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. Hannabella

Your Name. Breonna Martin

Your Institution. Harford Community College

Your email. brmartin@harford.edu

Additional emails. (for correspondence). jmadden@harford.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.

Below, we have summaraized the changes that we made to the auto-annotation, and then we note challenges that we ran into while annotating Hannabella. This was our first genome annotation, so we appreciate all feedback.

**Summary of feature changes:**

* Added features
* Stop @4994
* Deleted features
* Stop @477
* Stop @17295
* Stop @17633
* Stop @40995
* Changed starts
* Stop @1, New start @519
* Stop @2874, New start @3062
* Stop @5071, New start @5232
* Stop @34959, New start @32983
* Stop @39713, New start @38004
* Stop @44241, New start @42961
* Stop @51329, New start @50232

**Summary of protein functions (F:) called:**

1. Stop @589, F: DNA primase/polymerase
2. Stop @1277, F: DNA helicase
3. Stop @3065, F: DNA helicase
4. Stop @5071, F: ribbon-helix-helix DNA binding domain
5. Stop @5773, F: DNA polymerase I
6. Stop @7947, F: Cas4 family exonuclease
7. Stop @11548, F: AAA-ATPase
8. Stop @16444, F: DNA binding protein
9. Stop @17234, F: membrane protein
10. Stop @34959, F: portal protein
11. Stop @37993, F: minor tail protein
12. Stop @41005, F: minor tail protein
13. Stop @41783, F: endolysin
14. Stop @42270, F: membrane protein
15. Stop @48849, F: membrane protein
16. Stop @52924, F: terminase

**Summary of required proteins not found:**

* Major capsid protein
* Tail tube protein

**tRNA and tmRNA:**

* The complete genome FASTA file was run on two separate occasions using the following programs on the dates specified.
* Aragorn v1.2.41 on 3/07/22
* Result: no tRNAs or tmRNAs found
* TRNAscan-SE v. 2.0 on 3/107/22
* Result: no tRNAs found
* Aragorn v1.2.41 on 4/19/22
* Result: no tRNAs or tmRNAs found
* TRNAscan-SE v. 2.0 on 4/19/22
* Result: no tRNAs found

**Calls that required significant discussion**

* Gene 2 (Stop @589 )- 3 different starts by Starterator @1461, Glimmer @1263, GeneMark @1338. CP visible on GM to 1338. RBS scores and Starterator favored start @1461, but we did not choose that start because it had -184 bp overlap with Gene 3. We called the start @1263 by Glimmer because of the RBS scores and reasonable gap (+13 bp).
* Gene 7 (Stop @4994) - Small feature added in the gap. CP found on GM, RBS values Raw SD -1.74 Final -2.58, No ST calls length of 81 bp. Phages DB local protein blast- Arete\_gp8, Q:S 1:1, 100% identity and 5E-9 E-value, NCBIp BLAST- 100% identity and 2E-22 E-value, 100 sequences on NCBI, not a membrane protein, no high alignment on HHpred.
* Gene 9 (Stop @ 5773) matched mostly with hypothetical proteins on NCBI BLASTP, and PhagesDB Blast. When we ran this on HHPred, we found matches to Holliday junction revolvase, RuvC. However, we were unable to access many of the papers. What we could access (4EP4\_A), we didn't find that all of the functional residues were conserved so we didn't call the Holliday junction revolvase function. We may be mistaken. Here is a link for the paper: https://academic.oup.com/nar/article/41/1/648/1145164
* Gene 14 (Stop @9804) - only hypothetical protein matches with EM cluster phages on BLASTP. One significant HHPred hit 5ODJ\_A single-stranded DNA-binding protein, DNA replication Enterobacter phage. 95% prob. 75% align. This isn't an enzyme but we did look for conserved residues and couldn't find them all. We were not confident in calling this function. If we were to conservative in this decision, please let us know.
* Gene 16 (Stop @11548) – Glimmer call start @12744, RBS values Raw SD -7.267 Final -8.024 /w 57 GAP 12801-12744 and length 1197 bp. While GM calls 12504, RBS values Raw SD -1.462 Final -2.156 /w 297 GAP 12801-12504 and length 957 bp. At both lengths protein function AAA-ATPase is found. Start called by Glimmer @12744 was chosen over GeneMark for its length.
* Gene 26 (Stop @32986) is very long (13,465 bp), but there is no blast information for this gene and there wasn’t a protein function found. It is still unknown what the purpose of this gene would be. We attempted to blast small segments (~500 bp at a time) of this protein to check if it may have different protein functions but had no success.
* Gene 31 (Stop @ 39713) we had a lot of discussion around this protein. There were several hits with high alignments for proteins such as fibronectin, IGF-1, interleukin, Contactin-1.  None of those proteins seemed to be relevant to phages. While some of these are ECM or anchoring proteins, we couldn’t find hits for the collagen-like proteins to be able to call this a minor tail protein. We’d like to know if we missed something.
* Gene 33 (Stop @41783) we are calling this function “endolysin” since we were not able to find a second lysin in the genome. This function call is different from all other EM cluster phages, so we would like to make sure that it is the appropriate call based on the information in the approved functions list. Our phage infects *Microbacterium foliorum*.
* Gene 37 (Stop @44241) 3 different starts called by Glimmer @43093, GeneMark @42979, and Starterator @42961. The start called @42961 by Starterator was chosen because of the MA’s, the RBS scores, and this start provided the greatest length.

**There are several large gaps that we were unable to find CP or BLAST data to support filling:**

* Gap @ 10493-10605: viewing ST for gene 14 shows one start that would be able to extend the gene to start @ 10592 decreasing the gap to +12, GM output does not show additional coding potential for the start @10592, ST showed 3 MA’s for start @ 10493; Start @ 10493 was used as there is no evidence to support changing it; No coding potential was found between 10493-10605. Gap length 111 bp
* Gap @13274-13563: No coding potential found between 13274-13563; ST shows additional start for gene 17 @13322, @13346; GM did not show coding potential for either alternate start and RBS supports the gene 17 start. There is no evidence to support the presence of an additional gene in the gap @ 13274-13563. Gap length 288 bp
* Gap @15597 – 16444: No CP was seen between bp 15597 and bp16444; there was a potential gene start @ 15651 but no CP was found to support this start or to support the presence of a gene between bp 15597 and bp16444. CP was found @16296-16457 further blast searches prove low alignment; the gene was not called. Gap length 846
* Gap @17967- 18868: No coding potential found between 17967-18868; ST shows no additional start for gene 24. GAP length 900 bp
* Gap @19104 – 19502: there is a switch from reverse to forward. No coding potential found between 19104-19502; ST shows no additional start for gene 25 or for gene 26. GAP length 397 bp

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)

Y **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)

Y **Spreadsheet**

      Powerpoint

      Word document (must be easily searchable)

      Other: Describe.