Isolation and genome annotation of mycobacteriophage Thespis

David Z. Bushhouse '19, Eric K. Bowen '19, and Michael J. Wolyniak

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

INTRODUCTION

The HHMI Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) program is a discovery-based initiative aimed at giving undergraduate students an active research experience that is both engaging and financially feasible (Seaphages.org).

More than 600 of those phages have been isolated and sequenced as part of SEA-PHAGES classes across the country, and added to the Phamerator database, allowing comprehensive cross-genome analysis which has revealed clusters and subclusters of sequentially similar phages, and also indications that these clusters are actually points along a continuum of genetic diversity (Pope et al., 2015). In order to form more conclusions about the genetic characteristics of the mycobacteriophage population and its evolutionary history, more phages need be isolated and sequenced.

To date, Hampden-Sydney College, as a participant in the SEA-PHAGES program, has collectively isolated and contributed 21 phages to The Actinobacteriophage Database at phagesdb.org. Additionally, Genomics and Bioinformatics courses at Hampden-Sydney often incorporate the analysis of real phage genomes into group projects. This paper is about one of those phages, Thespis.

DNA sequencing is a powerful tool that allows researchers to see how exactly a genome is arranged. Viruses have very short sequences of DNA which allow for them to be sequenced very easily. However, elucidating the function of each gene would be a tedious process in the lab. Bioinformatics allows for the use of multiple databases and software programs to make predictions in order to understand the function of each gene in a specific genome. By comparing recently sequenced data to similar sequences with a known function, the function of the newly sequenced gene can be derived.

Annotation is the process of taking the predictions made by multiple programs and consolidating them into one main prediction that will be used in further study. Each program uses a different algorithm and will usually provide different answers to gene product length, start site location, and gene function. Because there is uncertainty as to which program provides the correct answer, it is up to the researcher to deduce the answers given the information provided from the programs as well as use their reasoning to determine which gene calls are correct.

The bacteriophage that was to be annotated, Mycobacteriophage Thespis, was found at the base of an American Beech tree and isolated by plaque assay and selective propagation during the summer of 2016. The virus was sequenced shortly after. Due to strong correlation, Thespis was initially believed to be similar to that of Mycobacterium phage Brusacoram. Software program gene-call predictions would determine how similar Brusacoram and Thespis truly are.

Principal Approaches

Isolation and Purification.

Soil samples were taken in three locations on the Hampden-Sydney Campus. Samples were collected into sterile 50mL conical tubes while GPS coordinates, air temperature, and soil composition were noted. The three sampling sites were (1) a streambed along the Wilson Trail, (2) the base of a tree in front of Cushing Hall, and (3) the base of a beech tree in front of Gilmer Hall (Table 1).

For each sample, approximately 1g of sample soil was mixed with enrichment medium and incubated at 37°C with shaking for 24h. Enriched samples were pelleted, and supernatant was removed from each sample and filter-sterilized using a luer-lock syringe and 0.22µm filter membrane attachment. The sample filtrates were stored at 4°C.

Frozen stocks of *M. smegmatis* were propagated on Luria agar plates (incubating at 37°C), and grown in a Middlebrook 7H9 complete medium incubated at 37°C with shaking for 48h. This procedure was repeated weekly to maintain liquid bacterial cultures no older than 96h.

The sample filtrates were serially diluted by a factor of 10 per dilution using phage buffer[3] (PB). Liquid *M. smegmatis* culture was transferred to the culture tubes containing the PB negative control and the 10^{-2}, 10^{-3}, and 10^{-4} dilutions of each sample filtrate for a 25-minute infection period.

Top Agar (TA) was freshly prepared before each plating and kept in a 55°C water bath to prevent solidification. Liquid TA was transferred to each culture tube after the 25-minute inoculation period, upon which the entire bacteria-filtrate-TA mixture was transferred to the surface of a Luria agar plate treated with CB and CHX. Any bubbles were popped using a sterile inoculating needle. After the TA solidified, plates were inverted and incubated at 37°C for 48h.

Plates were examined for the presence of plaques. Plaques were labeled with a permanent marker and carried on to the purification procedure. The PB negative control plate was also examined for the presence of plaques.

Every plating during the purification process was controlled positively and negatively using high titer lysates of mycobacteriophages Brusacoram and QuinnKiro, supplied by The College of St. Scholastica, and PB, respectively.

Suspected plaques from plaque screening were "picked": a pipet tip was depressed into the center of the plaque to the base of the TA layer. The end of the pipet tip was then submerged in PB in a microcentrifuge tube and gently tapped on the wall of the tube to dislodge any harvested phage. The samples were vortexed to suspend any potential phage.

Plates were marked into quadrants and labeled (One PB neg. control quadrant and three test quadrants). The entire plate surface was covered with a mixture of uninfected liquid culture and TA. After the TA solidified, putative phage suspensions were applied onto the centers of the respective test quadrants. After the samples had been absorbed by the agar, plates were inverted and incubated at 37°C for 48h. The presence of a spot in a test quadrant indicated that the picked putative plaque contained active plaque forming units (pfu's)—phage. The lack of a spot in a test quadrant indicated that the picked putative plaque was actually an air bubble or other agar-surface anomaly.

Web-pattern plates with a high concentration of phage were flooded with PB. After 2h at room temperature, the PB was aspirated with luer-lock syringes and filter-sterilized using 0.22μm filter membrane attachments. The collected lysates were stored at 4°C.

In order to determine the concentration of pfu's, or titer, in the harvested lysate, a wide range of dilutions were plated so as to gain the resolution necessary to manually count plaques. The 25-minute infection time and volume of TA used were the same as in the plaque screening stage. Plaques were counted manually to determine a titer of $2.26 * 10^{10} \frac{pfu}{mL}$.

The Wizard® DNA Clean-Up System was used to prep and extract DNA from 1mL of lysate. The DNA extraction was carried out twice. ethanol precipitation was used to concentrate the DNA, which was resuspended in ddH_20 using a 60°C heating block in 10 minute intervals. The sample was stored at -20°C until it was shipped to the Daniel Russell Lab at the University of Pittsburgh for sequencing. Sequencing results were downloaded from the Actinobacteriophage Database.

Annotation.

The annotation took place in the computer program DNA Master 5, which laid out a preliminary prediction for genes using the Glimmer gene prediction program. Glimmer uses an algorithm that determines which start site in the sequence is the correct one for each gene and displays the sequence until it hits a stop codon. DNA Master also has the ability to perform a Basic Local Alignment Search Tool (BLAST) search which compares a predicted gene sequence to other sequences. By matching a sequence of unknown function to similar sequences of known function, the function of the unknown sequence can be predicted.

For the annotation of each gene, the length and start site of the gene first had to be determined. This was done by comparing the Glimmer length to the approximations given by a second gene prediction program called GeneMark. GeneMark lays out predictions for gene length on all six reading frames with the frame that has a viable gene peaking above a threshold determined by the program's algorithm. The two programs are compared to ensure that there is some probable gene located at that point on the genome and that it was not a false positive. Figure 1 below shows an example of two reading frames within the GeneMark file in which the hash marks pointing up are potential start sites and the hash marks pointing down are the stop codons. By getting an estimate of where a gene is, the start site can be further pinpointed.

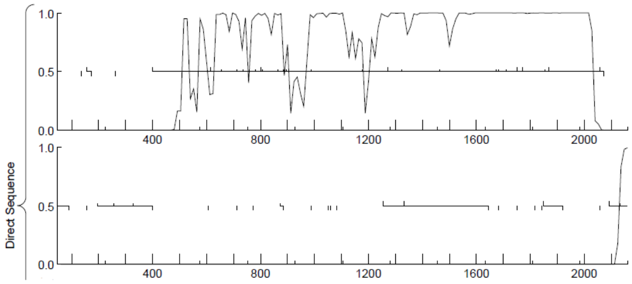

Figure 1: *Image from GeneMark file that shows an area where a gene is likely to be coded for. The second reading frame shows no real potential for coding.*

In DNA Master, the open reading frames (ORFs) readout is similar to GeneMark in that it shows where the Glimmer believes genes to be. An individual gene can be selected and the different viable start sites can be compared using a Shine-Dalgarno score. The Shine-Dalgarno (SD) score is based off the upstream sequence of the same name which signals ribosomal binding to an mRNA sequence. The highest SD score was usually the best starting site, but if a really large gap or overlap would result, usually the next highest scores are used because they still indicate strong potential for a start site to be the correct one. Figure 2 shows an example of the SD score window and the scores closest to zero are the best candidates for starting sites.

After determining the length and start sites of the gene, the next step was predicting putative function by running a BLAST search of the sequence. BLASTing the genes compares the sequence to other sequences with known function as well as gives a side-by-side comparison of the two genes. The correlation between the predicted sequence and the known sequence was given a score derived from an algorithm. The more similar the two genes are in terms of amino acids increases the likelihood of the specific gene having a similar function.

So overall, the process was to look for where potential genes could be using Glimmer and GeneMark. This was followed by determining the start site of the genes by finding the highest SD score and then running a BLAST comparison on the genes to predict putative functions.

Present Knowledge

Sequencing concluded that Thespis was a P1 subcluster mycobacteriophage with a 47,618 bp genome consisting of 78 genes. After analysis of the 78 original genes, gene 56 was deleted, and five genes had their start sites shifted: genes 25, 30, 32, 54, and 57. Because of the densely-packed nature of the Thespis genome, the shifts were usually under 75 base pairs. Looking at the BLAST comparison, there were many genes (59 out of 77) that were almost, if not completely, identical to the bacteriophage Brusacoram, but there were other gene products that were more similar to other bacteriophages of the same cluster. There were also instances of higher SD score start sites being turned down as the actual starting site due to large gaps or overlaps that would have resulted from such a shift. Overall, not many changes were made based off of the original predictions that were produced by Gilmer.

DISCUSSION

During the purification procedure, several experimental procedures gave cause for modification. For example, TA, which must be stored in a 55°C water bath to prevent solidification, often developed free-floating chunks of partially solidified agar, even when stored properly. These chunks were nearly indistinguishable from the rest of the liquid, but produced a bumpy surface with gaps in the bacterial lawn that were confused for genuine plaques. Upon discovering this, TA was no longer prepared in advanced and stored, but prepared immediately before use.

Also, there was considerable concern, after the subcluster placement was made by the Daniel Russel Laboratory, that cross-contamination had played a role in the discovery of Thespis. P1 is a fairly rare subcluster, containing only 20 other members at the publication of this article, and since a P1 subcluster phage, Brusacoram, was used as a positive control along with A3 subcluster phage QuinnKiro, there was concern that Thespis was a copy of Brusacoram.

However, after annotation, we are confident that Thespis cannot be a copy of Brusacoram, despite their many similarities. The extreme relation is most likely because the two phages are cousins of one another, though they have enough differences in gene coding that they cannot be the same species. How these differences are seen physically would require further testing between the individual genes.

CONCLUSION

Further testing would include EM imaging of the Thespis phage particle. RNA-seq transcriptome analysis of infected host cells would confirm gene-calls and illuminate gene expression patterns. We are happy to report the continuing contribution of Hampden-Sydney College to the SEA-PHAGES program.

REFERENCES

1)The Actinobacteriophage Database. (2016). *Mycobacterium phage Thespis*. Retrieved from: http://phagesdb.org/phages/Thespis/
2) Jacobs-Sera, D., Bowman, C. A., Pope, W. H., Russell, D. A., Cresawn, S. G., & Hatfull, G. F. (2014). *Annotation and Bioinformatic Analysis of Bacteriophage Genomes: A User Guide to DNA Master*.
3) Jordan, Tuajuanda C. et al. (2014) A broadly implementable research course in phage discovery

and genomics for first-year undergraduate students. *mBio*, *5*(1), e01051-13.
4) Klug, W. S., Cummings, M. R., Spencer, C. A., & Palladino, M. A. (2015). *Concepts of Genetics*. Boston: Pearson.
5) Pope, Welkin H. et al. (2015). Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *eLife, 2015*; 4:e06416.
6) Poxleitner, Marianne et al. (2016). *Phage Discovery Guide*. Chevy Chase, MD: Howard Hughes Medical Institute.
7) SEA-PHAGES. The SEA-PHAGES Program. Retrieved from seaphages.org.