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. Causa

Your Name. Joseph Stukey

Your Institution. Hope College

Your email. stukey@hope.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.

Causa is only the second member of one of the newest clusters – AI. Consequently, in general, it has a lower percentage of similar genes in our databases for comparison work. Many genes do NOT have significant or many times any BlastP hits. Further, the right arm genetic space has multiple sections that were fun but a challenge to annotate - those are listed below and in our Causa annotation notes slide file for possible review by the SMART team.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

original gene 8 (5211-5327) - No coding potential but ORF starts following gene 7 after a 1 base gap and does have a strong RBS (Z 2.855, Final -2.856). No significant NCBI BlastP hits. But a slightly longer version of this gene (all in same Pham 8108) has been called in about 10 phages. Short gene, suggest likely translated and therefore should be kept as a gene.

original gene 12 (7819-8010) - Although appears as an ORPHam (134839), it does have some (although chiefly atypical) coding potential. It also has a 4-bp overlap with an acceptable RBS (Z 1.884, Final -5.753). No significant NCBI BlastP hits. Kept as gene.

original gene 67 (53029-51033, REV) - can change start to 51081 based on strong CP up to this point in sequence (can we leave strong CP?); however, this change does not improve other metrics, is NOT identified in other annotated genes, and results in big reduction in RBS. Kept starting site as is based on preponderance of evidence - but still question strong CP through upstream sequence.

original gene 83 (61777-61986, REV) - initial auto-call leaves 80 bp gap, atypical CP exists in extended ORF region, can call start at 62976 resulting in 11-bp overlap and with equal RBS but no supporting BlastP results, all other data not informative. Decided to leave start at 61986.

original gene 100 (69837-70094, FWD) - auto-annotated start leaves 80 bp gap, small amount of CP left off but this start gives 1:1 BlastP alignment to StAugustine\_96; can move start up to 69789 but at large cost to RBS and still leaves 32-bp gap. Decided to leave start at 69387, not sufficient amount of evidence to change.

original gene 108 (73005-73199, FWD) - large gap with upstream gene. Gene can be extended with start 72933 and this will cover some missed CP. But this change comes at a large cost to RBS from Z 2,653 Final -4.128 to Z 0.467 Final -8.652. Further, see possible promoter-like sequence in this region. Extended start to 72933 to cover small amount of missing CP and gave 1:1 alignment with StAugustine\_105

original gene 112 (74457-74762, REV) - auto-call includes all of CP but leaves a gap of 22 bp. Some 1:1 BlastP alignments but others align to longer proteins. Extending the start generates a 4-bp overlap. New start RBS still strong but no BlastP 1:1 alignments. Starterator shows several adjustments but no good match to Causa gene. Decided to leave based on absence of CP in extended gene sequence and many 1:1 BlastP alignments with auto-call start. But I have to question whether a ribosome will use the earlier start?

original gene 146 (82940-83302, REV) - this start has a 1-bp overlap and largely aligns near 1:1 with BlastP hits, but it is missing some strong CP. Can extend start 5-codons upstream from 83317. This would capture more CP but also produce a 16-bp overlap. Most NCBI BlastP hits show this gene product to be slightly larger in alignments. This change would also bring in a slight decrease in RBS metric. Starterator not helpful. Decided to keep gene at original start (83302)

original gene 147 (82940-83302, REV) - another similar overlapping/CP gene case. This auto-called gene has a 22-bp overlap and captures CP. Can shorten the gene to 83496 but this would leave out CP and has significantly reduce RBS (original start Z 2.868 Final -2.845, start at 83496 Z 1.303 Final -6.063). BlastP and Starterator not helpful. Decided to keep this start to capture most/all of strong CP and retain strong RBS.

original gene 161 (88100-88282, REV) - auto-call has 4-bp overlap and covers most, but not all, of strong CP. No useful BlastP data (ORPHam gene). Strongest RBS (Z 2,612 FINAL -3.447). Missing CP can be covered moving start to 88306 but this would result in a 28-bp overlap and the poorest RBS (Z 0.824 FINAL -7.213). Decided to leave start at 88282 and with 4-bp overlap.

original gene 166 (89755-90015, REV) - auto-call has 4-bp overlap but misses some strong CP. RBS strong (Z 3.021 FINAL -2,584). NCBI BlastP hits data inconclusive - can be left at current gene size or some hits to longer proteins suggesting can be extended. Starterator not helpful. TO fully cover strong CP would need to move start to 90060, which also has good RBS (Z 2.201 FINAL -4.176) but wold result in an48-bp overlap. Decided to leave start at 90015.

original gene 173 (92449-92661, REV) - auto-call has 4-bp overlap but misses a little CP. Also has solid RBS (Z 2.192 FINAL -4.332). BlastP and Starterator not helpful. Can move start up to 92670 to cover all of CP. This position also has good RBS (Z 2.460 FINAL -4.454) but results in 13-bp overlap. Decided to leave start at current position 92661.

original gene 174 (92658-93026, REV) - almost identical situation as described for gene 173. auto-call has 1-bp overlap but misses a little CP. Also has acceptable RBS (Z 2.016 FINAL -5.213). BlastP and Starterator not helpful. Can move start up to 93071 to cover all of CP. This position also has better RBS (Z 2.526 FINAL -4.536) but results in 45-bp overlap. Decided to leave start at current position 93026.

original gene 181 (95339-95809, REV) and adjacent region (including alternative reading frame) up to gene 182 - 5'-end of original 181 has strange CP profile - 3'-end of gene has good CP but then there is a sharp break where CP goes to 0, then quickly rises before falling again to 0. The original start site is in the middle of the 5'-end CP spike (please examine GeneMarkS data). There is a large gap of of 280 bp to gene 183. There are additional upstream start sites (all with a bit better RBS) but it is a genome area that lacks any CP. There is strong CP in an alternative reading frame (same RF as gene 180) for which a gene can be called with approx coordinates 95599-96090 that covers all that CP (RBS: Z 1.946 FINAL -5.059). This CP is like that for gene 181 - 2 strong sections with a valley between. Further, this gene would overlap with original gene 181 by about 250-bp. This alternative reading frame gene is in our phagesdb database as an ORPHam gene StAugustine\_170 (the only other cluster AI phage and similar to Causa). No NCBI BlastP hits and no Starterator data. We decided to add this gene based on large stretch of strong CP even though it has a large overlap with genomic 5'-side gene (181). This added gene would have a 4-bp overlap with gene 182.

Several genes in this section really looks like one area of strong CP that has shifted into different reading frames.

original genes 183-185 - multiple auto-called genes (and all ORPHams) that overlap. All have some strong CP but genes 184 and 185 are the shortest in length. None of these genes have NCBI BlastP or Starterator data. All have acceptable RBS at auto-called start as well as alternative start sites. But auto-called genes 184 and 185 completely overlap other called genes that appear in the expected end-to-end arrangement in the genome. Decided to just keep original gene 183 as it is the largest and links up the upstream and downstream genes with short overlaps. We do not have any compelling reasons to include the additional alternative frame genes, despite covering some small sections of strong CP.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

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?

NA 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

YES DNA Master shorthand (previously used format)

YES Spreadsheet

YES 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)

      Spreadsheet

YES Powerpoint

      Word document (must be easily searchable)

      Other: Describe.