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. ScienceWizSam

Your Name. Dr. Hari Kotturi

Your Institution. University of Central Oklahoma

Your email. hkotturi@uco.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.

They are listed at the end.

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?

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

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

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 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:

YES PECAAN output

      DNA Master shorthand (previously used format)

      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.:

YES PECAAN output

      DNA Master shorthand (previously used format)

      Spreadsheet

      Powerpoint

      Word document (must be easily searchable)

      Other: Describe.

Gene 1367-2491- SMART members should look at this gene. Phages Synepsis and Tatanka (annotated close matches) call it Lysin A, however, there is no Lysin B like gene in the genome. Hence we think it should be considered as Endolysin, with N-acetylmuramoyl-L-alanine amidase domain. There is evidence for this domain from HHPred and CDD databases. Also, in the SEAPHAGES functional Assignments, it mentions that if we cannot locate Lysin B, we should call it as Endolysin.

Gene 2655-3056 – SMART members should look at this gene. We talked about this gene in class and decided to change the start site. We changed the start site from 2697 with the TTG start site leaving 38bp gap to 2655 with a -4 overlap, Z score of 2.2, a final score of -4.372, and GTG start site. This change gives a bigger reading frame and 100% alignment in NCBI BLAST.

Gene 2488-2658 – SMART members can look at this gene. We talked about this in class. Glimmer (2515) and GenMark (2488) have different starts. Starterator supports 2515. The start 2515 leaves a gap of 23 bases. We think the actual start site should be 2488 as it gives the longest ORF with a -4 overlap. NCBI-BLAST supports this longest ORF. The same longest ORF and over lap is seen in phages castorTroy and Breylor17.

Gene 11484-11783 - SMART members can look at this gene. We think it should be called a tail terminator as it has HHpred hit 5A21\_G: SPP1 gp17. Pjahe DiVitoJr in AU1 cluster also calls it tail terminator. However, other annotated phages in the cluster call it "hypothetical protein".

Genes 13264-14064 - These cluster members (AU) have two copies of this major tail protein as described on SEA-PHAGES AU cluster information page.

Gene 16800-17618 - We talked about this gene. We think it should be considered a minor tail protein. The closest match to our phage, Phage DiVitoJr, and Tatanka call it a hypothetical protein. We think it should be a minor tail protein as it is upstream of a tape measure (synteny), next to other minor tail proteins, and NCBIBLAST shows e value of e-4.4x-167 with Tenno. Also, the HHPred match shows a probability of 97% with bacteriophage T7 Tail fiber protein

Gene 17645-18610 - We think it should be considered a minor tail protein. As per Phamerator, this gene is present in ScienceWizSam and phage Tatanka. However, it is absent in its close relative Synepsis. Tatanka also calls it a minor tail protein. We think it should be a minor tail protein as it is upstream of a tape measure (synteny), next to other minor tail proteins, and NCBIBLAST match shows (% identity of 76.3407, % alignment of 84.2271, and 100% coverage) with e value of e-165 with Tatanka. Also, the HHPred match shows a probability of 97.9% and 28% coverage with Baseplate protein inLactococcus phage tp901-1

Gene 27293-28081 - We think this should be a hypothetical protein. Few members in the cluster designate it as a Minor tail protein. Phage Synepsis is the only one that calls it hypothetical protein. We also agree with this as there is no evidence or HHPred match with a phage protein. All HHpred matches are for human, mouse, or insect proteins.

Gene 28830-29222 - We think this should be a Holin. Few members in the cluster designate it as a hypothetical protein. Phage Synepsis is the only one that calls it Holin. We also agree with this function. There is evidence to consider it as holin. HHpred match 99.9% probability with phage\_holin\_DP1 with % coverage of 46.9. There is also published work on Pfam with this phage indicating it is holin. It also has transmembrane domains in TMHMM and SOSUI.

Gene 29731-29979 - We added this gene to DNAMaster. This ORF was only detected by GeneMark. GeneMark suggested start site is 29746, we changed it to 29731 as it gives a longer ORF with 1:1 Q:S match to other members in this subcluster (Breylor17, CastorTroy, and others).

Gene 30055-30372 - We added this gene to DNAMaster. This ORF was only detected by GeneMark. GeneMark suggested start site is 30070, we changed it to 30055 as it gives the longest ORF with 1:1 Q:S match to other members in this subcluster (Tatanka, Nightmare, and others). Enough evidence for membrane protein through SOSUI and TMHMM.

Gene 31063-31215 - We investigated a 393 gap between the genes as GeneMark shows some coding potential in this gap. We found a gene that is also present in another member of the same subcluster (Tenno and also a few members in AU2 subcluster) with e-value of -22. This makes a 153 amino acid protein. We are including all the evidence for this gene.

Gene 44180-44473 - Glimmer did not call this gene. Starterator agrees with GeneMark call. This gene is present in the same location in all of the members of this cluster.

Gene 44451-44792 - Glimmer and GeneMark did not call this gene. We found this gene in the gap between two genes and looking at other annotated members of the cluster (DevitoJr and Breylor17).

Gene 45061-45285 - This start site has been changed. Glimmer suggested start is 45148l eaving a gap of 95bp with Z score of 1.694 and a final score of -5.360. However, GeneMark, and starterator call it 45061. We changed it to 45061 which closes the gap with a better Z-score of 2.882 and a final score of -4.51. There is no match for this gene in phagesdb, NCBI, or HHpred. We can clearly see the coding potential in GeneMark.

Gene 46505-46750 - Glimmer and GeneMark did not call this gene. We found this gene in the gap between two genes and looking at other annotated members of the cluster (Synepsis, Tatanka and CastorTray).

Gene 53282-53524 - Glimmer and gene mark have different start sites. Starterator supports Glimmer as it gives the longest ORF with goof Z-score and Final score.

Gene 53858-53965 - This gene is not called by Glimmer and GeneMark. We found this gene in the gap between two genes and looking at other annotated members of the cluster (ElephantMan, Niktson, Gordon, and others).

Gene 54032-54202 - This gene is not called by Glimmer and GeneMark. We found this gene in the gap between two genes and looking at other annotated members of the cluster (ElephantMan, CapnMurica, Niktson, Gordon, and others).

Gene 54288-54656 - This gene is not called by GeneMark. Glimmer suggested start is 54447 with a Z-score of 1.5 and a final score of -6.4. This start site leaves a 244bp gap. We changed the start site to 54288 as it gives the longest ORF with a Z-score of 2.2 and final score of -5.2. This changed start site (54288) also matches (Q:S-1:1) the longer ORF in other members of this cluster in NCBI BLAST. The function is supported by HHPred evidence too apart from other members of the cluster calling it hydrolase.