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. **Hum25**
Your Name. Amanda Freise
Your Institution. UCLA
Your email. afreise@ucla.edu
Additional emails. (for correspondence).

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.
—----------

Very interesting singleton genome. First chunk, up to the tape measure protein, looks similar to cluster AY phages (especially EvePickles and Faja). Rest of genome is very mosaic. Most of the below is explanations for decisions made. I have colored purple the calls that I was really not sure about.

Gene 4: Start site 3 (3433bp) had 3 MA's in Starterator, but the next start site is the second of two tandem starts, so chose that one. --- Most members of pham called capsid maturation protease, but I couldn't find any data to support this in this gene.

Gene 13: Possible major tail based on synteny, but HHpred hits in support of this call are terrible (e-value 9.8), and most genes in pham do not call a function.

Genes @ stops 9039, 9374: Tail assembly chaperones. I think I have called them correctly, but it was tricky because there other pham members of the first gene (e.g. Faja) were much longer than the first gene in Hum25, and Hum25 didn't have the same slippery sequence. I found one upstream beginning at 8994 with the sequence: TTTTTTG, a frameshift found in phage P2 that goes from an FL to FF. Could use a double-check on this one!

Stop @ 9327 (reverse): May not be real. Reverse gene in middle of some forwards, but it does have good coding potential on GM-Self. 2 TMDs detected on DeepTMHMM. However: stop site overlaps with previous F gene stop @ 9374.Faja also has a R gene in this location, but it doesn't overlap with the TAC.

Stop @ 17019: Chose start site with -29 overlap (16292(because it didn't cut off any coding potential, and was suggested on Glimmer, Genemark, and Starterator.

Stop @ 17992: Likely holin gene. Good evidence in HHpred, CDD for holin. DeepTMHMM predicts two TMDs, but they are each only about 14-15 aa long. Gene is directly downstream of endolysin. Many members of pham call holin.

Stop @ 20152: Not sure on start site: the selected start (22092) begins the gene right where the CP on GM-self begins, but also produces a large gap. It **does** meet evidence requirements for this function. (the N-terminus should contain an acyltransferase domain and the C-terminus should contain a SGNH domain. 11 TMDs as predicted by DeepTMHMM)

Stop @ 22241: Added this gene based on small amt of CP on both Genemarks, and to help fill gap made by previous gene.

Stop @ 22392: Added this gene based on TINY amt of CP on Genemark-Self (almost invisible), and to help fill gap.

Stop @ 25774: Contains required helix-turn-helix in HHpred prediction.

Stop @ 26420: Contains required helix-turn-helix in HHpred prediction.

Stop @ 27827: Contains required helix-turn-helix in HHpred prediction.

Stop @ 28900: There are a strong HHpred hit and strong CDD hit to Erf protein, and many BLAST hits with reasonable e-values to ERF family protein too. But per this recent forum post, the sequence does not have the required residues to make this call. https://seaphages.org/forums/topic/5536/ All other HHpred hits (e.g. to HTH hits) have very poor e-values. Leaving as NKF.

stop @ 29445: Might be a Ku-like dsDNA break-binding protein based on good hits to Gam proteins, analogues of Ku. But example gene for Ku-like (Omega_206) definitely has strong hits to actual Ku protein. Left as NKF for now. See this forum post proposal. https://seaphages.org/forums/topic/4766/

Stop @ 30984: Possible "atypical" endonuclease with HNK and HNN motifs (per this forum post: https://seaphages.org/forums/topic/5505/). has good hits to HNH proteins in HHpred.

Stop @ 33336: Decent hits in HHpred to **RepA**, but may not be strong enough... HTH is an alternative call.

Stop @ 36400 (R): Tricky call. This is the start predicted by Glimmer, and it cuts off some of the coding potential for this gene, which is good (esp on GM-Self).. The start predicted by Genemark is 36556, and overlaps the following F gene dramatically (the following F gene also has good coding potential).

Stop @ 36614 (F): Maybe should be deleted to make room for upstream R gene? Does have CP, but not especially strong. The following gene (stop 36898) does have a potential -4 start site which suggests operon with this gene.

Stop @ 37182: Good CP on GM-self, none on GM-host. (relevant for next gene)
Stop @ 37354: Start chosen overlaps with preceding gene by 38bp, but next available start produces ORF that is way too short (63bp long). This gene has good CP on both GM host and self.

Stop @ 40183: Contains required helix-turn-helix in HHpred prediction.

Stop @ 40484: Added this gene based on some CP present in 4th frame on GM-self.

—----------

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?
Yes 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)?
No 8. Have you cleared your Draft_Blast data and have you re-Blasted the submitted DNA Master file?
I was unable to save my re-BLAST results within DNA master and have yet to figure out what the issue is. Debbie said it would be fine for now to note this here and submit.
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:

- X PECAAN output
- DNA Master shorthand (previously used format)
- Spreadsheet
- X 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.:

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