

Genome Annotation Submission Cover Sheet

Pre-QC Phage Genome Annotation Checklist

Phage Name:

Your Name:

Your Institution:

Your email:

Additional emails:
(For correspondence)

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

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?
2. Are all the genes "valid" when you click the "validate" button? *Section 9.3.2*
3. Have the genes been renumbered such that they go sequentially from 1 to the highest number? *Section 9.3.3*
4. Have all old BLAST hits been cleared, and all gene features reBLASTed? *Section 9.3.4*
5. Are the locus tags the phage name? *Section 9.3.3*
6. Has the Documentation been recreated to match the information in the feature table?
Section 1.4
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:

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

For the YourPhageName .dnam5 file:

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

GuuelaD Annotation Changes and Comments
SEA-PHAGES Program
Academic Years 2015 – 2016, 2016 – 2017
Dr. Michael Rubin

Auto Annotated Gene # 3

Start866 - Stop1069

Start Site Changed from 875 to 866

875 was called by Glimmer. It results in a 9 bp gap with previous gene.

866 was called by Genemark, supported by Starterator, it comprises the Longest ORF and it results in a 1 bp overlap with previous gene. It is also supported with better q1: t1 blast matches with other phages and has the best SD score among all the possible start sites.

Auto Annotated Gene # 16

Start10890 - Stop16019

Start site was changed from 10926 to 10890

10890 was called by Glimmer and Genemark, but not Starterator.

It doesn't cover all the coding potential

All blast matches agree with the 10890 start

Auto Annotated Gene # 23

Gene Deleted

- Only called by Glimmer and not Genemark.
- Very short length (123bps).
- No blast matches
- 37bp overlap with gene 24

Auto Annotated Gene # 28

Start25387 - Stop26337

Changed Start from 25438 to 25387

- Called by Genemark and Starterator,
- Covers all Coding Potential
- Gets better blast matches

Auto Annotated Gene # 33

Start28456 - Stop28722

Changed start from 28501 to 28456

- New start called by Genemark and Starterator,
- Covers all Coding Potential
- Gets better blast alignments
- Better SD score

Auto Annotated Gene # 35

Stop29142 -Start29429 Rev

Changed start from 29399 to 29429

- New start called by Genemark and Starterator,
- Covers all Coding Potential
- Gets better blast alignments
- Better (Best) SD score

Auto Annotated Gene # 45

Deleted gene

- Short length (189)
- Big Overlap with previous gene (97bp)
- Start site (35210) only called by Glimmer
- Very low SD score (-8.563)
- Alternative start site predicted by Genemark and Starterator results in a 111bp length
- Only 1 Blast match alignment

Auto Annotated Gene # 46

Start35386 - Stop35811

Changed Start from 35482 to 35386

- New start covers all Coding Potential
- New start results in better Blast alignments
- New start was causing an overlap with now deleted gene (45)
- New start now reduces a gap of 175bp to 79bp

Auto Annotated Gene # 47

Deleted gene

- Very Short length
 - No Blast matches
 - Not predicted by Genemark
- *Everything applies to both possible start sites

Auto Annotated Gene # 49

Start36643 - Stop37677

Changed start from 36841 to 36643

- Covers all Coding Potential
- Has a better SD score
- It results in better BLAST alignments (with function)

Auto Annotated Gene # 50

Start37658 - Stop37993

Changed start site from 37700 to 37658

- Called by Genemark and Starterator
- Covers more Coding Potential

- Has a better SD score
- It results in better BLAST alignments

Auto Annotated Gene # 52

Start38305 - Stop38433

This gene should be strongly considered for deletion:

- Very short length (129bp)
- Only one good Blast alignment.
- No putative function predicted.
- Other possible start sites increase length but don't result in better Blast results and cause excessive overlap with previous gene.

Auto Annotated Gene # 53

Start38470 - Stop38598

This gene should be strongly considered for deletion:

- Very short length (129bp)
- Only one good Blast alignment.
- No putative function predicted.
- No other possible start sites that would increase length or result in better Blast matches.

Both gene 52 and 53 were left during annotation because deletion of both consecutive genes would result in a 389 bp gap and no other possible ORFs were found in any of the frames in this segment.

Auto Annotated Gene # 61

Start43197 - Stop43373

The start site of this gene was changed from 43182 to 43197.

This is a peculiar case.

- Each of the three logarithms used calls for a different start site:
 - Glimmer – 41382
 - Genebank – 43245
 - Starterator – 43197
- Although the Glimmer call results in the longest ORF possible, any of the start sites results in an ORF below the length limit for careful scrutiny (200bps).
- Among the three, 41382 has the best Final SD Score (-3.699), and 43245 the worst one (-5.541).
- Only the Starterator start call results in a gene with perfect Blast alignments.

The chosen start site was based on the perfect Blast matches, Starterator prediction, and good SD score. A good case could still be made for choosing the start site predicted by Glimmer.

Auto Annotated Gene # 64

Start44524 - Stop44838

The start site of this gene was changed from 44557 to 44524

- Predicted by Genemark and Starterator.
- Lengthens the ORF of the gene (Includes al CP).
- Better (best) Final SD Score.
- Better Blast matches.

The next three changes (two deletions and an addition) were made at the 46431 – 47018 segment of the genome. These should be carefully evaluated.

Auto Annotated Gene # 69

Start46478 - Stop46879

Gene deleted.

- Only reverse gene in its segment (with forward genes immediately upstream and downstream of it).
- Starterator calls for a start site on a different frame, where there are no putative ORFs.
- Blast only results in one match, with a q1:s6 alignment and no determined function.
 - Although a downstream start site results in a better Blast alignment with the only match (and slightly longer ORF), it has a lower Final SD score and it was not predicted by any of the algorithms.
- Relatively short length (261).
- Gene causes big overlap with later included forward gene.

Auto Annotated Gene # 70

Gene deleted.

- Starterator doesn't seem to call the gene, assigning the GuuelaD_Draft_70 the start site identified for DNAM_71 on DNAMaster.
- Blast only results in one match.
- Relatively short length (234).
- Gene causes big overlap with later included forward gene.

Gene # 69 (Added)

Start46478 - Stop46879

New gene added at start 46478 – stop 46879

- Gene call results in perfect blast matches with predicted function
- Gene addition changes a 587 gap (after previous consecutive deletions) to a 47 gap with previous gene and 39 gap with next gene.
- Note: Gene call doesn't seem to be supported by Glimmer, Genemark or Starterator.
- Note: Other possible start site (46460) results in a slightly longer ORF without causing overlap but compromises the Blast matches.
- Start site that results in good blast matches (46478) has one of the lowest SD scores.

Auto Annotated Gene # 83

Start54093 - Stop54329

No changes made

Other possible start site called by Genemark (54033) represents a better choice in every aspect (best SD score, longer ORF, better Blast matches...), except for the fact that it causes a 52 overlap with previous gene. Based on this the original Glimmer call (54093) was still chosen.

Auto Annotated Gene # 85

Start54724 - Stop54822

****No changes made****

Despite its extremely short length, gene 85 was kept based on prediction by the three algorithms in use, and the fact that it doesn't cause excessive overlap with adjacent genes. Nevertheless, this gene should be carefully examined.

Auto Annotated Gene # 86

Start54813 - Stop55382

The start site of this gene was changed from 54879 to 54813.

- Covers all coding potential.
- Longer ORF
- Better Blast alignments

Auto Annotated Gene # 88

Start55897 - Stop56115

The start site of this gene was changed from 55921 to 55897.

- Covers all coding potential.
- Longer ORF
- Better Blast alignments
- Best Final SD Score

Auto Annotated Gene # 89

Deleted gene

Despite it being called by Genemark and Starterator, the gene was deleted because of its considerable short length (72 bps) and few to none Blast matches.

Auto Annotated Gene # 91

Start57444 - Stop57623

The start site of this gene was changed from 57483 to 57444.

- Longer ORF
- Better BLAST Matches
- Covers more coding potential
- Better Final SD Score

Auto Annotated Gene # 92

Start57623 - Stop58291

The start site of this gene was changed from 57620 to 57623.

- Longer ORF
- Better BLAST Matches
- Better Final SD Score

Auto Annotated Gene # 103

Start62093 - Stop62311

The start site of this gene was changed from 62132 to 62093.

- Supported by Genemark and Starterator
- Longer ORF
- Better BLAST Matches

Auto Annotated Gene # 116

Stop66529 -Start67245 Rev

The start site of this gene was changed from 67221 to 67245.

- Supported by Genemark and Starterator
- Longer ORF
- Better BLAST Matches

Auto Annotated Gene # 117

Stop67242 -Start67430 Rev

The start site of this gene was changed from 67319 to 67430.

- Supported by Starterator
- Longer ORF
- Better BLAST Match