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. Skinny

Your Name. Frederick Baliraine

Your Institution. LeTourneau University

Your email. FredBaliraine@letu.edu

Additional emails. N/A

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.

1. For the tRNAs, the detailed notes have been left in the notes section of the minimal file.
2. Start for feature 3 (1264-1449 rev). This gene was inserted, and it is an Orpham. No Coding potential in GenMarkS, and smeg, but there is significant CP in GeneMark-TB, and no significant BLAST hits; we think the selected start is the best.
3. Capsid maturation protease for feature 17 (7881-9041)? Though this gene hits “MuF-like minor capsid protein” in many phages in phagesDB, this is no longer an acceptable function name (Now “Hypothetical Protein”) according to the current (14 May 2022) Official SEA-PHAGES Functions List. We settled on the “capsid maturation protease” based on several factors. First, the gene hits the same hit, which is phageD29 gene 15 in HHPred just like the reference sequence (Sisi gp 4) and has synteny with phage Reindeer gp 14. Second, in phage Sisi, the capsid maturation protease is right downstream of the portal protein, as in the case for this gene in phage Skinny. Third, in phage D29, gp15 is also a capsid maturation protease according to phagesDB gene list. Besides, there are currently 24 hits to capsid maturation protease in phagesDB for this gene. Finally, the general synteny of phage structural genes according to the Resource Guide suggests genes coming in the following order: terminase, portal protein, capsid maturation protease, scaffolding protein (https://seaphagesbioinformatics.helpdocsonline.com/article-90), hence lending further support for the “capsid maturation protease” function.
4. Feature #27 (14636-21457; tape measure protein): Could not get BLAST data in DNA Master upon re-Blasting (gives “Status 8: Error”), but data obtainable on phagesDB BALSTp.
5. Feature #36 (30228-31595; lysin A): Could not get BLAST data in DNA Master upon re-Blasting (gives “Status 3: Error”), but data obtainable on phagesDB BALSTp.
6. Function for feature #40 (33191-33469). There is a "mixed bag" of "specific" functions for this gene, with many hits to NrdH-like glutaredoxin and glutaredoxin in phagesDB, with a few hits to Thioredoxin for this Skinny sequence. In view of the SEA-PHAGES forum post (<https://seaphages.org/forums/topic/5376/>) of 08 May 2022 entitled, "Function for subcluster A11 phage Gilberta (37505-37777 rev): Thioredoxin, NrdH-like glutaredoxin or glutaredoxin?" I have settled for the less specific function of "oxidoreductase," except if the QC team suggests otherwise.
7. The feature #57 at 39743-39913 bp was inserted because there is significant coding potential (CP) in this region in GeneMark-smeg, GeneMarkTB and GeneMarkS. CP is noted as “not covered” because the only plausible start at 39743 bp lies inside the CP plateau. Whereas there are no significant BLASTp hits in NCBI and phagesDb, tBlastx in NCBI BLAST using the nucleotide sequence gives significant hits with 7 phage genomes, the top hit being with phage TyDawg, q171: s39484, 100%, 5e-28, Accession number MN586024.1. Thoughts? We believe it is a real gene.
8. The feature #71 at 45287-45775 is an orpham; we changed the start from those called by GM and GL because they would give large (>170 bp) gaps and the selected start has a 58 bp overlap but has the best RBS score (Z= 3.047, Final score = -3.350) and q37:s1 with LilhomieP, gp68. Moreover, considering the CP from GeneMark-TB and S, the selected start is the one that well covers all the CP in TB & S. It also hits the HNH endonuclease, PDB, Geobacillus virus E2, 5H0M\_A , 74.85% alignment\*,99.21% probability, but the BLASTp hits to HHN endonucleases shows little identity.
9. “Incorrect length” error message crops up when validating the feature #73 at 47049 – 47393 bp. However, this is start was called by GeneMark and is the most annotated start, found in 85% of genes in the pham so far, called 82.4% of the time when present, and is part of an operon, with a 4 bp GTGA overlap, there are no stop codons in the sequence when using the selected start. Not sure why the error message appears.
10. The feature at 52415-52504 bp (#85) was inserted. It is an Orpham, detectable in starterator and phamerator, but had not been autocalled in DNAMaster. It has no CP in GenMark-smeg, very low CP in GeneMarkS, and CP in TB slightly below the 0.5 mark. It is a membrane protein (1 transmembrane domain found by TMHMM, and 1 transmembrane domain found by SOSUI), but has no significant BLAST alignments.
11. The feature at 56853-65996 (#97) is autocalled and has matching BLASTp hits q1:s1, 100% with a few phages in phagesDB, but there’s no coding potential in this area. In fact, the entire region from 55464-57955 bp has no CP seen in GeneMark-smeg, TB and S.
12. The feature at 59486-59743 (#109) overlaps with a tRNA downstream, but we believe that it should be considered as a gene as it could be a conserved gene. Evidence by Phamerator shows that it has been called across all three M subclusters for various phages even though it overlaps with a tRNA downstream. The autocalled start by Glimmer is the most annotated start, found in 100% of the pham members and called 92. 9% of the time when present.
13. Tried inserting a gene at 62150-62311 bp, but there is no coding potential and no hits in BLASTp; but hits on transmembrane domain in TMHMM and SOSUI, indicating a membrane protein, but deleted it for lack of coding potential or significant hits in BLASTp.
14. “Incorrect length” ”complement” error messages crops up when validating the feature #142 at complement (72509 – 73033 rev). However, this is start creates the shortest gap possible and has a better RBS score than the starts autocalled by Glimmer and Genemark, and gives q1:s1 with phage LilhomieP, gp136. Not sure why the error message appears.
15. The feature at 77761-77868 bp rev (#157) was inserted. Whereas there is no coding potential seen in GeneMark-semeg and TB, there is some CP, slightly below the 50% mark seen in GeneMakS (that's the reason I stated "no" for CP), and it is part of an operon with a TGATG stop-start overlap with the downstream gene, and an 8 bp overlap with the upstream gene. Moreover, TBBLASTX in NCBI of the nucleotide sequence shows several hits sith phage SirSheldon and TayDawg (q108:S77406), 100%. We think it is an Orpham, that’s why its amino acid sequence gives no significant hits in phagesDb or NCBI.

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**

Yes 2. Are all the genes ‘Valid” when you click the [Validation button](https://seaphagesbioinformatics.helpdocsonline.com/article-84)? **NO; see notes #8 & 13 above**

Yes 3. Are the genes (and matching LocusTag numbers) [sequential](https://seaphagesbioinformatics.helpdocsonline.com/article-77), starting with #1, counting by 1s.

Yes 4. Are the Locus Tags the “[SEA\_PHAGE NAME](https://seaphagesbioinformatics.helpdocsonline.com/article-77)” format? **YES**

Yes 5. Has the [documentation been recreated](https://seaphagesbioinformatics.helpdocsonline.com/article-86) from the Feature Table to match the latest file version? **YES**

Yes 6. Have tRNAs followed the [tRNA protocol](https://seaphagesbioinformatics.helpdocsonline.com/undefined), **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**

Yes 7. Has the [frameshift in the tail assembly chaperone](https://seaphagesbioinformatics.helpdocsonline.com/article-54) been annotated correctly (if applicable)? **YES**

Yes 8. Have you cleared your Draft\_Blast data and have you [re-Blasted](https://seaphagesbioinformatics.helpdocsonline.com/article-57) the submitted DNA Master file? **YES**

Yes 9. Has every gene been [described and supported in your Supporting Data file](https://seaphagesbioinformatics.helpdocsonline.com/article-44)? **YES**

Yes 10. Did you investigate ‘[gaps](https://seaphagesbioinformatics.helpdocsonline.com/article-31)’? **YES; where gaps were left unfilled, explanatory notes were made in the “Additional Notes” section of the gene before of downstream of the respectinve gaps.**

Yes 11. Did you [delete the genes](https://seaphagesbioinformatics.helpdocsonline.com/article-65) that you meant to delete? **YES**

Now, [make a profile of the file](https://seaphagesbioinformatics.helpdocsonline.com/article-64) you plan to send. (And you can save this file for [Review to Improve!)](https://seaphagesbioinformatics.helpdocsonline.com/untitled-18)

Yes 1. Have any duplicate genes been deleted? **YES**

Yes 2. Has the Notes field been cleared (using the automated buttons)? **YES**

Yes 3. Do the gene numbers and locus tags match? **YES**

Yes 4. Are the correct Feature\_Types correctly selected (most will be ORFs, but check that tRNAs and tmRNAs are correctly labeled)? **YES**

Yes 5. Do the function names in the Product field either match the official function list or say “Hypothetical Protein”? **YES**

Yes 6. Has the Function field been cleared (using the automated buttons)? **YES**

How are you documenting your gene calls in class? Choose any/all that apply:

      PECAAN output

Yes DNA Master shorthand (previously used format) **YES**

Yes 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) **YES**

Yes Spreadsheet

      PowerPoint

      Word document (must be easily searchable)

      Other: Describe.