'Broinformatics: The Study and Analysis of Cluster K Mycobacteriophage Genomic Data

James B. Hughes '14 and Michael J. Wolyniak

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

The objective of this project is to annotate and characterize the genomic data for mycobacteriophage Cheetobro. The Cheetobro phage was isolated by Drew Whitt in Hampden-Sydney, Virginia in 2011 and was sequenced at the Pittsburgh Bacteriophage Institute using Ion Torrent technology. Based on the genome sequence, Cheetobro was classified as a Cluster K mycobacteriophage and a member of the subcluster K4 along with phages Fionnbharth and Slarth. The annotation work was done using the DNA sequencing software DNA Master, which uses the programs Glimmer and Genemark to predict the individual protein coding genes within the genome based on common gene parameters such as start and stop codons, length, promoter regions, Shine-Dalgarno sequences, and coding potential. Once the predictions were made, human editorial skills were used to analyze each predicted gene and correct any mistakes made by the software. It was found that the Cheetobro genome was a double-stranded linear DNA genome 57,253bp in length with 92 predicted protein coding genes and a GC content of 68.0%. Corrections made during annotation included removing the falsely predicted gene 44 and altering the start sites of predicted genes 10, 15, 37, 43, 47, 57, 75, 76, and 87. Also during the annotation process it was found that Cheetobro possesses a remarkable similarity to mycobacteriophage Fionnbharth. In addition, the Tapemeasure and integrase genes were possibly identified preliminarily at genes 22 and 45 respectively along with Lysine tRNA from base pairs 30852 to 30927. The next step of this project would be to further characterize the Cheetobro genome by using protein BLAST searches through the NCBI protein database in order to identify the putative functions of each predicted gene and to compare the Cheetobro genome to other Cluster K phages using the bioinformatics software Phamerator. In addition, experimentation will be performed on the phage in order to characterize its host range to determine its potential for the medical field and to explore the importance of the tRNA found in the genome.

INTRODUCTION

Mycobacteriophage Cheetobro was isolated from an old mulch pile in Hampden-Sydney, Virginia in 2011 by Hampden-Sydney College student Drew Whitt. Once isolated and purified, Cheetobro was sent off to the Pittsburgh Bacteriophage Institute where on April 23, 2012 its genome was sequenced using Ion Torrent sequencing technology. After each phage genome is sequenced it is organized into a group of other phages based on how similar its genome is to the other phages in the group. These groups of genetically similar phages are known as clusters and within each cluster are subclusters of even more genetically similar phages. In this case Cheetobro was identified as a Cluster K phage and was placed in the subcluster K4 along with two other sequenced phages, Fionnbharth and Slarp. Using Electron Microscopy the morphotype of Cheetobro was identified as *Siphoviridae*. This means its morphology consists of an icosahedral capsid along with a long, flexible, and non-contractile tail, which is the common morphology among the Cluster K phages as found by Hatfull and colleagues (Pope *et al.*, 2011, p. 6).

Mycobacteriophage Cheetobro was isolated as part of the Science Eduaction Alliance phage research project sponsored by the Howard Hughes Medical Institute in order to construct the public online genomic database of isolated novel bacteriophages, the Mycobacteriophage Database, and to further research and knowledge in the field of bacteriophages. This interest taken to the study of phages and their genomic data has been sparked by the increasing threat of infectious bacterial strains becoming ever more resistant to treatment by antibiotics. As a result, this project and research is being conducted in the hope that phages may be utilized as an alternative treatment against bacterial infection, especially against that of the lethal and increasingly resistant *Mycobacterium tuberculosis*. So far the genomic information of 3,607 bacteriophages has been entered into this database by schools from around the country in the hopes of reaching this goal (The Mycobacteriophage Database, n.d.).

However, the focus of this individual project is directed toward the Cluster K phages and the novel phage Cheetobro. Currently there are 23 known Cluster K phages but not a whole lot of research has been done with them as a whole and not much is known about them collectively, though what is known about one K phage in particular shows some of the most promise toward the medical field. The most known and most studied Cluster K phage is mycobacteriophage TM4, which was isolated as a prophage from a strain of *Mycobacterium avium* in 1984. Research with TM4 has shown that the Cluster K pages have a broad host range that spans both fast growing and slow growing mycobacterium, which very

importantly includes *M. tuberculosis.* As a result, TM4 was the first phage to be used for shuttle plasmid construction and is still utilized for efficient gene delivery to *M. tuberculosis* (Pope *et al.*, 2011, p. 4). Research has also shown that TM4 tagged with enhanced green fluorescent protein can be an effective indicator of antibiotic resistance in strains of *M. tuberculosis* grown on antibiotic medium (Rondón *et al.*, 2011, p. 1838-1842). With the great potential shown by mycobacteriophage TM4 for benefitting fields of medicine and mycobacterial research, it is very likely that the other Cluster K phages possess the same potential as well.

The overall goal of this project is to characterize the genome and behavior of mycobacteriophage Cheetobro through bioinformatics and experimentation. This will be done first by identifying each gene within the genome using the sequencing software DNA Master. This software is unique because it utilizes the programs Glimmer and GeneMark to scan all six open reading frames (ORF) of the genome for possible genes. They do this by looking for the telltale signs of a gene such as gene start sites or start codons like ATG, GTG, and TTG along with possible stop sites or stop codons like TAG, TGA, and TAA. They also search upstream of possible genes for the possible presence of promoters, which facilitate the transcription of that gene. In addition, they analyze the amino acid products of the DNA to determine which regions have the greatest coding potential or the greatest potential to code for and produce a functional protein. Genemark also generates a spreadsheet displaying all of the coding potential on every ORF for the entire genome. The genes are also predicted by Shine-Dalgarno sequences, which are nucleotide sequences preceding a gene that serve as ribosomal binding sites for the process of translation or protein synthesis and are marked by the consensus sequence of nucleotides AGGAGG. These programs use these sequences to identify possible genes and assign each a Shine-Dalgarno score based on how well the sequence fits the consensus sequence and parameters of a Shine-Dalgarno sequence. Glimmer and Genemark then use these markers along with appropriate gene length and minimal overlap between genes to piece together and predict the genes of the genome. In addition, each program assigns a score to each predicted gene based on how well it fits the gene parameters. Once this is complete, DNA Master then searches for the closest match for each gene by performing a BLAST search for each gene through the Mycobacteriophage Database. A BLAST search takes the object of interest such as a gene and compares it to every gene stored within a database in order to find its closest match. Once the genes of the Cheetobro genome have been successfully identified, the putative function or encoded protein product of each gene will be analyzed by performing BLAST searches through the protein database of NCBI. In addition, the bioinformatics software Phamerator will be used to compare the entire Cheetobro genome side by side to any phage genome within the Mycobacteriophage Database in order to discern the differences and similarities between Cheetobro and other phages. Lastly, experimentation will be performed on genes or regions of interest within the Cheetobro genome that were identified during the annotation process in order to further characterize mycobacteriophage Cheetobro.

METHODS

The Fasta file of the completely sequenced Cheetobro genome was downloaded from the online Mycobacteriophage Database and was then loaded into the DNA Master sequencing software where it was processed and analyzed by Glimmer and Genemark in order to predict the genes of the genome. After DNA Master analyzed the genome and completed the gene predictions, human editorial skills were then utilize to analyze and confirm or correct the predictions made by the software by using the gene prediction data, BLAST information, and GeneMark coding potential spreadsheet. The genome was then processed through web-based tRNA seach program tRNAscan-SE in order to confirm any tRNA sequence identified by the Glimmer and Genemark programs in the genome and to identify any additional tRNA sequences missed by Glimmer and Genemark.

RESULTS AND DISCUSSION

The goal of the first part of the project to characterize the mycobacteriophage Cheetobro as reported by this paper was to complete the annotation of the phage's genome by identifying each potential protein coding gene in the genome using the predictions made by the DNA Master sequencing software. The genome of Cheetobro was found to be a double-stranded linear DNA genome 57,253 nucleotide base pairs (bp) in length with a guanine and cytosine (GC) nucleotide content of 68.0%. The genome was also found to have defined physical ends with 3' single-stranded complementary DNA

extensions that were 11bp in length. In addition, the genome was found to contain 92 protein coding genes. This initial genome data was found to be consistent with the other Cluster K phage genomes because as reported by Hatfull and colleagues, the Cluster K phages all together have an average length of 59,442bp with an average GC content of 67.1%. The Cluster K phages, with the exception of the K2 phages TM4 and Mufasa, all have the 11bp 3' terminal extensions as well. I was also reported by Hatfull that each Cluster K phage contains between 90 and 100 predicted protein coding genes, which is also the case for Cheetobro (Pope *et al.*, 2011, p. 9).

Originally in the DNA Master predictions, the Glimmer program had predicted 93 potential protein coding genes, but during the annotation process it was found that the Glimmer predicted gene 44 possessed little evidence of being gene at all. Glimmer called the start of gene 44 at base pair 33028 and the end at base pair 33507 giving it a length of 480bp. Glimmer also assigned it a relatively low gene potential strength of 6.63. However, Genemark did not call a potential gene within that region of the genome, and the Genemark coding potential spreadsheet did not display any coding potential for that ORF. Also the predicted gene 44 had no BLAST matches within Mycobacteriophage Database and its product sequence of amino acids did not have any protein BLAST matches in the NCBI protein database as well, which further discredited the prediction. As a result, it was determined that the predicted gene 44 was not protein coding gene and was falsely called by Glimmer. The gene was then removed from the DNA Master file.

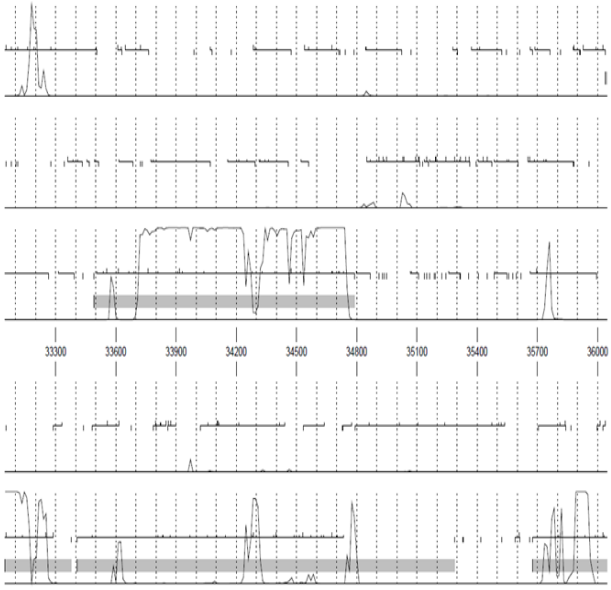

Figure 1: Above is an image of the Genemark spreadsheet, which demonstrates the protein coding potential for each gene. In this image it can be seen that the region predicted as gene 44 (base pairs 33028 to 33507) did not possess any reasonable coding potential.

It was also found during the annotation process that the predicted start codon sites for genes 10, 15, 37, 43, 47, 57, 75, 76, and 87 had been incorrectly predicted by Glimmer and Genemark and required a new start site to be determined using the evidence presented by DNA Master and the Genemark coding potential spreadsheet. For gene 10 the original Glimmer prediction called the start codon at base pair 3956 with a predicted strength of 19.82. This start site was shown not to include all of the coding potential in the ORF on the Genemark spreadsheet, had a Shine-Dalgarno score of 252, and caused the gene's amino acid product sequence to be out of line with its closest BLAST match, Fionnbharth gene product (gp) 10, by 9 amino acid residues. Genemark called the start for gene 10 at 3932 and this start contained all of the coding potential in the ORF, had a better Shine-Dalgarno score of 441, and it aligned perfectly with closest BLAST match. As a result, the predicted start of gene 10 was moved to base pair 3932. For gene 15 Glimmer predicted the start codon at base pair 10319 with a strength of 8.60, but the start site did not include all of the coding potential in the ORF on the Genemark spreadsheet, it had a Shine-Dalgarno score of 420, and it caused the gene product sequence to be out of line with the closest BLAST match, Fionnbharth gp 15, by 26 residues. Genemark called the start site for gene 15 at 10244, which contained all of the coding potential in the ORF, had a better Shine-Dalgarno score at 609, and fixed the product alignment with the closest BLAST match. With this the start site was changed to 10244. For gene 37 the original Glimmer prediction placed the start codon at base pair 30629 with a low strength of 3.31. This start also caused the product sequence of the gene to be out of alignment with its closest BLAST match, Fionnbharth gp 37, by 3 residues. Genemark placed the start site at 30635, which corrected the alignment issue. With this the start site for gene 37 was moved to base pair 30635. For gene 43 the start codon site was changed from the original Glimmer prediction at base pair 33056 with a strength of 14.50 to the Genemark predicted site at 33290. This was done because the 33290 start contained all of the coding potential in that ORF as seen on the Genemark spreadsheet and the start at 33056 did not, the 33290 start had a higher Shine-Dalgarno score at 609 while the 33056 had a score of 399, and the 33290 caused the product sequence to go from being 78 residues out of alignment with its closest BLAST match, Fionnbharth gp 44, to being 1 residue out of alignment. For gene 47 the original Glimmer predicted the start codon at base pair 35990 but gave it a low strength of 3.98. In addition, the start at 35990 caused the amino acid product sequence of the gene to be out of line with the closest BLAST match,

Fionnbharth gp 47, by 16 residues. Genemark called the start site at 36035, which caused the gene product to have perfect alignment with the closest BLAST match so the start codon was moved to base pair 36035. Like gene 47, the start codon sites for genes 57, 75, 76, and 87 were all moved because the original Glimmer predicted start sites caused the gene amino acid product sequences to be misaligned with the closest BLAST match. For gene 57, the start was moved from base pair 38921 to 38897 because it fixed a misalignment of 9 residues with its closest match, Fionnbharth gp 57. For gene 75, the start site was moved from base pair 47963 with a relatively low strength of 6.64 to base pair 47987 because it fixed a misalignment of 9 residues with the closest match, Fionnbharth gp 75. For gene 76, the start site was moved from base pair 48373 with a low strength of 2.80 to 48349 to fix a misalignment of 9 base pairs with its closest match, Fionnbharth gp 76. Lastly, the start site for gene 87 was moved from base pair 53160 to 53154 in order to correct a misalignment of 3 residues with the closest match, Fionnbharth gp 89.

In addition, during the annotation process it was noticed that the Cheetobro phage possessed an incredible similarity to its fellow K4 phage Fionnbharth. As seen in the DNA Master BLAST results for Cheetobro, the closest match for every gene in Cheetobro's genome, except for gene 4, is a gene from Fionnbharth. Though Fionnbharth has a larger genome of 58076bp and has 94 predicted protein coding genes, the remarkable similarity between the two appears to not only suggest that they will be extremely similar in characterization and behavior during infection and replication but also appears to suggest that both of these phages are possibly evolutionarily linked. However, not many conclusions about the characterization of Cheetobro can be drawn from Fionnbharth since little work has been done to characterize that phage as well. On another interesting note, the same similarity between Cheetobro and Fionnbharth cannot be said for the other known K4 phage Slarp because none of the genes of Slarp appeared as close matches to any of the genes of Cheetobro. The reason for this could be that Cheetobro and Slarp are not as close together evolutionarily as Cheetobro and Fionnbharth. Instead, the two phages may have originated from the same ancestor phage but they have diverged evolutionarily over time. This could explain why they were placed in the same subcluster but appear to be so different. Overall, this BLAST data seems to suggest that whatever characterizations of Cheetobro that will be made in the future can also be very much used to characterize Fionnbharth.

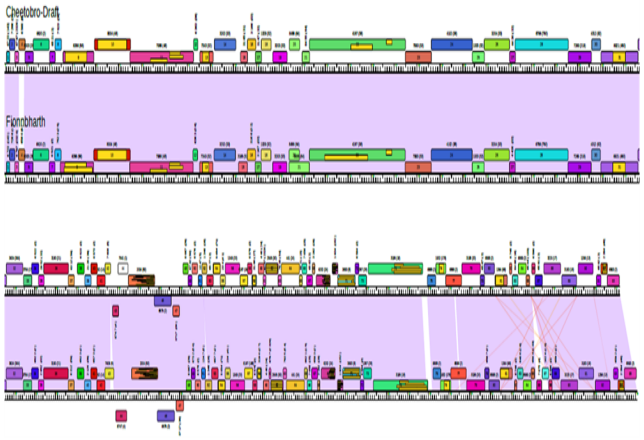

Figure 2: Above are Phamerator genome maps for the K4 phages Cheetobro and Fionnbharth. As seen in the image, Phamerator compares both genomes together and determines where the genomes are similar and where they are different. Areas of purple represent areas of conserved nucleotide sequences between the two and areas of white represent areas of no similarity.

Even though no putative functions or protein products of any of the Cheetobro genes have yet been explored or analyzed using Phamerator or protein BLAST searches using the NCBI protein database, the function of several genes have possibly already been preliminarily identified during the annotation process. The first gene to be possibly identified was the Tapemeasure gene. The Tapemeasure gene is the gene that codes for the long, flexible, non-contractile tail and it tends to be the longest gene in the genome of the phage. In the Cheetobro genome, the best candidate for the Tapemeasure gene was determined to be gene 22 because it is the longest gene in the genome with a length of 4173bp and the closest BLAST match for that gene is Fionnbharth gene 22, which is the identified Tapemeasure gene for that genome with a length of 4176bp. If gene 22 is indeed the Tapemeasure gene of the Cheetobro genome, then according to Hatfull and colleagues the Cheetobro Tapemeasure gene along with the Fionnbharth Tapemeasure gene are over 300bp longer than the K1 phage Tapemeasure genes and are some of the longest Tapemeasure genes among the Cluster K phages (Pope *et al.*, p. 9, 2011). The reason for this longer Tapemeasure gene and tail is unknown but it could play a role in facilitating penetration or infection into the bacterial host. Further research and experimentation on the matter will be required.

The other gene that has been possibly preliminarily identified is the gene coding for integrase. Integrase is an enzyme that is often encoded by phages because it serves to facilitate the recombination and integration of the phage genome into the bacterial host genome so that it may enter a dormant or lysogenic state in the form of what is known as a prophage. According to Hatfull and colleaques the integrase found in Cluster K genomes is of the tyrosine recombinase family. This integrase

gene along with the attP site, which is the region of the phage genome that initially binds to the host genome for recombination, make up the integration cassette with the attP site oriented 5'to the integrase gene. According to the Hatfull paper, in Cluster K phages these integration cassettes are located toward the center of the genome and are flanked by three genes of unknown function transcribed in the reverse or leftward direction (Pope *et al.*, p. 10, 2011). In Cheetobro only genes 43, 46, and 47 are transcribed in the reverse direction and they flank gene 45 in the same manner described with the integrase gene. With this information, it is highly likely that gene 45 is the integrase gene and that the region 5' to gene 45 between base pairs 32766 and 33528, which was previously the incorrect gene 44, could very possibly be the attP site.

Another interesting aspect of the Cheetobro genome was that Glimmer and Genemark both identified a region between genes 37 and 38 from base pairs 30852 to 30927 as tRNA. In fact it was identified as tRNA specific for the amino acid residue Lysine. To confirm that this sequence was indeed tRNA and to possibly locate other tRNA sequences, the Cheetobro genome was processed through the tRNA search tool tRNAscan-SE, and it was found that that sequence was correctly identified as Lysine tRNA and was the only tRNA within the genome. The search tool also generated an image of the predicted structure of the tRNA molecule, which can be seen in Figure 1 below. It was also found that Fionnbharth genome also possessed a Lysine tRNA roughly in the same location as the Cheetobro genome. However, according to the Hatfull paper most of the other Cluster K phage genomes like K1 phages Adephagia, Anaya, and Angelica possess tRNA very near to the left end of the genome around genes 6 and 5 and also that these phages usually possess Tryptophan tRNA (Pope *et al.*, 2011, p. 9). In this respect Cheetobro and Fionnbharth appear to be strikingly different from the rest of the Cluster K phages but why that is or to what difference would it make for both phages to have a different tRNA in a different location from the other Cluster K phages is difficult to say since little research can be found with regards to tRNA within phage genomes. It is likely that the tRNA is found in the genome because the amino acid associated with it possibly plays an important role in the structure of the phage and it requires a large quantity of that amino acid. With this, it can possibly be said that the K1 phages require a large quantity of Tryptophan tRNA in the viral structure while Cheetobro and Fionnbharth require a large quantity of Lysine, but more experimentation and research must be done into the purpose of these tRNAs in order to confirm or deny this speculation.

REFERENCES

The Mycobacteriophage Database. (n.d.). Retrieved from http://phagesdb.org/

Pope, W. H., Ferreira, C. M., Jacobs-Sera, D., Benjamin, R. C., Davis, A. J., DeJong, R. J., … Hatfull, G. F. (2011). Cluster K mycobacteriophages: Insights into the evolutionary origins of mycobacteriophage TM4. *PLoS ONE* 6: 1-22.

Rondón, L., Piuri, M., Jacobs Jr., W. R., de Waard, J., Hatfull, G. F., & Takiff, H. E. (2011). Evaluation of fluoromycobacteriophages for detecting drug resistance in *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* 49:1838-1842.

Rybmiker, J., Kramme, S., and Small, P. L. (2006). Host range of 14 mycobacteriophages in *Mycobacterium ulcerans* and seven other mycobacteria including *Mycobacterium tuberculosis* – application for identification and susceptibility testing. *J. Med. Microbiol.* 55: 37-42.

Xie, P., Wei, F. Y., Hirata, S., Kaitsuka, T., Suzuki, T., and Tomizawa, K. (2013). Quantitative PCR measurement of tRNA 2-methylthio modification for assessing type 2 diabetes risk. *Clin.Chem.* 59:1604-12.

Zaborske, J. M., Narasimhan, J., Jiang, L., Wek, S. A., Dittmar, K. A., Freimoser, F., … Wek, R. C. (2009). Protein Synthesis, Post-Translational Modification, and Degradation: Genome-wide Analysis of tRNA Charging and Activation of the eIF2 Kinase Gen2p. *J. Biol. Chem.* 284: 25254-25267.