
# Impact of Coral Reef-Derived Viruses on Host Carbon Metabolism: Insights from Genomic Sequencing

Jacob Fisher[1] and Ben Knowles[2]

1 BIG Summer Program, Institute for Quantitative and Computational Biosiences, UCLA

2 Department of Ecology and Evolutionary Biology, UCLA

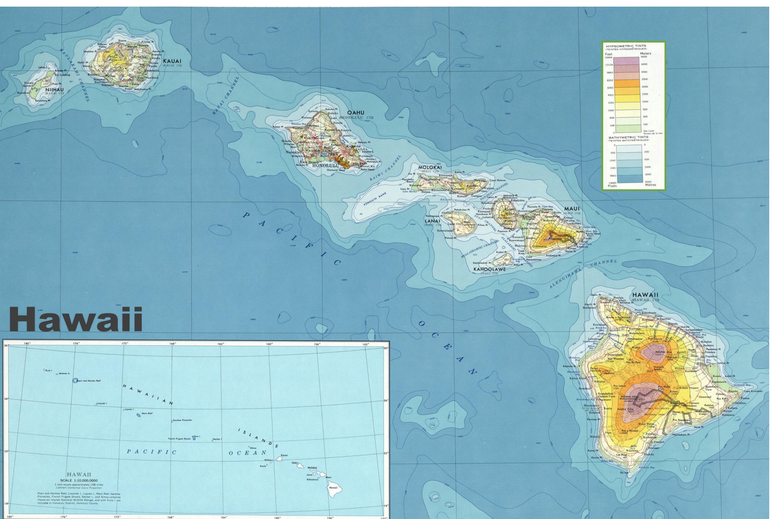
Hopeful Monsters

## Abstract

This research delves into the intricate relationships between marine viruses, derived from coral reefs, and the metabolic pathways of their hosts. Utilizing high-throughput genomic sequencing, we identify and analyze viral proteins that could modulate key enzymes involved in carbon metabolism, particularly focusing on the glycolytic pathway. The objective is to elucidate the strategies employed by these marine viruses to potentially redirect host metabolic processes to favor viral replication. Our initial analyses reveal that these viruses may engage in complex interactions with host cellular mechanisms to manipulate carbohydrate metabolism, suggesting a novel layer of viral influence on coral ecosystem health. Such insights are crucial for understanding the broader implications of viral presence in marine environments and for developing strategies to mitigate their effects on coral reef stability and resilience.

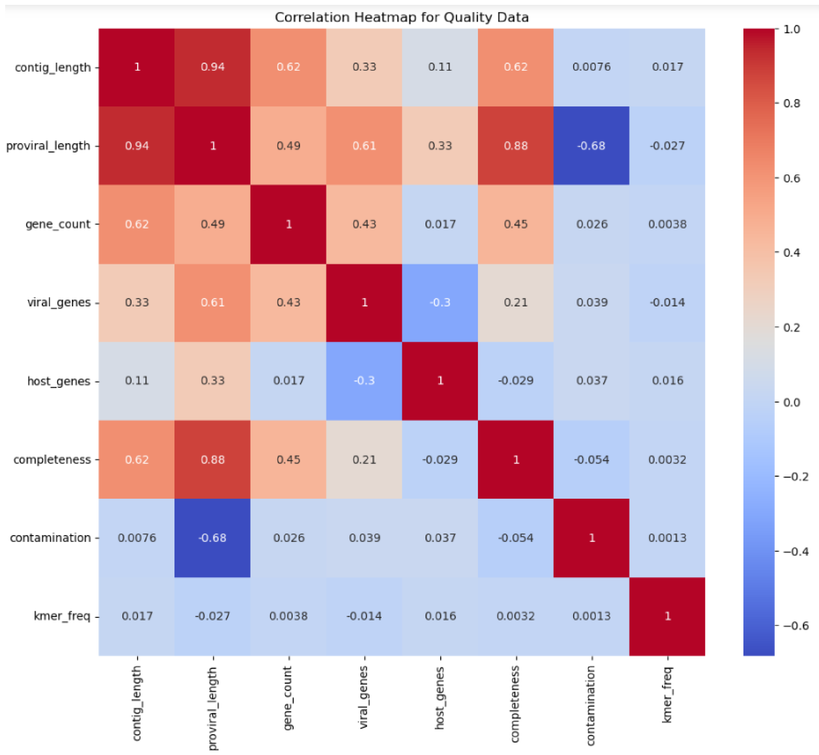
## Data Collection

Viral genomic data was collected from coral reefs in the central Pacific Ocean. Approximately 60 to 100 liters of seawater was concentrated to less than 500 milliliters using a 100 kDa tangential flow filter. The concentrated sample was then filtered through a 0.45 µm filter to remove bacteria, and 0.5% chloroform was added to destroy any residual cells. Samples were stored at 4°C until further processing. Purification of viruses was achieved using a cesium chloride step gradient, followed by DNase treatment to degrade contaminating DNA. Viral DNA was then extracted using the formamide/chloroform isoamyl alcohol technique. The purity of the extracted viral DNA was confirmed through PCR amplification with universal 16S rDNA primers. For sequencing, viral DNA was amplified using the Linker Amplified Sequencing Library method and sequenced on an Illumina MiSeq platform. Low-quality reads and human contaminants were removed during bioinformatics processing, with the cleaned sequences deposited in the MG-RAST database for analysis.



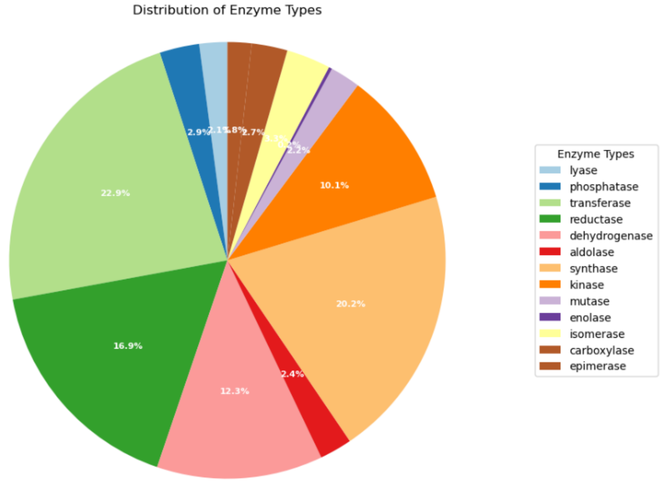

## Pipeline

Initial Data Collection and Preprocessing: Acquired single read DNA FASTA files and ran through MEGAHIT to produce viral contigs.
CheckV Analysis: The FASTA files were run through CheckV to assess the quality and completeness of the viral contigs.
Completeness Analysis: The completeness.tsv file generated by CheckV was analyzed to identify contigs with the highest completeness scores for further analysis.




Shown in this heatmap:
Strong Correlation between Contig Length and Proviral Length: This strong positive correlation indicates that contig length is closely associated with proviral length, suggesting that longer contigs may represent more complete or intact viral genomes. Correlation between Proviral Length and Completeness: This high correlation suggests that longer proviruses tend to have higher completeness scores, reinforcing the idea that length is a good proxy for assembly quality and completeness. Negative Correlation between Proviral Length and Contamination: This negative correlation indicates that longer proviruses tend to have lower contamination levels, which is a positive sign of assembly quality. Lower contamination means that the sequences are more likely to be accurate representations of viral genomes without extraneous material. Correlation between Gene Count and Completeness: A moderate positive correlation here suggests that contigs with a higher number of genes are more likely to be complete, supporting the idea that gene-rich contigs are of higher quality. Correlation between Viral Genes and Proviral Length: This suggests that as proviral length increases, the number of viral genes also increases. This could be indicative of more complex viral genomes being captured in longer contigs. Negative Correlation between Host Genes and Viral Genes: This negative correlation might indicate that sequences with more viral genes have fewer host genes, useful for differentiating between viral and host contamination or for identifying host-derived sequences.

Annotation and Functional Analysis with Prokka: Custom Database Configuration: Prokka was configured to use a custom database, specifically the carbohydrate-active enzyme (dbCAN) database, to annotate the viral contigs. Stepwise Annotation Process: Longest Contigs Annotation: The longest contigs were first selected based on sequence length and annotated using Prokka. Most Complete Contigs Annotation: Contigs with the highest completeness scores were then processed through Prokka to identify functionally relevant genes. All Contigs Annotation: Finally, all contigs were run through Prokka to ensure comprehensive annotation of potential carbohydrate metabolic enzymes. Post-Annotation Analysis: Enzyme Identification and Filtering: After Prokka annotation, the output files were examined for the presence of key enzymes involved in carbohydrate metabolism, such as kinases, phosphatases, transferases, and aldolases. Custom scripts were written and executed to filter and extract annotations related to these enzymes into separate files for downstream analysis.




## Future Interests

Functional Insights and Further Analysis with InterProScan: InterProScan Analysis: The annotated protein sequences (from the .faa files) will be further analyzed using InterProScan to identify functional domains and better understand the potential roles of the identified enzymes in viral interaction with host carbohydrate metabolism. Metabolic and protein modeling may also be done to potentially elucidate more interactions.

## Acknowledgements

BIG Summer
Ben Knowles and Hopeful Monsters Lab