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. Andre

Your Name. Kristen Clermont

Your Institution. La Salle University

Your email. clermont@lasalle.edu

Additional emails. (for correspondence). Sean McClory: mcclory@lasalle.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.

In phage Andre, we deleted two genes from our DNA Master auto-annotation and added four genes, one of which was called in the auto-annotation available on Phamerator and the other three not in either the DNA Master or Phamerator auto-annotation. Annotation of added genes relied heavily of synteny and identification on the Starterator report for the pham predicted through sequence homology. The start site auto-annotated in DNA Master was not selected as the manually annotated start site of eleven genes in the genome. In the column “MA also the most called start site in pham?” eight genes are marked as having a final manually annotated start site that is different from the most annotated start in the pham. The “Decision summary, start site” column summarizes for each gene how the final decision was made.

For review of the start-site annotation, fourteen start site calls are described in the ”Uncertainty on start site” column of the annotation notes spreadsheet as having desperate lines of evidence that lead to some level of uncertainty in the final call (highlighted in blue). The calls of the genes with stop codons at 3365 (gp 6), 13983 (gp 16), 41817 (gp 49), 47590 (gp 55), 55613 (gp 66), and 66292 (gp 99) rely heavily on position relative to the spike in coding potential in GeneMarkS graphs. The calls of the genes with stops sites at 45368 (gb 53), 49133 (gp 60), and 59981 (gp 75) rely heavily on reducing gaps. The calls of the genes with stops at sites 60955 (gp 78) and 64026 (gp 90) rely on guiding principle 12F in the bioinformatics guide stating that empirical evidence suggests choosing the second of two tandem start codons.

Per the cluster-specific forum post “Tail assembly chaperones?” we did not call the tail assembly chaperone in the B1 phage. https://seaphages.org/forums/topic/4518/?page=2

Five genes are called membrane proteins based on output from DeepTMHMM. Proteins that only have one transmembrane domain that is at the N-terminus were not called membrane proteins.

Some of the functional annotations we called are not high-consensus calls based on other functional annotations in the pham and/or synteny. We called the holin protein based off two different 96% probability Pfam hits (N-terminal and C-terminal) to a Pfam holin domain (PF16081, Phage\_holin\_7\_1).

The use of Pfam domains to call holins was recommended in the following forum post. https://seaphages.org/forums/topic/4726/

An additional minor tail protein was called at the end of the operon region containing numerous minor tail proteins. This was called based on very strong hit to recently-published 8JOU\_g in the PDB (99.96% probability, e-value 3.7e-28). This protein is in a complex making up the "Fiber I and fiber-tail-adaptor of phage GP4." One other member of this pham currently has a function annotated (P3MA\_26 is also called a minor tail protein). https://www.rcsb.org/structure/8JOU

The gene with the stop site 62977 (gp 85) is commonly annotated as an HNH endonuclease, however the protein product only has one histidine in its entire sequence, so this cannot be called. Feedback is requested on the choice to call this a DNA binding protein based on the fact that the hits in HHPred (endonuclease, CRISPR-associated endonuclease, mature reverse transcriptase) all have DNA binding function in common.

After review of HHPred hits of proteins currently annotated “DNA primase/polymerase/helicase” vs. “DNA primase/helicase” we observed a loose pattern. Genes called “DNA primase/helicase” more often have the gene architecture DNA primase domain (with HHPred hits to 2AU3\_A, 8DWJ\_A, and PF01807.24) + helicase domain. Genes called “DNA primase/polymerase/helicase” more often have the gene architecture bifunctional DNA primase/polymerase domain (with HHPred hits to 3M1M and PF09250.15) + helicase domain. Our gene has hits to the bifunctional DNA primase/polymerase domain so we are calling it “DNA primase/polymerase/helicase.”

Also, as a point of interest, gp 93 has several low e-value hits to EF-P/eEF5A structures, which may be of interest to people outside of the SEA program. This forum discussion is the last time it was discussed: https://seaphages.org/forums/topic/4758/. No function was called.

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](https://seaphagesbioinformatics.helpdocsonline.com/article-84)?

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?

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

N/A 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?

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

No 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 9. Has every gene been [described and supported in your Supporting Data file](https://seaphagesbioinformatics.helpdocsonline.com/article-44)?

Yes 10. Did you investigate ‘[gaps](https://seaphagesbioinformatics.helpdocsonline.com/article-31)’?

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

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 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)

      Spreadsheet

      Powerpoint

      Word document (must be easily searchable)

X Other: Describe. Class OneNote file

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.