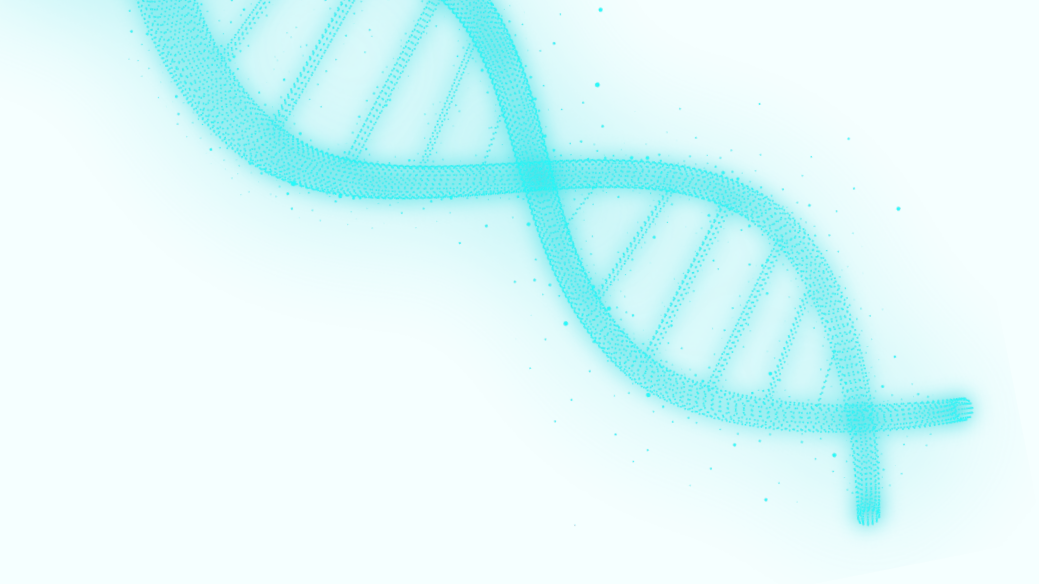
Genome Analysis: Assembly and Annotation of *E. faecalis*

By: Joseph Sada



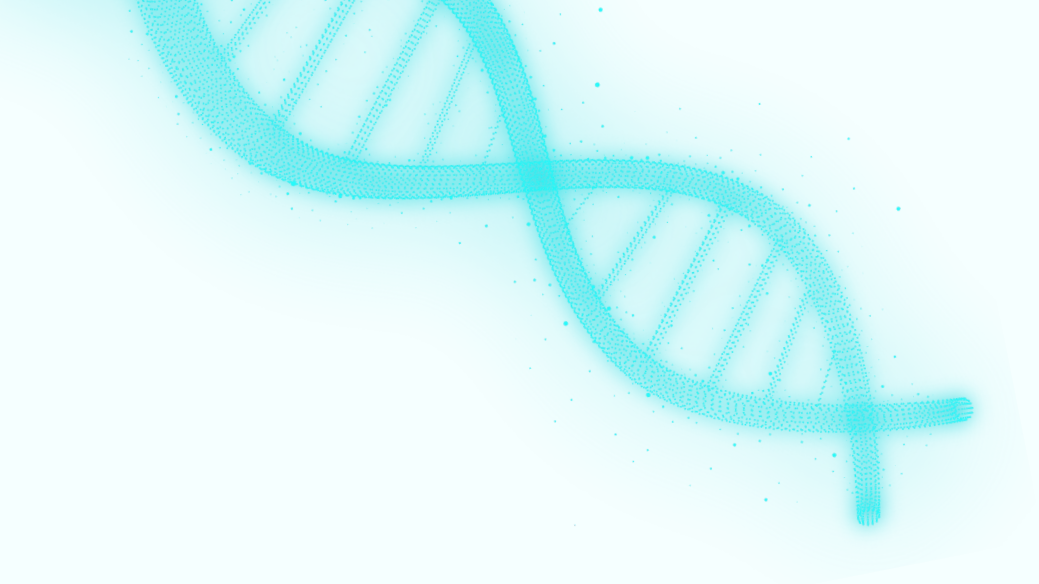
01

INTRODUCTION

INTRODUCTION

Purpose

I will be performing a genome analysis on *Enterococcus faecalis* by assembling and annotating the genome. The purpose of this is to identify genes that attribute to the bacterium's 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.



02

Methods/Results

QUAST v5.3.0 Quality Report

SPAdes v4.1.0 and ABySS used for genome assembly

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Assembly	scaffolds
# contigs (>= 0 bp)	16
# contigs (>= 1000 bp)	7
# contigs (>= 5000 bp)	7
# contigs (>= 10000 bp)	6
# contigs (>= 25000 bp)	5
# contigs (>= 50000 bp)	4
Total length (>= 0 bp)	2876090
Total length (>= 1000 bp)	2873642
Total length (>= 5000 bp)	2873642
Total length (>= 10000 bp)	2868555
Total length (>= 25000 bp)	2849700
Total length (>= 50000 bp)	2801918
# contigs	8
Largest contig	1424258
Total length	2874457
GC (%)	37.32
N50	675746
N90	287553
auN	953985.6
L50	2
L90	4
# N's per 100 kbp	13.92

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Assembly	assembly-scaffolds
# contigs (>= 0 bp)	2161
# contigs (>= 1000 bp)	223
# contigs (>= 5000 bp)	159
# contigs (>= 10000 bp)	102
# contigs (>= 25000 bp)	26
# contigs (>= 50000 bp)	4
Total length (>= 0 bp)	3176647
Total length (>= 1000 bp)	2809002
Total length (>= 5000 bp)	2631061
Total length (>= 10000 bp)	2228954
Total length (>= 25000 bp)	989695
Total length (>= 50000 bp)	277631
# contigs	268
Largest contig	91696
Total length	2842718
GC (%)	37.31
N50	19240
N90	5861
auN	24310.8
L50	46
L90	146
# N's per 100 kbp	198.65

SPAdes assembly is more fit based on these results.

- Larger contigs, larger N50 and N90, longer contigs, less errors

Barrnap → Bedtools → Blastn Results

- Barrnap 0.9 used to identify 16S rRNA sequences
- Bedtools v2.31.1 used to pull FASTA sequences from SPAdes through the gff file

select all *100 sequences selected* GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Enterococcus faecalis isolate 27725_1#12 genome assembly, chromosome: 1	Enterococcus faecalis	2878	11507	100%	0.0	100.00%	3055501	LR962267.1
Enterococcus faecalis isolate 27725_1#242 genome assembly, chromosome: 1	Enterococcus faecalis	2878	11512	100%	0.0	100.00%	2789770	LR962230.1
Enterococcus faecalis isolate 28157_4#255 genome assembly, chromosome: 1	Enterococcus faecalis	2878	11507	100%	0.0	100.00%	2912080	LR962771.1
Enterococcus faecalis strain S39-4 chromosome, complete genome	Enterococcus faecalis	2878	11512	100%	0.0	100.00%	2890864	CP088200.1
Enterococcus faecalis isolate 28157_4#381 genome assembly, chromosome: 1	Enterococcus faecalis	2878	11507	100%	0.0	100.00%	3142668	LR962526.1
Enterococcus faecalis strain VRE-WC031 chromosome, complete genome	Enterococcus faecalis	2878	11505	100%	0.0	100.00%	3005854	CP092576.1
Enterococcus faecalis isolate 26975_2#149 genome assembly, chromosome: 1	Enterococcus faecalis	2878	11507	100%	0.0	100.00%	3151849	LR962639.1
Enterococcus faecalis strain BM5 chromosome, complete genome	Enterococcus faecalis	2878	11501	100%	0.0	100.00%	2926129	CP173670.1
Enterococcus faecalis isolate 28157_4#117 genome assembly, chromosome: 1	Enterococcus faecalis	2878	11512	100%	0.0	100.00%	2868132	LR962474.1
Enterococcus faecalis strain EfsC29 chromosome, complete genome	Enterococcus faecalis	2878	11507	100%	0.0	100.00%	2901800	CP124950.1
Enterococcus faecalis strain 18-243 chromosome, complete genome	Enterococcus faecalis	2878	11512	100%	0.0	100.00%	2991600	CP065784.1
Enterococcus faecalis strain Z217-3 chromosome, complete genome	Enterococcus faecalis	2878	11512	100%	0.0	100.00%	2781050	CP159629.1
Enterococcus faecalis strain L11 chromosome, complete genome	Enterococcus faecalis	2878	11512	100%	0.0	100.00%	2809691	CP069185.1
Enterococcus faecalis isolate 27688_1#152 genome assembly, chromosome: 1	Enterococcus faecalis	2878	11512	100%	0.0	100.00%	2767840	LR962628.1
Enterococcus faecalis isolate 28157_4#36 genome assembly, chromosome: 1	Enterococcus faecalis	2878	11512	100%	0.0	100.00%	3135930	LR962807.1
Enterococcus faecalis isolate 26975_2#180 genome assembly, chromosome: 1	Enterococcus faecalis	2878	11512	100%	0.0	100.00%	3022375	LR962046.1
Enterococcus faecalis isolate 28099_2#233 genome assembly, chromosome: 1	Enterococcus faecalis	2878	11501	100%	0.0	100.00%	3068299	LR962512.1
Enterococcus faecalis isolate 28099_2#340 genome assembly, chromosome: 1	Enterococcus faecalis	2878	11512	100%	0.0	100.00%	2742529	LR962331.1
Enterococcus faecalis isolate 27688_1#305 genome assembly, chromosome: 1	Enterococcus faecalis	2878	11512	100%	0.0	100.00%	2946523	LR962296.1
Enterococcus faecalis isolate 28157_4#330 genome assembly, chromosome: 1	Enterococcus faecalis	2878	11512	100%	0.0	100.00%	2716367	LR962548.1
Enterococcus faecalis isolate 26975_1#182 genome assembly, chromosome: 1	Enterococcus faecalis	2878	11512	100%	0.0	100.00%	2888262	LR962036.1

Confirmed genome to be *Enterococcus faecalis*

DFAST and Prokka results

DFAST ver 1.3.6 and Prokka used for genome annotation

Here are the results:

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*.

FastANI results

I then performed a FastANI on my genome against two strains of *Enterococcus faecalis.*

This is to determine the average nucleotide identity. Here are the results:

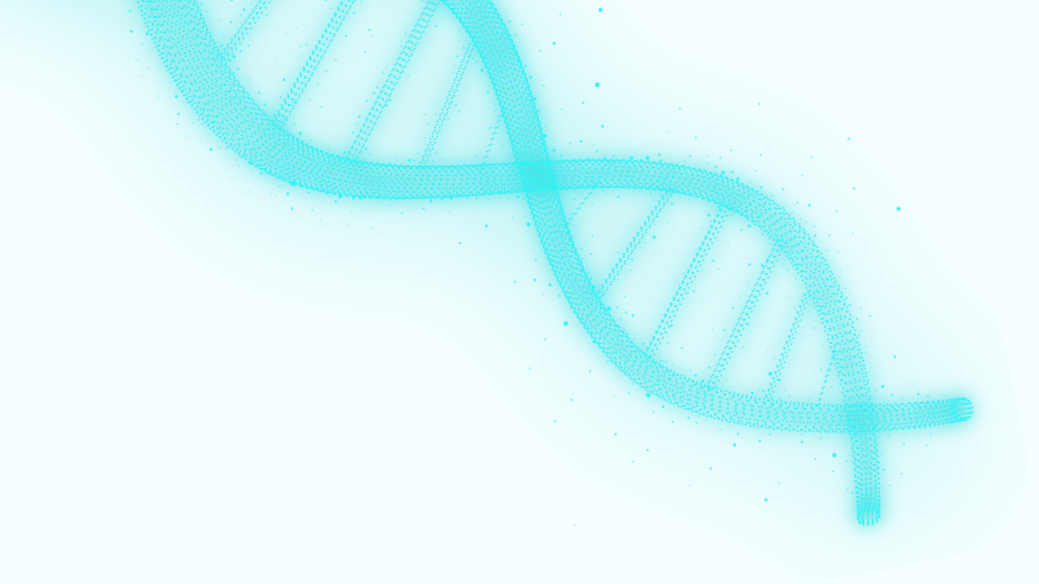
Column 1	Column 2	% Similarity	Fragments Matched	Total Quary Fragments
Scaffolds FASTA	VE14089 FASTA	98.850%	873	954
Scaffolds FASTA	VE18395 FASTA	98.845%	870	954

PathogenFinder2

I will use PathogenFinder2 to determine the pathogen capacity of my genome:

Module	Prediction
Neural network 1	0.8779
Neural Network 2	0.7822
Neural Network 3	0.8394
Neural Network 4	0.8623
Mean *(std)*	**0.8405** *(0.0363)*

This shows me that on a scale of 0-1, 0 being least pathogenic and 1 being most pathogenic, my genome has genes that are rated a 0.8405 (84.05%)



03

Conclusion

Conclusion

The purpose of this project was to perform a genome assembly and annotation on *Enterococcus faecalis*.

- I wanted to focus on genes that allow this bacterium to survive harsh environments.
  - Genes such as those that allow it to prevail in the gastrointestinal tract of humans and animals.
  - Also, genes that cause this bacteria to be a common cause for infection.
- After performing my genome assembly and annotation, I was able to find genes that directly contribute to this bacterium's pathogenicity.
- I also determined this bacterium to have a pathogen capacity of 0.8405.
- This means that this bacterium is very pathogenic and prevails in causing infection because of those specific genes discussed previously.