Genomic Annotation of Bosection6, a Bacteriophage that Infects *Mycobacterium smegmatis*

Robert H. Frazier ‘26 and Michael J. Wolyniak

Department of Biology, Hampden-Sydney College, Hampden-Sydney, VA 23943

Introduction

Bacteriophages, also known as phages, are viruses that infect bacterial hosts to replicate. Consisting primarily of a head and a tail, the only function of a phage is to reproduce, requiring bacteria as living hosts since phages cannot reproduce on their own. The head of a phage contains the genome, while the tail acts as the mechanism that binds and penetrates the host cell membrane. The secondary function of the tail is to be the pathway of DNA between the head of the phage to the cytoplasm of the host cell. Once inside the host cell, phages reproduce through two distinct cycles with varying degrees of host cell component influence. First, and most prominent, is the lytic cycle, where the phage commandeers the ribosomal host machinery to create replica phages based on the DNA of the original attacking phage. Second, and less frequent, is the lysogenic cycle, where the DNA of the original attacking phage inserts itself into the genome of the host cell, to then follow the lytic cell but only when the host cell replicates their own genome. The phages populate inside the host cell until the cell membrane cannot contain all of the phages, leading the cell to lysate or burst open, causing the cell to be killed and populating the environment with many new phages (Hatfull et al., 2022).

As there are phages for every type of bacteria, an estimated tenfold amount of phages to bacteria is thought to exist. This belief leads to phages potentially being the most abundant virus, in addition to being the most diverse organism or biological entity (Clokie et al., 2011). This exceptional diversity is shown visually from their phenotype, which has a range of heads and tails to follow the biological concept of “structure follows function”, since bacteria do not have uniform characteristics for phages to overcome (Ackermann, 2006). Where phages are even more diverse are their genomes, which can range from 3,000 to 500,000 base-pairs, often including many non-existential genes and sequence regions. When this trait is paired with the relatively minor number of genes that phages need to replicate, the diversity of phages increases dramatically. Yet, phages are limited to bacteria as the only external source of diversity, leading to the back-and-forth mechanisms of phages and bacteria. For bacteria, phages operate as vectors to drive evolution, where phages can easily incorporate and transfer the DNA of previous hosts to new hosts. On the other hand, phages use this DNA to protect themselves against new threats of bacteria, leading to a phage population often turning over quite quickly (Hatfull et al., 2011).

The most novel feature of phages is that the virus attacks and kills only bacteria specific to the individual phage, with the exception of phages often being able to infect species of bacteria that are very similar (Hatfull et al., 2022). This feature has given rise to what is known as “phage therapy”, where a specific phage is introduced to an organism to stop illness that the bacteria specific to the given phage causes. A notable example of this was when phages that undergraduate students of the SEA-PHAGES program found were used to treat the symptoms of a mycobacterial infection. The patient, a 15-year-old with cytosis fibrosis and a multitude of other health problems, was infected with *Mycobacterium abscessus*. To combat the infection, intravenous antibiotics were used, but only resulted in worsening symptoms. As a last resort, the health team turned to phage therapy and easily recreated phages that the SEA-PHAGES program found. Using three phages named Muddy, ZoeJ, and BPs, a cocktail of phages and antibiotics were given intravenously every 12 hours for 32 weeks at a 10^9 pfu concentration. As continued treatment was administered, the symptoms of illness began to decrease and the patient was discharged (Dedrick et al., 2019).

This example serves as the first known use of phage therapy to treat a human mycobacterial infection. As antibiotic resistant bacteria become more common, phage therapy continues to be the most evident candidate to replace antibiotics to stop human bacterial infections. For phage therapy to continue, further phages need to be identified and characterized (Hatfull et al., 2022). Alongside this need, the role of phages within the microbial world is a complex relationship, such that understanding their evolution and diversity includes constant research on new phages (Hatfull et al., 2011). The SEA-PHAGES program exists to allow undergraduate students to annotate phage genomes in an informative and instructive process. This project, using a unique system of the accurate procedures of the SEA-PHAGES program, allowed for the annotation of the genome of Bosection6, a previously identified and phenotypically characterized bacteriophage that infects *Mycobacterium smegmatis*. *M. smegmatis* was specifically used as the species’ DNA genome is over 90% identical to *M. tuberculosis*, a pathogen that is causing a public health crisis (Tuberculosis). From sequencing the genome, identifying the genes, and discovering the encoded products of a single phage, the medical and ecological facets of phages can be improved.

Materials and Methods

The project begins with the genome of Bosection6 sequenced at the Pittsburgh Bacteriophage Institute of the University of Pittsburgh. With the genome known, basic characteristics of the genome can be identified, such as the base-pair length that serves as the range where genes reside. To view the genome, the sequence is uploaded into a software program called DNA Master, a commonly used software for annotating genomes. The software is also the recommended program for annotating phages, making the program the basis for annotating Bosection6. DNA Master provides a platform on which to find open reading frames or ORFs, while having features that manipulate, add, and subtract ORFs. Best described as “candidate genes” that can be determined to be genes through different methods, the first step with the genome in DNA Master is to have the software auto-annotate the genome, to find some, but not all, of the ORFs. This feature saves time, due to the alternative being the manual search for ORFs between each start and end site of the genome, which also increases the probability for human error. The auto-annotation also predicts the direction that the ORFs transcribe, in either the 5’ or 3’ direction, which can be cross-examined when the genes are evaluated through BLAST (SEA-PHAGES, 2024).

To determine whether the ORFs that DNA master found could be considered real genes, the common gene prediction software GeneMark is utilized. DNA master and GeneMark use the same method of gene prediction, where the “*ab initio*” method is based upon intrinsic evidence within the genome, rather than comparative or “wet-lab” evidence, to predict the ORFs of a sequence. This includes finding and marking known components of an ORF, such as a promoter region that enhances transcription, and coding potential, where the specific arrangement of base-pairs within a region of a sequence leads to the region possibly being a gene. When these two components are combined, ORFs are created with relative start and end sites, with the region of the sequence having significant potential to encode a gene. As such, the DNA Master auto-annotation and GeneMark data can be cross-examined to decide whether the previously found ORFs can be considered real, with the ORFs that both software programs predict, with similar relative start and end sites, being considered genes (SEA-PHAGES, 2024).

After finding all genes, the next step is to decide on the start and end sites for each gene, which begins with comparing the predicted start sites of many software programs, including DNA Master, GeneMark, and Glimmer. Although GeneMark does not provide a confidence score in their generated start site, a DNA master score below 2.0 and a Glimmer score above 2.0 is considered confident. Since the software programs are known to not properly account for gaps between genes, all possible start sites between the upstream gene and the start sites that the software programs predict are evaluated through BLAST. The start site that most closely matches the same gene in another phage is labelled as the chosen start site. Although BLAST is treated as the final predictor of the start site, the program does not account for mutations within the genes; to account for this, the software program Starterator is utilized. Allowing the chosen start site to be evaluated against many start sites for the same gene, rather than just a single start site when using BLAST, and also having the chosen start site within the range of the start sites for the same gene results in the chosen start site agreeing with the other start sites on where the gene begins. DNA Master and GeneMark are both known to be very accurate for the end site, so the end site was not evaluated in a manner similar to the start site. This leaves the two software programs to always have the same end site, which is treated as the chosen end site (SEA-PHAGES, 2024).

With the start and end sites known, the encoded product of the genes can now be found. More specifically, the comparative method of determination is utilized at two different levels of the central dogma of biology, specifically the DNA sequence that forms the genes and the protein structure/sequence that the genes encode. At the DNA sequence level, genes with matching nucleotide sequences in other organisms or biological entities are found through BLAST, and genes with matching nucleotide sequences in other phages are found through PhagesDB. At the protein level, genes with matching protein structures are found through Hhpred, and genes with matching protein sequences are found through BLAST, the same search engine used as before but for a different type of sequence. As BLAST, PhagesDB, and Hhpred are search engines, finding matching sequences or structures to those of Bosection6 logically results in the same encoded product. The E-value, or expected value of a function of Bosection6 randomly matching the function of another phage, must be less than 1-50% for comparative evidence to be genuine. As the DNA sequence is usually more accurate than the protein structure or sequence when using comparative evidence, having matching results from BLAST and PhagesDB is used for the final product of what the genes encode for, while matching results from Hhpred and BLAST is used at supporting evidence (SEA-PHAGES, 2024).

The role of comparative analysis is further used when looking at the genes as a whole through another software program called Phamerator. Allowing the entire genome of Bosection6 to be compared as a whole to the genome of other phages, Phamerator displays the biological concept of conservation, where certain features that construct an organism or biological entity can be seen to not change through evolution. The software program also permits Bosection6 to be viewed as a complete biological entity, rather than just a graph in DNA Master or individual genes through search engines, which contributes to other facets of the annotation. This includes whether genes known to be required for a phage to exist are present within the genome, while determining which genes or region of the sequence show genetic diversity. Therefore, Bosection6 can be compared to a phage with a similar number of genes and base-pair length named Butters, for whether the genes of Bosection6 matched the genes of Butters and matched their position within the gene order of their respective genomes. While looking at Bosection6 as a whole, any overlaps or shared base-pairs of genes are examined to within an acceptable length of 50 base-pairs, while any gaps or noncoding spaces between genes that are greater than 100 base-pairs are examined through BLAST, Hhpred, and Phamerator to not contain or identify with other genes (SEA-PHAGES, 2024).

Results

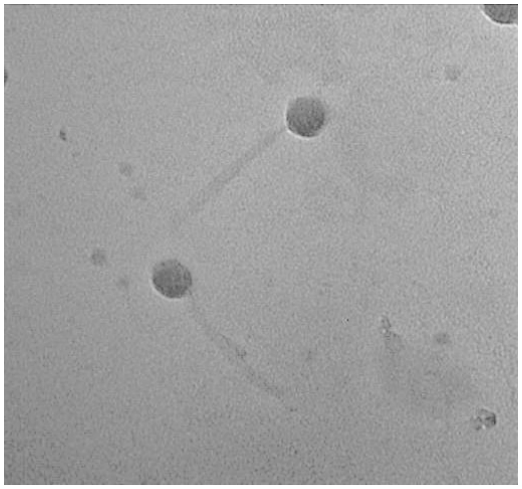

Figure 1: Bosection6 under an electron microscope at 36,000x magnification

Figure 2 displays the sequenced Bosection6 genome and DNA Master auto-annotation. The sequencing discovered that Bosection6 has 43,412 base-pairs, with a 13-base 3' overhang of "CCCGCCGCAATGG". The genome also has 66% guanine-cytosine gene content and 33% adenosine-thymine gene content, with sequence similarity that PhagesDB labels Bosection6 as belonging to cluster N of all sequenced phages. The DNA Master auto-annotation resulted in 66 ORFs within the genome of Bosection6, which are shown below. The short vertical lines represent start sites, and the long vertical lines represent end sites. The green horizontal boxes represent the ORFs that encode in the 5' direction while the red horizontal boxes represent the ORFs that encode in the 3' direction.

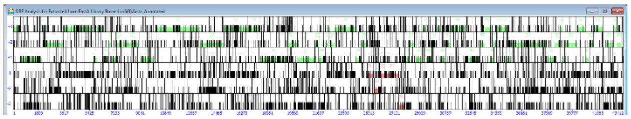

Figure 2: Sequenced Bosection6 genome

Table 1 displays the direction of the ORFs and whether the ORFs appeared in the GeneMark data. The "Genemark" column shows whether the ORFs that DNA Master predicted appeared on the GeneMark graph with similar relative start and end sites, the "Direction" column shows the direction of the genes after the directions that DNA Master predicted were cross-examined when the genes were evaluated through BLAST.

Gene	Direction	Genemark
1	Forward	Does appear on gene mark graph
2	Forward	Does appear on gene mark graph
3	Forward	Does appear on gene mark graph
4	Forward	Does appear on gene mark graph
5	Forward	Does appear on gene mark graph
6	Forward	Does appear on gene mark graph
7	Forward	Does appear on gene mark graph
8	Forward	Does appear on gene mark graph
9	Forward	Does appear on gene mark graph
10	Forward	Does appear on gene mark graph
11	Forward	Does appear on gene mark graph
12	Forward	Does appear on gene mark graph
13	Forward	Does appear on gene mark graph
14	Forward	Does appear on gene mark graph
15	Forward	Does appear on gene mark graph
16	Forward	Does appear on gene mark graph
17	Forward	Does appear on gene mark graph
18	Forward	Does appear on gene mark graph
19	Forward	Does appear on gene mark graph
20	Forward	Does appear on gene mark graph
21	Forward	Does appear on gene mark graph
22	Forward	Does appear on gene mark graph
23	Forward	Does appear on gene mark graph
24	Forward	Does appear on gene mark graph
25	Forward	Does appear on gene mark graph
26	Forward	Does appear on gene mark graph
27	Forward	Does appear on gene mark graph
28	Forward	Does appear on gene mark graph

29	Forward	Does appear on gene mark graph
30	Forward	Does appear on gene mark graph
31	Reverse	Does appear on gene mark graph
32	Reverse	Does appear on gene mark graph
33	Reverse	Does appear on gene mark graph
34	Reverse	Does appear on gene mark graph
35	Reverse	Does appear on gene mark graph
36	Forward	Does appear on gene mark graph
37	Forward	Does appear on gene mark graph
38	Forward	Does appear on gene mark graph
39	Forward	Does appear on gene mark graph
40	Forward	Does appear on gene mark graph
41	Forward	Does appear on gene mark graph
42	Forward	Does appear on gene mark graph
43	Forward	Does appear on gene mark graph
44	Forward	Does appear on gene mark graph
45	Forward	Does appear on gene mark graph
46	Forward	Does appear on gene mark graph
47	Forward	Does appear on gene mark graph
48	Forward	Does appear on gene mark graph
49	Forward	Does appear on gene mark graph
50	Forward	Does appear on gene mark graph
51	Forward	Does appear on gene mark graph
52	Forward	Does appear on gene mark graph
53	Forward	Does appear on gene mark graph
54	Forward	Does appear on gene mark graph
55	Forward	Does appear on gene mark graph
56	Forward	Does appear on gene mark graph
57	Forward	Does appear on gene mark graph
58	Forward	Does not appear on gene mark graph
59	Forward	Does appear on gene mark graph
60	Forward	Does appear on gene mark graph
61	Forward	Does appear on gene mark graph
62	Forward	Does appear on gene mark graph
63	Forward	Does appear on gene mark graph
64	Forward	Does appear on gene mark graph
65	Forward	Does appear on gene mark graph
66	Forward	Does not appear on gene mark graph

Table 2 displays the first half of the information collected for determining the start sites of the genes. The "DNA master auto-annotation" column shows the start site that DNA Master predicted, the "Chosen start site" column shows the start site that is most probable, and the "Genemark" column shows the start site that GeneMark predicted.

Gene	DNA master auto-annotation	Chosen start site	Genemark
1	rbs of -4.585 for 64 nuc	rbs of -4.585 for 64 nuc	Does not align with gene mark start of 97
2	rbs of -4.122 for 501 nuc	rbs of -4.122 for 501 nuc	Does not align with gene mark start of 504
3	rbs of -4.335 for 2068 nuc	rbs of -4.699 for 2029 nuc	Does not align with gene mark start of 2050
4	rbs of -3.733 for 2249 nuc	rbs of -3.733 for 2249 nuc	Does not align with gene mark start of 2342
5	rbs of -2.681 for 3538 nuc	rbs of -2.681 for 3538 nuc	Does align with gene mark start of 3538
6	rbs of -2.505 for 4372 nuc	rbs of -2.505 for 4372 nuc	Does not align with gene mark start of 4384
7	rbs of -3.219 for 5799 nuc	rbs of -5.134 for 5730 nuc	Does align with gene mark start of 5730
8	rbs of -5.867 for 5956 nuc	rbs of -5.867 for 5956 nuc	Does align with gene mark start of 5956
9	rbs of -5.351 for 6438 nuc	rbs of -5.351 for 6438 nuc	Does align with gene mark start of 6438
10	rbs of -4.467 for 6794 nuc	rbs fo -4.467 for 6794 nuc	Does align with gene mark start of 6794
11	rbs of -6.418 for 6975 nuc	rbs of -6.418 for 6975 nuc	Does align with gene mark start of 6975
12	rbs of -3.279 for 7343 nuc	rbs of -3.279 for 7343 nuc	Does align with gene mark start of 7343
13	rbs of -2.297 for 7877 nuc	rbs of -2.297 for 7877 nuc	Does align with gene mark start of 7877
14	rbs of -3.314 for 9036 nuc	rbs of -3.314 for 9036 nuc	Does align with gene mark start of 9036
15	rbs of -5.672 for 9517 nuc	rbs of -3.314 for 9036 nuc	Does not align with gene mark start of 9541
16	rbs of -7.604 for 10123 nuc	rbs of -7.604 for 10123 nuc	Does align with gene mark start of 10123
17	rbs of -3.709 for 13268 nuc	rbs for -3.709 for 13268 nuc	Does align with gene mark start of 13268
18	rbs of -3.319 for 14989 nuc	rbs of 3.319 for 14989 nuc	Does align with gene mark start of 14989
19	rbs of -5.045 for 16714 nuc	rbs of -5.045 for 16714 nuc	Does align with gene mark start of 16714
20	rbs of -4.578 for 17556 nuc	rbs of -4.578 for 17556 nuc	Does align with gene mark start of 17556
21	rbs of -2.214 for 19550 nuc	rbs of -2.214 for 19550 nuc	Does align with gene mark start of 19550
22	rbs of -3.189 for 20223 nuc	rbs of -4.159 for 20220 nuc	Does not align with gene mark start of 20223
23	rbs of -3.917 for 21446 nuc	rbs for -3.917 for 21446 nuc	Does align with gene mark start of 21446
24	rbs of -2.175 for 21744 nuc	rbs of -2.175 for 21744 nuc	Does align with gene mark start of 21744
25	rbs of -4.337 for 22169 nuc	rbs of -5.307 for 22166 nuc	Does not align with gene mark start of 22169
26	rbs of -4.351 for 22348 nuc	rbs of -4.351 for 22348 nuc	Does align with gene mark start of 22348
27	rbs of -5.161 for 22601 nuc	rbs of -7.656 for 22565 nuc	Does not align with gene mark start of 22601
28	rbs of -5.352 for 22834 nuc	rbs of -5.352 for 22834 nuc	Does align with gene mark start of 22834
29	rbs of -4.337 for 24339 nuc	rbs of -4.337 for 24339 nuc	Does align with gene mark start of 24339
30	rbs of -6.155 for 24740 nuc	rbs of -6.155 for 24740 nuc	Does align with gene mark start of 24740
31	rbs of -4.134 for 25577 nuc	rbs of -4.134 for 25577 nuc	Does not align with gene mark start of 25565
32	rbs of -8.928 for 25968 nuc	rbs of -8.319 for 26040 nuc	Does not align with gene mark start of 25986
33	rbs of -6.577 for 26405 nuc	rbs of -6.577 for 26405 nuc	Does align with gene mark start of 26405
34	rbs of -7.274 for 27410 nuc	rbs of -7.274 for 27410 nuc	Does not align with gene mark start of 27311
35	rbs of -4.701 for 27661 nuc	rbs of -6.354 for 27823 nuc	Does not align with gene mark start of 27787
36	rbs of -5.376 for 27921 nuc	rbs of -6.161 for 27873 nuc	Does not align with gene mark start of 27921

37	rbs of -3.363 for 28151 nuc	rbs of -3.363 for 28151 nuc	Does align with gene mark start of 28151
38	rbs of -5.092 for 28647 nuc	rbs of -5.092 for 28647 nuc	Does align with gene mark start of 28647
39	rbs of -2.196 for 29134 nuc	rbs of -2.196 for 29134 nuc	Does align with gene mark start of 29134
40	rbs of -4.905 for 29385 nuc	rbs of -4.905 for 29385 nuc	Does align with gene mark start of 29385
41	rbs of -4.905 for 29741 nuc	rbs of -4.905 for 29741 nuc	Does align with gene mark start of 29741
42	rbs of -4.618 for 30103 nuc	rbs of -4.618 for 30103 nuc	Does align with gene mark start of 30103
43	rbs of -5.721 for 30372 nuc	rbs of -5.721 for 30372 nuc	Does align with gene mark start of 30372
44	rbs of -4.861 for 30752 nuc	rbs of -4.861 for 30752 nuc	Does align with gene mark start of 30752
45	rbs of -6.764 for 31093 nuc	rbs of -6.764 for 31093 nuc	Does align with gene mark start of 31093
46	rbs of -4.927 for 32109 nuc	rbs of -4.927 for 32109 nuc	Does align with gene mark start of 32109
47	rbs of -5.532 for 33170 nuc	rbs of -5.532 for 33170 nuc	Does align with gene mark start of 33170
48	rbs of -4.135 for 33529 nuc	rbs of -4.135 for 33529 nuc	Does align with gene mark start of 33529
49	rbs of -3.250 for 33918 nuc	rbs of -3.250 for 33918 nuc	Does align with gene mark start of 33918
50	rbs of -4.641 for 34163 nuc	rbs of -4.641 for 34163 nuc	Does align with gene mark start of 34163
51	rbs of -3.812 for 34375 nuc	rbs of -3.812 for 34375 nuc	Does align with gene mark start of 34375
52	rbs of -5.411 for 34777 nuc	rbs of -5.411 for 34777 nuc	Does align with gene mark start of 34777
53	rbs of -3.633 for 35094 nuc	rbs of -3.633 for 35094 nuc	Does align with gene mark start of 35094
54	rbs of -3.549 for 35309 nuc	rbs of -3.549 for 35309 nuc	Does align with gene mark start of 35309
55	rbs of -4.781 for 38193 nuc	rbs of -4.820 for 38055 nuc	Does not align with gene mark start of 38193
56	rbs of -4.516 for 38687 nuc	rbs of -4.583 for 38504 nuc	Does align with gene mark start of 38504
57	rbs of -2.645 for 38905 nuc	rbs of -2.645 for 38905 nuc	Does align with gene mark start of 38905
58	rbs of -3.854 for 39045 nuc	rbs of -3.854 for 39045 nuc	Not present on gene mark graph
59	rbs of -3.447 for 39206 nuc	rbs of -3.447 for 39206 nuc	Does align with gene mark start of 39206
60	rbs of -3.904 for 40096 nuc	rbs of -3.904 for 40096 nuc	Does align with gene mark start of 40096
61	rbs of -5.959 for 40695 nuc	rbs of -5.959 for 40695 nuc	Does align with gene mark start of 40695
62	rbs of -2.253 for 40940 nuc	rbs of -2.253 for 40940 nuc	Does align with gene mark start of 40940
63	rbs of -4.399 for 41269 nuc	rbs of -4.399 for 41269 nuc	Does not align with gene mark start of 41275
64	rbs of -5.390 for 42146 nuc	rbs of -5.390 for 42146 nuc	Does not align with gene mark start of 42227
65	rbs of -4.213 for 42694 nuc	rbs of -4.213 for 42694 nuc	Does align with gene mark start of 42694
66	rbs of -4.477 for 42917 nuc	rbs of -4.477 for 42917 nuc	Not present on gene mark graph

Table 3 displays the second half of the information collected for determining the start sites of the genes. The “Glimmer” column shows the start site that Glimmer predicted, the “Blast according to chosen start site” shows how the “Chosen start site” from Table 2 correlates with the start site of the same gene in other phages through BLAST, and the “Agreement in Starterator” column shows whether the “Chosen start site” started within the range of other start sites for the same gene in other phages through Starterator.

Gene	Glimmer	Blast according to chosen start site	Agreement in Starterator
1	Score of 12.13 for 64 nuc	Matches Charlie gp1 q1:s1 100% 6-99	Significant agreement
2	Score of 15.56 for 501 nuc	Matches Xeno gp2 q1:s1 100% 0	Significant agreement
3	Score of 10.3 for 2068 nuc	Matches Charlie gp3 q1:s1 100% 1-40	Significant agreement
4	Score of 10.82 for 2249 nuc	Matches Charlie gp4 q1:s1 100% 0	Significant agreement
5	Score of 12.2 for 3538 nuc	Matches Xeno gp5 q1:s1 100% 0	Significant agreement
6	Score of 17.65 for 4372 nuc	Matches Charlie gp6 q1:s1 100% 0	Significant agreement
7	Score of 10.99 for 5799 nuc	Matches Charlie gp7 q1:s1 100% 2-46	Significant agreement
8	Score of 11.44 for 5956 nuc	Matches Charlie gp8 q1:s1 100% 3-112	Significant agreement
9	Score of 12.33 for 6438 nuc	Matches Charlie gp9 q1:s1 100% 5-79	Significant agreement
10	Score of 13.14 for 6794 nuc	Matches Charlie gp10 q1:s1 100% 1-37	Significant agreement
11	Score of 8.24 for 6975 nuc	Matches Charlie gp11 q1:s1 100% 1-83	Significant agreement
12	Score of 10.54 for 7343 nuc	Matches Charlie gp12 q1:s1 100% 2-97	Significant agreement
13	Score of 13.66 for 7877 nuc	Matches Charlie gp13 q1:s1 100% 0	Significant agreement
14	Score of 15.94 for 9036 nuc	Matches Charlie gp14 q1:s1 100% 2-127	Significant agreement
15	Score of 11.58 for 9517 nuc	Matches Charlie gp15 q1:s1 100% 0	Significant agreement
16	Score of 9.82 for 10123 nuc	Matches Carcharodon gp16 q1:s1 100% 0	Significant agreement
17	Score of 11.66 for 13268 nuc	Matches Charlie gp17 q1:s1 100% 0	Significant agreement
18	Score of 11.14 for 14989 nuc	Matches Charlie gp18 q1:s1 100% 0	Significant agreement
19	Score of 13.76 for 16714 nuc	Matches Xeno gp19 q1:s1 100% 0	Significant agreement
20	Score of 14.49 for 17556 nuc	Matches Charlie gp20 q1:s1 100% 0	Significant agreement
21	Score of 2.73 for 19550 nuc	Matches Charlie gp21 q1:s1 100% 3-151	Significant agreement
22	Score of 6.4 for 20223 nuc	Matches Charlie gp22 q1:s1 100% 0	Significant agreement
23	Score of 10.27 for 21446 nuc	Matches Charlie gp23 q1:s1 100% 1-65	Significant agreement
24	Score of 8.88 for 21744 nuc	Matches Charlie gp24 q1:s1 100% 5-93	Significant agreement
25	Score of 12.01 for 22169 nuc	Matches Carcharodon gp25 q1:s1 100% 4-34	Significant agreement
26	Score of 4.87 for 22348 nuc	Matches Charlie gp26 q1:s1 100% 3-36	Significant agreement
27	Score of 17.43 for 22601 nuc	Matches Charlie gp27 q1:s1 100% 1-51	Significant agreement
28	Score of 10.33 for 22834 nuc	Matches Charcharodon q1:s1 gp28 100% 0	Significant agreement
29	Score of 15.37 for 24339 nuc	Matches Charlie gp29 q1:s1 100% 2-84	Significant agreement
30	Score of 11.76 for 24740 nuc	Matches Charlie gp30 q1:s1 100% 2-81	Significant agreement
31	Score of 10.93 for 25577 nuc	Matches Xeno gp30 q1:s1 100% 8-89	Significant agreement
32	Score of 5.36 for 25986 nuc	Matches Charlie gp32 q1:s1 100% 1-82	Significant agreement
33	Score of 7.2 for 26405 nuc	Matches Charlie gp33 q1:s1 100% 3-73	Significant agreement
34	Score of 7.42 for 27410 nuc	Matches Charlie gp34 q1:s1 100% 0	Significant agreement
35	Score of 2.22 for 27661 nuc	Matches Xeno gp34 q1:s1 100% 5-93	Significant agreement
36	Score of 5.85 for 27921 nuc	Matches Carcharodon gp38 q1:s1 100% 3-61	Significant agreement
37	Score of 13.93 for 28151 nuc	Matches Charlie gp37 q1:s1 100% 2-85	Significant agreement
38	Score of 11.86 for 28647 nuc	Matches Charlie gp39 q1:s1 100% 1-89	Significant agreement
39	Score of 7.77 for 29134 nuc	Matches Charcharodon gp41 q1:s1 100% 1-51	Significant agreement
40	Score of 9.2 for 29385 nuc	Matches Aggie gp40 q1:s1 100% 1-60	Significant agreement

41	Score of 5.15 for 29741 nuc	Matches Phrann gp45 q1:s1 100% 9-78	Significant agreement
42	Score of 12.13 for 30103 nuc	Matches Carcharodon gp44 q1:s1 100% 5-54	Significant agreement
43	Score of 6.06 for 30372 nuc	Matches Carcharodon gp45 q1:s1 100% 1-87	Significant agreement
44	Score of 11.65 for 30752 nuc	Matches Carcharodon gp46 q1:s1 100% 3-72	Significant agreement
45	Score of 11.4 for 31093 nuc	Matches Carcharodon gp47 q1:s1 100% 0	Significant agreement
46	Score of 12.95 for 32109 nuc	Matches Phrann gp49 q1:s1 100% 0	Significant agreement
47	Score of 8.51 for 33170 nuc	Matches Phrann gp50 q1:s1 100% 7-83	Significant agreement
48	Score of 11.55 for 33529 nuc	Matches Charlie gp50 q1:s1 100% 8-88	Significant agreement
49	Score of 10.66 for 33918 nuc	Matches Charlie gp51 q1:s1 100% 3-53	Significant agreement
50	Score of 10.39 for 34163 nuc	Matches Charlie gp52 q1:s1 100% 1-41	Significant agreement
51	Score of 14.21 for 34375 nuc	Matches Piper gp72 q1:s1 100% 1-92	Significant agreement
52	Score of 6.25 for 34777 nuc	Matches Piper gp73 q1:s1 100% 1-68	Significant agreement
53	Score of 9.44 for 35094 nuc	Matches Charlie gp55 q1:s1 100% 6-45	Significant agreement
54	Score of 8.1 for 35309 nuc	Matches Charlie gp56 q1:s1 100% 0	Significant agreement
55	Not calculated	Matches Charlie gp57 q1:s1 100% 2-105	Significant agreement
56	Score of 7.03 for 38687 nuc	Matches Charlie gp58 q1:s1 100% 8-95	Significant agreement
57	Score of 2.95 for 38905 nuc	Matches Charlie gp59 q1:s1 100% 1-25	Significant agreement
58	Score of 2.08 for 39045 nuc	Matches Charlie gp60 q1:s1 100% 1-31	Significant agreement
59	Score of 4.55 for 39206 nuc	Matches Charlie gp61 q1:s1 100% 0	Significant agreement
60	Score of 10.54 for 40096 nuc	Matches Carcharodon gp65 q1:s1 100% 5-138	Significant agreement
61	Score of 3.85 for 40695 nuc	Matches Aggie gp62 q1:s1 100% 4-55	Significant agreement
62	Score of 7.78 for 40940 nuc	Matches Aggie gp63 q1:s1 100% 3-74	Significant agreement
63	Score of 10.36 for 41269 nuc	Matches Charlie gp64 q1:s1 100% 4-159	Significant agreement
64	Score of 11.23 for 42146 nuc	Matches Phrann gp65 q1:s1 100% 7-127	Significant agreement
65	Score of 10.53 for 42694 nuc	Matches Carcharodon gp70 q1:s1 100% 1-44	Significant agreement
66	Score of 0.2 for 42917 nuc	Matches Carcharodon gp71 q1:s1 100% 7-47	Significant agreement

Table 4 displays the information collected to determine the encoded product of the genes. The "BLAST nucleotide" column shows the function that BLAST found for matching nucleotide sequences in other organisms or biological entities, the "PhagesDB" column shows the function that PhagesDB found for matching nucleotide sequences in other phages, the "Hhpred protein" column shows the function that Hhpred found for matching protein structures, and the "BLAST protein" column shows the function that BLAST found for matching protein sequences.

Gene	BLAST nucleotide	PhagesDB nucleotide	Hhpred protein	BLAST protein
1	hypothetical protein	function unknown	no match	no match
2	terminase, large subunit	terminase, large subunit	terminase, large subunit	terminase, large subunit
3	hypothetical protein	function unknown	no match	no match
4	portal protein	portal protein	portal protein	portal protein
5	capsid maturation protease	capsid maturation protease	no match	capsid maturation protease
6	major capsid protein	major capsid protein	major capsid protein	major capsid protein
7	hypothetical protein	function unknown	no match	no match
8	head-to-tail adaptor	head-to-tail adaptor	head-to-tail adaptor	head-to-tail adaptor
9	head-to-tail stopper	head-to-tail stopper	head-to-tail stopper	head-to-tail stopper
10	hypothetical protein	function unknown	no match	no match
11	minor tail protein	function unknown	no match	minor tail protein
12	head-to-tail adaptor	tail terminator	head-to-tail adaptor	head-to-tail adaptor
13	major tail protein	major tail protein	major tail protein	major tail protein
14	tail assembly chaperone	tail assembly chaperone	no match	tail assembly chaperone
15	tail assembly chaperone	tail assembly chaperone	no match	tail assembly chaperone
16	tape measure protein	tape measure protein	tape measure protein	tape measure protein
17	minor tail protein	minor tail protein	no match	minor tail protein
18	minor tail protein	minor tail protein	no match	minor tail protein
19	minor tail protein	minor tail protein	no match	minor tail protein
20	minor tail protein	minor tail protein	no match	minor tail protein
21	hypothetical protein	function unknown	no match	no match
22	minor tail protein	minor tail protein	no match	minor tail protein
23	hypothetical protein	function unknown	no match	no match
24	hypothetical protein	function unknown	no match	no match
25	hypothetical protein	function unknown	no match	no match
26	hypothetical protein	function unknown	no match	no match
27	hypothetical protein	function unknown	no match	no match
28	endolysin	lysin A	no match	endolysin
29	holin	holin	holin	holin
30	minor tail protein	function unknown	no match	minor tail protein
31	antitoxin in toxin/antitoxin system, HicB-like	antitoxin in toxin/antitoxin system, HicB-like	antitoxin in toxin/antitoxin system, HicB-like	antitoxin in toxin/antitoxin system, HicB-like
32	membrane protein	function unknown	no match	membrane protein
33	hypothetical protein	function unknown	no match	no match
34	tyrosine integrase	tyrosine integrase	tyrosine integrase	tyrosine integrase
35	immunity repressor	immunity repressor	immunity repressor	immunity repressor
36	immunity repressor	excise	immunity repressor	immunity repressor
37	hypothetical protein	function unknown	no match	no match
38	hypothetical protein	function unknown	no match	no match

39	hypothetical protein	function unknown	no match	no match
40	hypothetical protein	function unknown	no match	no match
41	hypothetical protein	function unknown	no match	no match
42	hypothetical protein	function unknown	no match	no match
43	WhiB family transcription factor	WhiB family transcription factor	WhiB family transcription factor	WhiB family transcription factor
44	hypothetical protein	function unknown	no match	no match
45	RecE-like exonuclease	RecE-like exonuclease	RecE-like exonuclease	RecE-like exonuclease
46	RecT-like DNA pairing protein	RecT-like DNA pairing protein	RecT-like DNA pairing protein	RecT-like DNA pairing protein
47	hypothetical protein	function unknown	no match	no match
48	Holliday junction resolvase	Holliday junction resolvase	Holliday junction resolvase	Holliday junction resolvase
49	thioredoxin	NrdH-like glutaredoxin	thioredoxin	thioredoxin
50	hypothetical protein	function unknown	no match	no match
51	hypothetical protein	function unknown	no match	no match
52	helix-turn-helix DNA binding domain	helix-turn-helix DNA binding domain	no match	helix-turn-helix DNA binding domain
53	hypothetical protein	function unknown	no match	no match
54	DNA methyltransferase	DNA methyltransferase	no match	DNA methyltransferase
55	hypothetical protein	function unknown	no match	no match
56	HNH endonuclease	HNH endonuclease	no match	HNH endonuclease
57	hypothetical protein	function unknown	no match	no match
58	hypothetical protein	function unknown	no match	no match
59	hypothetical protein	function unknown	no match	no match
60	hypothetical protein	function unknown	no match	no match
61	hypothetical protein	function unknown	no match	no match
62	hypothetical protein	function unknown	no match	no match
63	hypothetical protein	function unknown	no match	no match
64	hypothetical protein	function unknown	no match	no match
65	hypothetical protein	function unknown	no match	no match
66	HNH endonuclease	HNH endonuclease	no match	HNH endonuclease

Table 5 displays information on which genes show conservation or diversity. Specifically, the "Phamerator" column shows how the genes of Bosection6 match to the genes of another phage called Butters, for either gene or position. Matching for gene results in the gene of Bosection6 matching the gene of Butters, while matching for position results in the gene of Bosection6 and gene of Butters having the same place within their respective gene orders.

Gene	Phamerator
1	Matches Butters for gene and position
2	Matches Butters for gene and position
3	Matches Butters for gene and position
4	Matches Butters for gene and position
5	Matches Butters for gene and position
6	Matches Butters for gene and position
7	Matches Butters not for gene but for position
8	Matches Butters for gene and position
9	Matches Butters for gene and position
10	Matches Butters for gene and position
11	Matches Butters for gene and position
12	Matches Butters for gene and position
13	Matches Butters for gene and position
14	Matches Butters for gene and position
15	Matches Butters for gene and position
16	Matches Butters for gene and position
17	Matches Butters for gene and position
18	Matches Butters for gene and position
19	Matches Butters for gene and position
20	Matches Butters not for gene but for position
21	Matches Butters not for gene but for position
22	Matches Butters not for gene but for position
23	Does not match Butters for gene or position
24	Does not match Butters for gene or position
25	Matches 25 gene of Butters but not for position
26	Matches Butters not for gene but for position
27	Matches Butters not for gene but for position
28	Matches Butters not for gene but for position
29	Matches Butters not for gene but for position
30	Matches 29 gene of Butters but not for positon
31	Does not match Butters for gene or position
32	Does not match Butters for gene or position
33	Matches 36 gene of Butters but not for position
34	Does not match Butters for gene or position
35	Does not match Butters for gene or position
36	Matches 39 gene of Butters but not for position
37	Matches 40 gene of Butters but not for positon
38	Does not match Butters for gene or position
39	Does not match Butters for gene or position
40	Matches the 41 gene of Butters but not for position
41	Does not match Butters for gene or position
42	Matches 47 gene of Butters but not for position
43	Matches 48 gene of Butters but not for position
44	Does not match Butters for gene or position
45	Matches 50 gene of Butters but not for position
46	Matches 51 gene of Butters but not for position
47	Does not match Butters for gene or position
48	Matches 52 gene of Butters but not for position
49	Matches 53 gene of Butters but not for position
50	Matches 54 gene of Butters but not for position
51	Does not match Butters for gene or position
52	Does not match Butters for gene or position
53	Does not match Butters for gene or position
54	Does not match Butters for gene or position
55	Does not match Butters for gene or position
56	Does not match Butters for gene or position
57	Matches 58 gene of Butters but not for position
58	Does not match Butters for gene or position
59	Matches 62 gene of Butters but not for position
60	Matches 61 gene of Butters but not for position
61	Matches 62 gene of Butters but not for position
62	Does not match Butters for gene or position
63	Matches 63 gene of Butters but not for position
64	Matches 64 gene of Butters but not for position
65	Matches 65 gene of Butters but not for position
66	Matches 66 gene of Butters but not for position

Table 6 displays the final results of the genomic annotation of Bosection6. The "Gene" column shows the identifying number of the gene, the "Is gene real" column shows whether the DNA master auto-annotation and GeneMark data agree that the gene exists, the "Start" column shows the start site of the gene, the “End” column shows the end site of the gene, and the “Product” column shows the encoded product of the gene.

Gene	Is gene real	Start	End	Product
1	Yes	64	504	hypothetical protein
2	Yes	501	2036	terminase, large subunit
3	Yes	2029	2235	hypothetical protein
4	Yes	2249	3538	portal protein
5	Yes	3538	4344	capsid maturation protease
6	Yes	4372	5664	major capsid protein
7	Yes	5730	5963	hypothetical protein
8	Yes	5956	6441	head-to-tail adaptor
9	Yes	6438	6782	head-to-tail stopper
10	Yes	6794	6988	hypothetical protein
11	Yes	6975	7346	minor tail protein
12	Yes	7343	7762	head-to-tail adaptor
13	Yes	7877	8935	major tail protein
14	Yes	9036	9566	tail assembly chaperone
15	Yes	9036	9942	tail assembly chaperone
16	Yes	10123	13266	tape measure protein
17	Yes	13268	14989	minor tail protein
18	Yes	14989	16698	minor tail protein
19	Yes	16714	17556	minor tail protein
20	Yes	17556	19550	minor tail protein
21	Yes	19550	20191	hypothetical protein
22	Yes	20220	21437	minor tail protein
23	Yes	21446	21742	hypothetical protein
24	Yes	21744	22169	hypothetical protein
25	Yes	22166	22348	hypothetical protein
26	Yes	22348	22533	hypothetical protein
27	Yes	22565	22837	hypothetical protein
28	Yes	22834	24342	endolysin
29	Yes	24339	24743	holin
30	Yes	24740	25108	minor tail protein
31	Yes	25577	25182	antitoxin in toxin/antitoxin system, HicB-like
32	Yes	26040	25675	membrane protein
33	Yes	26405	26064	hypothetical protein
34	Yes	27410	26439	tyrosine integrase
35	Yes	27823	27416	immunity repressor
36	Yes	27873	28154	immunity repressor
37	Yes	28151	28537	hypothetical protein
38	Yes	28647	29075	hypothetical protein
39	Yes	29134	29388	hypothetical protein
40	Yes	29385	29744	hypothetical protein
41	Yes	29741	30106	hypothetical protein
42	Yes	30103	30375	hypothetical protein
43	Yes	30372	30755	WhiB family transcription factor
44	Yes	30752	31096	hypothetical protein
45	Yes	31093	32103	RecE-like exonuclease
46	Yes	32109	33173	RecT-like DNA pairing protein
47	Yes	33170	33532	hypothetical protein
48	Yes	33529	33921	Holliday junction resolvase
49	Yes	33918	34166	thioredoxin
50	Yes	34163	34378	hypothetical protein
51	Yes	34375	34776	hypothetical protein
52	Yes	34777	35097	helix-turn-helix DNA binding domain
53	Yes	35094	35312	hypothetical protein
54	Yes	35309	37933	DNA methyltransferase
55	Yes	38055	38504	hypothetical protein
56	Yes	38504	38905	HNH endonuclease
57	Yes	38905	39048	hypothetical protein
58	Yes	39045	39206	hypothetical protein
59	Yes	39206	40099	hypothetical protein
60	Yes	40096	40698	hypothetical protein
61	Yes	40695	40943	hypothetical protein
62	Yes	40940	41272	hypothetical protein
63	Yes	41269	41934	hypothetical protein
64	Yes	42146	42697	hypothetical protein
65	Yes	42694	42939	hypothetical protein
66	Yes	42917	43225	HNH endonuclease

Discussion and Conclusions

In conclusion, this project resulted in the genome of Bosection6 being sequenced and all genes being found within a certain level of confidence. The phage has a genome length of 43,412 base-pairs, which is slightly greater than the average genome length of phages belonging to cluster N at about 42,000 base-pairs (PhagesDB, 2024). More specifically, the genome length falls within the range of medium sized phage genomes, which is about 40,000 to 45,000 base-pairs (Dislers et al., 2020). The 13-base 3' overhang of "CCCGCCGCAATGG" is also not uncommon for phages to contain, but at times critical for the packaging of DNA inside the head of phages (Byrd et al., 2005). In terms of the genome content, the phage had a lower guanine-cytosine base pair content than the host bacterium, where the 66% g-c base pair content of Bosection6 is about 1.4% lower than the g-c base pair content of *M. smegmatis* (Almpanis et al., 2018). This follows a common principle of viruses, where phages usually have a lower g-c base pair content than their host (Baloni et al., 2015).

Within the genome of Bosection6, the DNA Master auto-annotation found 66 ORFs, which is lower than the average amount of about 69 genes for phages of cluster N (PhagesDB, 2024). This falls within the lower end of the number of genes for phages with double-stranded DNA, which is usually about 50 to 200 genes (Hatfull et al., 2020). DNA Master also correctly predicted the directions that the ORFs are transcribed, whereas the genes were evaluated through BLAST to have the same directions. When cross-examining the 66 ORFs that DNA master found to the GeneMark data, all ORFs, except #58 and #66, of the auto-annotation appeared in the GeneMark data, with similar relative start and end sites, to have all ORFs considered as genes. Since all other factors, such as BLAST, PhagesDB, and Hhpred, signaled ORFs #58 and #66 to in fact be real, it was decided to label the ORFs as genes. This is a practice used in other genomic annotations, that when only one factor signals an ORF to not be real, when many other factors do, the majority vote is chosen (SEA-PHAGES, 2024). This method of reasoning was extended to gene #55, which Glimmer could not generate a start site for. The GeneMark data also does not contain any ORFs that did not appear from the DNA Master auto-annotation.

The start sites of the DNA Master auto-annotation were more than often accurate, where the chosen start sites for 55 of the 66 genes followed the start sites of the auto-annotation after examination. Displaying the accuracy of DNA Master, the start sites consistently matched a start site that either GeneMark or Glimmer generated, while always matching the same gene but in a different phage when evaluated through BLAST. Yet, 11 start sites of the auto-annotation did not completely consider the gap between genes and were not replaced with a chosen start site found upstream. Although the chosen start sites for 9 of the 11 genes did not match the start site generated from either GeneMark or Glimmer, BLAST was most helpful to learn that the replacement "chosen start sites" match the same gene in another phage. In fact, all the genes of Bosection6 did not just closely match the start sites of the same gene in other phages, but had exactly matching start sites, leading to more definitive results. Additionally, all the chosen start sites had significant agreement in Starterator having the factor of mutation be accounted for, leaving the chosen start sites to become the definitive start sites. The end sites that DNA Master and GeneMark generated matched, leading to the chosen end sites becoming the definitive end sites. All genes had a DNA Master and Glimmer score above 2.0, except for gene #66 which had a Glimmer score of 0.8, but we continued with it since Glimmer did in fact generate a score.

All genes were found to have an encoded product, with 35 having a known function, while 31 could only be identified to the extent of being a hypothetical protein. Having a hypothetical protein as an encoded product does not diminish the confidence in the gene being real, as the function has simply not been determined with wet-lab evidence. Instead, using BLAST and PhagesDB to find matching nucleotide sequences showed that there was significant enough dry lab research involving the 31 genes of Bosection6 that had hypothetical proteins, leading to confidence when using the label for encoded products. Comparably, there was a great amount of evidence for the 35 genes with a known encoded product, since BLAST and PhagesDB found matching nucleotide sequences, in addition to BLAST finding matching protein sequences. Of the 35 genes, Hhpred could only find matching protein structures for 16 genes, which is not significant due to the protein structure being the least likely to find matching results. Furthermore, all the searches had the matching encoded product with an E-value less than 1-50%, or a value of 0%, demonstrating a 100% match. When looking at the genes as a whole, all of the encoded products needed for a phage to exist are within the genome, with the genes that encode for the required encoded products mostly towards the 5' end of the phage, and the genes that are not required but display genetic diversity towards the 3' end of the phage. (SEA-PHAGES, 2024).

Along with this subject, the conservation of the required encoded products is shown in the results of Phamerator, where the first 19 genes of Bosection6 match the first 19 genes of Butters for both gene and position, except for gene #7. Most of the genetic diversity is in the back half of the genome, where all 17 of the genes belonging to Bosection6 that do not match Butters for both gene and position are in the last half of the genome. When looking at the genes in relation to one another, there were 32 instances of genes overlapping, which occurred mostly with the genes in the last 2/3 of the genome, with no overlap exceeding the acceptable length for a phage. This number of instances of gene overlap is not uncommon, as phages have many genes within a relatively small genome. The overlap primarily stayed within the reading frame of the genome, such as base pairs being shared in multiples of 3, while multiples of 7 were occasionally shared. This blatant change in reading frames was furthered investigated, to discover that the reading frames of the overlapping genes usually matched the reading frames of overlapping genes of Butters. On the other hand, there were 27 instances of a gap between genes, with only 6 gaps being worthy of review. After analysis with BLAST, Hhpred, and Phamerator, there was no indication of the 6 gaps being contained or identifying with a gene. The gap lengths were relatively small in comparison to other phages, with the largest gap of Bosection6 being 212 base-pairs being considered quite minor for a phage (SEA-PHAGES, 2024).

Future Directions

Lastly, the genes are entered into a DNA Master file with the sequence of Bosection6, with their start/end sites and encoded products, to be the annotated genome of the phage. The annotated genome is sent to the SEA-PHAGES faculty for quality control, before being entered into PhagesDB, the database for phage genomes, and GenBank, the database for genomes of all organisms or biological entities.

REFERENCES

Ackermann, Hans-Wolfgang. “5500 Phages examined in the electron microscope”. Archives of Virology, vol. 152, no. 2, pp. 227-243, 2006. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/17051420/.

Almpanis, Apostolos, et al. “Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages”. Microbial Genomics, vol. 4, no. 4, pp. 000168, 2018. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/29633935/.

Baloni, Priyanka, et al. “Complete Genome Sequences of a Mycobacterium smegmatis Laboratory Strain (MC2 155) and Isoniazid-Resistant (4XR1/R2) Mutant Strains”. Genome Announcements, vol. 3, no. 1, pp. 01520-14, 2015. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/25657281/.

Byrd, Alicia, et al. “Increasing the length of the single-stranded overhang enhances unwinding of duplex DNA by bacteriophage T4 Dda helicase”. Biochemistry, vol. 44, no. 39, pp. 12990-12997, 2005. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/16185067/.

Clokie, Martha, et al. “Phages in nature”. Bacteriophage, vol. 1, no. 1, pp. 31-45, 2011. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/21687533/.

Dedrick, Rebekah, et al. “Engineered bacteriophages for treatment of patient with a disseminated drug resistant Mycobacterium abscessus”. Nature Medicine, vol. 25, 2019, pp. 730-733. National Library of Medicine, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6557439/.

Dislers, Andris, et al. “Motley Crew: Overview of the Currently Available Phage Diversity”. Frontiers in Microbiology, vol. 29, no. 11, pp. 579452, 2020. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/33193205/.

Hatfull, Graham, et al. “Bacteriophages and their genomes”. Current Opinion in Virology, vol. 1, no. 4, pp. 298-303, 2011. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/22034588/.

Hatfull, Graham, et al. “Identification of mycobacteriophage toxic genes reveals new features of mycobacterial physiology and morphology”. Scientific Reports, vol. 10, no. 1, pp. 14670, 2020. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/32887931/.

Hatfull, Graham, et al. “Phage Therapy for Antibiotic-Resistance Bacterial Infections”. Annual Review of Medicine, vol. 73, 2022, pp. 197-221. Annual Review, https://www.annualreviews.org/doi/full/10.1146/annurev-med-080219-122208#_i2.

“PhagesDB: The Actinobacteriophage Database”. The Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science, 2024. https://phagesdb.org/.

“SEA-PHAGES Bioinformatics Guide”. The Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science, 2024. https://seaphagesphagediscoveryguide.helpdocsonline.com/home.

“Tuberculosis”. World Health Organization, 2024. https://www.who.int/news-room/fact-sheets/detail/tuberculosis.