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. **DoobyDoo**

Your Name. **Pam Connerly**

Your Institution. **Indiana University Southeast**

Your email. **pconnerl@iu.edu**

Additional emails. (for correspondence). **erueschh@iu.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.

* We are calling Feature 10 (stop 8910) in DNA Master a Metalloprotease. It has a JAMM protease motif (ExnH xHx7Sx2D ). This was confirmed by cdshaffer in the forums: <https://seaphages.org/forums/topic/5672/>. The motif has been added to the function list, for metalloprotease.
* For Feature 19 (stop 13758) in DNA Master. HHpred has matches to both minor capsid protein and minor tail protein. Both BLAST and Phamerator indicate that this could be a minor capsid protein. We have called it a minor capsid protein, but are concerned about the HHPred hits calling it a tail protein.
* Feature 66 in DNA Master (Feature 65 (on Phamerator), stop 45445, has some of the features of a RecA-like DNA Recombinase, but not all of them. We checked the powerpoint listed in the forum. Our cluster (DV), falls under the category of not having a RecA recombinase but calling it anyway. Does this mean that we should not call it a RecA recombinase, or that we should call it a RecA, even though it is not a true RecA recombinase? On the spreadsheet, we have listed it as a RecA-like DNA recombinase, but are not sure if it should be called that.
* For feature 68 (stop 47155), we are a bit unsure about the function. BLAST and Phamerator suggest an oxidoreductase, but when we search on HHPred, we get six matched above 90% to an UDP-galactopyranose mutase. I did some research on UDP-galactopyranose mutases and found this quote: "Although UGMs catalyze a non-redox reaction, the flavin must be in the reduced state for activity and the exact role of the flavin in this reaction is controversial...." <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0032918>. There is no UDP-galactopyranose mutase in the function list, so we are not sure what to call the function. I have listed the function as an oxidoreductase since UDP-galactopyranose mutase is not on the function list.
* For feature 73 (stop 49551), there are no BLAST hits to support a function, but HHPred data suggests this could be a DNA primase. I labelled a DNA primase due to data from HHPred.
* Feature 74 (stop 50243) could be specifically a dehydrogenase or the more general oxidoreductase. The only dehydrogenase on the function list was UDP-Glucose dehydrogenase. I found no evidence that this was specifically a UDP-Glucose dehydrogenase in HHPred, so will list the function as the more general oxidoreductase.
* Feature 76 (stop 50619) has sizeable overlap of 53 bp. GeneMark very clearly shows strong coding potential for 76 in a different reading frame from that of 75 (stop 50561) and 77 (stop 50702). Any downstream alternative start for 76 (stop 50619) that would result in a smaller overlap would also result in lost coding potential.
* Feature 77 (stop 50702) also has a sizeable overlap of 55. The downstream start site of 50616 would have an overlap of 4, but the BLAST hits for that start are weak with the top hit having a 1:18 alignment and a second hit with 1:1 alignment but an e value of 2e-8. Starterator also favored the 50565 call with 4 manual annotations. There was some disagreement about this one, but we are going with 50565 based on BLAST and Starterator data. There appears to be a lot of variation in start calls for other DV phages for 76 (stop 50619) and 77 (stop 50702), plus a large gap after 77 – perhaps something interesting going on here, but not decipherable using only bioinformatic tools.
* Three auto-annotated features were deleted: Reverse feature 14800 – 14683, Reverse feature 35460 – 34861, and Reverse feature 50947 – 50831. These features are highlighted in light red in the Annotation Notes spreadsheet.
* Overall, we added 7 Features that were not auto-annotated in DNA Master: Feature 16 (12613 – 12729), Feature 22 (14637 – 14825), Feature 38 (31855 – 32049), Feature 58 (39646 – 39795), Feature 62 (42100 – 42261), Feature 70 (47822 – 48034), and Feature 96 (65719 – 65934). These features are highlighted light green in the Annotation Notes spreadsheet and are located at the bottom, after the auto-annotated features.

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?

**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)?

**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:

      PECAAN output

      DNA Master shorthand (previously used format)

**X** Spreadsheet

      Powerpoint

**X** 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.