The Annotation of the Complete Genome of the Mycobacterium phage Inverness

NICHOLAS SHELTRA[1], RAND DEARDORFF[1], SHARELLE BAILEY[1], EMMA DASILVA-MARTINEZ[1], SHANNON DICKEY[1], BETTINA EVANS[1], ADDIE MEHL[1], SHARON GUSKY[1]*

[1]*STEM Department Connecticut State Community College: Northwestern Campus, Winsted, CT*

**sharon.gusky@ctstate.edu*

Abstract: The gene annotation of the Mycobacterium phage Inverness was performed to establish certain genetic characteristics and qualities of the phage. Our research was designed to investigate the potential usefulness of this phage for medical purposes as part of the SEA-PHAGES project. Due to the decline in effectiveness of antibiotics when treating bacterial diseases, demand for alternative therapies and treatment has grown substantially. Phages have the potential to meet this demand. The genome of Inverness was found to be 68,264 base pairs in length, to possess a GC content of 66.5%, and to contain 99 protein-coding genes. Based on nucleotide similarities, the Mycobacterium phage Inverness was placed into cluster B and subcluster B1. The 33 genes with identifiable functions had an almost even split between rightward (49.49%) and leftward (50.51%) oriented genes. No putative function could be identified for 66 genes. No tRNAs were found to be present within the genome for this specific phage. It should also be noted that, among those genes with identified functions, two codes for lysins, which are proteins that kill bacteria cells. This suggests potential future opportunities to use this phage as a treatment for certain cases of antibiotic resistant infections.

Keywords: bacteriophage, gene annotation, bioinformatics

Introduction

The Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) Program is a course-embedded research program where students search for new bacteriophages in soil samples, perform a variety of laboratory techniques, and complete a complex genome annotation. The goals of the project are twofold: first, to increase undergraduate student interest and retention in the biological sciences, and second, to identify phages that can be used to treat antibiotic resistant infections [1]. The SEA-PHAGES project is jointly run by the Howard Hughes Medical Institute's Science Education division and Graham Hatfull's laboratory at the University of Pittsburgh. Participation in the SEA-PHAGES project has been shown to increase student retention and to influence career choices [1]. Students at Connecticut State Community College, Northwestern campus, participated in the SEA-PHAGES program while taking a Molecular Genetics course. As part of this course, students performed the structural and functional annotation of the Mycobacterium phage Inverness to determine its potential to treat antibiotic resistant infections.

Antibiotic resistant infections occur when mutation of the infecting bacterium prevents its destruction by antibiotics. Infections with antibiotic resistant organisms can be difficult or impossible to treat. According to the World Health Organization, 1.27 million people died in 2019 due to infections with antibiotic resistant organisms [2]. Bacteriophages have proven successful as a last resort in treating antibiotic resistant infections and are being tested in clinical trials [3].

Phages, also known as bacteriophages, are viruses that infect and replicate within bacterial cells. They are the most prevalent biological agent on Earth and are found everywhere in the environment [4]. The size, shape, and genetic structure of phages exhibit remarkable diversity [4]. All phages are made up of a nucleic acid genome covered in a capsid protein shell which protects the genetic information and facilitates its transfer to host cells. Many phages have tails that are used to deliver the genome into the host cell. The ability of phages to lyse bacterial cells makes them potentially effective in treating patients with antibiotic resistant infections [5]. To understand the potential for a phage to be used to treat patients, the genome of the phage must be

DOI: 10.5281/zenodo.12993820

annotated. Gene annotation involves the comprehensive process of detecting and characterizing genes within a genome. It starts by using computational gene prediction tools to identify potential genes based on genomic features. Structural annotation involves identifying the boundaries of the genes and functional annotation involves assigning putative functions to gene products. Gene annotation is crucial for understanding a phage's genetic traits, which must be considered when evaluating a phage for the possible treatment of antibiotic resistant infections. For a phage to be used as treatment "the phage genomes should not include any genes known or suspected to be toxic" [3].

Methods

Obtaining the Sequence of the Mycobacterium phage Inverness

The work to identify, isolate, purify, extract, and sequence DNA was not performed as part of this research project. However, a brief description of those processes is described here.

The Mycobacterium phage Inverness was collected from a bag of Miracle-Gro® potting soil obtained in Fort Collins, Colorado, USA. It was isolated, purified, and amplified by Sean Anderson of Rocky Mountain High School in Colorado, as part of the Phage Hunters Integrating Research and Education (PHIRE) program [6], using *Mycobacterium smegmatis mc^2 155* as a host [7].

DNA extracted from the phage was sent to the Pittsburgh Bacteriophage Institute for sequencing. The Pittsburgh Bacteriophage Institute completed sequencing on December 20th, 2020, using Illumina Sequencing, with an approximate shotgun coverage of 511. The shotgun method of sequencing involves breaking the genome up into pieces, sequencing each piece and then reconstructing the entire genome. The shotgun coverage number, in this case, 511, describes the average number of reads that align to the reference database.

The sequence information was used to create the FASTA file. The FASTA file containing the text-based sequence of the nucleotides for the phage genome was uploaded into the PhagesDB database by the SEA-PHAGES project administrators [8].

Annotation Process

The FASTA file was obtained from the PhagesDB database and loaded into DNA Master v5.0.2, a gene exploration and annotation tool used to predict the probable genes in the sequence [9]. The Mycobacterium phage Inverness genome sequence was also run through the evidential programs contained within The Phage Evidence Collection And Annotation Network (PECAAN) version 20221109 [10]. PECAAN was utilized to compile data from several other databases for comparison [10].

Within PECAAN, gene start and stop recommendations came from the Gene Locator and Interpolated Markov ModelER (Glimmer) system v3.02b, along with the GeneMark v4.28, and Starterator v1.2 systems [11,12]. GeneMark was also utilized for determining coding capacity [11].

This information along with the Z-score, gap or overlap between genes, the final score, and coding capacity were used to select the best starting position for each gene.

The Z-score provides the standard deviation of a score when compared to the best scores from all possible start positions in the genome. The Z-score provides the standard deviation of a score when compared to the best scores from all possible start positions in the genome. DNA Master and PECAAN produce Z-scores for the various possible starting positions of each predicted gene. While the exact values of the Z-scores vary for each gene, the best Z-score is one that is closest to 2. The starting position selected for each gene is based on selecting a position that has a calculated Z-score that is closest to 2.

In addition, the Genemark map was used as an effective visual reference for the regional coding potential of prospective gene candidates, and Actinobacteriophage Phamerator, version 567, was used to compare related phages in subcluster B1 [13].

Gene functions were determined by comparing the protein sequences for each gene to previously annotated genomes. PECAAN provided protein analysis recommendations from BLASTp v2.13.0., the Protein database and Non-Redundant Protein Sequences database, and HHPred v2.08 as well as NCBI_Conserved_Domains (CD) databases [14,15]. These databases are used by the PECAAN algorithms to detect similarities between proteins. The evidence selected was based on evaluating probability ratings and e-values, which measure the significance between sequences.

The Transmembrane Helices: Hidden Markov Model, TMHMM, provided evidence on transmembrane protein predictions [16]. Transmembrane proteins play a role in controlling the lysis of bacteria after infection by a bacteriophage. Phages with transmembrane proteins have the potential to kill bacteria and, therefore, to be used to treat antibiotic resistant infections.

TRNAscan and Aragorn programs were used to look for the presence of tRNAs [17,18]. The full annotation from PECAAN was run through DNA Master to create the minimal file suitable for submission to the PhagesDB website following SEA-PHAGES protocols.

A quality control check was performed before the complete genome was submitted to GenBank by the SEA-PHAGES' administrators who check to make sure that the guiding principles for gene annotation were followed and that all the functions call are allowable functions [19].

Results and Discussion

The Mycobacterium phage Inverness has a 68,264 base pair long genome with a GC content of 66.5% with a circularly permuted genome end character. Genomes with circularly permuted genomes form circular molecules upon injection into the host. These circular molecules can be used for replication of the phage. The genome annotation identified 99 protein-coding genes, with a nearly equal distribution between the rightwards (49.49%) and leftwards (50.51%) genes.

Thirty-three genes were assigned putative functions, but no functions could be identified for the remaining 66 genes.

Proteins involved in capsid structures are clustered between genes #9 and #13. Proteins involved in the tail structure are located between genes #18 and #41. The tape measure protein was identified as being coded for by gene #28. This gene determines tail length. Long-tailed phages have a tape-measure protein gene consisting of 2,000 or more base pairs.

The tape measure protein gene in the Mycobacterium phage Inverness is 5,976 base pairs long and codes for 1,991 amino acids, indicating that it has a long tail. Figure 1 shows the size of the tape measure protein gene when compared to the genes for the tail assembly chaperone and the minor tail protein.

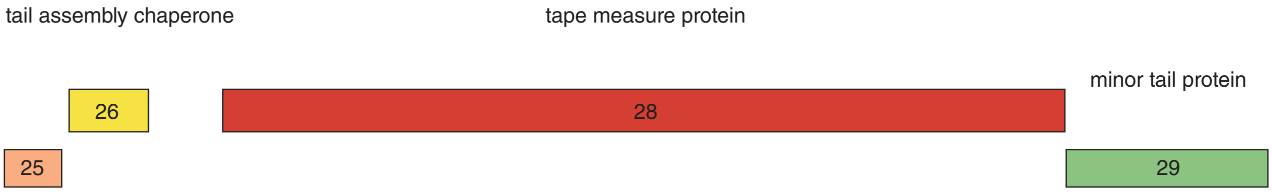

Fig. 1: This figure, taken from the Phamerator Database [13], shows the relative size of the tape measure protein gene (#28) when compared to the tail assembly chaperone genes #25 and #26 and the minor tail protein #29. The reverse gene #27 is not shown.

Genes assigned putative functions include those coding for structural proteins like RuvC-resolvase and for enzymes involved in DNA replication and packaging like DNA helicase, DNA primase, and HNH endonuclease.

DNA Helicase is the enzyme responsible for the break of the hydrogen bonds between DNA strands during the process of DNA replication. DNA primase is also used during DNA replication and is responsible for adding primers or starting sequences to the strands. This allows DNA polymerase to add new nucleotides to the growing nucleotide strands. HNH endonuclease is a small DNA-binding protein that can cleave the covalent bonds between nucleotides in a strand.

Genes that code for Lysin A and Lysin B were identified as #48 and #49, respectively. Lysins are enzymes that destroy bacteria cell walls. This leads to the death of the bacteria cells and indicates that this phage has the potential to be used to treat antibiotic resistant infections.

No tRNAs were found, indicating that the phage uses the host tRNAs in the translation process.

DOI: 10.5281/zenodo.12993820

The complete genome annotation can be found in the National Library of Medicine's GenBank Database under accession number OR159656 [20].

The full list of genes with identified functions can be found in Table 1.

Table 1 Genes with Identified Function

Gene Number	Length in basepairs	# of Amino Acids	Product
1	567	188	Adenylate Kinase
2	1788	595	Terminase
6	555	184	RUVC-like Resolvase
8	1905	634	Portal Protein
9	2574	857	Major Capsid and Fusion Protein
10	435	144	HNH Endonuclease
12	1746	581	Major capsid Hexamer Protein
13	804	267	Major Capsid Pentamer Protein
15	309	129	Holin
18	801	266	Major Tail Protein
20	777	258	Queuine tRNA Ribosyltranferase
22	738	245	Head-to-Tail Adapter
25	423	140	Tail Assembly Chaperone
26	564	187	Tail Assembly Chaperone
28	5976	1991	Tape Measure Protein
29	1434	477	Minor Tail Protein
30	1113	370	Minor Tail Protein
31	2256	751	Minor Tail Protein
32	1347	448	Minor Tail Protein
33	1161	386	Minor Tail Protein
41	408	135	Tail Fiber
45	216	71	Helix-turn-Helix Binding Domain Protein
46	525	174	Helix-turn-Helix Binding Domain Protein
48	1329	442	Lysin A
49	1356	451	Lysin B
51	1404	476	Exonuclease
52	1704	567	DNA Helicase
57	2748	915	DNA Primase
59	1860	619	DNA Replicase
66	180	59	Ribbon Helix-turn-Helix Binding Domain Protein
68	699	232	DNA Binding Protein
82	303	100	HNH Endonuclease

DOI: 10.5281/zenodo.12993820

Conclusion

The Mycobacterium phage Inverness was assigned to cluster B and subcluster B1 based on the nucleotide similarities to other phages in the clusters. Like other subcluster B1 phages, Mycobacterium phage Inverness infects mycobacterium and has a GC content of 66.5% and a base pair length of 68,264 [21]. Based on its gene content, it is predicted to be of the siphovirus morphotype. Phages with this morphotype have long, flexible tails that are non-contractible and have heads that are hexagonal and icosahedral.

The capsid size, head, and tail length of Mycobacterium Phage Inverness are unknown since electron microscopy was not performed. However, it was determined that gene #12 codes for a hexamer major capsid protein, confirming that the head is hexagonal. The tape measure gene (#28) was found to be 5,976 base pairs long, confirming that the tail is long.

The Mycobacterium Phage Inverness contains two genes that code for lysins. Lysin A is coded for in gene #48, and Lysin B is coded for in gene #49. Lysins disrupt the complex structures of bacteria cells, leading to rapid cell lysis [22]. This action kills the bacteria cells and offers a promising solution to combat infections caused by antibiotic-resistant strains of Mycobacterium.

Acknowledgements. We acknowledge support from the Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES), Howard Hughes Medical Institute, Chevy Chase, MD, including support from S.M.R. Gurney, Claire Reinhart, D.A. Russell, Deborah Jacobs-Sera, Vic Sivanathan, and Grahm Hatfull. The National Science Foundation (NSF) Advanced Technology Education Grant under award #1081062 provided research support for the authors.

Disclosures. The authors declare no conflicts of interest.

References

[1] D. I. Hanauer, et al., "An inclusive Research Education Community (iREC): Impact of the SEA-PHAGES program on research outcomes and student learning," *Proc Natl Acad Sci*, vol. 114, no. 51, pp. 13531-13536, Dec. 2017, doi: 10.1073/pnas.1718188115

[2] World Health Organization, "Antibiotic resistance," WHO Fact Sheets. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance

[3] G. F. Hatfull, R. M. Dedrick, and R.T. Schooley, "Phage Therapy for Antibiotic-Resistant Bacterial Infections," *Annu Rev Med*, vol. 73, pp. 197–211, Jan. 2022, doi: 10.1146/annurev-med-080219-122208.

[4] L. M. Kasman, and L.D. Porter, "Bacteriophages," in StatPearls [Internet], StatPearls Publishing, 2022. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK493185/

[5] G. F. Hatfull, "Mycobacteriophages: From Petri Dish to patient," *PLoS Pathog*. Jul. 2022, doi: 10.1371/journal.ppat.1010602.

[6] Hatfull Lab, "Phage Hunters Integrating Research and Education (PHIRE)." [Online]. Available: https://hatfull.org/courses

[7] M. Poxleitner, W. Pope, D. Jacobs-Sera, V. Sivanathan, and G. Hatfull, "Phage discovery guide," Howard Hughes Medical Institute. [Online]. Available: https://seaphagesphagediscoveryguide.helpdocsonline.com/home

[8] D. A. Russell, and G. F. Hatfull, "PhagesDB: The actinobacteriophage database," *Bioinformatics*, vol. 33, no. 5, pp. 784–786, Mar. 2017, doi: 10.1093/bioinformatics/btw711.

[9] W. H. Pope, D. Jacobs-Sera, "Annotation of bacteriophage genome sequences using DNA Master: an overview," *Methods Mol Biol*, vol. 1681, pp. 217–229, 2018, doi: 10.1007/978-1-4939-7343-9_16.

DOI: 10.5281/zenodo.12993820

[10] C. A. Rinehart, B. L. Gaffney, J. R. Smith, and J. D. Wood, “PECAAN: Phage Evidence Collection and Annotation Network,” Western Kentucky University Bioinformatics and Information Science Center. [Online]. Available: https://discover.kbrinsgd.org/login

[11] A. V. Lukashin, and M. Borodovsky, “GeneMark.hmm: new solutions for gene finding,” *Nucleic Acids Res*, vol. 26, no. 4, pp. 1107–1115, Feb. 1998, doi: 10.1093/nar/26.4.1107.

[12] M. Pacey, “Starterator guide,” University of Pittsburgh. [Online],
Available: https://seaphages.org/media/docs/Starterator_Guide_2016.pdf

[13] S. G. Cresawn, M. Bogel, N. Day, D. Jacobs-Sera, R.W. Hendrix, and G. F. Hatfull, “Phamerator: a bioinformatic tool for comparative bacteriophage genomics,” *BMC Bioinformatics*, vol. 12, no. 1, p. 395, Oct. 2011, doi: 10.1186/1471-2105-12-395.

[14] A. Marchler-Bauer et al.,“CDD: NCBI’s conserved domain database,” *Nucleic Acids Res* vol. 43, no. D1, pp. D222–D226, Jan. 2015, doi: 10.1093/nar/gku1221.

[15] J. Söding, A. Biegert, and A. N. Lupas, 2005, “The HHpred interactive server for protein homology detection and structure prediction,” *Nucleic Acids Res*, vol. 33, no. suppl_2, pp. W244–W248, Jul. 2005, doi: 10.1093/nar/gki408.

[16] N. Chaturvedi, S. Shanker, V. K. Singh, D. Sinha, P.N. Pandey, “Hidden markov model for the prediction of transmembrane proteins using MATLAB,” *Bioinformation*, vol. 7, no. 8, pp. 418–421, 2011, doi: 10.6026/97320630007418.

[17] D. Laslett, and B. Canback, “ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences,” *Nucleic Acids Res*, vol. 32, no. 1, pp. 11–16, 2004, doi: 10.1093/nar/gkh152.

[18] T. M. Lowe, and S.R. Eddy, “tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence,” *Nucleic Acids Res*, vol. 25, no. 5, pp. 955–964, Mar. 1997, doi: 10.1093/nar/25.5.955.

[19] M. Poxleitner, W. Pope, D. Jacobs-Sera, V. Sivanathan, and G. Hatfull, “Guiding Principles of Bacteriophage Gene Annotation,” SEAPHAGES Bioinformatics Guide, Howard Hughes Medical Institute. [Online]. Available: https://seaphagesbioinformatics.helpdocsonline.com/home

[20] S. A. Bailey, et al., “Mycobacterium phage Inverness: The Complete Genome,” National Library of Medicine: GenBank. [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/OR159656

[21] G. F. Hatfull, “The secret lives of mycobacteriophages,” *Adv Virus Res*, vol. 82, pp. 179–288, 2012, doi: 10.1016/B978-0-12-394621-8.00015-7.

[22] C. Ghose, C.W. Euler, “Gram-Negative Bacterial Lysins,” *Antibiotics*, vol. 9, no. 2, Art. no. 2, Feb. 2020, doi: 10.3390/antibiotics9020074.

DOI: 10.5281/zenodo.12993820