Actinobacteriophage Genome Annotation Submission Cover Sheet

Phage Name: **Maravista**
Your Name: **Kathleen Cornely**
Your Institution: **Providence College**
Your email: **kcornely@providence.edu**
Additional emails (for correspondence): **cpelagal@friars.providence.edu**

Describe any issues or specific genes that you would like to highlight for the QC reviewer. This includes any genes that you had questions about or received help with or that warrant further inspection in the QC review process. Include those genes that you deliberated on and/or want to strongly advocate for. If you contacted SMART, workshop facilitator, or a buddy school for help, please document:

Notes written by Charles Pelagalli, who took the lead on this annotation:

I deleted the following genes from the draft annotation of Maravista (i.e., the original auto-annotated file created from Maravista's FASTA), with reasoning listed below:

- gp25 (3' 25412 - 5' 25278 REV)
    - This gene was absent in every other F1 phage genome, forming a pham that included Maravista and only one other member (Nebkiss, an X cluster phage). Additionally, there were no meaningful BLAST hits for this gene, nor any apparent coding potential on the GeneMark map for Maravista (consequently, was not called by GeneMark)
- gp45 (5' 34303 - 3' 34536 FWD)
- gp46 (5' 34502 - 3' 34639 FWD)
    - Neither of these genes had any BLAST hits, and they formed huge overlaps with adjacent genes, in addition to having zero coding potential on any GeneMark map (self, smeg, M. tuberculosis). They were also both FWD genes surrounded by an island of REV genes.
- gp69 (5' 44521 - 3' 44724 FWD)
    - This gene had no coding potential and was not called by GeneMark. BLAST hits were numerous, but all had poor alignment (< ~ 50%), with the exception to a gene in phage Bubbles123, which had 100% alignment. All BLAST hits were to HNH endonuclease. As it turns out, this gene is part of a pham (87759) present in only six other phages, five of which are drafts, with the exception being Bubbles123. Given the lack of coding potential and poor alignment, I made the decision to delete this gene, as it also failed the requirement of having an H-N-H across a 30 aa sequence.
- gp72 (3' 47092 - 46970 5' REV)
    - This was a REV gene sandwiched by a sea of FWD genes with little to no gap for transition. It had zero BLAST hits, no coding potential, and was not called by GeneMark.
- gp82 (3' 51109 - 5' 50762 REV)
    - Another REV gene surrounded by FWD genes, with substantial overlap with surrounding genes. This orpham had no BLAST hits, no coding potential, and was not called by GeneMark.

I added the following genes:

- 5' 42918 - 43388
    - This gene was added on account of a coding potential spike present in the ~ 42900 - 44000 bp range. It also forms a -4 bp gap with the previous gene, as well as giving a -4 bp gap for the following gene. The top BLAST hits had ~ 85% alignment with 100% identity to HNH endonucleases. HHpred gave similar results, with 85% probability for many top hits to HNH endonuclease. The aa sequence did contain an H-N-H sequence but did not exceed 30 aa. I highlight this gene especially for QC review.
- 5' 54447 - 54593
    - This gene was included in the original draft, despite being included on Phamerator with a pham (84861), so I had to re-add it manually.

- 5’ 38339 - 3’ 383431
    - Same story as above, except this gene was replaced with an orpham that I then deleted.
- 5’ 52700 - 3’ 52969
    - This was another gene I added, although there was not substantial coding potential to support it. It did, however, have many blast hits, albeit with poor alignment.
- 5’ 59351 - 3’ 60061
    - Another gene I added, this one appearing at the end of the genome. This gene was added on account of many other F1 cluster phages including a pham (98524) at the end of their genomes, with HNH endonuclease called for the function. BLAST hits were modest, but the H-N-H sequence did not extend for > 30 aa.

Genes that deserve further attention:
- gp40 in the final submitted version (3’ 32125 - 5’ 30905 REV)
    - This was by far the gene most deliberated on in this annotation. BLAST and HHpred hits led me to call this gene as a transposase, but the requirement to call a gene as a transposase is to identify a transposon in the genome. We consulted the forums for this gene (SEA-PHAGES | Identify transposons (seaphages.org)), and ultimately came to suspect that one of the many endonucleases called in the genome could be the transposase, given that a transposase is by definition an endonuclease (both cleave the phosphodiester bonds of DNA).

Please record yes/no for each of the questions below. In the submitted DNA Master file (Yes/No):

Yes 1. Does the genome sequence in your submitted DNA Master file match the nucleotide fasta file posted on phagesDB (same number of bases, no N bases, etc.)?
Yes 2. Are all the genes ‘Valid” when you click the Validation button?
Yes 3. Are the genes (and matching LocusTag numbers) sequential, starting with #1, counting by 1s.
Yes 4. Are the Locus Tags the “SEA_PHAGE NAME” format?
Yes 5. Has the documentation been recreated from the Feature Table to match the latest file version?
N/A 6. Have tRNAs followed the tRNA protocol, **COPYING** tRNA-AMINOACID type (DNA equivalent of the anti-codon) from Aragorn output - tRNA-Gln(ctg) - AND the ends been adjusted to match the Aragorn output?
Yes 7. Has the frameshift in the tail assembly chaperone been annotated correctly (if applicable)?
Yes 8. Have you cleared your Draft_Blast data and have you re-Blasted the submitted DNA Master file?
Yes 9. Has every gene been described and supported in your Supporting Data file?
Yes 10. Did you investigate ‘gaps’?
Yes 11. Did you delete the genes that you meant to delete?

Now, make a profile of the file you plan to send. (And you can save this file for Review to Improve!)

Yes 1. Have any duplicate genes been deleted?
Yes 2. Has the Notes field been cleared (using the automated buttons)?
Yes 3. Do the gene numbers and locus tags match?
Yes 4. Are the correct Feature_Types correctly selected (most will be ORFs, but check that tRNAs and tmRNAs are correctly labeled)?
Yes 5. Do the function names in the Product field either match the official function list or say “Hypothetical Protein”?
Yes 6. Has the Function field been cleared (using the automated buttons)?

How are you documenting your gene calls in class? Choose any/all that apply:
✓ DNA Master shorthand (previously used format)

What is the file type (sort) submitted for QC to document your gene calls? Choose only one.:
✓ DNA Master shorthand (previously used format)