

# A genetic investigation on the function of gene 12:

## Faith Holt, Department of Biological Science College of Coastal Georgia

### Abstract:

The intricate world of viral genomes continues to yield fascinating insights into the fundamental processes of life. Understanding the function of each gene within a virus is paramount for deciphering its replication strategies, host interactions, and evolutionary trajectory. This study focuses on Gene 12 in the novel phage, LilTerminator. The probable coding potential of this region further implies that Gene 12 may encode a small protein or enzyme essential for viral activities. Notably, the genomic context reveals that Gene 12 overlaps with neighboring genes, hinting at a potential collaborative role in the assembly of viral tail structures. Furthermore, bioinformatics analyses, including BLAST and HHPred, have proposed that Gene 12 encodes minor tail and capsid proteins. Therefore, treating Gene 12 as NKF underscores the importance of experimental investigation to ascertain its true biological role. This report aims to analyze the existing information regarding Gene 12, explore the implications of its unknown function within the context of viral genomics, gene overlap, bioinformatics predictions, regulatory complexities, and potential therapeutic applications.

### Methods:

1. SSC (Stop/Start Coordinates): This tool identifies the exact positions of start and stop codons in a nucleotide sequence, which is essential for delineating protein-coding regions within a gene.
2. CP (Coding Potential): CP evaluates whether a DNA sequence can encode a protein by analyzing it for open reading frames (ORFs) and assessing their likelihood of translation into functional proteins.
3. SCS (Start Choice Source): SCS focuses on the sequence elements that dictate the selection of start codons during translation. It examines regulatory motifs and upstream sequences that guide ribosomes to the appropriate start site.
4. ST (Statorator): This computational tool models and visualizes biological states over time, aiding researchers in understanding how various factors influence cellular processes.
5. Blast-Start: This initial phase of the BLAST algorithm allows users to input a query sequence to search databases for similar sequences, providing alignment scores and percent identity to infer potential gene functions.
6. Gap: In sequence alignment, gaps represent missing nucleotides or amino acids. They are introduced to enhance alignment accuracy by accommodating insertions or deletions, with penalties applied to minimize excessive gaps.
7. LO (Lowest ORF): This tool identifies the smallest open reading frame (ORF) within a sequence that can be translated into a protein. It is crucial for pinpointing the minimal coding regions of a gene.
8. RBS (Ribosome Binding Site): The RBS is a specific sequence in mRNA that the ribosome recognizes to initiate translation. Its efficiency significantly impacts protein production levels.
9. F (Function): This term refers to the biological role or activity of a gene or protein. Understanding a gene's function is vital for elucidating its contribution to cellular processes and overall organism behavior.
10. SIF Blast (Sequence Information Files Blast): SIF Blast consists of output files generated from BLAST searches that provide detailed information about sequence alignments, similarity scores, and annotations of matched sequences.
11. Sif HHPred (Hidden Markov Model-based Protein Structure Prediction): This tool predicts protein structures and functions using Hidden Markov Models. It compares query sequences with known protein structures to infer potential functions.
12. Sif Syn (Synteny Information Files): Sif Syn provides insights into synteny, which refers to the conservation of gene order on chromosomes across different species. This information aids researchers in understanding evolutionary relationships and the functional conservation of genes.
13. Sif Mem (SIF Membrane Protein Prediction): Sif Mem encompasses tools designed to predict membrane proteins from sequence data. These proteins are crucial for understanding cellular signaling and transport mechanisms, playing significant roles in various biological processes.

### Acknowledgements:

The Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science. (2017).

Bioinformatics Guide. Seaphagesbioinformatics.com

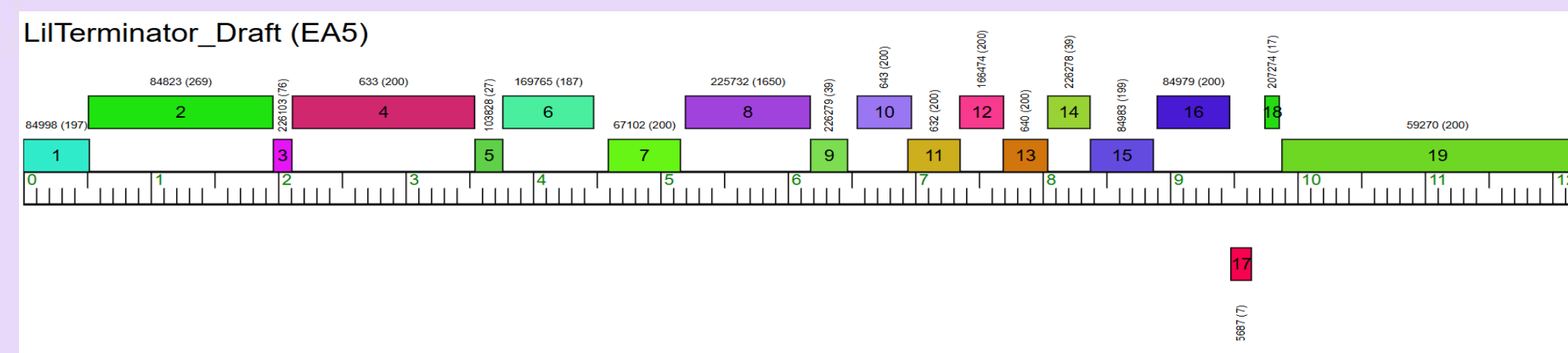
<https://seaphagesbioinformatics.helpdocsonline.com/home>

### Introduction:

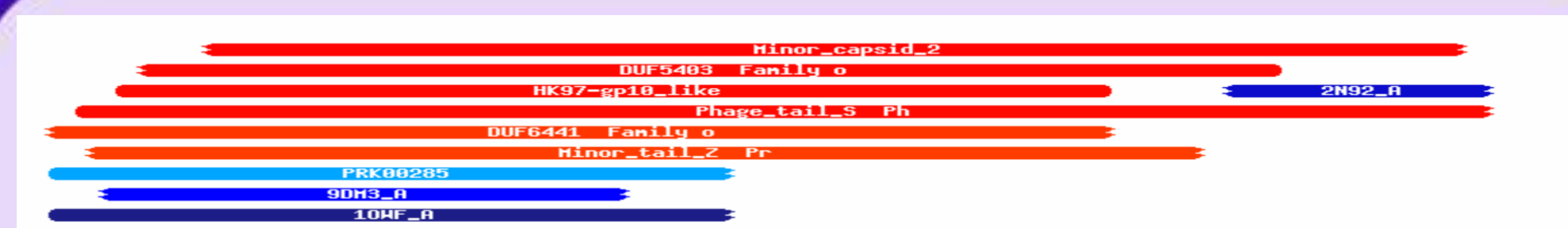
Bacteriophages, or phages, are viruses that specifically infect bacteria and are among the most abundant organisms on Earth. Since their discovery over a century ago, they have significantly impacted microbiology, virology, and biotechnology. The Phage Program at the University of Pittsburgh Department of Medicine seeks to harness these unique viruses for applications such as phage therapy, agricultural pest control, and environmental management, especially as antibiotic-resistant bacteria become more prevalent. Phages offer a promising alternative treatment by targeting harmful bacteria without affecting beneficial ones, thereby reducing side effects and minimizing the risk of resistance. This program focuses on isolating, characterizing, and applying phages to clinically relevant bacterial strains to develop innovative treatment options. My project specifically aims to elucidate the function of Gene 12 in the phage LilTerminator (Figure 1), and justify the conclusions drawn from this analysis.



[Figure 1. Electron Micrograph of a Lil Terminator Bacteriophage Virion.]



[Figure 2. This linear map illustrates the arrangement of predicted genes within the genome of bacteriophage LilTerminator\_Draft (EA5) from Phamerator. Each colored box represents a distinct gene, numbered sequentially from 1 to 19. The length of each box is proportional to the size of the corresponding gene in base pairs, with the approximate size indicated in parentheses above each gene. The horizontal scale at the bottom provides a reference for the genomic position.]



[Figure 3. HHPred output showing probable matches to the crystalline structure of Gene 12 in LilTerminator.]

Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
1	PF11114.13	; Minor_capsid_2 ; Minor capsid protein	99.6	1.8e-14	82.84	7.7	89	105
2	PF17395.7	; DUF5403 ; Family of unknown function (DUF5403)	99.34	1e-11	70.05	6	82	92

[Figure 4. HHPred Results. This table presents the significant alignment results of gene 12 against the Pfam database]

### Results and Discussion:

Figure 2 shows the distance between Gene 12 and the known tape measure protein, Gene 19. The tape measure protein is central to the tail assembly pathway, specifically length determination. In phage, it is typical for several minor tail proteins to be found just downstream of the large tape measure protein. This is the case in LilTerminator, as we identified 2 minor tail proteins downstream of gene 19. While there is one hit to a phage tail protein in the HHPred output for Gene 12 (Figure 3), synteny does not support the placement of a minor tail protein at this location in LilTerminator. The second hit in the HHPred output (Figures 3 and 4) are for DUF. "DUF" stands for Domain of Unknown Function. This indicates that this gene belongs to a protein family where the precise function has not yet been experimentally determined. The most significant hit (based on E-value and probability) for the sequence is PF11114.13, which is identified as a Minor capsid protein (Figures 3 and 4). The very low E-value ( $1.8 \times 10^{-14}$ ) and high probability (99.6%) strongly suggest that gene 12 likely belongs to this family of minor capsid proteins. However, results from PhagesDB BLAST and Phamerator do not support this function call.

In Figure 4, The Sequence analysis strongly indicates that gene 12 belongs to the Pfam family PF17395.7 (DUF5403; Family of unknown function), exhibiting a high probability (99.34%) and a significant E-value ( $1 \times 10^{-11}$ ). Given that proteins in the DUF5403 family lack any experimentally determined functional characterization, the primary functional annotation of the gene remains uncertain. Therefore, classifying gene 12 as having "No Known Function" is a justifiable conclusion based on this analysis.

### Conclusion:

Based on the sequence analysis, a significant alignment was observed with the Pfam family PF17395.7, characterized as "DUF5403; Family of unknown function," exhibiting a high probability of 99.34% and a notable E-value of  $1 \times 10^{-11}$ . The very definition of a "Domain of Unknown Function" signifies a current lack of established biochemical or structural roles within biological systems. Consequently, this strong bioinformatic association directly supports the classification of gene 12 as having no definitively known function. While other less significant alignments might exist, the strong match to a protein family entirely lacking functional characterization provides a clear rationale for maintaining "NKF". Until more experimental evidence emerges, the "No Known Function" annotation accurately reflects the current state of scientific understanding for this protein.