Monterey Bay Aquarium
Research Institute

The search for viruses in the ctenophore, *Bolinopsis infundibulum*

Ethan Ramsey, University of Nebraska-Lincoln

Mentor: Steve Haddock

Summer 2021

Keywords: Viruses, ctenophore, genome, DeepVirFinder, VirSorter2

ABSTRACT

Viruses are abundant in the ocean and found to infect a wide variety of species, but there is still much we don't know about their place in the ocean and how they interact with certain species. One such group are ctenophores which are a group of gelatinous organism important to the marine food web. This project for the MBARI 2021 summer internship seeks to identify double stranded DNA viruses within the genome of the ctenophore *Bolinopsis infundibulum*. To accomplish this two machine learning based viral identification programs, DeepVirFinder and VirSorter2, were used to search the target genome for viral sequences. The sequence contigs identified by these programs were then searched using BLAST in an attempt to identify the virus the program was calling. DeepVirFinder called 2537 contigs as viral and VirSorter2 identified 33 contigs with 24 contigs overlapping between the two programs. The BLAST search of the top results from these two programs matched to a large range of organisms. One contig, contig 474, did match to a siphoviridae virus although the match only represent 30 nucleotides. The matches to seemingly unrelated organisms could possibly explained by shared motifs or possibly viral sequences imbedded in the genomes of both organisms. Further work is need

to closely examine these results and confidently identify a virus associated with *B. infundibulum*.

INTRODUCTION

Viruses are found all over Earth in almost every environment including the oceans. Researchers estimate that there are more viruses on Earth than stars in the universe. There are an estimated 6 x 10^{22} stars in the universe (Manojlović, 2015) and over 10^{30} viruses are estimated to live in the oceans alone (Suttle, 2013). Unfortunately, marine viruses are not as terribly well studied and there is still much to learn as they are often thought to be a large area of undiscovered genetic diversity (Paul & Sullivan 2005). Many different organisms have been found to be infected with viruses. Everything ranging from whales, fish, crustaceans, mollusks, and bacteria have been found infected with viruses (Munn 2006). Viruses have also been found to infect various cnidarians such as corals where they can act as a pathogen or symbionts where they can actually confer beneficial traits to the host which are often known as auxiliary metabolic genes (Ambalavanan et al., 2021). Another important group of marine organisms are ctenophores. Ctenophores are a group of bioluminescent gelatinous organisms which are found around the world and at almost all depths. They are also major predators of plankton and other small marine organisms making them important to the overall marine food web. Despite being such an important group to the marine ecosystem there have been few studies looking at viral interactions in ctenophores. This project for the MBARI 2021 summer internship seeks to examine the interaction between viruses and ctenophores. This project will specifically try to identify double stranded DNA viruses or related sequences within the genome of the ctenophore, *Bolinopsis infundibulum*. *B. infundibulum* was chosen because little information is known about how viruses interact with this species. Also, they are important to the marine food chain as predators and as prey to other organism such as larger ctenophores. Finally, a high quality genome of this species had recently been assembled and readily available in the Haddock lab.

MATERIALS AND METHODS

VIRAL IDENTIFICATION PROGRAMS

This project utilized a *B. infundibulum* genome which had been assembled to the chromosome level and was about 277 megabases separated into contigs **(**Schultz, et al. in prep). Several programs were utilized in an attempt to find viral sequences within this genome. Two k-mer based machine learning programs, DeepVirFinder (Ren et al. 2019) and VirSorter2 (Guo et al. 2021), were used to identify sequences which could possibly be of viral origins. All files were stored on a Linux based server which allowed for remote access via the MBARI VPN and all programs and commands were run in Bash Powershell. These viral programs identification programs were Python based and downloaded to the lab server. DeepVirFinder was downloaded using the following command: `git clone https://github.com/jessieren/DeepVirFinder`.

VirSorter2 was downloaded with the following commands:

```
conda create -n vs2 -c conda-forge -c bioconda "python>=3.6" scikit-learn=0.22.1
       imbalanced-learn pandas seaborn hmmer==3.3 prodigal screed ruamel.yaml
       "snakemake>=5.18,<=5.26" click mamba
conda activate vs2
git clone https://github.com/jiarong/VirSorter2.git
cd VirSorter2
pip install -e .
virsorter setup -d db -j 4
```

The *B. infundibulum* genome was then used as a dataset for both programs. Two cutoff lengths were used for DeepVirFinder, 1500 bases and 500 bases, and a cutoff length of 1500 bases was used for VirSorter2. The search commands for DeepVirFinder were the following:

```
conda activate dvf
  python  dvf.py  -i  /data/user/virus/data/boli_contigs_masurca_output.fasta  -o
/data/user/virus/results/boli_deepvirfinder -l 1500 -c 4
  python  dvf.py  -i  /data/user/virus/data/boli_contigs_masurca_output.fasta  -o
/data/user/virus/results/boli_deepvirfinder -l 500 -c 4
```

The search commands for VirSorter2 were the following:

```
screen -x Boli_VirSorter2
  conda activate vs2
  virsorter run -w /data/user/virus/results/boli_virsorter.out -i boli_contigs_masurca_output.fasta --min-length 1500 -j 4 all && mail -s "VirSorter done..." ramsey_ethan@hotmail.com <<< "Finished"
```

Upon obtaining results from both programs the results files were sorted based on score. The DeepVirFinder results file was sorted using the following command:

```
sort -k 3 -nr boli_contigs_masurca_output.fasta_gt1500bp_dvfpred.txt > sorted_deepvirfinder_1500_pred.txt
```

The VirSorter2 results file first had to have its columns rearranged which was accomplished using a Python script, columnsorter.py. The commands used to sort the Virsorter2 results file are the following:

```
columnsorter.py final-viral-score.tsv > columnsort-final-viral-score.tsv
  sort -k 3 -nr columnsort-final-viral-score.tsv > sorted-final-viral-score.tsv
```

The two results files were then combined in R Studio to identify contigs which were called by both programs. A graph was also constructed to compare the scores from both programs for these shared contigs, visualized in Figure 2.

BLAST SEARCHES

In order to try to identify a virus within the *B. infundibulum* genome BLAST searches were performed based on the results of the two viral identification programs. The two results files were kept as separate entities because the way results were displayed were slightly different between the two programs. DeepVirFinder gave a score for the entirety of every contig within the search dataset while VirSorter2 extracted segments of the contig which it identified as possibly being viral. Because of this difference only the top 250 contigs from the DeepVirFinder results and all the sequences identified by VirSorter2 were searched using BLAST. The sequences for the contigs from DeepVirFinder to be searched using BLAST were extracted using a Python script, getAinB.py. The commands used to extract the sequences were the following:

```
cut -f1 sorted_deepvirfinder_1500_pred.txt | head -n 250 | paste -d"," -s –
```

```
getAinB.py
contig3050,contig3026,contig1392,contig1593,contig1779,contig2423,contig1342,co
ntig785,contig1579,contig656,contig2606,contig2056,contig85,contig1899,contig38
56,contig2394,contig526,contig444,contig1951,contig2842,contig201,contig359,con
tig166,contig2114,contig1897,contig1550,contig174,contig1644,contig46,contig362
,contig1127,contig256,contig549,contig1191,contig1172,contig3429,contig537,cont
ig2066,contig1654,contig778,contig255,contig429,contig285,contig2080,contig3353
,contig2888,contig2746,contig722,contig2710,contig1987,contig849,contig2822,con
tig353,contig3443,contig24,contig2495,contig3772,contig766,contig3408,contig176
9,contig2688,contig3824,contig2734,contig501,contig158,contig3350,contig1416,co
ntig394,contig952,contig3175,contig1445,contig791,contig3280,contig2054,contig2
427,contig128,contig3607,contig1729,contig2289,contig2820,contig3253,contig2146
,contig474,contig2652,contig3519,contig658,contig1068,contig95,contig1252,conti
g1509,contig438,contig269,contig2704,contig607,contig2862,contig2309,contig1882
,contig2562,contig1559,contig235,contig2297,contig3167,contig1438,contig3243,co
ntig1573,contig2237,contig313,contig1138,contig3038,contig2768,contig241,contig
962,contig762,contig3105,contig2458,contig3693,contig3638,contig3703,contig3182
,contig1096,contig3447,contig488,contig2788,contig2576,contig1525,contig3570,co
ntig248,contig3863,contig2043,contig2610,contig761,contig1489,contig2728,contig
2959,contig1624,contig400,contig1911,contig2038,contig1581,contig3612,contig354
6,contig2440,contig1668,contig1238,contig1331,contig463,contig2216,contig2647,c
ontig2609,contig2887,contig2360,contig1524,contig2724,contig2574,contig1251,con
tig1590,contig624,contig2861,contig414,contig938,contig1483,contig891,contig148
8,contig1244,contig2623,contig3627,contig1813,contig2175,contig820,contig1545,c
ontig2351,contig1804,contig296,contig1294,contig437,contig1422,contig514,contig
1856,contig2346,contig3246,contig486,contig673,contig3214,contig2602,contig3069
,contig2662,contig2058,contig2174,contig2846,contig3764,contig652,contig2451,co
ntig3007,contig2736,contig2277,contig1310,contig3711,contig3707,contig868,conti
g3767,contig3723,contig54,contig184,contig3227,contig1351,contig751,contig2350,
contig777,contig3841,contig2664,contig60,contig2968,contig357,contig3692,contig
2656,contig616,contig2762,contig3689,contig1810,contig2068,contig132,contig3431
,contig1115,contig1052,contig1862,contig108,contig1883,contig415,contig2400,con
tig836,contig2773,contig3752,contig2803,contig1600,contig156,contig428,contig12
67,contig546,contig1520,contig3520,contig2667,contig1131,contig1363,contig3684,
contig2007,contig1925,contig702,contig793,contig2457
../../data/boli_contigs_masurca_output.fasta>fasta_files/deepvir_results_top250
.fasta
```

The command used to BLAST these sequences was the following:

```
blastn -query deepvir_results_top250.fasta -num_threads 24 -db nt -outfmt "6
stitle std length evalue" -max_target_seqs 5 > ../blastresults_top250.txt
```

VirSorter2 assembled a fasta file containing the sequences it identified as viral, and this file was used as a BLAST query using the following command:

```
blastn -query final-viral-combined.fa -num_threads 24 -db nt -outfmt "6 stitle
std length evalue" -max_target_seqs 5 > blastresults_final-viral-
combined.fa.txt
```

BLAST RESULTS ANALYSIS

To better analyze the results of the BLAST searches and get a better sense of the representation of certain groups the results were broken up into taxonomical groups. A Python script, genbanknames.py, was used to extract the taxon group for each BLAST result. A list of each ascension number was extracted from the BLAST results files and run through the Python script to obtain the taxon groups. These taxon groups were then sorted and organized to show the number of times each group was called using the following commands:

```
sort virtaxon_chunk1.tsv | uniq -c > sortedvirtaxon_chunk1.tsv
  cut -f1 sortedvirtaxon_chunk1.tsv | uniq -c > virtaxon_grouponly.tsv
```

These files were then used to create Treeplot figures comparing the groups called. The program R Studio was used to construct all graphs and figures in this project.

RESULTS

DeepVirFinder identified 2537 contigs with a p-value less than 0.05, although a score was given for all contigs. VirSorter2 identified only 33 contigs as being a possible viral match. Figure 1 visualizes the significant contigs identified by DeepVirFinder and the overlapping contigs from VirSorter2. The contigs which were given scores by both programs were compared in Figure 2 and it was found that contig 2581 had the highest score given by both groups. The BLAST search of the top 250 contigs from DeepVirFinder had one match to a virus at contig 474. The match was to a bacteriophage in the group siphoviridae and the represented 30 nucleotides. The other BLAST results were matches to many different organisms. The BLAST search for the sequences extracted by VirSorter2 had no matches to viruses but matched to a large variety of other organisms similar to the DeepVirFinder BLAST results. The full range of taxon groups for both programs is visualized in Figure 3 and 4 respectively.

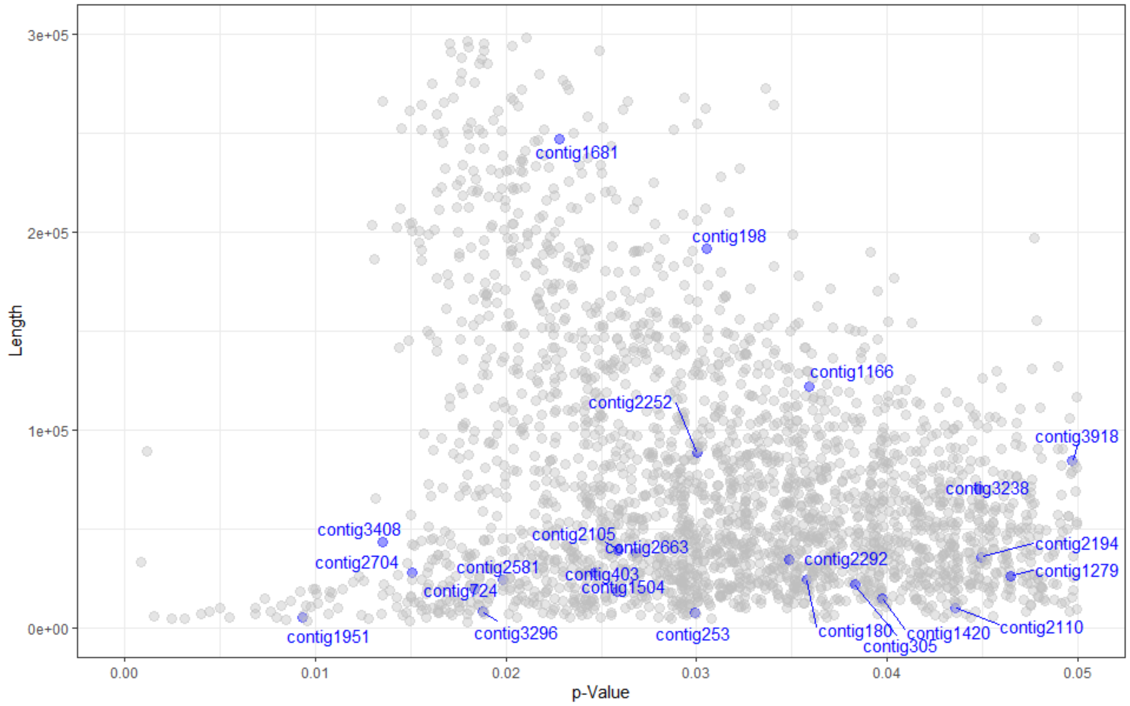

Figure 1. The graph depicts all contigs which according to the program DeepVirFinder were a significant viral match with each point representing an individual contig (n=2537). The y-axis shows the length of the contig, and the x-axis shows the p-value of the contig with more significant matches on the left of the graph. The labeled points highlighted in blue represent contigs which were also called by VirSorter2 (n=24).

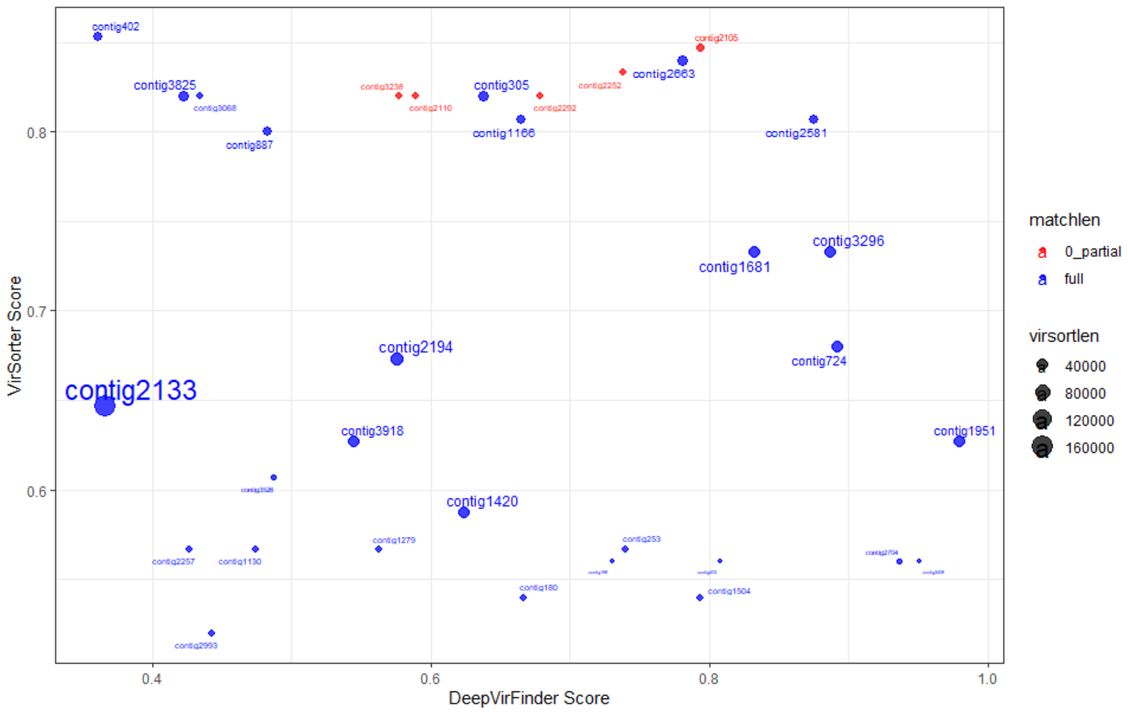

Figure 2. Compared scores of contigs called by both VirSorter2 and DeepVirFinder (n=33). The y-axis depicts the score given by VirSorter2 and the x-axis depicts the score given by DeepVirFinder. The size of the point represents the length of the contig following the legend to the right of the graph. The color of the point also illustrates the if the match was to a full or partial match to a virus with blue points representing full matches and red points representing partial matches.

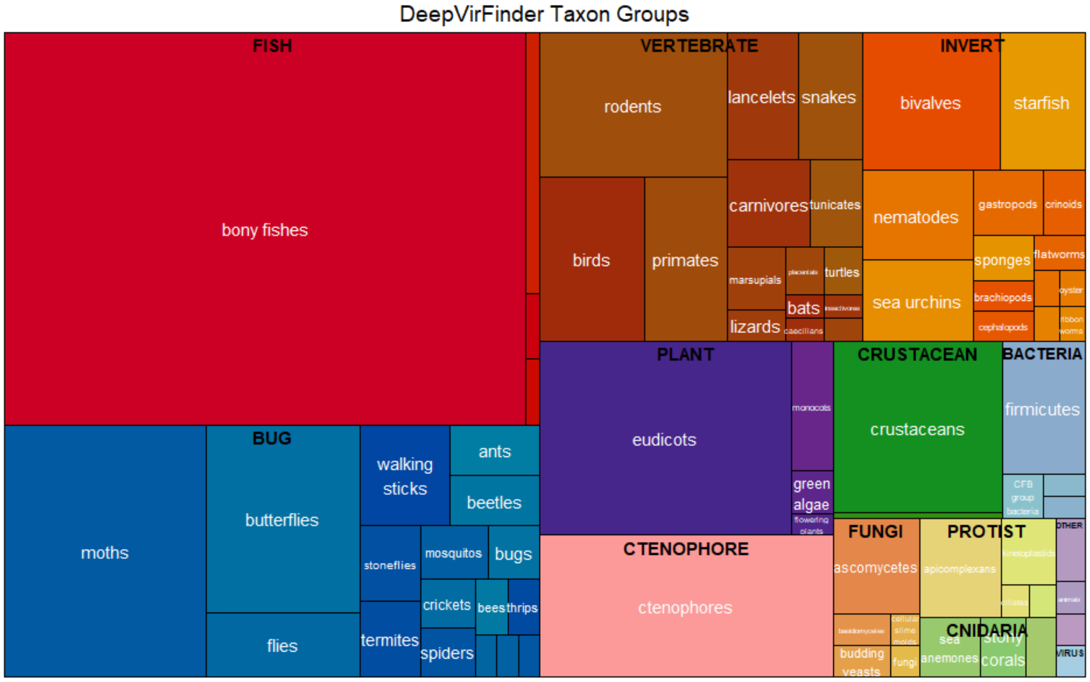

Figure 3. Treeplot representing the taxonomical groups called by a BLAST search of the top 250 matches from the DeepVirFinder program. This plot represents 769 BLAST results. Similarly colored blocks represent subgroups, labeled in white, which fall under the larger taxonomical group which is labeled in black. Fish was the most represented group followed by bugs (insects and arachnids) and then vertebrates. The single viral is visualized in the bottom right corner.

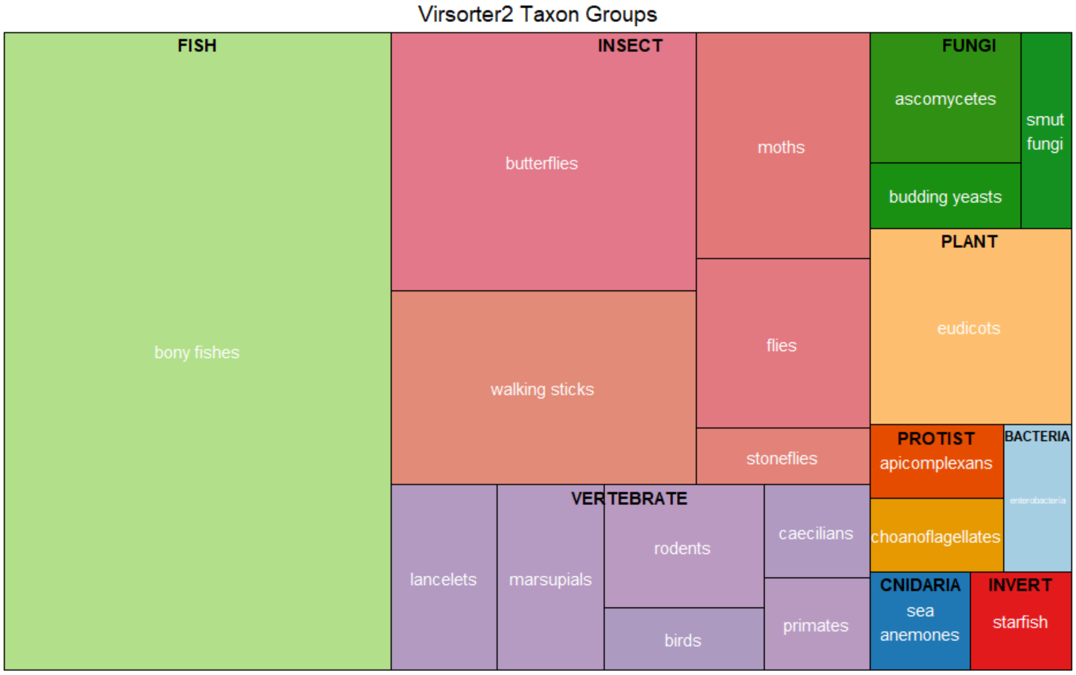

Figure 4. Treeplot representing the taxonomical groups called by a BLAST search of the sequences extracted by the program VirSorter2 (n=69). Similarly colored blocks represent subgroups, labeled in white, which fall under the larger taxonomical group which is labeled in black. Fish was the most represented group followed by insects and then vertebrates.

DISCUSSION

According to the virus identification programs there were a large number of significant matches to viral sequences although when these sequences were searched using BLAST hardly any viruses were called. One likely explanation is simply that not all of the contigs which the virus identification programs called as possibly being a virus actually contain a full viral sequence. It is possible that there is a common sequence or motif present in the ctenophore and virus which is what is being identified. This could also explain the diversity of the BLAST searches of these contigs as there are likely common motifs between many organisms and *B. infundibulum*. An interesting trend present in the taxon groups is that fish were the most prominent group identified in the BLAST search and there were a large amount and diversity of other marine organisms present as well. It is possible that there is a marine virus or a group of viruses which is present in all of them and that is the region matching between them. It could also be that there are viral sequences embedded in some of the genomes of other organisms which were called in the BLAST search and there are viral motifs which are matching between the two sequences.

Finally, the match to a siphoviridae virus at contig 474 was not a great match as it only represented 30 nucleotides, but it is still an exciting find. Siphoviridae are a family of bacteriophage which have been found in the ocean before (Wang et al., 2018). There are also several groups of marine viruses which share genetic characteristics with siphoviridae viruses (Paul & Sullivan, 2005). The siphoviridae called in the BLAST search was isolated from humans although it is likely that there would be some sequence similarity between related viruses. It would also be logical that a bacteriophage would be identified from a ctenophore because bacteria have been known to live within ctenophores and more specifically several bacterial genomes were identified during the assembly of the *B. infundibulum* genome (Daniels & Breitbart, 2012). Because they have been found in the ocean and bacteria are known to live with ctenophores it would be logical that this type or a related type of bacteriophage would be present in *B. infundibulum* although a specific virus was not identified.

CONCLUSIONS/RECOMMENDATIONS

Unfortunately, during this summer internship project we were unable to identify any specific viruses associated with *B. infundibulum*, although the viral match at contig 474 and trends within the taxon groups of the BLAST results are promising. This suggests that closer examination of these contigs could reveal viral sequences within the *B. infundibulum* genome. A possible next step could be to break the contigs up into smaller segments which might improve the chances of finding a short viral sequence from a BLAST search. Also, the BLAST search on the DeepVirFinder results could be expanded to include more than the top 250 results as there were many more contigs with p-values less than 0.05. Another possible step to improve this search would be to develop a specific training database made up of exclusively marine viruses to increase the chances of identifying a marine virus using DeepVirFinder or VirSorter2. The marine virus dataset used in a study by Gregory et al. could be modified to be a marine virus training set for these two programs (2019). Other approaches besides the bioinformatics approach could also be used to identify viruses associated with ctenophores. One way would be to utilize microbiology-based techniques such as plaque assays to isolate viruses from ctenophore samples which would be beneficial to identifying viruses which were filtered out during the preparation of the *B. infundibulum* genome. If a virus were to be identified from *B. infundibulum*, the nature of their relationship could be examined further such as if the virus is a pathogen of the ctenophore or is it conferring some beneficial trait. Also, it would be interesting to determine if this virus is specific to just *B. infundibulum*, ctenophores in general, or found in many different marine organisms. There is still much left to discover regarding this project, but hopefully more will be discovered about viruses and ctenophores as time goes on.

ACKNOWLEDGEMENTS

The MBARI Summer Internship Program is generously supported through a gift from the Dean and Helen Witter Family Fund and the Rentschler Family Fund in memory of former MBARI board member Frank Roberts (1920-2019) and by the David and Lucile Packard Foundation.

Also, special thank you to my mentor Steve Haddock and his lab. I could not have done this project without their support, expertise, resources, and patience. They taught me so much during the short time I worked on this project.

Finally, thank you to George Matsumoto, Megan Bassett, Lyndsey Claassen, and the other Summer 2021 MBARI interns for their support and advice during this internship. It was an invaluable addition to this experience.

REFERENCES:

Ambalavanan, L., Iehata, S., Fletcher, R., Stevens, E.H., Zainathan, S.C., (2021). A Review of Marine Viruses in Coral Ecosystem. *Journal of Marine Science and Engineering*. **9** (7): 711. https://doi.org/10.3390/jmse9070711

Daniels, C., Breitbart, M., (2012). Bacterial communities associated with the ctenophores *Mnemiopsis leidyi* and *Beroe ovata*. *FEMS Microbiology Ecology*. **82** (1): 90-101. https://doi.org/10.1111/j.1574-6941.2012.01409.x

Gregory, A.C., Zayed, A.A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., Dimier, C., Domínguez-Huerta, G., Ferland, J., Kandels, S., Liu, Y., Marec, C., Pesant, S., Picheral, M., … Sullivan, M.B., (2019). Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*. **177** (5): 1109-1123. https://doi.org/10.1016/j.cell.2019.03.040

Guo, J., et al, (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*. **9** (1): 37. https://doi.org/10.1186/s40168-020-00990-y

Manojlović, L.M., (2015). Photometry-based estimation of the total number of stars in the Universe. *Applied Optics*. **54** (21): 6589-6591. https://doi.org/10.1364/AO.54.006589

Munn, C.B., (2006). Viruses as pathogens of marine organisms - from bacteria to whales. *Journal of the Marine Biological Association of the United Kingdom*. **86** (3): 453-467.

Paul, J.H., Sullivan, M.B., (2005). Marine phage genomics: what have we learned?. *Current Opinion in Biotechnology*. **16** (3): 299-307. https://doi.org/10.1016/j.copbio.2005.03.007

Ren, J., et al, (2019). DeepVirFinder: Identifying viruses from metagenomic data by deep learning. University of South California. https://github.com/jessieren/DeepVirFinder

Suttle, C.A., (2013). Viruses: unlocking the greatest biodiversity on Earth. *Genome*. **56** (10): 542-544. https://doi.org/10.1139/gen-2013-0152

Wang, Z., Hardies, S.C., Fokine, A., Klose, T., Jiang, W., Cho, B.C., Rossmann, M.G., (2018). Structure of the Marine Siphovirus TW1: Evolution of Capsid-Stabilizing Proteins and Tail Spikes. *Structures*. **26** (2): 238-248. https://doi.org/10.1016/j.str.2017.12.001