Final Report: *Enterococcus faecalis*

Purpose: I will be performing a genome analysis and annotation on *Enterococcus faecalis* to identify genes that attribute to its survival in harsh environments. I will also focus on the genes that allow this bacterium to prevail in the gastrointestinal tract of humans and animals causing it to be a common cause of hospital-acquired infection.

Methods:

All assembly and annotation will be performed through command line.

First, I will perform an assembly using both SPAdes and ABySS.

* To run SPAdes v4.1.0, I will enter the code:

spades.py -1 joey\_1.fastq.gz -2 joey\_2.fastq.gz -o spadesout

* Next, to run ABySS, I will enter the code:

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

Now that I have completed my assembly through SPAdes and ABySS, I will check the quality of each using QUAST v5.3.0.

* To run QUAST for SPAdes, I will enter the code:

quast.py spadesout/scaffolds.fasta -o quastspades

* To run QUAST for ABySS, I will enter the code:

quast.py abyssout/assembly-scaffolds.fa -o quastabyss

After this, I will run Barrnap 0.9 to identify 16s rRNA sequences to identify what species our genome is from.

* To do so I will run the code

barrnap --kingdom bac spadesout/scaffolds.fasta > rRNAsequences.gff

I will then use bedtools v2.31.1 to pull the FASTA sequence from my SPAdes assembly through the gff file.

* I will be using this code to do so:

bedtools getfasta -fi spadesout/scaffolds.fasta -bed rRNAsequences.gff -fo rRNAsequences.fasta

Now that I have created a fasta file for my rRNA sequences I will use blastn to identify the best match to my 16S rRNA

After confirming my species to be *Enterococcus faecalis*, I will perform an annotation of my genome using DFAST ver 1.3.6 and Prokka. Prokka will be done through galaxy (as stated above all other analysis will be done through command line).

* To initiate the annotation of the SPAdes assembly using DFAST, I will be using the code:

dfast --genome spadesout/scaffolds.fasta --out dfast\_output

* Protein annotation with Prokka galaxy version 1.14.6+galaxy1

I will find two closely related genomes from NCBI and run a fastANI to determine the average nucleotide identity. I originally used *Enterococcus faecium* and *Enterococcus durans.* However, these species had to low of an ANI. I then decided to use two similar strains Enterococcus faecalis, the two strains I chose were VE18395 and VE14089.

* To run FastANI, I will use the code:

fastANI -q spadesout/scaffolds.fasta --rl neighbors.txt -o salmonellaneighbors.txt

Lastly, I will use PathogenFinder2 to determine pathogen capacity of my genome

* I will use the PathogenFinder2 website and upload my scaffolds.fasta file from SPAdes to determine this.

Results:

After running SPAdes and ABySS assembly on my genome, I then ran QUAST to check the quality.

Here are the results for SPAdes:A screenshot of a computer

AI-generated content may be incorrect.

Here are the results for ABySS

A screenshot of a computer

AI-generated content may be incorrect.

Based on these results, I have determined the SPAdes assembly to be more fit. Although SPAdes has fewer Contigs, we can see that they’re much larger (with SPAdes being 1,424,258 bp compared to ABySS being 91,696 bp). Also, the N50 and N90 are much larger meaning that fewer, longer contigs make up the bulk of our genome. There are also less errors present in the SPAdes assembly meaning fewer gaps and better quality.

I then used barrnap to find the 16S rRNA of my genome and used blastn to confirm my species to be Enterococcus faecalis.

Here is the result from blastn:

A screenshot of a computer

AI-generated content may be incorrect.

To perform a genome annotation, I used DFAST and Prokka. A genome annotation will help us identify functional parts of a DNA sequence and their functional role. As stated in the purpose, I want to analyze genes that play a role in the species survival in harsh environments and virulence factors. Here is what I found:

|  |  |  |  |
| --- | --- | --- | --- |
| Gene | BP length | Product | Function |
| asa1 | 3,891 | Aggregation substance | Promotes bacterial aggregation, which defined as the formation of things into a cluster. This allows the bacteria to facilitate plasmid transfer and increase adherence to surfaces like the host cells and extracellular matrix proteins. |
| gelE | 1,530 | Gelatinase | Involved in cleaving of misfolded surface proteins, reducing pheromone levels, affecting chain length and degrading fibrin. All of these contribute to the ability of this bacterium to spread and interact with the environment (humans and animals) |
| htrA | 1,299 | Serine protease | Plays a vital role in the bacterium’s ability to cause disease. This allows the bacteria to degrade host tissue and increase infection and colonization. This specific gene will work with genes like gelE as a key virulence factor in *E. faecalis.* |

To determine average nucleotide identity, I chose two strains of *Enterococcus faecalis* as stated above. Here are my results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column 1 | Column 2 | % Similarity | Fragments Matched | Total Quary Fragments |
| Scaffolds FASTA | VE14089 FASTA | 98.850% | 873 | 954 |
| Scaffolds FASTA | VE18395 FASTA | 98.845% | 870 | 954 |

After performing PathogenFinder2 to determine the pathogen capacity of my genome, here are the results:

A screenshot of a graph

AI-generated content may be incorrect.

This shows me on a scale of 0-1 my genome contains genes that are 0.8405 pathogenic