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: MulchExplorer
Cover sheet written by: Palina Hancharonak, Jennifer Fleury, Dinalis Jones, Narjiss Haouam, and Kristen Clermont
Your Institution: LaSalle University and Arcadia University
Your email: Kristen Clermont - clermontk@arcadia.edu
Additional emails. (for correspondence). Sean McClory: mcclory@lasalle.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.

Summary of annotation of gene locations and start sites

In Phage MulchExplorer, we deleted five genes from DNA Master auto-annotation (stop sites 4896 (R), 29634 (R), 32190 (F), 38436 (R), 47058 (R), and 53483 (R)). Annotation of deleted genes relied heavily on the absence of coding potential and interferences with nearby genes with high coding potential. Relative to the class's DNA Master auto-annotation we added 8 genes, many of which were already in the Phamerator auto-annotation (stop sites 32032 (R), 38671 (F), 44593 (F), 44797 (F), 44924 (F), 47274 (F), 49082 (F), and 53652 (F)).

Relative to the DNA Master auto-annotation, we changed 6 start sites. Auto-annotation consistently selected 55592 as the start site for the gene ending in 55798 (F), but we selected 55586. For genes ending in 4330 (F), 24102 (F), and 7353 (F) we changed the DNA Master auto-annotated start site to that selected by GeneMark in DNA Master. For the genes with stop sites at 35035 (R) and 47080 (F) only Glimmer annotated the gene in DNA Master, and we replaced the auto-annotated start site.

The gene with a stop site at 38074 (F) is a gene with a relatively small pham which would often indicate that it is not a real gene. However, strong evidence using GeneMark graphs from PhagesDB, synteny with similar genes, strong BLAST hits, and a 1000 bp long open reading frame supports that it is a real gene.

The tail assembly chaperone (gp 13 and gp 14) has a proposed programed frame shift. The sequence "RAESDSK" is where the frameshift possible have occurred. This was determined by using sequence from genes 13 and 14 of Kimberlium. Slippage occurs at "GAAA" in the nucleotide sequence.

Summary of biologically-relevant annotations

MulchExplorer contains all general viral functions such as terminase, HNH endonuclease, portal, capsid maturation protease, scaffolding protein, major and minor capsid protein, head-to-tail adaptor, head-to-tail stopper, tail terminator, endolysin A and B, and holin. MulchExplorer also contained common functions characteristic of the Siphoviridae morphotype like tail assembly chaperone, tape measure protein,

tailspike, tail tube, formerly called major tail protein, and numerous minor tail proteins including Dit, baseplate hub, and tail wing brush proteins.

According to PhagesDB, MulchExplor is a temperate phage. To support this, a cro protein was identified which is responsible for regulating the life cycle of a phage by switching between lysogenic and lytic states. This allows for additional immunity against super infection. HHpred did not show the presence of a cro protein among top hits. However, HHpred results from this gene with HHpred results when searching Che9_47, the standard example of a cro protein according to the Official Function List, retrieved the same top hit. This evidence supports the presence of the cro protein which was expected due to the presence of an immunity repressor within the genome. SEA-PHAGES | Official Function List

W annotated five helix-turn-helix DNA binding proteins. The function of this protein utilizes a structural unit found in various proteins particularly those involved in DNA binding. It is a common motif in transcription. DNA binding proteins are crucial in phages when it comes to gene regulation and aiding phages in bypassing bacterial immune systems.

Functional annotations edge cases for review

We gave two Siphoviridae tail **proteins annotations that are not currently on the approved function list**. These are "tail wing base, D-ala-D-ala carboxypeptidase" and "tail wing brush."

(1.) The name on the approved function list, "minor tail protein, D-ala-D-ala carboxypeptidase," does not take into account that this minor tail protein can be more specifically annotated based on Krista Freeman's Bxb1 structures in PDB. In HHpred, there is 100% probability and 66% coverage hit to 9D93_Sa from Krista Freeman's new Bxb1 structure ("tail wing base"). In Bxb1, there are three copies of the tail wing base, but they do not interact. AlphaFold folding of this MulchExplorer protein as a monomer yields a structure that aligns well with a Bxb1 tail wing base using Chimera (RMSD between 247 pruned atom pairs is 1.106 angstroms). The name "tail wing base, D-ala-D-ala carboxypeptidase" would be consistent with both the current naming and the naming in the Bxb1 publication.

(2.) Likewise, we believe we have strong evidence based on similarities with a Bxb1 protein that one of the other minor tail proteins is a tail wing brush. The gene ending at 26211 has a 99.5% probability and 99.6% coverage hit via HHPred to 9D93_Qc (tail wing brush) from Krista Freeman's Bxb1 structure. In Bxb1, there are three copies of the tail wing brush, but they do not interact. AlphaFold folding of this MulchExplorer protein as a monomer yields a structure that aligns well with a Bxb1 tail wing brush using Chimera (RMSD between 229 pruned atom pairs is 0.596 angstroms). The most appropriate naming depends on the current consensus, but we wanted to suggest this level of specificity.

The gene ending at 8272 is a tail tube protein, formerly known as a major tail protein. The Approved Function List lists both "tail tube" and "tail tube protein." Both annotations apply equally well.

Genes with a stop site at 30263 (F), 31719 (F), 31829 (R) were identified as a membrane protein based on the results of DeepTMHMM. Proteins that only contain one transmembrane protein domain located at the N-terminus were not called membrane proteins. The gene that ends at 30263 (F) is in a pham of 829 members, most of which are annotated as NKF and some as minor tail protein (for which we did not find evidence). It has only been annotated as a "putative membrane protein" twice and never directly as a "membrane protein." The gene ending in 31719 (F) is in a pham with 258 members and the gene ending at 31829 (R) is in a pham with 55 members. **In both cases all of the members of the pham are annotated as having no known function. However, DeepTMHMM results suggest that "membrane protein" is appropriate in each of these cases.**

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):

Y 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.)?
Y 2. Are all the genes 'Valid" when you click the Validation button?
Y 3. Are the genes (and matching LocusTag numbers) sequential, starting with #1, counting by 1s.
Y 4. Are the Locus Tags the "SEA_PHAGE NAME" format?
Y 5. Has the documentation been recreated from the Feature Table to match the latest file version?
NA 6. Have tRNAs followed the tRNA protocol, **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?
We think so. 7. Has the frameshift in the tail assembly chaperone been annotated correctly (if applicable)?
Y 8. Have you cleared your Draft_Blast data and have you re-Blasted the submitted DNA Master file?
Y 9. Has every gene been described and supported in your Supporting Data file?
Y 10. Did you investigate 'gaps'?
Y 11. Did you delete the genes that you meant to delete?

Now, make a profile of the file you plan to send. (And you can save this file for Review to Improve!)

Y 1. Have any duplicate genes been deleted?
Y 2. Has the Notes field been cleared (using the automated buttons)?
Y 3. Do the gene numbers and locus tags match?
Y 4. Are the correct Feature_Types correctly selected (most will be ORFs, but check that tRNAs and tmRNAs are correctly labeled)?
N 5. Do the function names in the Product field either match the official function list or say "Hypothetical Protein"? [These cases are listed above in the text of the cover letter]
Y 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.:

X PECAAN output (formatted as a Word document, with annotations of interest highlighted)
DNA Master shorthand (previously used format)
Spreadsheet
PowerPoint
Word document (must be easily searchable)
Other: Describe.