Bioinformatic Characterization of Y Cluster Mycobacteriophage Bipper

Gatt,S.M., Osking,Z.B., Barcellona,C.M., Dozier,K.D., Faust,J.M., Fedrick,A.J., Gagliardi,L.E., Gleason,P.S., Gomez,E.A., Ho,Q.A., Hoffman,A.M., Jenkins,M., Jones,M.J., Lang,J.F., Lequay,S.M., Mars,P.J., Mtchedlidze,N., Paul,L.M., Pica,A.N., Robison,M.D., Rodriguez,D., Rosales,K.A., Saravis,L.E., Sisson,B.M., Tan,A.L., Voltaire,R., Warner,M.H., Bradley,K.W., Asai,D.J., Michael, S. Isern,S.

Florida Gulf Coast University, Fort Myers, FL 33965 USA

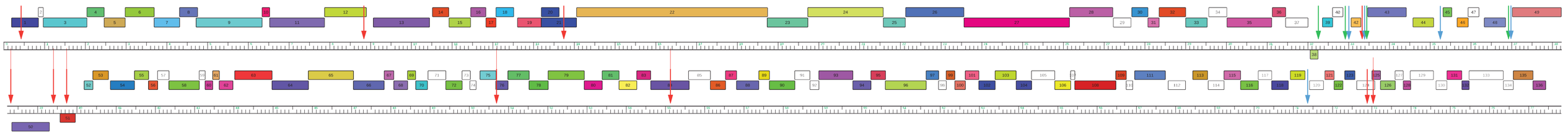

FIGURE 1. Promoter, Terminator, and Repeat Locations. The predicted locations of Sigma-70 promoters and rho-independent terminators in relation to the ORFs found in Bipper's genome are depicted. Promoters are indicated by green arrows, while terminators are red. The most common repeat found in Bipper's genome by MEME is labeled in blue. Genes with white backgrounds indicate orphams.

Abstract

Our Virology class at Florida Gulf Coast University adopted mycobacterium phage Bipper from the University of Pittsburgh in the fall of 2015. Bipper was considered a singleton when we began the annotation process. We found that Bipper had 135 open reading frames (ORFs) and one methionine (cat) tRNA. Fifty seven percent of its phams were orphams. Synteny was followed with the structural genes. There was a striking abundance of long stretches of overlapping ORFs. As many as 8 consecutive ORFs overlapped by exactly 4 bp. A stretch of 36 ORFs [gp75 to gp110] encompassing 16634 bp only had a single small gap (2 bp). About 6% of ORFs contained transmembrane domains. The smallest protein had no known function and was 108 bp long, while the tape measure protein was the longest at 4,686 bp. The start site preferred by Bipper was ATG. This site was used 55% of the time while GTG was used 44%. Only one ORF used TTG as the start site. The average GC% content was 67.3% and was almost identical to that of Mycobacterium smegmatis, 67.4%, suggesting the two have co-evolved for a significant period of time. We found several inter- esting repeat elements using MEME. One motif was particularly long (50 bp) and had 5 occurrences in locations correlating to large gaps in the genome. Seven putative Sigma-70 promoters were found throughout its genome, two of which were directly upstream of an operon; while six rho-independent tran- scriptional terminators were predicted within open reading frames and six more were in noncoding regions. Bipper's attachment site for integration was located in its immunity repressor gene. The 35 bp sequence corresponded to its host's attachment site at the 3' end of M. smegmatis Arg tRNA, as expected for phage with tyrosine integrases. Overall, we found that Bipper displayed many of the common characteristics shared by bacteriophage that infect M. smegmatis. Bipper was published in GenBank on 22 March 2016. Its accession number is KU728633.1.

Introduction

Bacteriophage can be found nearly anywhere on earth, leading to exceptional genetic diversity and many opportunities for discovery of new phage. The genetic diversity of phage makes the characterization of their genomes difficult and has already led to the creation of many different clusters and sub clusters (68 identified clusters for Actinobateriophage alone). While the diversi- ty of phage genomes makes their characterization difficult, it also makes it a great exercise in comparative genomics. Through the use of accessible Bioin- formatic tools phage genome annotation can be conducted quickly and reli- ably, allowing for the growth of our knowledge base on what seems to be an impossibly diverse group.

Taking raw sequencing data and transforming it into a final annotation requires a suite of Bioinformatic software. NCBI Blastp, which is used to compare the primary sequence of an unknown gene with a database of known pro- teins. "Blasting" unknown viral genes allows for a function call to be made for them without the use of in vitro resources. With every annotated genome sub- mitted the breadth of the NCBI database grows and creates the possibility for more accurate function calls to be made in the future. This concept applies to many of the tools used in the annotation process, making all of the work done as part of the SEA-PHAGES initiative significant in a greater context. Annotat- ing a greater number of phage genomes will also eventually enable us to draw conclusions about bacteriophages as a whole. Having more data on phages could bring about a deeper understanding about the phage-host relationship, diversity, origins, and the answers to any number of unknown questions.

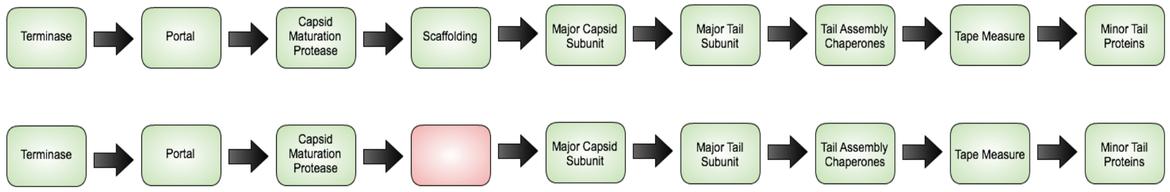

FIGURE 2. Synteny. Bipper appears to follow bacteriophage synteny for virion structure and assembly genes. Function calls could be made for all but scaffolding protein, which was not found anywhere in Bipper's genome. All of these proteins appear in the first 27,216 bp of the genome except for one minor tail protein that is found 73,223 bp in.

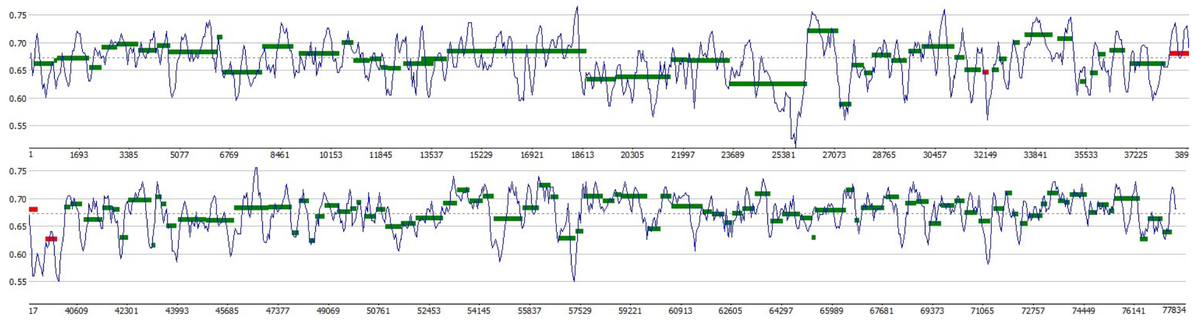

FIGURE 3. GC Content. Bipper had an overall GC content of 67.3% throughout its genome. This is almost identical to that of its host Mycobacterium Smegmatis mc^122_155 which has a GC% content of 67.4%, suggesting a significant period of co-evolution between phage and host. The figure above shows the distribution of GC% content throughout the genome with ORFs marked in green (forward coding) and red (reverse coding). The ORFs are distributed throughout areas of high and low GC content so there does not appear to be a relationship between GC% content and coding potential.

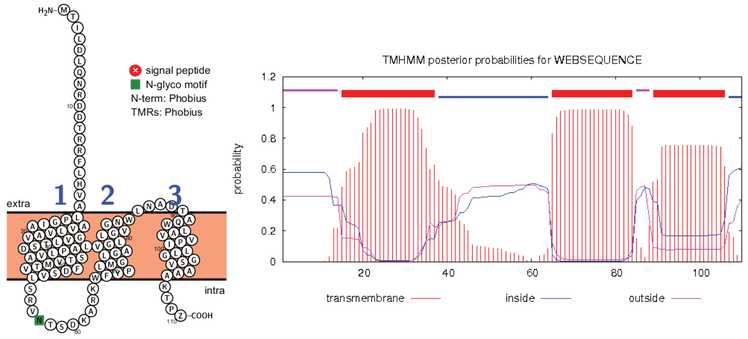

FIGURE 4. Transmembrane Domains. There were eight proteins found that contained transmembrane domains. Most of these had NKF with the exception of gp36 (holin) and gp 50(integrase). holin was predicted to contain three distinct transmembrane domains. The figure to the left illustrates the transmembrane domains for this gene, generated by Protter. To the right, the three domains are shown by TMHMM as they were predicted according to the probability of the amino acid sequence being inside or outside of a cell as well as throughout its membrane. While holin had the most TM domains, the gene immediately downstream of it, as well as gp53, had two do- mains.

	ATG	GTG	TTG
Structural	5.88%	2.21%	0%
Non-Structural	13.24%	13.24%	0%
NKF	33.82%	30.15%	1.47%
Total	52.94%	45.59%	1.47%

FIGURE 5. Start Site Preference. This table reflects the start site preferences for Bipper's genome. ATG was the overall preferred start site as well as the most used for genes identified as structural or no known function. Genes identified as non-structural were equally distributed between ATG and GTG. The only TTG start site used is in ORF 81 and no function could be attributed to this gene at the time of annotation.

Methods

- DNA Master was used to initially predict open reading frames throughout Bipper's genome.
- NCBI BLASTp, HHPred, and PhagesDB BLASTp were used to predict the function of each gene.
- A combination of NCBI, HHPred, PhagesDB, and GeneMark's graphical output of coding potential was used to predict the starting site for each gene.
- TMHMM was used to predict the number of transmembrane domains in each gene product.
- Aragorn helped in the annotation of the single tRNA.
- DNA Master was also used to visualize the changes in GC content and predict sigma-70 promoters.
- NCBI was used to find integration sites by BLASTing the entirety of the M. smegmatis and Bipper genomes against each other.
- Arnold was used to predict the number and location of rho-independent terminators.
- MEME was used to find repeat motifs.

Conclusion

- A total of 135 ORFs were annotated and one met (cat) tRNA was found.
- Bipper follows bacteriophage structural gene synteny with the exception of scaffolding protein, which could not be found. All of these genes are found in the first 35% of the genome except for a single minor tail protein at gp126.
- GC% content is very similar to M. smegmatis. This suggests that Bipper has been infecting and coevolving with M. smegmatis for a significant period of time.
- ATG is the most frequently used start codon with about a 53% frequency. GTG was the second most frequently used at roughly 46%. TTG was used only twice (1.47%) and no function was able to be assigned to either of these genes. There was not any clear correlation between the start site preference of a gene and its function.
- A total of eight proteins containing transmembrane domains were predicted. The only two with a called function are ORF's 36 (holin) and 50 (integrase). The other six proteins with predicted transmembrane domains are distributed throughout the genome.
- The program MEME found a 50 bp motif that is found mostly in large gaps in the genome. This repeat could be significant because the majority of Bi- pper's genome consists of overlapping genes. There are only 11 gaps in the genome greater than 150 bp, and 4 of these contain the motif found by MEME.
- The integration sites were found to be within Bipper's immunity repressor gene and M. Smegmatis's Arg tRNA. This is to be expected for phage with tyrosine integrases such as Bipper.
- 2 of the 7 predicted promoter were located directly upstream of an operon; whereas 6 of 12 predicted terminators were found in the gaps between genes.

Acknowledgements

Thank you to Dr. Sharon Isern and Dr. Scott Michael for all of their support and guidance. We would also like to thank Dr. Welkin Pope for letting us adopt Bipper and the SEA-PHAGES program.