Genome Annotation Submission Cover Sheet

Pre-QC Phage Genome Annotation Checklist

Phage Name:	Craff
Your Name:	Joyce Stamm and Ann Powell
Your Institution:	University of Evansville
Your email:	js383@evansville.edu
Additional emails: (For correspondence)	ap96@evansville.edu

Please check each box indicating completion of each task. Annotation Guide section #'s indicated

- [x] 1. Does the genome sequence in your final contain the same number of bases and is it the same as the posted sequence on phagesdb.org?
- [x] 2. Are all the genes "valid" when you click the "validate" button? *Section 9.3.2*
- [x] 3. Have the genes been renumbered such that they go sequentially from 1 to the highest number? *Section 9.3.3*
- [x] 4. Have all old BLAST hits been cleared, and all gene features reBLASTed? *Section 9.3.4*
- [x] 5. Are the locus tags the phage name? *Section 9.3.3*
- [x] 6. Has the Documentation been recreated to match the information in the feature table? *Section 1.4*
- [x] 7. Have tRNA ends been adjusted with web-based Aragorn and/or tRNAscan SE? *Section 9.5.3-4*

8. For the items below, generate a genome profile, and review the following. *Section 11.3*

For the YourPhageName_CompleteNotes.dnam5 file:

- [x] a. Have any duplicate genes (or any with the same stop coordinate?) been removed?
- [x] b. Does every gene have **one and only one** complete set of Notes (see fig 12.2 in the Annotation Guide)?
- [x] c. Do the functions in the Notes match the official function list?
- [x] d. Is the function field EMPTY for all features?
- [x] e. Do the notes contain the initial Glimmer/GeneMark data from the autoannotation?

For the YourPhageName .dnam5 file:

- [x] a. Have any duplicate genes (or any with the same stop coordinate?) been removed?
- [x] b. Is the Notes field empty for all the features with no known function?
- [x] c. Do the function names in the Notes match the official function list, when applicable?
- [x] d. Is the function field EMPTY for all features?

9. Describe any issues or specific genes that you were unable to satisfactorily resolve, and warrant further inspection in the Quality Control review.

See next page.

Here are the regions of concern:

CRAFF_9: This had many BLAST hits with capsid morphogenesis protein, which is not on the approved list, so we entered NKF.

CRAFF_20: This had many BLAST hits with queuine tRNA ribosyltransferase, which is not on the approved list. Since this enzyme is a glycosyltransferase, that's what we entered.

CRAFF_39: SSC: 36244-36612 (FWD). There is another possible start at 36238. We picked this start because most of the 1:1 matches are with the shorter ORF.

CRAFF_61: This had BLAST hits to genes with a number of different names, including replicative helicase, DNA helicase, DNA primase/helicase, DNA primase/polymerase. We entered DNA primase/polymerase based on the top HHpred hits, but there were also good hits for DNA helicase.

CRAFF_65: SSC: 55543-55695(REV). There are two adjacent starts. We called the second one because we were told that the evidence suggests the second start. However, the first start will give a 4 bp overlap.

CRAFF_66: SSC: 55695-56132 (REV). We called an ORF with a 32 bp gap so as not to truncate conserved coding sequence. The starterator suggested start at 56051 is called 81% of the time, but mostly in draft sequences. The start we called is called 12.9% of the time, but not in any draft sequences.

CRAFF_68: SSC: 56580-56789 (FWD). This is an ORF at that was not called by Glimmer or GeneMark.

CRAFF_69: SSC: 56888-57742 (FWD). This gene start was called by Glimmer at 57011, but we included an additional 123 bp upstream so as not to truncate the ORF. The start we called is annotated only 10.9% of the time, but all in non-draft sequences.

CRAFF_77: SSC: 61072-61401 (FWD). This gene start was called by Glimmer and Starterator at 61165. But we called the start at 61072, which is called only 2.8% of the time (all in non-draft sequences), so as not to truncate the ORF. This is also the start called in the closely related phage Orion. However, this start gives a 35 bp overlap, which could be problematic..

CRAFF_78: SSC: 61683-61477 (REV): There are two adjacent starts. We called the second start.

CRAFF_89: SSC: 64764-64543 (REV): There are two adjacent starts. We called the second start, this start also gives a 4 bp overlap with the previous gene.

Thanks!