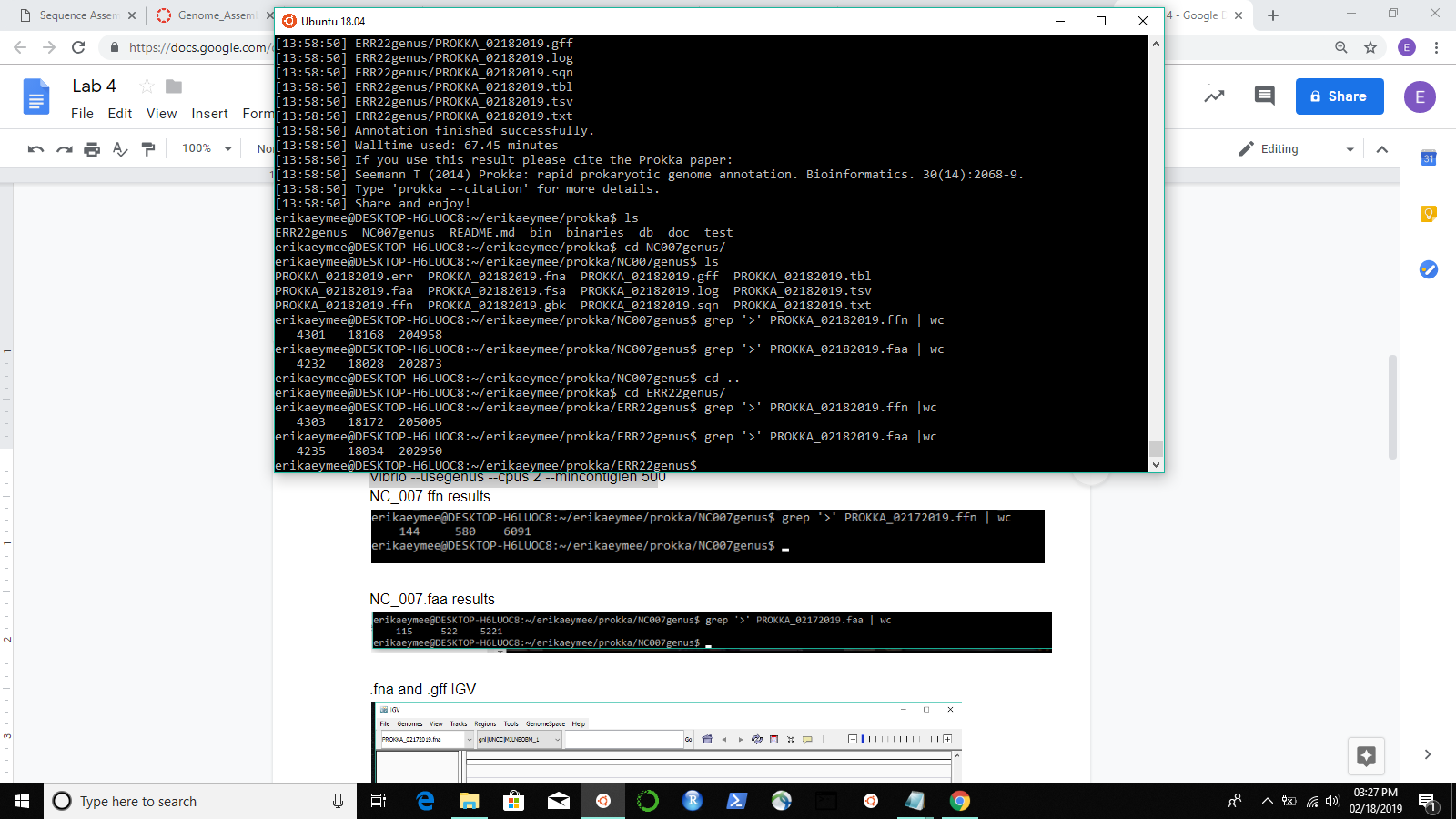


.fna and .gff IGV

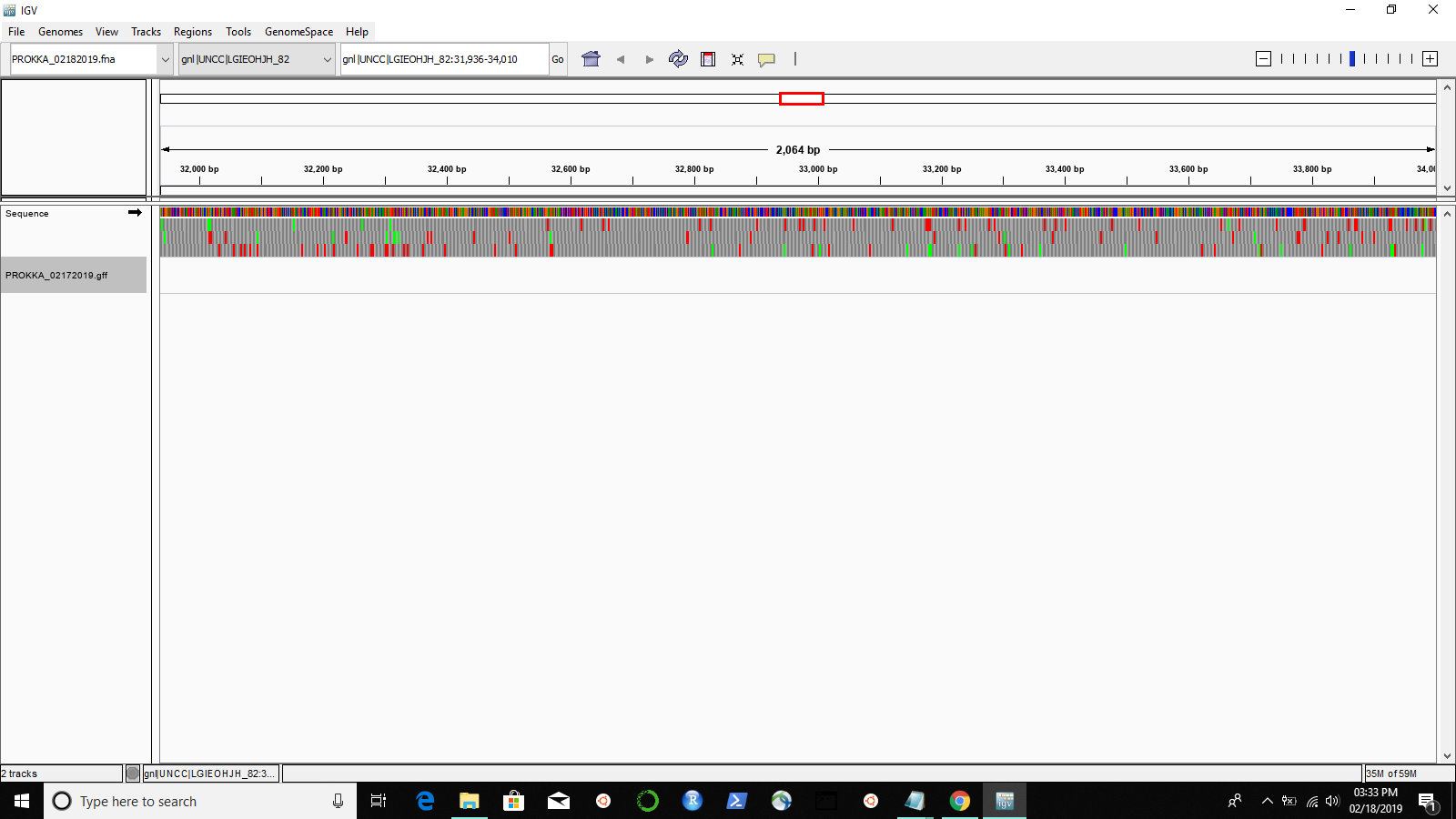


bin/prokka ~/ERR22contigs.fasta --outdir ERR22genus --force --compliant --centre UNCC --genus Vibrio --usegenus --cpus 2 --mincontiglen 500

ERR0022.ffn and ERR0022.faa results

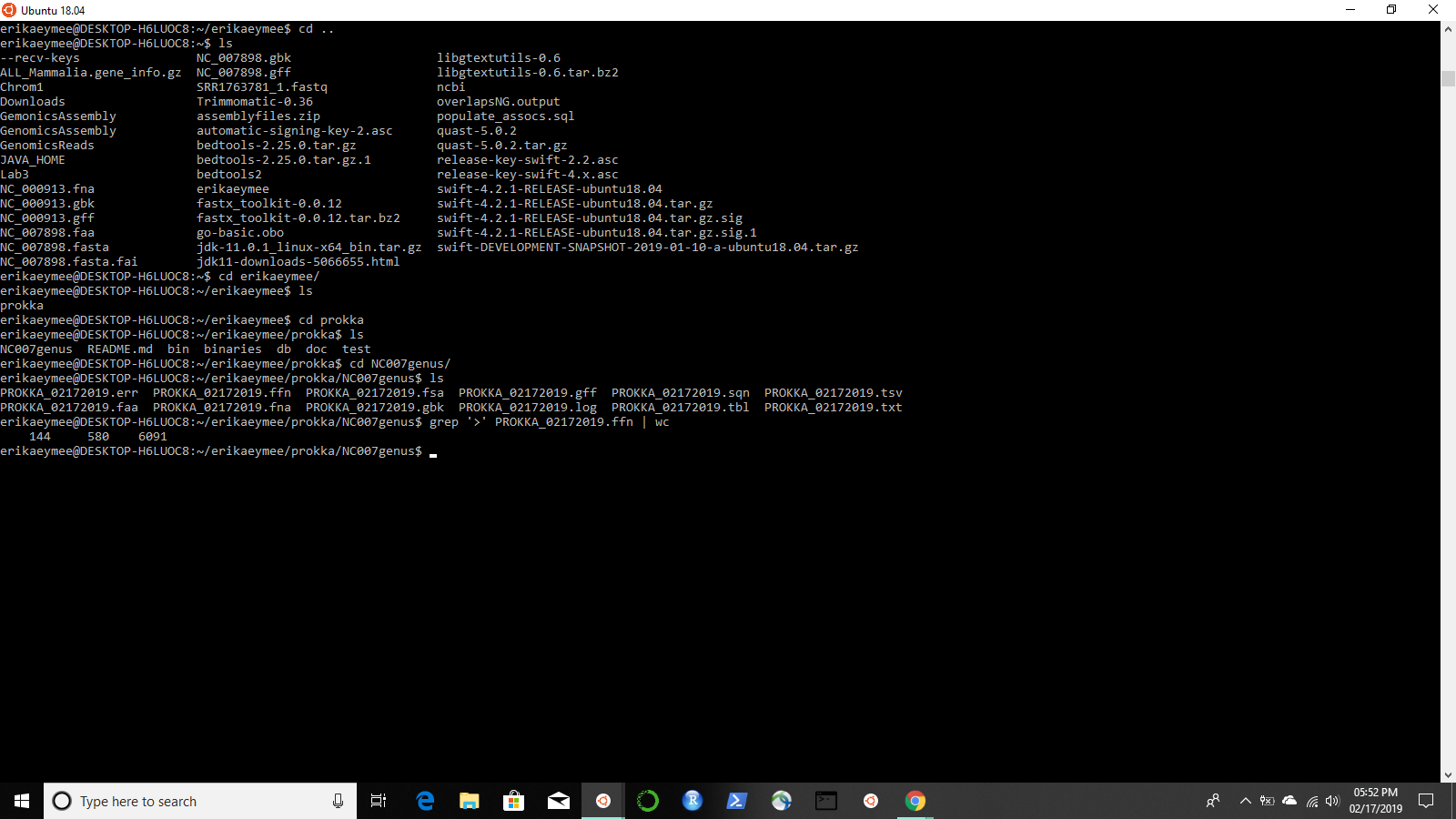


.fna and .gff IGV

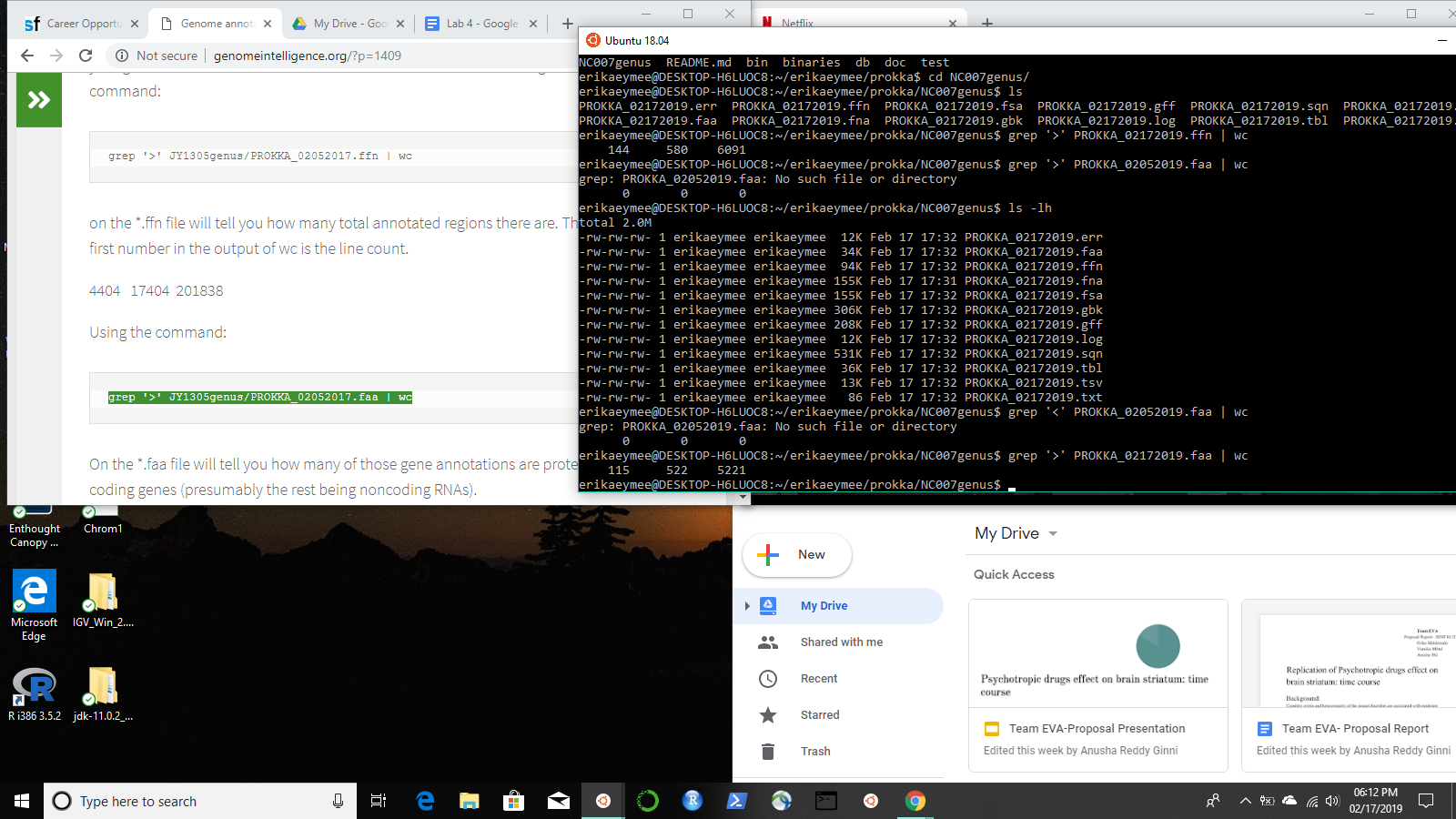


bin/prokka ~/NC\_007898.fasta --outdir NC007genus --force --compliant --centre UNCC --genus Vibrio --usegenus --cpus 2 --mincontiglen 500

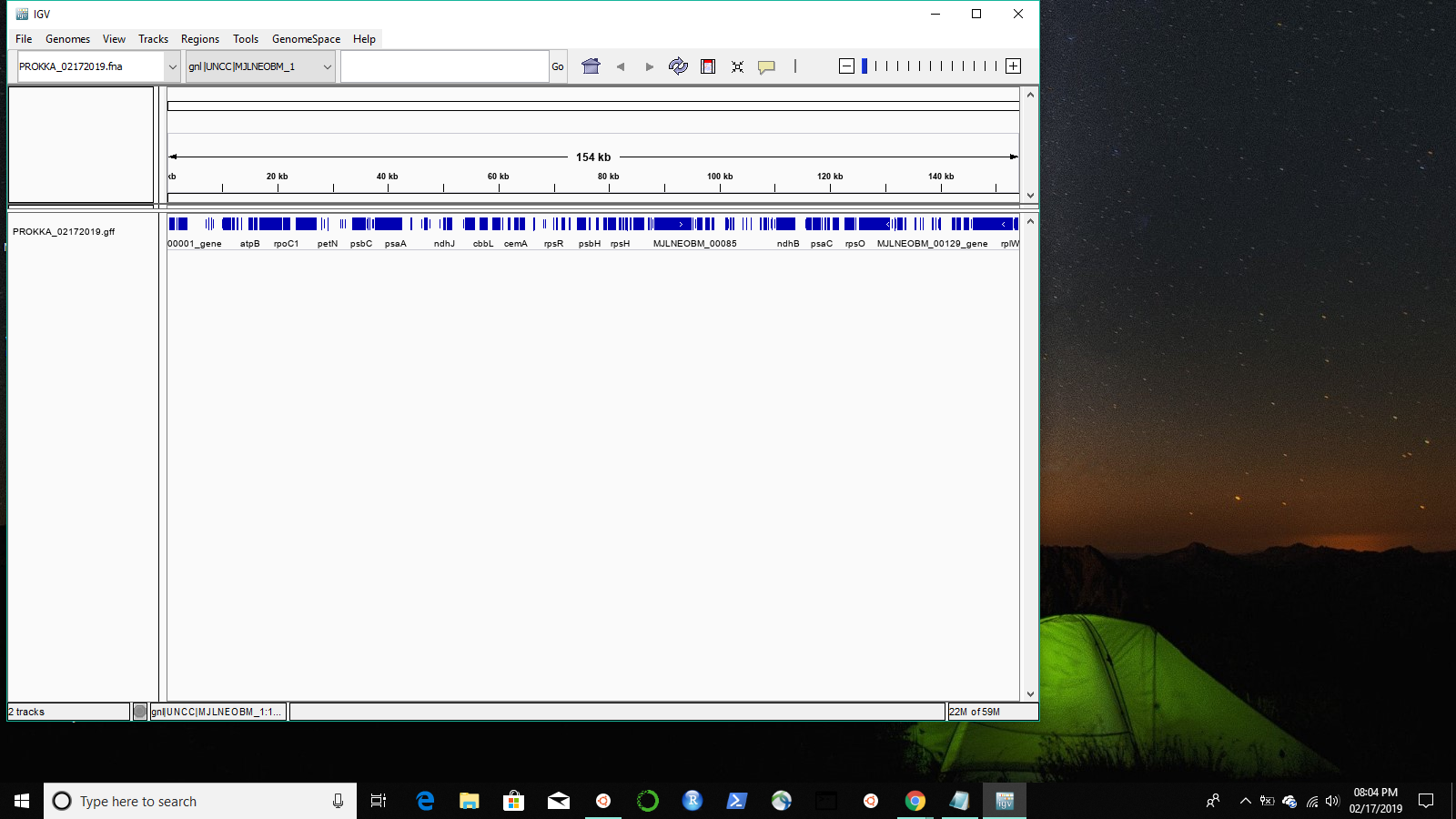
NC\_007.ffn results



NC\_007.faa results

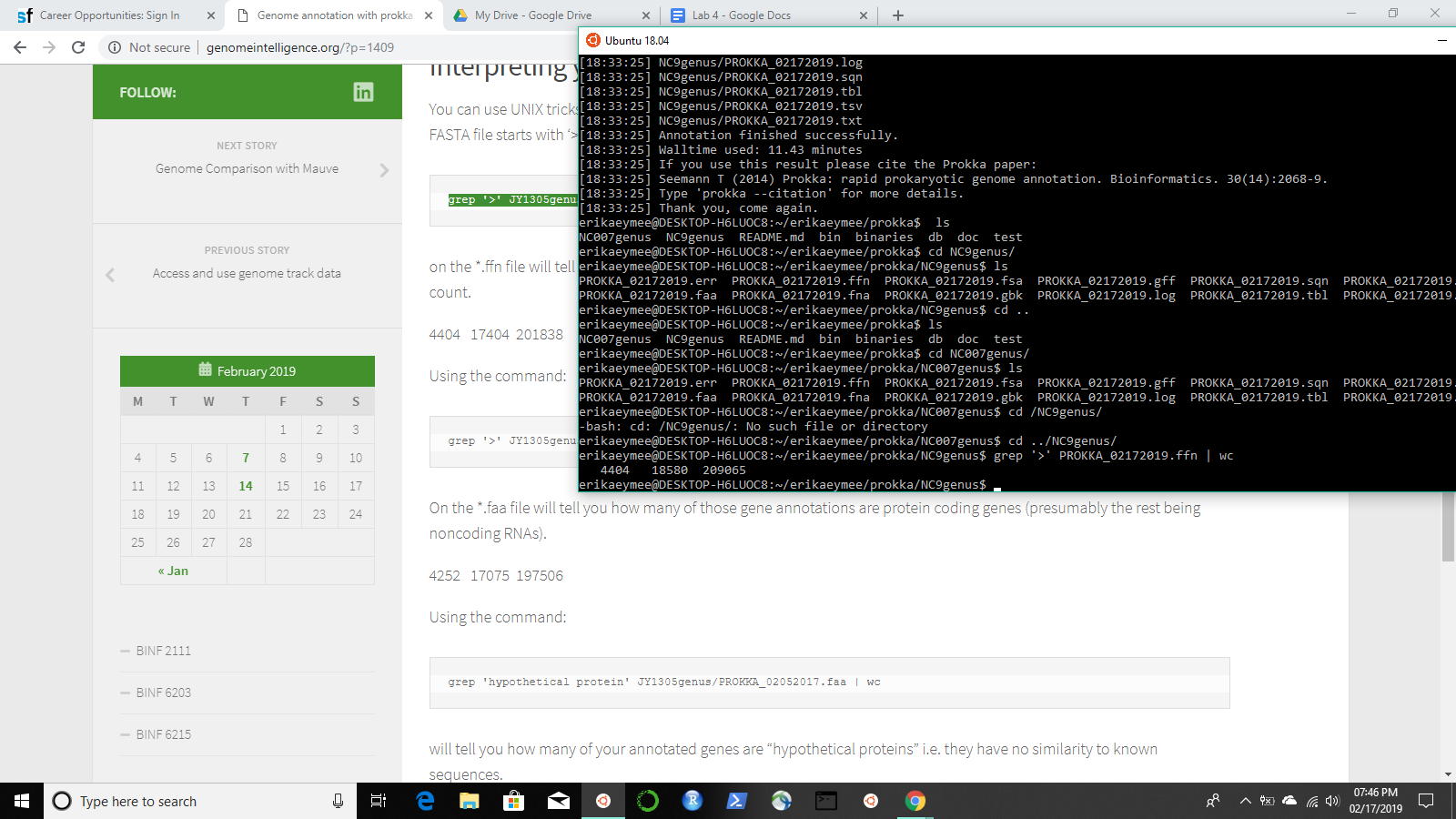


.fna and .gff IGV

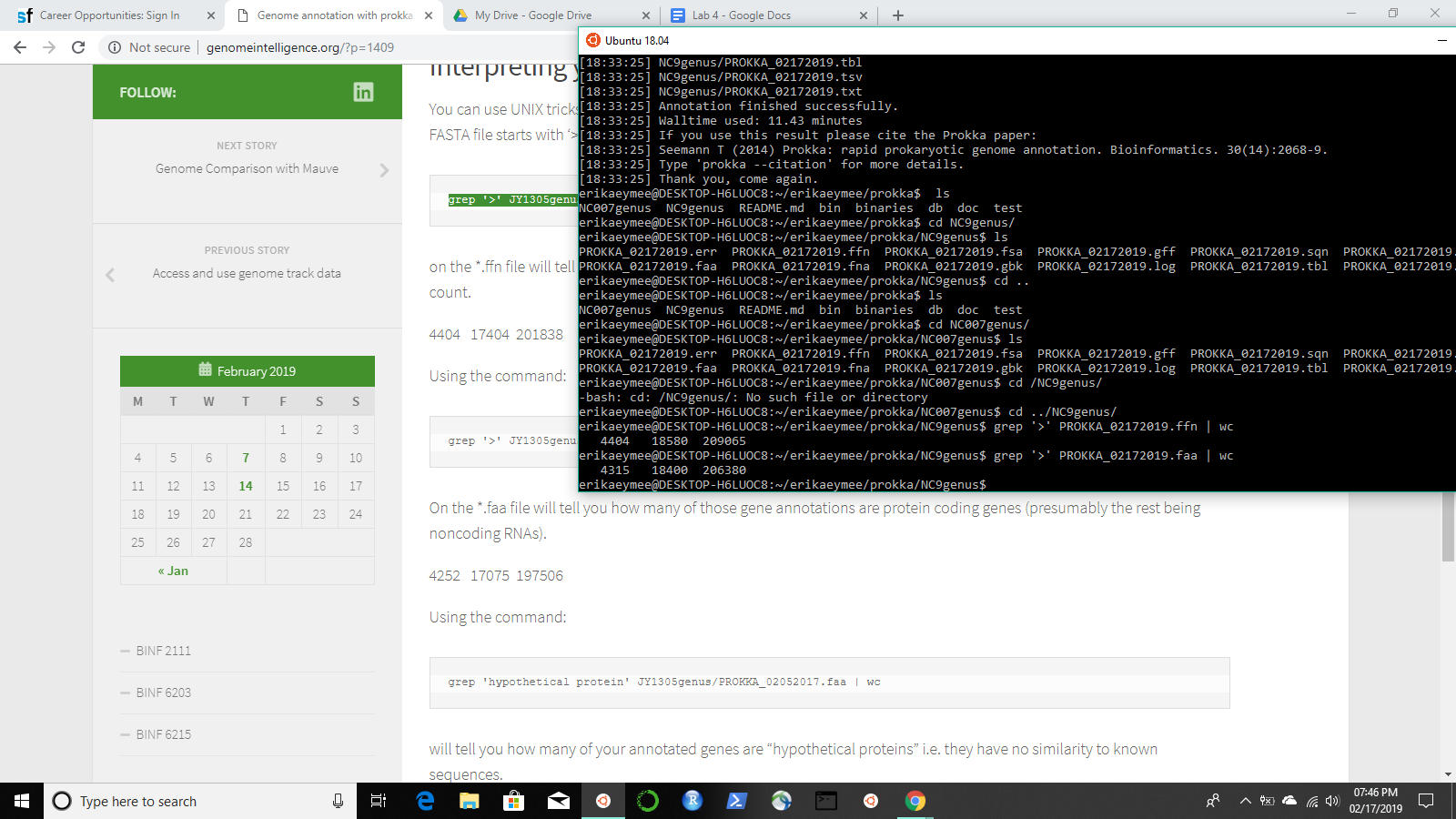


bin/prokka ~/NC\_000913.fna --outdir NC9genus --force --compliant --centre UNCC --genus Vibrio --usegenus --cpus 2 --mincontiglen 500

NC\_0009.ffn results



NC\_0009.faa results



.fna and .gff IGV

