Drashti Mehta
Genomics Lab 2

Genome annotation and comparison:

Introduction:

Whole genome annotation is a process that is used to find the interested genomic DNA sequences and then label them with meaningful information. In this assignment, we will be using PROKKA for carrying out genome annotation.

Method:

Prokka has to be operated on the cluster by loading the mambaforge module. After loading the module, prokka has to be installed. Prokka runs 5 more programs using the sequence database searches to identify the products of the gene that they predict. The programs include:

- Prodigal (finds coding sequences)
- HMMER (finds HMM matches — the HAMAP database is available in the default installation but you could add others if you want them)
- RNAmmer (finds ribosomal RNAs)
- Aragorn (finds tRNAs)
- SignalP (finds signal peptides)
- Infernal (finds noncoding RNA)

We run prokka on the contig files for the output of the spade. We use prokka for annotating the contig files, naming the files, making the annotations compliant with the standards of NCBI, and overwriting it from the previous iterations of the default output directory.

Result and discussion:

This process does not take a long time to run, it is pretty quick. The files are generated as below:

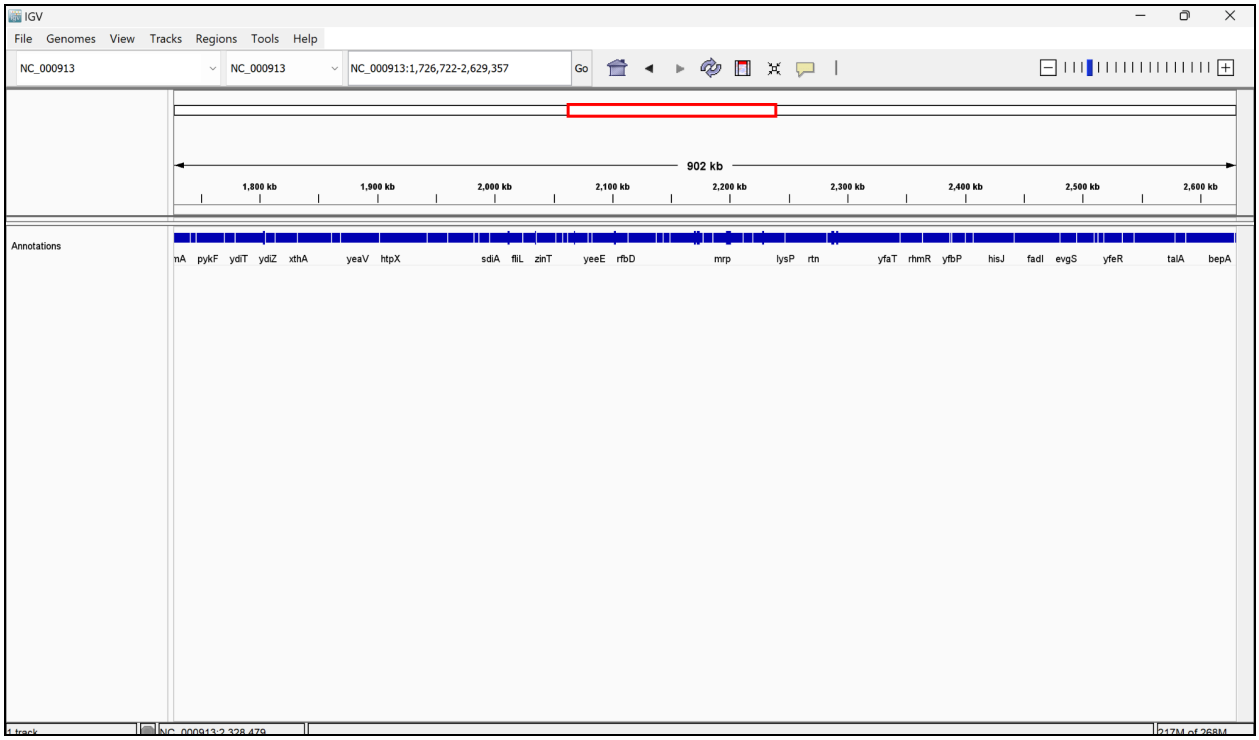
```
[dmehta12@gal-i1 Lab1-2023]$ ls prokka_8613/
PROKKA_03102023.err  PROKKA_03102023.fsa  PROKKA_03102023.sqn
PROKKA_03102023.faa  PROKKA_03102023.gbk  PROKKA_03102023.tbl
PROKKA_03102023.ffn  PROKKA_03102023.gff  PROKKA_03102023.tsv
PROKKA_03102023.fna  PROKKA_03102023.log  PROKKA_03102023.txt
```

From this, we will consider the .fna, .ffn, .faa, and .gff files. The next step is to count the number of annotations created for the genome of interest. Here, each annotation line in the fasta file starts with '>' and so we will use the grep command along with wc to count. The .ffn file tells the total annotated regions and the .faa file tells those gene annotations which are protein-coding genes. Some of the examples are attached below:

```
[dmehta12@gal-i1 Lab1-2023]$ grep '>' prokka_8613/*.ffn | wc
   4306   18245  205813
[dmehta12@gal-i1 Lab1-2023]$ grep '>' prokka_ComLib/*.ffn | wc
   4318   18286  206268
[dmehta12@gal-i1 Lab1-2023]$ grep '>' prokka_pac/*.ffn | wc
   5211   21191  241449
[dmehta12@gal-i1 Lab1-2023]$ grep '>' prokka_pac_both/*.ffn | wc
   5258   21415  243477
[dmehta12@gal-i1 Lab1-2023]$ grep '>' prokka_pac_both/*.faa | wc
   5164   21202  240579
[dmehta12@gal-i1 Lab1-2023]$ grep '>' prokka_pac/*.faa | wc
   5117   20981  238565
[dmehta12@gal-i1 Lab1-2023]$ grep '>' prokka_ComLib/*.faa | wc
   8451   22329  578536
[dmehta12@gal-i1 Lab1-2023]$ grep '>' prokka_8613/*.faa | wc
   4236   18097  203687
[dmehta12@gal-i1 Lab1-2023]$ grep 'hypothetic protein' prokka_8613/*.faa | wc
      0       0       0
```

To visually see the contigs and the annotations, we will be using the IGV browser. This helps us to select the contigs, zoom in and see the locations of individual genes.

Drashti Mehta
Genomics Lab 2



The annotated area is around 2815k bp compared to the reference with around 4640k bp.

```
[dmehta12@gal-i1 Lab1-2023]$ grep '>' refseq1.gff3 | wc
      5      64    1997
[dmehta12@gal-i1 Lab1-2023]$ grep '>' prokka_8613/*.gff | wc
    132     132    2532
[dmehta12@gal-i1 Lab1-2023]$
```