Going to the Gap for more than genes

Authors: Ethan A Thibault, Maxwell T Dorman, Sarai L Smith, Jackson R Foley, Ashleigh Gibula, Corey A Halliday, Taaniel Kiidli, Sarah S Knowles, Hannah Melotto, Ashley A Moineau, Savannah E Sojka, Ashley N Soucy, Sally D Molloy*, Keith W Hutchison*
The Honors College, and the *Department of Molecular and Biomedical Sciences, University of Maine, Orono, ME

Bacteriophage (phage) are the most numerous and diverse biological entities on Earth. This diversity provides a reservoir of information about not only gene and genome structure and function but about mechanisms of evolution. This year we explored the genomes of two mycobacteriophage, Phaja and Pippin. Phaja is a Cluster E phage with a genome size of 75685bp. Pippin is a Cluster A1 phage with genome 52034bp in length. To explore these two genomes we focused on the gaps. One gap visible on Phamerator is created by the tRNA genes. Phaja carries two, both encoding Gly tRNAs. They do not reflect alternative codon use when comparing Phaja to *Mycobacterium smegmatis.* Rather they are codons used at a higher frequency by Phaja suggesting a need for a larger tRNA pool. The mismatch of codon frequency of use with the host, suggests that M. smegmatis may not be Phaja's natural host. Another gap we explored was one created by an approximately 500bp insertion immediately upstream of the tapemeasure protein gene in Phaja. The insertion contains two overlapping reading frames. BLAST analysis shows that both reading frames have the potential to encode an endonuclease. The endonuclease has been annotated in other phage. However, the structure suggests the possibility of a reading frame shift. It is intriguing that that the structure occurs immediately downstream of the tail chaperone ORFs with their reading frame shift. Phaja also contains an endonuclease/recombinase that it shares with a small subset of the Cluster E phage, downstream of gp94. Most of the E cluster phage have a methylase at this position. The endonuclease is a candidate for horizontal gene transfer. Gaps are usually mined for promoter sequences and in Pippin we explored the promoter region of the putative repressor protein. Pippin forms lysogens at a very low frequency and the plaque morphology is essentially clear. The peptide is 100% conserved with the putative repressor of BxB1 as is the sequence in the gap. The area has a weak promoter based on sequence and the same sequence is found in other cluster A1 phage suggesting there is another reason for the clear plaques.

Phaja originated in a host other than *Mycobacterium* and needs the tRNA for GGA in order to replicate in *Mycobacterium*.

Table 1. Comparison of the frequency of Glycine codons per 1000 codons in all the open reading frames of Phaja and various *Mycobacterium* species. Similarities within a 3.5 range to Phaja's codon use are highlighted in yellow and the two codons recognized by Phaja's tRNAs are highlighted in green. Codon bias alone does not account for the need for the phage tRNAs.

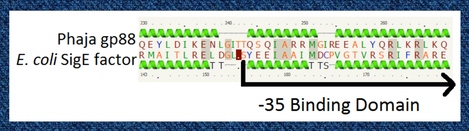
	Phaja	*M. smegmatis*	*M. tuberculosis*	*M. abscessus*	*M. bovis*	*M. laprae*	*M. africanum*
Codon	codon/1000	codon/1000	codon/1000	codon/1000	codon/1000	codon/1000	codon/1000
GGG	16.33	14.8	19.1	17.02	19.09	15.1	18.78
GGA	18.67	8.6	10	12.29	10.08	11.62	9.89
GGT	18.63	16.7	18.8	18.21	18.08	23.35	17.94
GGC	32.31	47.7	50.4	41.63	54.95	34.25	53.26

- Is the need for the two glycine tRNAs because the phage have more glycines?
- Could the level of gene expression impact codon bias?

Phaja's gp88 encodes a sigma factor.

Figure 6. HHPred analysis of Phaja's gp88. Black stars represent a helix-turn-helix domain, or DNA binding domain. Sigma factors typically contain two helix-turn-helix domains that correspond to the −10 and −35 box of the promoter sequence.

Figure 7. Structural and amino acid alignment of Phaja's gp88 with the beginning of the −35 binding domain of the sigma-e factor in *E. coli* (4).



- Gp88 was identified as a possible sigma factor by structural alignment not sequence alignment.
- How is it that so little sequence identity can lead to functional identity.
- What does this indicate about the promoter sequence and structure?

Five cluster E phage, including Phaja, encode an endonuclease in place of a methyl transferase.

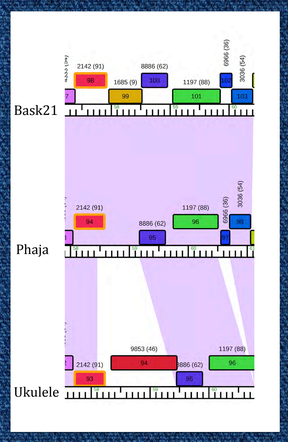


Figure 1. Genomic comparison of cluster E phage Bask21, Phaja, and Ukulele. Maps were generated using Phamerator. Boxes above the ruler indicate rightward-transcribed genes, colored according to pham membership with the pham number indicated above the gene and the number of phamily members in parentheses. Purple shaded regions indicate areas of high nucleotide sequence similarity (>95%). Pham 9520 (gp94 of Ukulele) putatively functions as a DNA methyltransferase, Pham 1685 (gp100 of Bask21 and gp94.1 of Phaja) putatively function as an endonuclease VII.

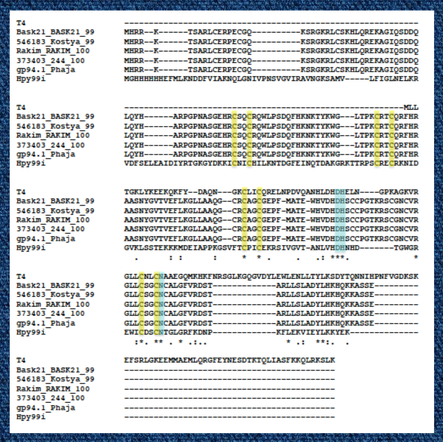


Figure 2. Multiple alignment of the Holliday junction resolvase of phage T4, *Helicobacter pylori* J99 type II restriction endonuclease (Hpy99I) and the putative Endonuclease VII proteins from cluster E phage Bask21, Kostya, Rakim, 244, and Phaja. Conserved cysteine dyads involved in chelating Zn^{2+} are highlighted in yellow (1). Conserved residues of the active site are highlighted in blue (1).

- Does the cluster E endonuclease function as a restriction endonuclease or as a resolvase?
- Many phage encode a methyl transferase to counter bacterial restriction systems.
- Is it a coincidence that the endonuclease replaces a methyl transferase?

Phaja has a 500 bp insertion with two overlapping ORFs, both encoding endonuclease-related proteins.

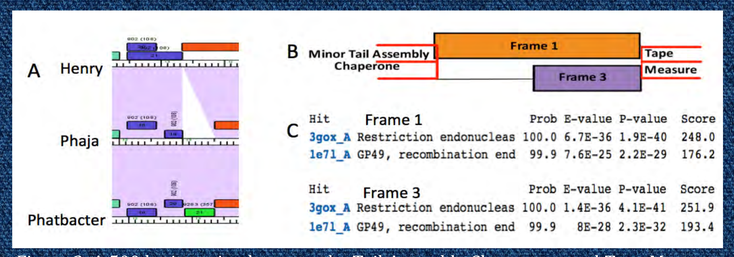


Figure 3. A 500 bp insertion between the Tail Assembly Chaperones and Tape Measure is highly conserved with two possible ORFs. A) Phamerator output showing alignment of Phaja to Henry and Phatbacter (insert is found in 15 Cluster E phage); B) Schematic representation of the two overlapping ORFs contained within the insertion. Top = reading frame 1. Bottom = reading frame 3; C) HHpred output for the peptides encoded by Reading Frames 1 and 3. Both frames align to the same list of endonucleases. The top alignment is to restriction endonuclease HPY99I, found in *H. pylori*.

Major Tail Assembly Chaperone ... . . .GAGGACCCAAACTCCTAG

Figure 4. Location of the potential slippery sequence (CCCAAAC) in the tail assembly chaperone genes. A putative tail assembly chaperone slippery sequence is located 4 nts upstream of the termination codon for the major chaperone. A -1 frame shift would produce the minor chaperone. The endonuclease does not have this sequence.

Frame 1 ...GGC GCA AGG CGG CCG CTG TGG CAT CTG CTC
Frame 3 ... GAGGCGCAAGGCGGCCGCT GTG GCA TCT GCT

Figure 5. An internal Shine-Dalgarno sequence could lead to a frame shift. Weis, et al. (3) reported on a frame shift in E. coli that uses an internal SD sequence. The underlined SD-like sequence together with the downstream GTG could produce the -1 frame shift needed to shift from Frame 1 to Frame 3.

- Use of reading frame 3 in the inserted endonuclease would require a 240 nt ribosome read through, or an alternative promoter or a frameshift. Is this possible?
- There is no slippery sequence, could it be done using an internal Shine-Dalgarno sequence?
- Is the endonuclease needed?

Searching gaps for ORFs, promoters and other structures also reveals conflicts in annotation.

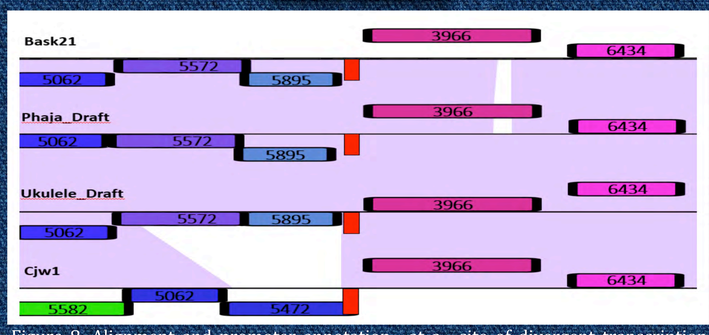


Figure 8. Alignment and promoter annotation at a site of divergent transcription (Phaja [73770-75170]). A strong leftward promoter was identified at site indicated by red vertical bar.

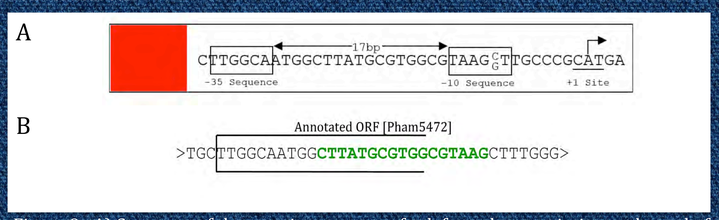


Figure 9. A) Sequence of the putative promoter for leftward transcription at the end of the Cluster E phage right arm. B) Location of the promoter sequence relative to the start site for gp142 [pham 5472] in Cjw1. The promoter is shown in green. The 5' end of the transcript is boxed. Direction of transcription indicated by the ">" symbol.

- Genome annotation follows rules e.g ORF length and coding potential. Filling gaps can find areas of conflict. If the predicted promoter is indeed a promoter then:
- Leftward transcription in Cjw1 of this operon is potentially blocked, or…
- Translation of the ORF is initiated at a start codon downstream of the predicted start codon.

THE UNIVERSITY OF MAINE 1865
MAINE INBRE IDeA NETWORK OF BIOMEDICAL RESEARCH EXCELLENCE
Grant 8P20GM103423-12
SEA PHAGES
HHMI The Hatfull Laboratory

1. Sokolowska M, Czapinska H, Bochtler M. (2009) Crystal structure of the ββα-Me type II restriction endonuclease Hpy99I with target DNA. Nucleic Acids Res. 1–2.
2. Labrie SJ, Samson JE, Moineau S. (2010) Bacteriophage resistance mechanisms. Nat Rev Microbiol. 8:317–327.
3. Weiss RB, Dunn DM, Dahlberg AE, Atkins JF, Gesteland RF. (1988) Reading frame switch caused by base-pair formation between the 3' end of 16S rRNA and the mRNA during elongation of protein synthesis in Escherichia coli. EMBO J. 7:1503–1507.
4. Gruber TM, Gross CA. (2003) Multiple Sigma Subunits and the Partitioning of Bacterial Transcription Space. Annu. Rev. Microbiol. 57:441-466.