Annotation of the GemG Bacteriophage

Amanda Gregg, *Northwestern Connecticut Community College (currently at Central Connecticut State University)*

Renee Brenckman, Xavier Rodriguez, Elijah Lovejoy and Professor Sharon Gusky - Northwestern Connecticut Community College

Jianna Calcinari, Kate Hass, Loiuse Leonard, Skylar Robinson, and Logan Wilson - Torrington High School

Supported by NSF ATE Grant Number: 1801062

INTRODUCTION

According to the CDC, each year more than 2.8 million Americans are diagnosed with antibiotic resistant infection and over 35,000 people die.[1] Bacteriophages can be used to treat antibiotic resistant bacterial infections.

A bacteriophage is a virus that infects specifically bacteria by hijacking a bacterial cell's mechanisms and components in order to produce more of itself, resulting in the lysis of the host cell. Our work focuses on studying the GEMG bacteriophage genome. We determined where genes are within the newly discovered GEMG genome and identified their function; This is gene annotation. This project is part of the SEAPHAGES project, a student-sourcing undergraduate research program that is jointly administered by Graham Hatfull's group at the University of Pittsburgh and the Howard Hughes Medical Institute's Science Education division.

METHODS

During the spring, we identified the genes that make up GemG's genome and identified their functions using various bioinformatic software programs. To identify the start of a gene, DNA Master's BLAST feature was used to approximate the start and end of a gene and its length. To verify these results, PhagesDB's Starterator compared starts from DNA Master with other bacteriophages in the same cluster. To identify the function of a gene, students verified information provided by DNA Master by submitting the gene's protein sequence into the database HHPRED. Information was documented using PECAAN. The following summer, Torrington high school students worked with us to identifying the presence of promoter regions and conserved repeats using DNA Master and Softberry annotation software and the PhagesDB database. We investigated the function of small sections of sequences within the generated chart that visibly shared the alignments in the DNA sequences of different phages.

RESULTS

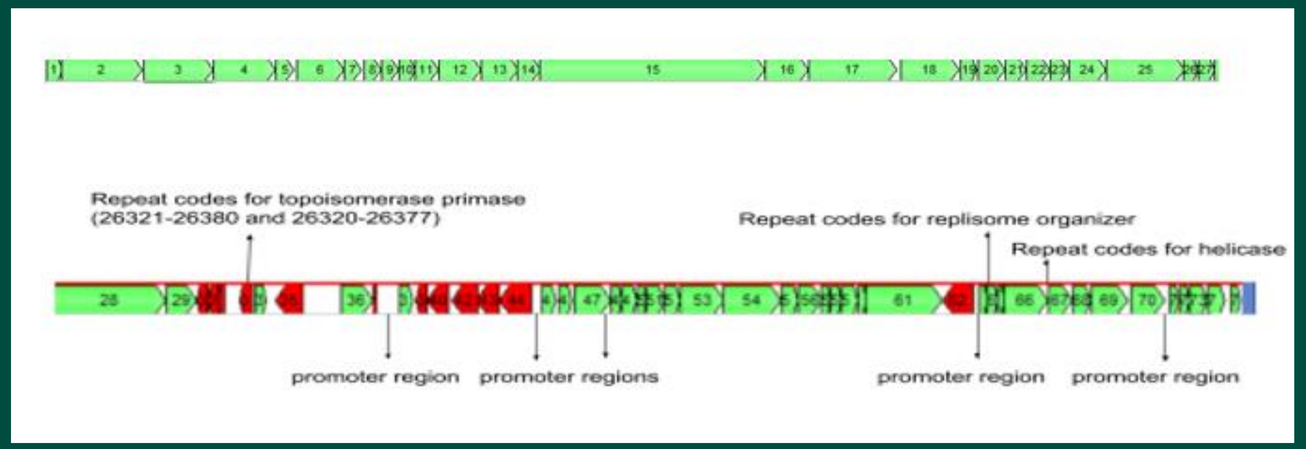
Number of Genes Found	**75 genes**
Number of Gene Functions Identified	**51 genes**
Number of Genes with Unknown Functions	**24 genes**
Number of Promoter Regions Identified in the 23,001–46,185 bp region	**5 promoters**
Number of conserved repeats Identified in the 23,001–46,185 bp region	**6 repeats**

DISCUSSION

Among the 37 genes studied we could only validate five promoter regions. They were found upstream of genes 38, 45, 48,63, and 71. Conserved repeats were located by genes 28, 33, 34, between genes 66 and 67, 68 and 69. The significance of repeats in genes 28 and 34 were undetermined, but 2 repeats located within gene 33 that codes for Topoisomerase primase, an enzyme that assists in the removal of supercoiling in DNA.

We determined that the genome of GemG contains 75 genes and we identified the function for 51 of the genes. The **annotated genome** was submitted to GenBank and published in September under Accession# MT818424.1

Research this summer resulted in the identification of 5 **promoter regions** and 6 **conserved repeats** within the second half of its genome.



Discussion: Implications and Impact

Gordonia Phage GemG's genome was fully annotated this spring and submitted to GenBank for review the following July and was recently published on September 29th, 2020. The ability to contribute to bacteriophage research was made possible by the Howard Hughes Medical Institute, a non-profit research organization located in Chevy Chase, Maryland. The SEAPHAGES project is a student-sourcing undergraduate research program founded by Graham Hatfull's group at the University of Pittsburgh and the Howard Hughes Medical Institute's Science Education division. The purpose of this project is to discover new bacteriophages through established microbiology techniques and characterize them through the genetic annotation and bioinformatic analysis of their genome. There are still many unknowns in the realm of phage research. With each annotation project, however, more connections and discoveries can be made which may be the basis for life-changing technology and medical breakthroughs. Through continued research, it's very possible the development of highly intricate, effective medicines and gene therapies from bacteriophage can be made available in the near future.

Discussion: Limitations of the project

Limitations were also experienced while using the software. DNA master called 91 promoter regions in total between base pairs 23,001 and 46,185. Between DNA Master and the results from Softberry's BPROM program, only 5 promoter regions could be verified. These promoter regions lie upstream of genes 38, 45, 48, 63, and 71. Using the BLAST feature in DNA master served as a secondary assessment to validate our findings. DNA Master has a genome sequencing accuracy of 80%. It's possible that BPROM sifted through this 20% error margin and found what was accurately generated by DNA Master. Another potential limitation of this project involved the range in education. Participating college students have taken more biology classes than the participating high schoolers, so this might have affected our research. However, this issue was managed by clearly by stating our goals and providing introductory material to the high school students at the beginning of this project, explaining the biology behind the microorganism we are studying and how we will study them.

References:

1. Antibiotic/Antimicrobial Resistance (AR / AMR). Atlanta (GA): Centers for Disease Control and Prevention; 2019 [updated 2020 June 18; cited 2020 Sept. 1]. https://www.cdc.gov/drugresistance/biggest-threats.html