Faculty Guide - Resources for Submission of Quality Phage Genome Annotations

## Overview

The ever-decreasing cost of full genome sequencing and the development and widespread availability of genome annotation programs have increased genome annotation accessibility. Auto-annotation programs, such as Glimmer and GeneMark, are often used for gene prediction, but manual curation of genomes can improve gene calling and detect nuances in the genome. Phage genomes are particularly suited for manual curation because the genomes tend to be small (<250kbp). In the HHMI SEA-PHAGES courses, manual curation of phage genomes provides students an opportunity to participate in authentic research while learning genetics, bioinformatics, proteomics, and data science.

One of the goals of the SEA-PHAGES program is the submission of a high-quality genome annotation to GenBank. To ensure the best quality annotation is achieved by the end of the term, it is important to establish early best practices for data acquisition, interpretation, and note-keeping. The submission of a high-quality annotation starts with the student’s work. By providing students with tools to accurately and completely interpret the data from prediction programs, faculty should receive a nearly complete genome annotation by the end of the term thereby minimizing the time faculty spend checking annotations prior to GenBank submission.

This set of resources provides tools for students and faculty to use to ensure the highest quality phage genome annotation is submitted at the conclusion of the term.

### Summary of Resources

* Faculty Guide - Resources for Submission of Quality Phage Genome Annotations
* Student Genome Annotation Worksheet
* Student tRNA Annotation Worksheet
* Genome Annotation Summary Spreadsheet
* Student Group Quality Control Checklist
* Whole-genome Starterator Output File Instructions

###

### Learning Objectives

After completing this module, students should be able to:

1. Use a suite of bioinformatics programs to determine whether an ORF should be called a gene
2. Analyze data from a suite of bioinformatics programs to make a determination of the most likely start codon for a gene
3. Explain why the start codon selected for a gene is the most likely start codon
4. Find and annotate tRNAs in phage genomes
5. Use a suite of bioinformatics programs to assign a putative function to a called gene
6. Explain why an assigned function is appropriate for a particular putative gene.
7. Evaluate various elements of a typical annotated phage genome to identify inconsistencies and potential areas of improvement (putative gene calls, frameshifts, large gaps, membrane proteins, putative function assignments)

After completing this module, faculty should be able to:

1. Actively involve students in the genome quality control/peer reviewing process.
2. Use the DNA Master Frames window, the genome comparison tool in DNA Master, and/or Starterator to quickly identify putative gene calls and/or significant gaps that warrant further review.
3. Quickly evaluate called functions for all putative genes in a phage genome.
4. Create correctly formatted complete notes and minimal DNA Master files when using PECAAN as a supplementary annotation tool.

### Development of Students as Scientists

Phage genome annotation work has many opportunities for supporting students as developing scientists. By using the resources provided in this set, students have a set of tools that empowers them to complete a challenging project while working collaboratively with their peers. By nature of the work, students will face ambiguity while analyzing data from multiple bioinformatics projects but the tools provided, along with instructor mentorship, will provide a format for them to work through the ambiguous data to make the best gene start decisions possible.

###

### Implementation Notes

#### Student Resources

Four resources are provided to assist students in making high-quality genome annotations. Students are encouraged to work collaboratively to ensure all bioinformatics programs have been properly evaluated for each putative gene. Implementation suggestions using the provided student resources are outlined below.

Preparing for Annotation (before your genomes are returned)

* + **Use the Student Genome Annotation Worksheet** to have the entire class work through annotating the same gene, assigning each section as homework (individual or group) and then discussing choices during the following class, with the entire class coming to a consensus on the gene call. You may want to use a practice genome for this assignment while you are waiting to receive your final genome sequences. Consider posting a final copy of the template for the practice gene so that students have a reference for how the template should be completed.
	+ **OPTIONAL: Use the Student tRNA Annotation Worksheet** to have the entire class work through annotating the same tRNA gene and discuss as a class if you plan to cover tRNA annotation in your course.

During Annotation of a Novel Bacteriophage Genome:

* + **Use the Student Genome Annotation Worksheet** to have teams or individual students annotate assigned sections of the bacteriophage genome. You may want to assign regions of the genome (bp1 - bp1000, for example) instead of specific gene numbers, since gene numbers depend on auto annotations and may differ among various programs. Assigning regions of the genome also may remind students that they are responsible for checking gene gaps in this area. Students can use these sheets to keep track of their rationale for making changes, and you can provide feedback on these sheets to make sure students are on track with the annotation process. You can also require students to share these documents with each other during the peer quality control genome review process, since the SEA notes format may not completely capture the annotation thought process.
	+ **Use the Genome Annotation Summary Spreadsheet** to keep track of the changes students make to their genome. Gene coordinates can be obtained from a Genome Profile (see Video Tutorial: [Creating a Genome Profile in DNA Master](https://pitt.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=bbd11644-6cf2-4eca-9f35-ac0c00ef7176)) and copied into the template. The template is coded with pre-filled dropdown boxes to indicate whether or not students changed a start site and for peer reviewers to select a gene review status (checked-no change; checked-change start site, checked-change function; checked-change start site and function). This worksheet is available for the entire class to help them keep track of what changes have been made to the auto annotation. Genes can be marked for deletion in the notes field, and genes can be added by inserting rows into the document.

Peer Quality Control Review of the Genome

* **Use the Student Group Quality Control Checklist**to help students guide their own review process of their annotations. This sheet can also be used to help faculty summarize changes made to the genome. This sheet can be completed by each annotation group and then provided to peer reviewers.
* **Assign peer quality control reviewers in the Genome Annotation Summary Spreadsheet** and direct students to enter all verification data in this spreadsheet.

#### Faculty Resources

The number one way to reduce the amount of time a faculty member must spend checking a phage genome and preparing a phage genome for GenBank submission is to establish a process for students to present the most complete phage genome annotation. Once student work is complete, it is prudent for faculty to spend 1-2 hours checking the quality of the annotation prior to PhagesDB submission using one or more of the tools described below.

##### Mechanics of using PECAAN (optional)

SEA-PHAGES faculty use PECAAN in many different ways in their courses, ranging from PECAAN serving as the primary tool that students use to annotate a genome to only a source of data for faculty review of student annotations. Regardless of how you decide to use PECAAN in your course, we do recommend that you generate PECAAN pages for your institution’s phages to help you in the quality control review of your students’ work. Having access to data from multiple sources in one page will help to drastically reduce the amount of time you spend gathering data to review student annotations before submission to PhagesDB.

* [**Upload your phage into PECAAN (7:54)**](https://pitt.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b7d76085-c608-48bb-858a-ac010114641f) if you are using this tool in your course. SEA-PHAGES faculty use PECAAN in many different ways, ranging from the primary student annotation platform to a faculty-only resource used for quality control. Regardless of how you decide to use PECAAN in your course, we do recommend that you use it to aid in your quality control review of your students’ annotation work.
* [**Create a DNA Master file from PECAAN (9:25)**](https://pitt.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=88be7083-c942-4e21-97f4-ac010116edae) if you used PECAAN in your course for students to annotate a genome. This video tutorial will demonstrate how you can transfer a complete PECAAN annotation into DNA Master. Note that some additional formatting in DNA Master is required to generate the complete notes file for submission to PhagesDB.

##### Tools for Organizing Student Annotation Work

* **Use the Genome Annotation Summary Spreadsheet** to quickly summarize the changes students made to the genome annotation and to facilitate peer quality control reviews of other students’ work.

 Video Tutorial: [Creating a Genome Profile in DNA Master](https://pitt.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=bbd11644-6cf2-4eca-9f35-ac0c00ef7176) (2:50)

##### Tools for Checking Start Sites

* **Use the whole genome Starterator output file** to identify start sites that differ from the most commonly called starts in Starterator. *Please note that you will need to have installed Virtual Box and a recent* [*HHMI SEA-PHAGES Virtual Machine*](https://seaphages.org/faculty/information/#bioinformatics) *to generate the whole genome Starterator output file.*

Video Tutorial: [Whole-genome Starterator Output File Video Tutorial](https://pitt.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3a272092-836f-4e19-9447-ac0c00f1ef98) (5:22)

Written Instructions: **Whole-genome Starterator Output File Instructions**

Video Tutorial: [Using Whole Genome Starterator Reports to Evaluate Student Annotations](https://pitt.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=d4a54809-94f9-421f-8f28-ac0c00f8b4a1) (8:21)

* **Use the Genome Comparison Tool in DNA Master** to visually identify putative gene calls with start sites that differ from genes called in other, similar phages. First, identify the names and GenBank accession numbers of 2-3 annotated phages similar to the phage you are annotating. If your phage is a singleton, you may consider selecting 1-2 most similar phages for comparison or consider using a different listed tool.

 Video Tutorial: [Genome Comparison Tools in DNA Master Video Tutorial](https://seaphages.org/video/64/) (8:30)

Written Instructions: [Genome Comparison Tools in DNA Master Instructions](https://seaphagesbioinformatics.helpdocsonline.com/article-89)

* **Use the Frames Window in DNA Master** to scan putative gene calls and identify start sites that may need additional review (large overlaps, significant gaps)

Video Tutorial: [Checking Student Genome Annotations: Start Sites](https://pitt.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=acefa3d7-d475-45cd-a177-ac01011fa81e) using the Frames WIndow in DNA Master (16:11)

##### Tools for Checking Function Assignments

For many phage genomes, it will not be possible to assign a function to a significant proportion of the putative gene calls. Yet, there are some common functions to a putative gene call in nearly all phage genomes. Prior to the start of the term, faculty are encouraged to review the information posted in the [Bioinformatics Guide: Predicting Phage Gene Functions](https://seaphagesbioinformatics.helpdocsonline.com/article-39) for an overview of the data required to support function predictions.

Please note that the evidence required to support assignment of a function to a putative gene call varies depending on the function. Faculty and students should reference the [SEA-PHAGES Official Function List](https://seaphages.org/blog/2017/10/30/official-function-list/) for the most up-to-date information regarding function names and important notes as well as the [Functions Present in (Almost) All Phage Genomes](https://seaphagesbioinformatics.helpdocsonline.com/article-91) to review the annotation of essential genes in the genome.

* For every phage genome, you should be able to identify:
* terminase - at least 1, at most 2
* portal - only 1. If more than one gene appears to be a portal protein, see the [Bioinformatics Guide: Portal and head-to-tail connectors](https://seaphagesbioinformatics.helpdocsonline.com/portal-and-head-to-tail-connectors) for additional guidance
* capsid maturation protease - at least 1
* major capsid protein - most genomes have 1
* major tail protein - only 1
* tail assembly chaperone - usually 2
* tape measure protein - longest gene, only 1
* minor tail proteins - downstream of tape measure, 1-3kb in length, 4-6 per genome
* lysin - at least 1
	+ Lysin A, at least 1 (if no Lysin B is located, annotate as “endolysin”)
	+ Lysin B - must align to cutinase domain, may not be present
* The following are functions that are likely present, but you should not be alarmed if these functions were not identified in your phage genome
	+ HNH endonuclease – often the first gene, last gene, or near the terminase, but may be present in other locations in the genome.
	+ scaffolding protein – may or may not be present, if present 1 copy. Near major capsid protein subunits
	+ Holin– 100aa protein with multiple transmembrane domains
* Confirm that any putative gene assigned the function ‘membrane protein’ is supported by data from **two** different bioinformatics programs. For additional guidance evaluating membrane proteins, see the [Bioinformatics Guide: Using TMHMM and SOSUI to predict membrane proteins](https://seaphagesbioinformatics.helpdocsonline.com/article-102).
* Check to see if the phage genome has any tRNAs. Reference the [Bioinformatics Guide: Predicting tRNA and tmRNAgenes](https://seaphagesbioinformatics.helpdocsonline.com/article-40) to ensure tRNAs have been identified and trimmed in accordance with up-to-date guidance from the SEA-PHAGES team. The checklist below is intended to serve as a guide but is subject to change.
	+ Aragorn (v1.1) --run during auto-annotation in DNA Master--- finds the basic tRNA genes, however, frequently the ends of the genes are not correct.
	+ Aragorn (v1.2.38) (http://130.235.46.10/ARAGORN/ ) finds the basic tRNA genes, and adjusts the ends of the genes correctly. *Use this program to trim the tRNA to the correct length.*
	+ tRNA ScanSE 2.0 (http://trna.ucsc.edu/tRNAscan-SE/ ) finds non-canonical tRNAs