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. Guinevere

Your Name. Rivka Glaser, Victoria Herrera

Your Institution. Stevenson University

Your email. rglaser@stevenson.edu

Additional emails. (vherrera2@stevenson.edu).

Please note: Victoria is my student who did the annotation as part of an independent research course. We included her name here as the primary name, but we can change it if only faculty names are needed as correspondents.

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. Note that the gene numbers in excel do not reflect the added gene between 41 and 42. The DNA Master file was updated to reflect this change. The added gene as well as the gene that we switched frames for are at the bottom of the excel sheet.
2. When renumbering the genes in DNA Master, the new gene (feature 44) did not move to its proper location in the features list even after multiple attempts. This gene is still properly numbered and labeled.
3. For feature 2 (5’ 91, 3’ 306), I am unsure about the start site. Blast results are conflicting between different matches, and Starterator equally supports either site. The RBS and Z scores for 318 were better than 306. Because site 306 would result in an overlap of 4, which is a common pattern in phage genome, I ended up calling this site.
4. For feature 8 (5’ 4963, 3’ 6216), I am unsure about the functional call. There were several high-quality matches to Clp proteases, so I went with the call for capsid maturation protease, which was also a common functional call for this pham.
5. For feature 9 (5’ 6219, 3’ 6581), I am also unsure about this functional call. There was a high-quality match to a capsid fiber protein, so I made the call for a head fiber protein assuming “capsid” was the equivalent of “head”. I noted that genes in this pham are predicted to be capsid decoration proteins or scaffolding proteins, which did not have any matches in HHPred.
6. Features 18 and 19 are included in a -1 programmed frameshift between frames +1 and +2. The slippery sequence was identified within GGGGGAAGACACCG with the underlined G being the site repeated (bp 10506). The frameshift results in the amino acid sequence GRHRL.
7. For features 27 (5’ 19991, 3’ 21094), and 28 (5’ 21094, 3’ 22041), I called both functions minor tail proteins. 27 had a high-quality match to a minor tail protein. I was unsure what percentage of glycine would make something glycine rich, which was something the approved functional calls list said to look for. Both of these genes were predicted to be minor tail proteins and had high quality matches to a variety of tail proteins. I was unsure if these were the proper calls.
8. Feature 40 (5’ 28559, 3’ 28924), was auto called as being in the +1 frame but did not have phage matches on NCBI or DNA Master. The only alignments found were to draft FF phages. Removing the gene would result in a large gap, so other frames were investigated for coding potential. The -3 frame had coding potential, so the area was blasted, which found an alignment to FF phage Popper. The start site was unclear for this gene, but I called the one that resulted in the least drastic misalignment since no site would resolve the alignment with Popper.
9. For feature 43 (5’ 29637, 3’ 29972), I am unsure about the start site. For now, I went with the site called by glimmer (the gene was not called by GeneMark). There were many matches to other FF phages, but each had different alignments. Starterator had 2 manual annotations for two different sites, but those sites had poor RBS and Z scores. The only site with good scores were not called in Starterator. The functional call was also tricky for this gene because there was a high quality match to the Lambda CI Repressor. Analysis of the domains revealed that the gene lacked the c terminal dimerization domain, but contained the helix-turn-helix DNA binding domain. I referred to the forums, which suggested to call this gene as a helix-turn-helix DNA binding domain as a conservative call, though some argue it could be an immunity repressor.
10. I decided to add feature 44 (5’ 30166, 3’ 30381) because analysis of FF phages on phamerator showed that Guinevere was missing a gene seen in other phages. The +3 frame indeed had coding potential present in an ORF. Blasting this site resulted in high matches to MerR-like helix-turn-helix DNA binding domain proteins. There was only one start site present, and it aligned with Zaheer on NCBI. The addition of this gene maintains a large enough gap between the previous gene (feature 43) for the transition from reverse to forward. This gene had a high-quality match to a MerR-like DNA binding protein on HHPred. Analysis of the secondary structure showed the presence of 3 helix-coil motifs and a coil-coil, which were two things the forums said to look for before making this functional call.
11. For feature 65 (5’ 38393, 3’ 38869), I was unsure about the start site. No blast data was available from NCBI and DNA Master. PhagesDB had a match to Sashimi, but was not aligned. Because this gene was an orpham, no Starterator data was available to help this call. The site 38399 had better Z and RBS scores but would result in a gap of 3. The original start site created an overlap of 4, which is commonly seen in phage genomes. Because of the overlap, I decided to keep the autoannotated call.
12. Feature 66 (5’ 38873, 3’ 40036) was predicted to be an HNH endonuclease. This gene had a high quality match to HNH endonuclease and contained a HKH pattern across a 30 amino acid span. The forums suggested that this HKH pattern should still be called an HNH endonuclease. I am confident in this call, however, feature 71 (5’ 41765, 3’ 42061) also had a high quality match to an HNH endonuclease on PhagesDB, NCBI, and HHPred. There was an HNH and HKH pattern found within a 30 amino acid span. I was unsure if there could be two HNH endonuclease genes in a single genome and I was not able to find any information on the forums regarding this. Additionally, this gene is an orpham, so I am unsure what to make the functional call.
13. We have had trouble deciding which of the three tRNA’s in Guinevere to eliminate. Currently, there are no annotated FF phages that possess three tRNA’s. The three tRNA’s in Guinevere are located in the same positions as the other phages. Some annotated phages have the first two tRNA’s while others have only the last two tRNA’s. tRNA 1 (5’ 753, 3’ 836) is further away from the other two tRNA’s, but the presence of this tRNA and tRNA 2 (5’ 27504, 3’ 27577) in other genomes suggests that the distance might not be far enough to consider eliminating. tRNA 3 (5’ 34735, 3’ 34809) has an infernal score of 33.5, which is under the preferred threshold. It is unclear whether tRNA 1 or tRNA 3 should be eliminated. It was also noted that some other FF phage drafts possess three tRNAs (GoodLuckBabe, IsHungry, and Julie).

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)

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.