

FINAL GENOME ANALYSIS

Ciara F. Blanco
BIOL 4810 – Bioinformatics

Introduction

Bacteria: *Vibrio cholerae*

Sample Description: O1

SRR: [SRR33049458](https://www.ncbi.nlm.nih.gov/srr/SRR33049458)

Background: Two specific serogroups of *Vibrio cholerae* are associated with causing cholera—O1 and O139.


Question: Will pathogenic *Vibrio* species have higher Average Nucleotide Identity (ANI) results with my *V. cholerae* sample than non-pathogenic *Vibrio* species?

Methods (1/3)

1. Genome Assembly with SPAdes (v4.1.0) and ABySS (v2.3.7)
2. Quality Check with QUAST (v5.3.0)


SPAdes QUAST Report

Assembly	scaffolds
# contigs (≥ 0 bp)	837
# contigs (≥ 1000 bp)	55
# contigs (≥ 5000 bp)	25
# contigs (≥ 10000 bp)	19
# contigs (≥ 25000 bp)	17
# contigs (≥ 50000 bp)	17
Total length (≥ 0 bp)	4456675
Total length (≥ 1000 bp)	4034084
Total length (≥ 5000 bp)	3972412
Total length (≥ 10000 bp)	3925738
Total length (≥ 25000 bp)	3901578
Total length (≥ 50000 bp)	3901578
# contigs	476
Largest contig	638239
Total length	4295388
GC (%)	47.69
N50	246717
N90	72121
auN	325175.3
L50	5
L90	17
# N's per 100 kbp	6.98



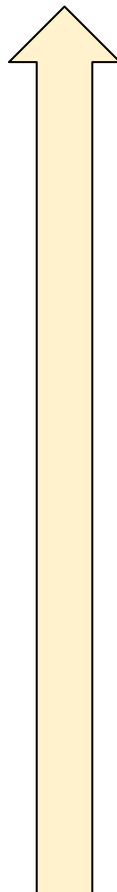
ABySS QUAST Report

Assembly	assembly-scaffolds
# contigs (≥ 0 bp)	1001
# contigs (≥ 1000 bp)	149
# contigs (≥ 5000 bp)	113
# contigs (≥ 10000 bp)	85
# contigs (≥ 25000 bp)	51
# contigs (≥ 50000 bp)	26
Total length (≥ 0 bp)	4166685
Total length (≥ 1000 bp)	4008622
Total length (≥ 5000 bp)	3900163
Total length (≥ 10000 bp)	3697469
Total length (≥ 25000 bp)	3169209
Total length (≥ 50000 bp)	2250260
# contigs	170
Largest contig	172801
Total length	4021995
GC (%)	47.50
N50	57580
N90	12537
auN	65738.4
L50	22
L90	79
# N's per 100 kbp	42.14



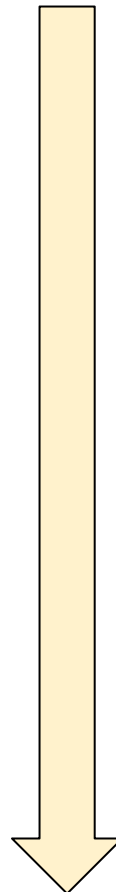
SPAdes QUAST Report

	scaffolds
# contigs (≥ 0 bp)	837
# contigs (≥ 1000 bp)	55
# contigs (≥ 5000 bp)	25
# contigs (≥ 10000 bp)	19
# contigs (≥ 25000 bp)	17
# contigs (≥ 50000 bp)	17
Total length (≥ 0 bp)	4456675
Total length (≥ 1000 bp)	4034084
Total length (≥ 5000 bp)	3972412
Total length (≥ 10000 bp)	3925738
Total length (≥ 25000 bp)	3901578
Total length (≥ 50000 bp)	3901578
# contigs	476
Largest contig	638239
Total length	4295388
GC (%)	47.69
N50	246717
N90	72121
auN	325175.3
L50	5
L90	17
# N's per 100 kbp	6.98



ABySS QUAST Report

	assembly-scaffolds
# contigs (≥ 0 bp)	1001
# contigs (≥ 1000 bp)	149
# contigs (≥ 5000 bp)	113
# contigs (≥ 10000 bp)	85
# contigs (≥ 25000 bp)	51
# contigs (≥ 50000 bp)	26
Total length (≥ 0 bp)	4166685
Total length (≥ 1000 bp)	4008622
Total length (≥ 5000 bp)	3900163
Total length (≥ 10000 bp)	3697469
Total length (≥ 25000 bp)	3169209
Total length (≥ 50000 bp)	2250260
# contigs	170
Largest contig	172801
Total length	4021995
GC (%)	47.50
N50	57580
N90	12537
auN	65738.4
L50	22
L90	79
# N's per 100 kbp	42.14



Methods (2/3)

3. Use NCBI to download *Vibrio* spp. files.

Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa

Vibrio Enter one or more taxonomic names

Filters

Download

Select columns

35,259 Genomes

Rows per page

20

1-20 of 35,259

<input type="checkbox"/>	Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action
<input type="checkbox"/>	ASM222426v1	GCA_002224265.1	GCF_002224265.1	Vibrio vulnificus NBRC 15645 = ...	ATCC 27562 (strain)	NCBI RefSeq Submitter	...
<input type="checkbox"/>	ASM155841v2	GCA_001558415.2	GCF_001558415.2	Vibrio fluvialis	ATCC 33809 (strain)	NCBI RefSeq Submitter	...
<input type="checkbox"/>	ASM145625v1	GCA_001456255.1	GCF_001456255.1	Vibrio natriegens NBRC 15636 ...	ATCC 14048 (strain)	NCBI RefSeq Submitter	...
<input type="checkbox"/>	ASM636435v1	GCA_006364355.1	GCF_006364355.1	Vibrio furnissii	FDAARGOS_777 (st...	NCBI RefSeq Submitter	...
<input type="checkbox"/>	ASM2434735v1	GCA_024347355.1	GCF_024347355.1	Vibrio chagasii	LMG 21353 (strain)	NCBI RefSeq	...
<input type="checkbox"/>	ASM2434731v1	GCA_024347315.1	GCF_024347315.1	Vibrio atlanticus	CECT 7223 (strain)	NCBI RefSeq	...

Feedback

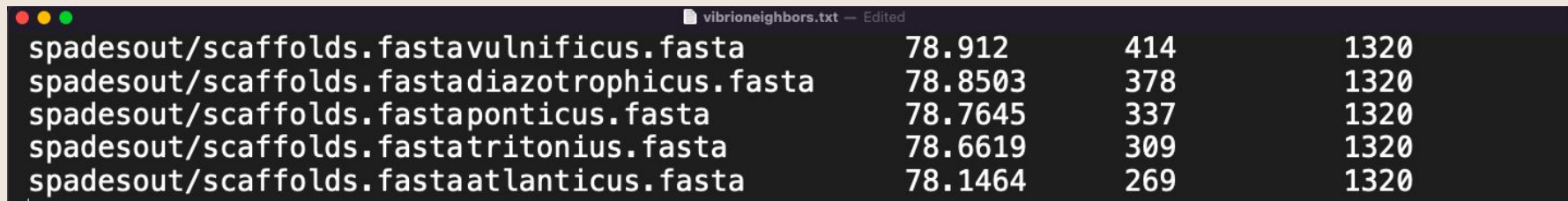
Methods (3/3)

4. Determine Pathogenicity with PathogenFinder2 (v0.4.1)

<i>Vibrio</i> spp.	PathogenFinder2 Results	Categorization
<i>V. vulnificus</i>	0.9501	Human Pathogenic
<i>V. diazotrophicus</i>	0.7999	Human Pathogenic
<i>V. ponticus</i>	0.7405	Human Pathogenic
<i>V. tritonius</i>	0.6404	Human Pathogenic
<i>V. tasmaniensis</i>	0.4500	Human Non Pathogenic
<i>V. rumoiensis</i>	0.3899	Human Non Pathogenic
<i>V. penaeicida</i>	0.3347	Human Non Pathogenic
<i>V. atlanticus</i>	0.2794	Human Non Pathogenic
<i>V. cholerae</i>	0.9765	Human Pathogenic

Results

5. Measure Overall Similarity Between Genomes with fastANI (v1.34)



A screenshot of a terminal window with a dark background. The title bar shows three colored dots (red, yellow, green) and the text "vibrioneighbors.txt - Edited". The terminal displays the output of a fastANI command, showing similarity percentages, the number of aligned regions, and the total number of regions for five different Vibrio species scaffolds.

spadesout/scaffolds.fastavulnificus.fasta	78.912	414	1320
spadesout/scaffolds.fastadiazotrophicus.fasta	78.8503	378	1320
spadesout/scaffolds.fastaponticus.fasta	78.7645	337	1320
spadesout/scaffolds.fastatritonius.fasta	78.6619	309	1320
spadesout/scaffolds.fastaatlanticus.fasta	78.1464	269	1320

<i>Vibrio</i> spp.	PathogenFinder2 Results	Categorization	fastANI Results
<i>V. vulnificus</i>	0.9501	Human Pathogenic	78.912
<i>V. diazotrophicus</i>	0.7999	Human Pathogenic	78.8503
<i>V. ponticus</i>	0.7405	Human Pathogenic	78.7645
<i>V. tritonius</i>	0.6404	Human Pathogenic	78.1464
<i>V. tasmaniensis</i>	0.4500	Hu. Non Pathogenic	N/A
<i>V. rumoiensis</i>	0.3899	Hu. Non Pathogenic	N/A
<i>V. penaeicida</i>	0.3347	Hu. Non Pathogenic	N/A
<i>V. atlanticus</i>	0.2794	Hu. Non Pathogenic	78.1464

Conclusion

The fastANI algorithm is designed to remove distant genomes (Jain et al., 2018).

Given that 3/4 of the non-pathogenic *Vibrio* spp. were removed from the fastANI results, it is likely they share less similarity to the *V. cholerae* sample than the pathogenic *Vibrio* spp.

References

- Ferrer Florensa, A., Almagro Armenteros, J. J., Kaas, R. S., Clausen, P. T., Nielsen, H., Rost, B., Aarestrup, F. M. (2025). Whole-genome prediction of bacterial pathogenic capacity on novel bacteria using protein language models, with PathogenFinder2. bioRxiv, 2025-04.
- Jain, C., Rodriguez-R, L.M., Phillippy, A.M. *et al.* High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 9, 5114 (2018).
<https://doi.org/10.1038/s41467-018-07641-9>