Bioinformatics Final Report

James Macak

12/02/2018

Abstract

Vibrio albensis (*Vibrio cholerae biovar albensis*) *ATCC 14547* is a species of *bacteria* found in freshwater that emits bioluminescence. It was determined that freshwater bioluminescence is produced (emitted), upon studying the *lux* operon, and outside of this research the regulatory system operon within this strain, by means of a system that contains a branch of 4,5-dihydroxy 2,3-pentanedione (DPD) as inducer and a sector of the *luxO* repressor gene – can become damaged upon destruction of two nucleotides. *V. albensis ATCC 14547* is also co-species with *V. cholerae* which can cause *cholera* in humans. This is a potentially deadly virus if left untreated. There are many species under the *genus Vibrio* that are present in saltwater when there is a minimum of 2% NaCl. However, *V. albensis ATCC 14547* is the only species that if found in freshwater. Because of these traits, various species of *Vibrio* are found all over the oceans, ubiquitous by beaches and other places people enjoy water sports. This provides all the more reason to sequence this gene and understand its ramifications to discover ways of counteracting these traits.

Introduction

In this project I compared *V. albensis ATCC 14547*'s IMG results to those I found using Megahit, as well as, three other genes under the same genus, these are: *V. cholerae ATCC 14035, V. cholerae 116-117b,* and *V. caribbenthicus ATCC BAA-2122.* The lineage of *V. albensis ATCC 14547* and where it falls in the tree of life is as follows: *domain of bacteria, phylum of proteobacteria, class of gammaproteobacteria, order of vibrionales, family of vibrionaceae, genus of vibrio,* and *species of vibrio albensis.*[i] *V. albensis ATCC 14547* is known for being the only species within the *genus Vibrio* to produce bioluminescence in freshwater. *V. cholerae*, and other species within the *Vibrio genus* often are only capable of producing bioluminescence in saltwater. Additionally, they require a minimum of 2% NaCl (Sodium Chloride) to be present to emit luminescence. To more accurately describe this process, functionality from the *lux* operon causes an emittance of *halotolerant favodoxin* (FP390 or the P-flavin binding protein), rather than producing light. This was determined through the observation of a system in which a branch of *4,5-dihydroxy 2,3-pentanedione* (DPD), used as an inducer, and *luxO* (the repressor gene for the lux operand) becomes damaged when two of the nucleotides are deleted.[ii]

Upon doing further research into *V. albensis ATCC 14547* and different species of *Vibrio*, I found that *V. cholerae* can cause cholera. Cholera is a severe diarrheal sickness that without proper treatment, can be life threatening (death can occur in under four hours) and can be picked up by drinking contaminated water. *V. cholera* is found on the shore of most oceans; this fits into the constraint in which it requires a minimum of 2% NaCl. However, there are two serogroups of toxigenic bacterium *Vibrio*, O1 or O139. Of the two groups, O1 is the more dangerous as it can breakout in an epidemic. This is not the case with O139 – it is the less dangerous of the two. *V. albensis ATCC 14547* and *V. cholerae* falls into the latter. Both of these species when observed under the constraint of O139 are only found off the shores of Asia. While these can cause a severe diarrheal disease, which is much less severe than the *choleria* disease, they do not pose a risk to humans in large quantities as *Vibrio* species under the O1 serogroup. Simply put, *V. albensis ATCC 14547* does not pose the threat of an epidemic outbreak.[ii] Since the early 2000s, there is are approximately 40 reports of individuals becoming infected by non-O1 or non-O139 *Vibrio* per year.

It is critically important that scientists sequence these genes in order to better understand the threat they pose to society. This research can be used to fight the potentially deadly consequences that are caused by various *Vibrio* species. There are many important metabolic pathways that exist within this gene, but I chose three notable ones for this report: flagellar assembly, bacterial chemotaxis, and glycolysis and gluconeogenesis. There are detailed pathways of each of these metabolisms within the 'Results' section of this report. But, here is a brief summarization each of these. Flagellar Assembly describes, in low-level detail, the "biological macromolecular nanomachine for locomotion."[iiii] Synonymous with this name is modality, which describes the methods of which this species is able swim around in water. Bacterial Chemotaxis is an interesting trait for *V. albensis ATCC 14547* to possess, as this allows it to direct its movement based on chemicals in its environment.[iiiii] Glycolysis and Gluconeogenesis are, in fact, the first metabolic pathway that is encountered when studying *V. albensis ATCC 14547*. It is a very ancient pathway (a sequence of reactions) that converts (metabolizes) a single molecule of glucose into two molecules of pyruvate with a focus of producing two molecules of ATP (adenosine triphosphate).[iiiiii] This project is designed to assemble and analyze a genome from a single cell.

Methods

In this project a variety of tools have been used to find and analyze various data sets. These include tools hosted and operated on Synergy: Megahit, CheckM, ARB, Muscle, RAxMLx, and inkscape. As well as, a select handful of tools hosted online: https://www.arb-silva.de/, http://img.jgi.doe.gov/, and http://blast.ncbi.nlm.nih.gov/Blast.cgi. I will briefly summarize these tools in order. Megahit is designed as a single-node tool that handles large genome assemblies. CheckM allows for the testing of the accuracy, completeness, and contamination of single cells (this is observable in the 'Results' portion of this report. ARB is a package of tools that provide help when sequencing genomes within databases, and analysis, this is then passed into Muscle which will perform an alignment on the sequence. Inkskape is a visual editor that allows the customization of trees and charts that one can generate. RAxML was used in the creation of the phylogenetic tree, which visually displays where the species is located. The ARB-silva database provided detailed datasets on RNA sequences, this is not limited in size. IMG is a government hosted website that stores huge amounts of information on various genomes and metagenome datasets. Lastly, BLAST (basic local assignment search tool) was used to find similar regions with different genomes.

V. albensis ATCC 14547 (taxon ID number: 2545555863) was isolated from fish within the Elbe River in the Czech Republic.[iiiiiii] The assembly was achieved through the program named Megahit (as described above) – the steps have been recorded in a separate text document (Final Notebook), see for exact details on how this was achieved. *V. albenesis ATCC 14547* (this time with taxon ID number: 2545555863) has been uploaded to the JGI's Integrated Microbial Genome (IMG) website and was used as reference and comparison (see Table 1. in the 'Results' section of this report. This genome was annotated and assembled through the IMG pipeline. I calculated N50 through the following formula.

$$\sum_{n=1}^{sequence\ len-1} n_{\mathrm{t}} = n_{\mathrm{n}} + n_{\mathrm{n+1}}$$

$$n_{\mathrm{h}} = \frac{n_{\mathrm{t}}}{2}$$

Then iterate through the list of sequence lengths, from smallest to largest, summing with each new index until n_{h} > current index, at which point, once true you have the calculated N50 value. There was an error with CheckM that impeded me from calculating Genome Completeness on my account. However, these numbers were calculated and are reported in Table 1. under the ‘Results’ section of this report. Average Nucleotide Identity (ANI) was reported from IMG, but is not available when self-performing Megahit. As stated above, the phylogenetic tree in this report was generated using ARB, and several other tools (see above), by means of importing the fasta file containing the *V. albensis ATCC 14547* sequence and aligning it before observing the graph. This sequence was selected for me, as the previous sequence I selected was not idea for this assignment. These files were downloaded from IMG and were manually worked on within ARB. I did not have to manually align the sequence as ARB has a tool that allows one to do this alignment by issuing a series of commands. Then the sequence is reviewed to ensure there are no errors before continuing on to the next step of generating the phylogenetic tree.

Results

This was an interesting project to partake in, and I have found some interesting conclusions about the methods used in Megahit to those found on IMG (*V. albensis ATCC 14547, V. cholerae ATCC 14035, V. cholerae 116-17b,* and *V. caribbenthicus ATCC BAA-2122*). Beginning with Assembled Genome size, with respect to Table 1. we can see a relatively small deviation of starting size: 3,973,024 bp, 3,931,204 bp, 4,026,422 bp, 4,087,587 bp, and 4,410,536 bp. However, *V. albensis ATCC 14547* from Megahit does place in the fourth largest position, which is useful to know when looking at the longest, average, and total scaffold data points. Listed now are the longest scaffolds, respectively: 256,607 bp, 561,255 bp**,** 394,436 bp, 363,762 bp, and 248,284 bp. Again, this genome places fourth based on a scale of largest to smallest. When looking at average scaffold length, there is a noticeable difference in size between *V. albensis* calculated from Megahit and the rest from IMG; listed respectively: 16,554 bp, 281,589 bp, 198,170.5 bp, 182,427 bp, and 124,248 bp. This means that while the scaffold size is similar, from Megahit, we can see that the individual pieces are much smaller than those found on IMG. N50 is found below, in Table 1. and displays similar trends as those found in average scaffold size. Genome completeness has been calculated to be 100% across the board, both from Megahit and IMG. The 16s rRNA percentage content displays a decrease in overlap from left to right, when reading Table 1. Similarly, ANI% follows this trend, with a noticeable similarity between *V. cholerae ATCC 14035* and *V. cholerae 116-17b*. This signifies even spacing when looking at the tree. Genes annotated displays a steady increase when moving across the table to the right. GC% is the final row to look at, the values are within a few percent of eachother across the board, save for *V. caribbenthicus ATCC BAA-2122* which is recorded at 42%. This is an additional representation of the distance from *V. albensis ATCC 14547*.

Table 1. Comparison of IMG statistics between *V. albensis ATCC 14547, V. cholerae ATCC 14035, V. cholerae 116-17b,* and *V. caribbenthicus ATCC BAA-2122*. This table also compares *V. albensis ATCC 14547*'s IMG statistics to those obtained from Megahit.

Organism Name	Vibrio albensis ATCC 14547	Vibrio albensis ATCC 14547	Vibrio cholerae ATCC 14035	Vibrio cholerae 116-17b	Vibrio caribbenthicus ATCC BAA-2122
Genome accession number	2545555863	2545555863	2556921658	2740892244	649990029
Assembly Type	Megahit	IMG	IMG	IMG	IMG
Assembled genome size	3,973,024 bp	3,931,204 bp	4,026,422 bp	4,087,587 bp	4,410,536 bp
Longest scaffold (or contig)	256,607 bp	561,255 bp	394,436 bp	363,762 bp	248,284 bp
Average scaffold	16,554 bp	281,589 bp	198,170.5 bp	182,427 bp	124,248 bp
Total scaffold	240	62	63	70	126
Genome completeness	100%	100%	100%	100%	100%
N50	100,081	2,086,407	2,207,107	2,097,286	2,265,169
16s rRNA identity %	N/A	100%	99%	99%	93%
ANI %	N/A	100%	98%	98%	71%
# of genes annotated	3,486	3,624	3,748	3,779	4,084
# of hypothetical proteins	1,254	574	631	584	1,176
GC content	46%	48%	47%	48%	42%

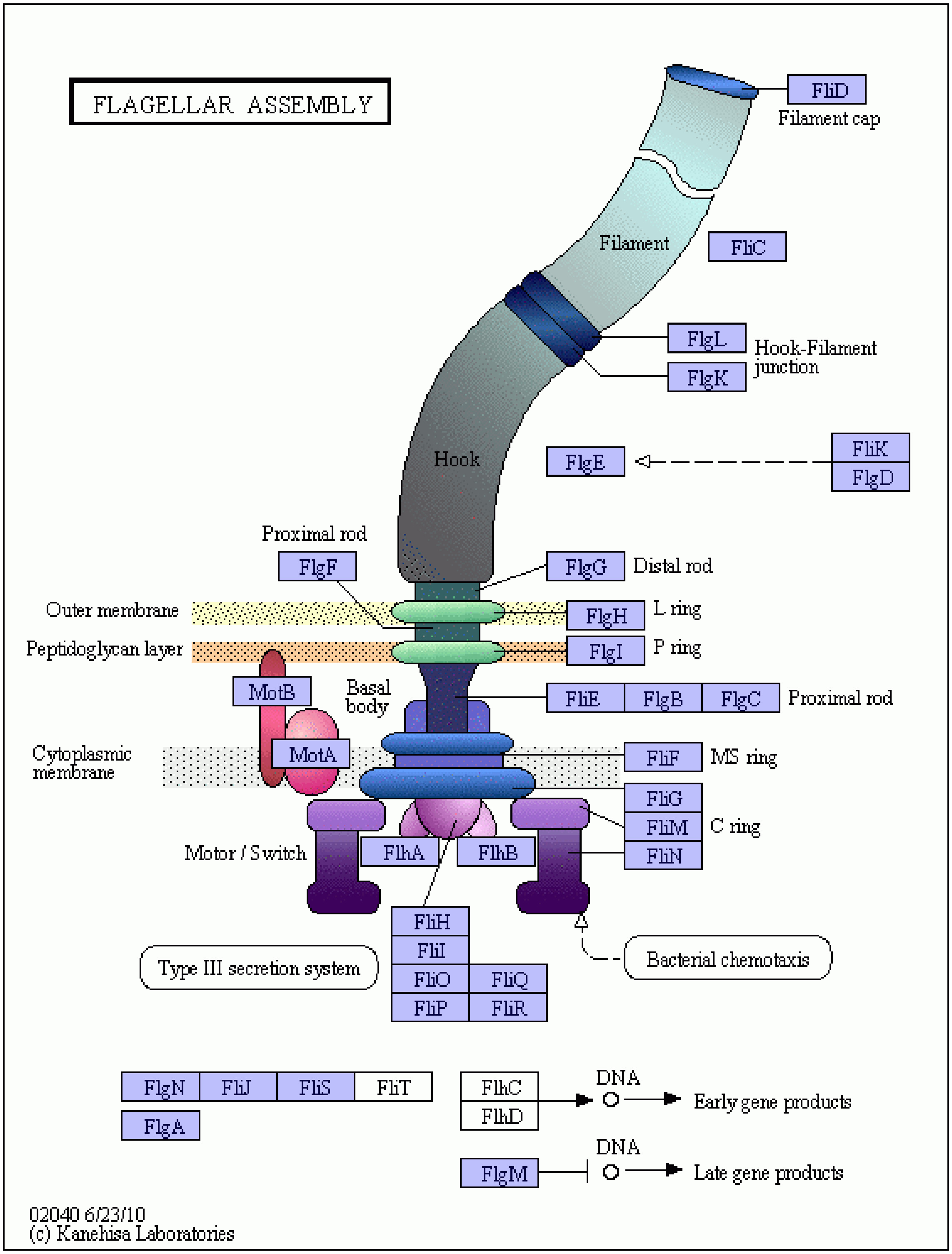

Figure 1. The visualization of the motility pathway (Flagellar Assembly) in *V. albensis ATCC 14547*. Items marked purple represent present genes. Higher amounts of present genes signify increased levels of completion, thereby representing more ability to complete the metabolic pathway.

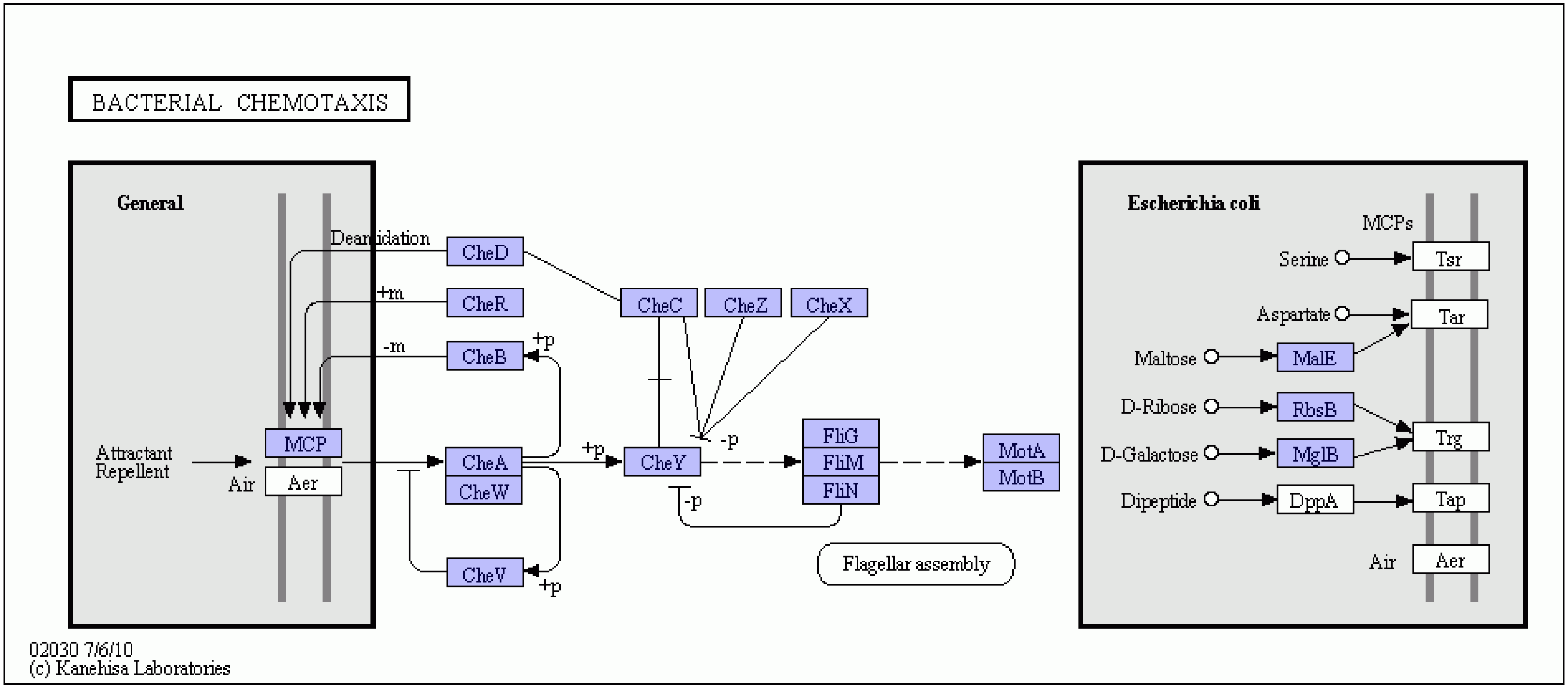

Figure 2. Bacterial Chemotaxis found in *V. albensis ATCC 14547*. The genes that are present are highlighted in purple.

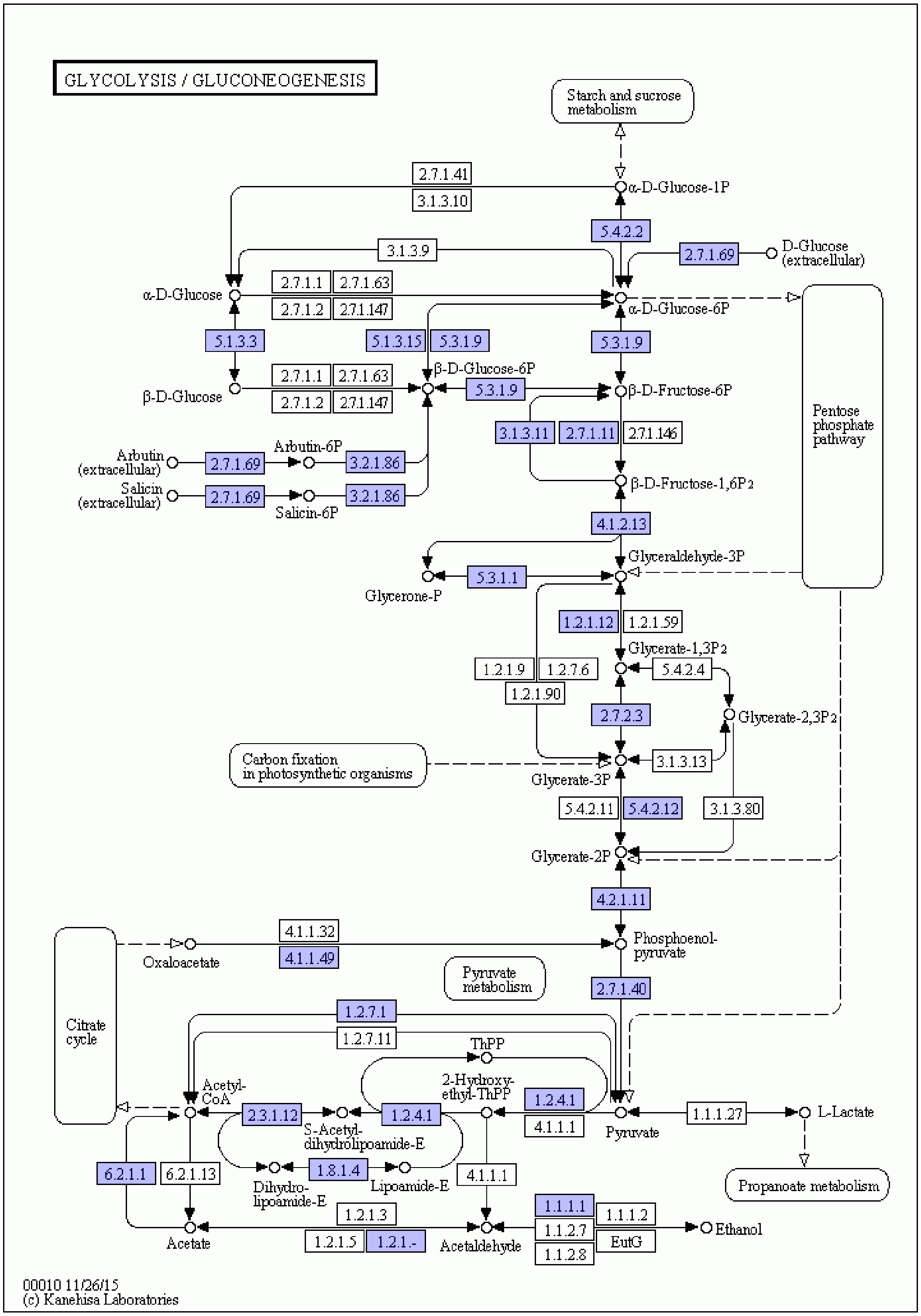

Figure 3. Glycolysis/Gluconeogenesis pathway found in *V. albensis ATCC 14547*. Items in purple are used in the metabolic pathway. This figure shows the proteins/enzymes used in the pathway, along with the genes involved.

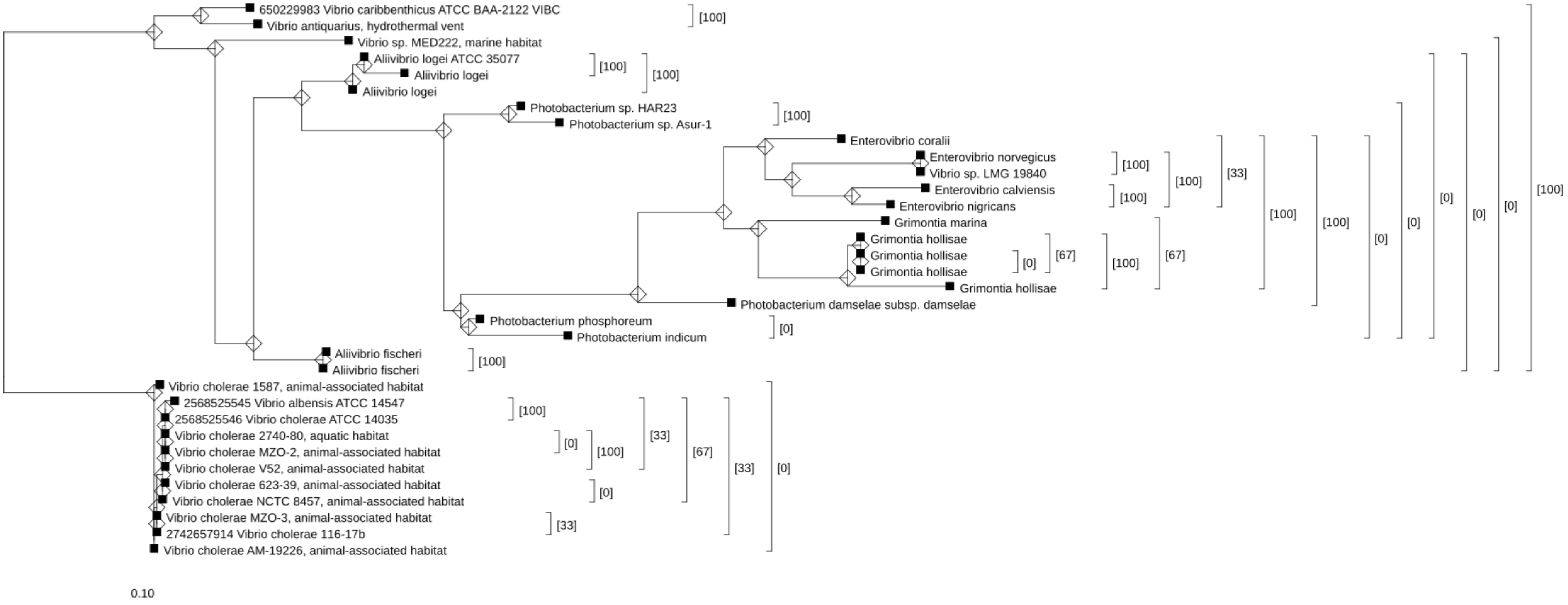
650229983 Vibrio caribbenthicus ATCC BAA-2122 VIBC
Vibrio antiquarius, hydrothermal vent
Vibrio sp. MED222, marine habitat
Aliivibrio logei ATCC 35077
Aliivibrio logei
Aliivibrio logei
Photobacterium sp. HAR23
Photobacterium sp. Asur-1
Enterovibrio coralii
Enterovibrio norvegicus
Vibrio sp. LMG 19840
Enterovibrio calviensis
Enterovibrio nigricans
Grimontia marina
Grimontia hollisae
Grimontia hollisae
Grimontia hollisae
Grimontia hollisae
Photobacterium damselae subsp. damselae
Photobacterium phosphoreum
Photobacterium indicum
Aliivibrio fischeri
Aliivibrio fischeri
Vibrio cholerae 1587, animal-associated habitat
2568525545 Vibrio albensis ATCC 14547
2568525546 Vibrio cholerae ATCC 14035
Vibrio cholerae 2740-80, aquatic habitat
Vibrio cholerae MZO-2, animal-associated habitat
Vibrio cholerae V52, animal-associated habitat
Vibrio cholerae 623-39, animal-associated habitat
Vibrio cholerae NCTC 8457, animal-associated habitat
Vibrio cholerae MZO-3, animal-associated habitat
2742657914 Vibrio cholerae 116-17b
Vibrio cholerae AM-19226, animal-associated habitat
[100]
[100]
[100]
[100]
[100]
[100]
[100]
[33]
[100]
[100]
[0]
[67]
[100]
[67]
[0]
[0]
[0]
[0]
[0]
[0]
[100]
[0]
[100]
[100]
[100]
[0]
[100]
[33]
[67]
[33]
[0]
[0]
[33]
0.10

Figure 4. Phylogenetic associations of various species under the genus *Vibrio*, based on the 16S rRNA gene. Species of interest is *V. Albensis ATCC 14547*. Percentages represent bootstrap values calculated from 100 replicates.

Bibliography

Primary Sources

[i] https://www.uniprot.org/taxonomy/140100

[ii] https://academic.oup.com/jb/article-abstract/139/3/471/1043566?redirectedFrom=PDF

[iii] https://www.cdc.gov/cholera/non-01-0139-infections.html

[iiii] http://www.pnas.org/content/104/17/7116

[iiiii] https://jb.asm.org/content/182/24/6865

[iiiiii] https://www.ncbi.nlm.nih.gov/books/NBK21150/

[iiiiiii] https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=2545555863

Secondary Sources

https://www.uniprot.org/taxonomy/1408477

http://www.microbiologyresearch.org/docserver/fulltext/micro/134/6/mic-134-6-1699.pdf?expires=1543784761&id=id&accname=guest&checksum=AC25851F472EF847D4ABBD7D1F918109

https://www.ncbi.nlm.nih.gov/pubmed/16567412