Actinobacteriophage Genome Annotation Submission Cover Sheet

This Cover Sheet will accompany each genome's annotation file(s) submission and succinctly describe the work that your students and you have done. This document ensures that the work done was as complete and thorough as it could be. Most important to the QC reviewer, denote where the trouble spots were in your annotation and how they were resolved.

Phage Name. Soos
Your Name. Pam Connerly
Your Institution. Indiana University Southeast
Your email. pconnerl@ius.edu
Additional emails. (for correspondence). erueschh@ius.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.

- Our DNA Master file contains 87 features. Feature numbers listed here are FINAL Feature numbers, not original feature numbers.
- Feature 4 (3205-3693) - We contacted Debbie Jacobs-Sera for confirmation that the start should be 3205 based on the 4 bp overlap and more coverage of the coding potential.
- Feature 26 (12596-13972) - There was no conclusive evidence for calling the start. However, consultation with Debbie Jacobs-Sera suggested a call for the start at 12596 because it covers all coding potential, has the smallest gap and the starterator data suggests that the appropriate start would be 12593. Since this is the second start in a row we then selected 12596.
- Feature 27 (13969-14922) - There was no conclusive evidence for calling the start. However, consultation with Debbie Jacobs-Sera suggested a call for the start at 13969 since there is a 4bp overlap.
- Feature 39 (22170-22328) - NCBI BLAST matches with Clawz 1:1 as a hypothetical protein. BLAST through DNA Master results in a range check error and partial BLAST results.
- Features 40 (22417-24207) and 41 (24289-25056) - These BLASTed to major capsid hexamer and major capsid pentamer proteins respectively. However, there was no acceptable HHPred evidence to support a function. We therefore called these hypothetical proteins due to the lack of HHPred evidence.
- Features 50 (29331-29867) and 51 (29927-30148) - There was one hit on HHPred that suggested that the function of 50 might be a tail assembly chaperone. However, there is no evidence that 51 is the other part of the tail assembly chaperone nor any indication on the GeneMark output that there is a frameshift for this feature. There is a gap of 59 between feature 50 and 51. We left the function of these two features as hypothetical proteins as we didn't feel there was enough evidence to conclude that this was a tail assembly chaperone with a frameshift.
- Feature 59 (39254-40327) - We left the function of this feature as a hypothetical protein. Both BLAST and HHpred point toward a tail protein, but there is no indication as to which type of tail protein. There is a BLAST hit pointing to a collegen-like protein, but it is unknown if it is in the syntenic region. Without further evidence of what type of tail protein, we did not call a specific protein.

- Feature 61 (41060-43834) - There were a large number of blast hits and HHpred results not listed in the spreadsheet that mention hydrolase as a function. However, the function list says not to call a Lysin A without having a Lysin B in the genome. There is a paper (Pollenz et a. 2022; PLOSOne) about Gordonia endolysins in which Clawz calls Lysin A. There is a clear homolog by BLAST and synteny, so we called the Lysin A. It may be interesting and significant that there appears to be a Lysin A without a Lysin B in the genome.

- Feature 62 (43831-44238) - There are multiple BLAST hits for holins. DeepTMHMM indicates there are two transmembrane domains. The function lists states that the presence of at least 2 transmembrane domains and the gene being adjacent to the endolysin as acceptable evidence to a holin. This feature has a nearly identical sequence to the Clawz holin (feature 66) and synteny. Although Clawz 66 is annotated as a holin, Pollenz et al. 2022 refer to it as a holin-like gene.

- Feature 83 (55576-55950) - This gene was added in the gap between features 82 and 84. Both Genemark S and G. terrae show coding potential and there are BLAST hits to multiple phage HNH endonuclease genes.

- Feature 85 (56431-56631) - This gene does not BLAST. Gives a range check error. BLASTing directly through NCBI reveals no hits, but a phagesdb.org BLAST reveals a 1:1 BLAST hit with Sting_draft_90.

- Possible feature 49191-49346 not included in annotation. Evidence is not conclusive to call gene but may be of interest when more data becomes available. GeneMark S and GeneMark *G. terrae* outputs show small blip of weak CP. PhagesDB BLAST results only included a Q1:S1 start, 100% alignment and similarity, and 5E-26 with Sting_Draft_75 orpham. RBS data had a Z-Score= 1.508 and Final Score= -6.147. Possible gene would take up 156 bp of a 328 bp gap between genes 70 (48164-49015) and 73 (49343-50062) where a random reverse gene was removed. Additionally, if the feature was added, gene 73(49343-50062) would have a 4 bp overlap with the new upstream gene (49191-49346).

- Note on item #8: All features with start sites changed during manual annotation were re-BLASTed individually (and double checked). We have had great difficulty with BLASTing the full DNA Master file.

- For the authors list, we have an author with a last name of De La Paz. Since it is three separate words, we were unsure of how to list it. We included it in the last name column as De La Paz, but are concerned about it being truncated.

Please record yes/no for each of the questions below. If further explanation is needed, please add this item to the above box.

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 (no tRNAs) 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?

n/a 7. Has the frameshift in the tail assembly chaperone been annotated correctly (if applicable)?

n/a (see note) 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!)

n/a 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:

PECAAN output
DNA Master shorthand (previously used format)
X **Spreadsheet**
Powerpoint
Word document (must be easily searchable)
Other: Describe.

What is the file type (sort) submitted for QC to document your gene calls? Choose only one.:

PECAAN output
DNA Master shorthand (previously used format)
X **Spreadsheet**
Powerpoint
Word document (must be easily searchable)
Other: Describe.