# **Table of Contents**

How	v to use this guide	5
1 I	ntroduction to DNA Master	7
1.1		
1.2		
1.3		
1.4	-	
	alysis programs running within DNA Master	
	1.4.1 Glimmer	
1	1.4.2 GeneMark	9
1	1.4.3 Aragorn	10
1.5	5 Setting Preferences	
	1.5.1 Set Default Translation Table & Template Insertion (Local Settings Tab)	
	1.5.2 Set color preferences	
	1.5.3 Set start codon choices	
	1.5.4 Set default values for BLAST searches	
	<ul> <li>1.5.5 Choose a default location for saving files</li> <li>1.5.6 Finishing up your Preference settings</li> </ul>	
1.6		
1.0		
1.7		
1.0		
2 P	Provisional Cluster assignment of your phage	17
2.1	l Overview	
2.2		
2.3		
<b>э</b> т	monting your phase general coguence into DNA Master	22
	mporting your phage genome sequence into DNA Master	
3.1		
3.2		
3.3		
3.4	Reverse-complementing your sequence	
4 P	Performing and viewing a rapid automated annotation of your ge	nome 27
4.1		
4.2		
4.3	-	
4.4		
	4.4.1 Viewing the documentation	
	4.4.2 Viewing features in the Feature Table	
Z	4.4.3 Viewing the sequence in the Sequence tab	
4	4.4.4 Viewing ORFs in the Frames window	34
4.5	0	
4.6	6 Re-opening an archived (saved) file	

5	Gathering additional information for refining your annotation	
	5.1 Generating a six-frame translation	
	5.2 Generating a provisional genome map in DNA Master	
	5.3 Generating a graph of coding potential using GeneMark	
6	Phamerator & other Tools to assist with annotation	
	6.1 Phamerator	
	6.1.1 Overview	
	6.1.2 Why Phamerator is useful to you at this stage of your annotation	
	6.1.3 How did my genome get into Phamerator already?	51
	6.1.4 Making Phamerator maps	
	6.1.5 Understanding and using the genome maps made by Phamerator	
	<ul><li>6.1.6 Viewing nucleotide sequence similarities in Phamerator</li><li>6.1.7 Other Phamerator features</li></ul>	
	6.1.8 Saving Phamerator maps	
	6.2 Starterator	
	6.2.1 Overview	
	6.2.2 Interpreting Starterator results:	
7	Cuiding Dringiples of Destanianhage Conome Appetation	()
7	Guiding Principles of Bacteriophage Genome Annotation	
	7.1 Overview	
	7.2 The Guiding Principles	63
8	Gene by gene: evaluating and improving your draft annotation	
	8.1 Overview	
	8.2 Button-pushing mechanics reserved for Section 9	
	8.3 Decision Tree for evaluating the draft annotation	
	8.4 Evaluating protein-coding gene calls	
	8.4.1 Is the designation of this ORF as a gene well-supported?	
	8.4.2 Is the called start site for this gene the best possible choice?	72
	8.5 Checking gaps in the draft annotation for uncalled genes	77
	8.6 Finding and refining tRNA and tmRNA genes	
	8.7 Completing your annotation refinement	78
9	The mechanics of making changes to your annotation	
	9.1 Overview	
	9.2 Making common changes to your annotation	
	9.2.1 Deleting a gene	81
	9.2.2 Adding a gene	
	9.2.3 Changing the start site for a gene	
	9.3 Common steps to take after making changes	
	9.3.1 Posting changes	
	9.3.2 Validating your annotation	
	9.3.3 Renumbering & formatting annotated features	
	9.3.4 Re-BLASTing a gene	
	9.4 Making less common changes to your annotation 9.4.1 Annotating programmed translational frameshifts	
	9.4.1 Annotating programmed translational namesints	

9.4.3 Annotating wrap-around genes	95
9.5 Predicting tRNA and tmRNA genes	95
9.5.1 Running web-based Aragorn (version 1.2.36)	95
9.5.2 Running tRNAscan-SE (version 1.23)	97
9.5.3 tRNA secondary structure and end determination	
9.5.4 Entering a tRNA in DNA Master	
9.5.5 Identifying and annotating tmRNA genes	
9.6 Documenting your gene calls	
10 Assigning gene functions	105
10.1 Overview	
10.2 Using bioinformatic tools to assign gene function	
10.2.1 BLASTP	
10.2.2 Conserved Domain Database (CDD)	
10.2.3 HHpred	
10.3 Other ways to assign gene function	
10.3.1 Synteny	
10.3.2 Prior functional assignments	
10.3.3 Phamerator	
11 Merging and checking annotations	117
11.1 Merging overview	
11.2 Merging multiple annotations into a single file	
11.3 Checking an annotation	
12 Submitting final files for review and GenBank submission	
12.1 Details of your final DNA Master (.dnam5) file	
12.2 The second "minimalistic" Final .dnam5 file	125 176
12.2 The second minimalistic Final unants file	
-	
12.4 Details of your cover sheet	

# How to use this guide

Once you have a finished phage genome sequence, you are ready to make predictions as to the locations and functions of the tRNA-coding and protein-coding genes. This guide will provide step-by-step instructions as to how to do this.

There are several different ways you can use this guide.

- Begin at **Section 1**, and proceed section by section through the entire guide. This approach will give you a complete understanding of the entire process of annotation and how each of the programs involved works. It's a lot of information, but hopefully you'll emerge from the other side far more knowledgeable about genes and gene calling.
- If you've already used the **DNA Master Quick Start Guide** to create an automated annotation, you can jump in at **Section 5**, and proceed from there. You'll be skipping some basics, but you can always refer back to relevant sections if needed.
- If you're eager to get straight to gene calling, you can perform an automated annotation using the **DNA Master Quick Start Guide** or **Section 4** of this guide, then proceed to **Section 8** which covers how to refine your automated annotation. References back to previous sections are provided so that you'll be able to locate all the information you need.
- If you're already an experienced annotator, and all you want to know is how to push the correct buttons to modify gene calls in DNA Master, **Section 9** is for you. It's an à-la-carte section of "How-To" functions.
- Finally, even if you're accustomed to using a different program to annotate phage genomes, you can use the Guiding Principles described in **Section 7.2** to see how we think about making the best possible gene calls in phage genomes.

# A NOTE ON CLASSROOM PRAGMATICS

If you have a group of students annotating a single genome there are several different ways of organizing this activity. Assuming you have a class of around 20 students, there are two main considerations.

- 1. It works well for students to work in pairs, if possible using two computer stations. One of these can be set up to run DNA Master, while the other is set up to run Phamerator, as well as having other files (such as a six-frame translation) open.
- 2. You can organize students or groups of students such that:
  - All students annotate all of the genome. Upon completion, student groups (e.g. 5 groups of 4 students each) can each lead a discussion on a segment of the genome (i.e. 20% of it) aimed at resolving any differences found by different groups. The data are then compiled into a single DNA Master file.
  - Groups of students (e.g. 5 groups of 4 students) annotate a different segment of the genome (e.g. ~20%), followed by merging of the five DNA Master files into a single composite file. Instructions are provided in Stage 9 for doing this.

There are of course many other configurations and operational means of accomplishing your annotation. But it is helpful to keep in mind that the goal should be that all participants understand the full genomic context of the phage genome once the annotation is completed.

# AN IMPORTANT NOTE ABOUT THIS GUIDE'S SYNTAX

In this guide, we will refer to menus and submenus as follows. If the command is:

## File → Open → Archived DNA Master file

this means that you should click on the **File** menu at the top, scroll down to the sub-menu (**Open**), and select the sub-sub-menu (**Archived DNA Master file**) that appears.

le Manager Tools '	Window I	Help	
New	Ctrl+N	Mit Frighest Fri	Piler
Open		Archived DNA Master file     Ctrl+All	:+0
Import Merge Close	Ctrl+W	GCG-Formatted File Ctrl+O GenBank-Formatted File	
Close All		Previously downloaded NCBI file	
Save as DNAM5 File Export GCG file Export Apollo XML File	Ctrl+S Ctrl+E	Entrez ASN. 1 Sequence File Entrez XML Sequence File Apollo XML File Ctrl+All	:+A
Export Split Files		FastA Multiple Sequence File Ctrl+Al	:+F
Autoparse	a. L. a	Phylip Multiple Sequence File	
Preferences Utilities	Ctrl+P	Sequence from Accession Number	الدر ال

Tabs will be indicated by brackets, and sub-tabs will be shown by double brackets.

File → Preferences [Local Settings] [[Colors]]

File Edit Genome DN	A Taala	Window Help	DNA Master Preferences	
		window Help	Internet Local Settings Automation Phylogeny Timed Events Miscellaneous	<u>0</u> K
New Open	Ctrl+N	134	Directories GCG Export New Features Translation Shine Dalgarno Colors Validation	Revert
Open Import		Sec. 201	Click to specify Map colors Click to specify Six Frame colors	heven
Merge		and the second	Reverse Forward 🛆 CDS Frame 1	<u>C</u> ancel
Close	Ctrl+W	171 17 11 12	IRNA CDS Frame 2	6 L.
Close All		NE CORNE	rRNA CDS Frame 3	Apply
Save as DNAMS File	a. 1. a	1111111111	tmBNA CDS Frame 4	Export
Save as DNAMS File Export GCG file	Ctrl+S Ctrl+E	12. 1 1 1	URNA CDS Frame 5	
Export GCG file Export Apollo XML File	CUITE	121		
Export Split Files		5- 1071 M		
		12	PseudoORF Forward RNA	
<ul> <li>Autoparse</li> </ul>		13-4 · · · · ·	PseudoRNA Reverse RNA	
Preferences	Ctrl+P	E Cale State	-10 Site	
Otilicios		1.	-35 Site	
Quit	Ctrl+Q	13 1 1 1	Click to specify Frame Plot colors	
Product 1	PP 1	Start 1	Torreinstor	
Service -	1112	14.14 A	Frame 2	
-the first	-it		Repeat Frame 3	
- The state	and the second se	THE .	Binding Site	
			Oligonucleotide	

# 1 Introduction to DNA Master

## 1.1 DNA Master overview

The key program you will use in your genome annotations is **DNA Master**. DNA Master is a DNA sequence editor and analysis package that combines, analyzes, and displays data from a variety of DNA analysis programs, including GeneMark, Glimmer, Aragorn, BLAST, and HHPred. It organizes and collates all of these data into various tables and forms and saves it a single file with a **.dnam5** extension.

## **1.2** Installation

This guide assumes that you have installed DNA Master and can open the program successfully. If this is not the case, please install DNA Master before continuing with this guide. System requirements and installation instructions are at <a href="http://phagesdb.org/DNAMaster/">http://phagesdb.org/DNAMaster/</a>. DNA Master will need to be updated IMMEDIATELY!!! to complete the final step of the installation. To update, Open the program, go to Help, then click on Update DNA Master. The program requires a restart to accept all changes.

## 1.3 Quick Start Guide

The **DNA Master Quick Start Guide** is a useful tool when you are using DNA Master for the very first time and just want a quick look at basic functions. However, all parts of the Quick Start Guide are covered in more detail in this guide, so you may choose to use the Quick Start Guide as a future reference or a teaching tool.

## 1.4 DNA Master program structure

The various files, tables, and databases that DNA Master uses are a little complex, but a general understanding of the structure is important and will help prevent lost work.

#### The Feature Table

There are two important places DNA Master stores information about a genome annotation. The first, called the **Feature Table**, contains information about each feature (usually a gene) in a genome, including name, position, length, protein sequence, BLAST results, function, notes, etc. The Feature Table is a representation of the data stored for each feature. It is not a simple flat file, but rather a dynamic compilation of data. Within DNA Master, the data in the Feature Table for a particular genome can be viewed by going to the "**Features**" tab. When you **Post**\* changes to your annotation, like changing a start position or adding a gene, you're altering the Feature Table.

\* See **Section 9.3.1** for more on the importance of **Posting** changes.

#### The Documentation

The second place DNA Master stores information is the **Documentation**, accessible via the Documentation tab. This text contains much of the same information as is present in the Feature Table, but in a less human-friendly and more computer-readable format. Note that not all of the information from the Feature Table is contained in the Documentation Tab (e.g., amino acid sequence and BLAST hits are not present).

#### Interaction between the two

The Feature Table interacts with the Documentation as shown in **Figure 1.1**.

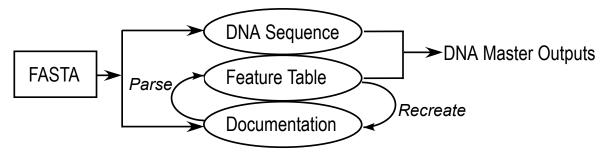


Figure 1.1

There are two functions—accessible through the Documentation tab—that control the interaction between the Feature Table and the Documentation:

**Parse** takes the contents of the Documentation and uses them to **OVERWRITE** the **Feature Table**. Parsing is done automatically by DNA Master when a genome is autoannotated, but thereafter should be used rarely if ever. The danger is that you'll have posted data to the Feature Table that are not included in the documentation, and then when you Parse, those data will be lost.

**Recreate** takes the contents of the Feature Table, and uses them to **OVERWRITE** the **Documentation**. This will update the Documentation with changes you've posted, and thus serves as a helpful backup of some of your data. You will want to do this often.

#### **IMPORTANT TO REMEMBER:**

Using **Parse** may overwrite user-inputted data, and thus Parsing may be **harmful**.

Using **Recreate** will store some user-inputted data in a new location, and thus it's **helpful**.

As a safety feature, **Recreate** the **Documentation** often. If your .dnam5 file gets corrupted, you can use Documentation to build a new file. Frequent recreation of the Documentation seems to improve the stability of the overall file, and so we recommend recreation prior to saving the file each time.

## Analysis programs running within DNA Master

As noted above, DNA Master runs a collection of programs that can assist in annotation and analysis of your phage genome. The following is a brief explanation of some of the key programs that DNA Master will be running for you, and some of their stand-alone versions that you will be using.

## 1.4.1 Glimmer

Glimmer (version 3.02) is a program that predicts the coding potential of open reading frames

(ORFs) based on a profile of frequently used codons found in the longest ORFs in the genome. DNA Master is set by default to use Glimmer in a heuristic way, meaning that it searches for potential coding regions (such as in long open reading frames) and then applies the nucleotide codon biases in those ORFs to search for other potential ORFs with similar biases. As such, it is not dependent on the use of externally defined parameters to determine coding potential. Glimmer also recognizes the use of TTG in addition to ATG and GTG as translation initiation (i.e. start) codons. It has very good predictive power for genes.

You will typically use Glimmer as a program that will run when you request DNA Master to perform an auto-annotation of your phage genome sequence and you will not be required to run it directly.

If you'd like to run Glimmer directly, it is available as a stand-alone program and is webaccessible at:

http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer\_3.cgi

#### 1.4.2 GeneMark

GeneMark also predicts genes within a genome sequence based on codon usage and codon potential. There are multiple forms of GeneMark tools developed by Mark Borodovsky and his colleagues that are useful to predict the genes in phage genomes. We will use at least two variations of this program. The ways to access this suite of programs has been altered in 2014 due to hacker misbehaviors. You will find that some of the previous versions are no longer available.

**Heuristic Version included in DNA Master: GeneMark.hmm** (version 2.0) provides a similar functionality to Glimmer. Its algorithms are different, however, and the joint use of Glimmer and GeneMark is a powerful combination for gene prediction. As with Glimmer, DNA Master runs GeneMark automatically within the Auto-Annotation function. Within DNA Master, the GeneMark that is run uses the heuristic model, in that it learns from the given genome what the codon usage preferences are in the longest ORFs and then applies this model to predict the remainder of the genes. GeneMark also takes into account potential ribosome binding sites when predicting gene start positions. This version of GeneMark will recognize TTG starts, importing them into DNAMaster.

**Heuristic Version to obtain a graph of coding potential: GeneMarkS** (Self training version for genomes >50 kb) is available on line at http://exon.gatech.edu/GeneMark/genemarks.cgi.

This version is similar to the one incorporated in DNA Master, however, you can obtain a graph of the coding potential by running the program through the website. This way, you can "see" what the numbers of the auto-annotation look like.

GeneMark with pre-trained model parameters: GeneMark.hmm (Version 3.25) is available online at <u>http://exon.gatech.edu/GeneMark/gmhmmp.cgi</u>.

In this internet browser-accessible version, the gene predictions are made using a codon usage model built from a previously annotated organism. GeneMark has many bacterial models available, and so for bacteriophage we pick a model based on the host organism. For the mycobacteriophage isolated on *M. smegmatis* mc155, we can use *Mycobacterium smegmatis* mc155 or one of the *Mycobacterium tuberculosis* strains listed.

This web version contains two key features that are useful for phage genome annotation:

• It allows you to specify the codon usage model from a bacterial host to use for gene

prediction, rather than generating a new model heuristically. A codon usage model for *Mycobacterium smegmatis* is available and can be selected to generate gene predictions in the phage genome based on the host's codon preferences. This sometimes allows you to find smaller genes that are not called during heuristic scans, but are likely to be reliable gene calls because they share codon preferences with the host. We refer to this output as the "**GeneMark-Smeg**" or "**GeneMark-TB**" output.

• It provides a graph (as a .pdf) of the gene predictions and coding potential. This is especially useful when you are determining gene starts.

A graph of the heuristic model can also be generated for comparisons. Use this version of the website <u>http://opal.biology.gatech.edu/GeneMark/heuristic\_hmm2.cgi</u>

Note:

- Refer to Sean R. Eddy's paper "What is a hidden Markov model?" to contemplate the math behind GeneMark and Glimmer . Nature Biotechnology 22, 1315-16 (2004).
- Both Glimmer and GeneMark heuristic versions only use a **random sample** of the ORFs to generate results, so outputs are not strictly reproducible.
- The GeneMark.hmm version in DNA Master will include TTG starts, the graph from the prokaryote model (Version 3.25) does not.

#### 1.4.3 Aragorn

**Aragorn** is a program for finding tRNAs and tmRNAs. Aragorn (version 1.1) can be run directly within DNA Master, although it is also accessible as a stand-alone program at:

http://130.235.46.10/ARAGORN/

The version of Aragorn available online is newer than the version embedded within DNA Master. It is **important to run the updated web-based version of Aragorn** (version 1.2.33.c.) in addition to the DNA Master version because it is better at determining the correct ends of tRNAs and because the version within DNA Master has a specific set of parameters that differ from the default. In addition, another tRNA predictor, tRNAscan-SE, is utilized to find additional tRNA predictions. Please refer to **Section 9.5** when you evaluate your tRNAs in your genome.

## **1.5** Setting Preferences

In general, setting preferences in DNA Master is a matter of opening the Preferences Window, making changes, and applying these changes. There are **five important preferences that you MUST set** before continuing with this guide. They are described in the next five subsections.

To get to the Preferences Window, select:

#### File → Preferences

You will see a dialog box with a series of tabs (Internet, Local Settings, ...) each of which has another set of sub-tabs associated with it.

## 1.5.1 Set Default Translation Table & Template Insertion (Local Settings Tab)

Changing this setting ensures you are using the correct translation tables for phages. Select:

#### File → Preferences [Local Settings] [[New Features]]

- From the Default Translation Table dropdown menu, select 'Bacteria and Plant Plastid Code'. Make sure that the boxes marked 'Add New Features to Documentation', and 'Add New Features to Feature Table' are both checked.
- Choose 'Insert template into notes during autoannotation'. Add the following codes to the text box: SSC: CP: SD: SCS: Gap: Blast: LO: ST: F: FS: ST:. These codes are explained in Section 9.6.
- Click 'Apply'. Note that the dialog box remains open.



Figure 1.2

#### 1.5.2 Set color preferences

You can select display colors for genes and tRNAs in various visual representations of your genome. The colors we recommend below are our preferences, and are used in most of the screenshots in this guide. You can select any colors you like, but note that if you use different colors, exported six-frame translations may not be properly viewable in Microsoft Word.

To set your colors to our recommended values, go to:

#### File → Preferences [Local Settings] [[Colors]]

Then set the values as shown below.

- Click on the colored box you want to change.
- A dialog box pops up with the color options.
- Click on the **color** of choice and then click **OK**.
- Continue to the next color.
- Don't forget to click 'Apply' to save changes.

CDS Frame 1	Yellow	CDS Frame 4	Gray
CDS Frame 2	Pink	CDS Frame 5	Light Green
CDS Frame 3	Light Blue	CDS Frame 6	Light Red

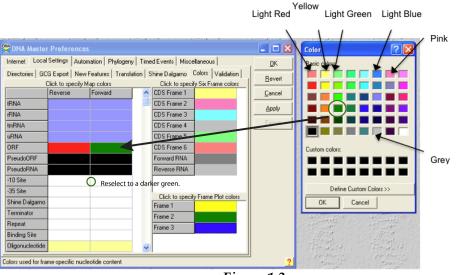


Figure 1.3

## 1.5.3 Set start codon choices

Because TTG is used as a translation initiation (start) codon in mycobacteriophage genomes – albeit rarely – you must make sure DNA Master recognizes it. To do so, go to:

## File → Preferences [Local Settings] [[Translation]]

- All boxes must be checked, as shown in **Figure 1.4** below.
- Click 'Apply'

DNA Master Preferences					
Internet Local Settings Automation Phylogeny Timed Events Miscellaneous	<u>o</u> k				
Directories   GCG Export   New Features   Translation   Shine Dalgarno   Colors   Validation   Default Start Codons	<u>R</u> evert				
☑ Use ATG start codons	<u>C</u> ancel				
✓ Use GTG start codons					
Use TTG start codons					
✓ Translate Cued Start Codons as Methionine					
Default Stop Codons					
✓ Use TAA stop codons					
✓ Use TGA stop codons					
I✓ Use TAG stop codons					

Figure 1.4

## 1.5.4 Set default values for BLAST searches

DNA Master can run batch BLAST searches and store the results for subsequent viewing. There are several settings relating to BLASTing inside DNA Master that may be helpful. Our suggestions are shown in **Figure 1.5**. Get to the BLAST menu by going to:

#### File → Preferences [Internet] [[Blast]]

- Set your preferences.
- Click '**Apply**' to save changes.
- Query timing may need to be adjusted depending on time of day that your send your BLAST request. See Section 4.5 for additional details

🞇 DNA Master Preferences		
Internet Local Settings Automation I	Phylogeny   Timed Events   Miscellaneous	<u> </u>
Email REBase NCBI Gene Predict	ion Blast	Revert
Communicating with Remote BLAST Ser	ver	
QBlast Server	http://blast.ncbi.nlm.nih.gov/Blast.cgi	<u>C</u> ancel
☑ Include "ignore self" terms when initia	ting BLAST form	Apply
Number of hits to request from server	100 🗢	Export
Number of hits expected	10 🗢	'
Saving BLAST results to Local Database	•	
Save hits with E-Values smaller than 10E	3 🔹	
🔽 Save a minimum number of	1 thits, regardless of E-Value	
🔽 Save a maximum number of	100 🚖 hits, regardless of E-Value	
Query Timing when performing Genome	BLAST	
Delay between multiple BLAST queries	4 seconds	
Send requests before retrieving for	40 🚖 seconds	
Limit queries to a maximum number of	20 🔶 pending queries	
Limit wait time for BLASt results to	10 💓 minutes	

Figure 1.5

#### 1.5.5 Choose a default location for saving files

DNA Master generates a number of files when it runs. It's good practice to create a dedicated DNA Master archiving folder, then direct DNA Master to use it. To do so, go to:

#### File → Preferences [Local Settings] [[Directories]]

- Click the 'Browse' button next to the 'Archive to...' field.
- Select your archiving folder, or create a new one.
- Click '**Apply**' to save.

DNA Master	Preferences					
Internet Local 9	ettings Automation Phylogeny Timed Events Miscellaneous		<u>0</u> K			
Directories GCG	Directories GCG Export New Features Translation Shine Dalgarno Colors Validation					
Databases	C:\Program Files\DNA Master\DMDB\	Browse	<u>C</u> ancel			
Helper Programs	C:\Program Files\DNA Master\Helper\	Browse	Apply			
Archive to	C:\Documents and Settings\Debbie\Desktop	Browse	Export			
Download to	ownload to C:\Program Files\DNA Master\Downloads\					
Manuals	C:\Program Files\DNA Master\Doc\	Browse				
BLAST Libraries		Browse				

Figure 1.6

## **1.5.6** Finishing up your Preference settings

Once you have finished setting your DNA Master preferences:

- Click the 'OK' button.
- Click '**Yes**' in the dialog box that asks if you want to save changes.

The Preferences Window will close.

## 1.6 Getting help

Help files and tutorials are available within DNA Master for many of its functions. Help is always available by clicking on the yellow ? button at the lower right corner of every window, or through the '**Help**' menu.

🕎 Extracted from FastA Library Timshel.fasta								
Overview Features References Sequence Documentation								
Sort By Index 💌 🔳	Name	Start	Stop		Description Sequence Product Regions Blast Context			
Select Features Direct SQL	▶ 1	408	704		Name 1 GenelD			
Tune is All	2	743	1177					
	3	1254	1577		Type CDS <u>GI</u>			
Name like	4	1567	2319		Start 408 Locus Tag DNAM1			
GenelD =	5	2345	3565		Stop 704 Regions	ī		
Locus like	6	3592	4380		Length 297 🔳 Tag	-1		
Start >	7	4377	5294		Direction Forward			
		5374	5449			-1		
Length	8	5547	6899		Translation Table Undefined	1		
Regions >	9	6896	7309		EC Number	٦		
% GC <	10	7306	8283			3]]		
CAI >	11	8479	10005		Product	-		
EC# like	12	10002	11459		gp1			
	13	11456	12469			1		
Product like	14	12520	13020		Function	-		
Function like	15	13055	14014			1		
FeatureID =	16	14080	14268			11		
Tag like	17	14276	14650		Notes	4		
				ᆂ	Original Glimmer call @bp 408 has strength 2.35; GeneMark	1		
Hide Ignored Features			<u> </u>	1	Calls start at 257			
Select All Features	Insert D	elete Pos	t Valida	ate		Ц		
	1 - 50000	P	osition : 4	8037	Controls >> Map 🔽 Map >> Controls			
BIA 5 617 8 840 11 124		25 V	26 27	V.	130 Vaa MUNIKAA MAKAA KUK KUKAK KE KUMA KUKATU	L.		
89 Features Live 53278 32 ?								
Live			_		33270	-		

Figure 1.7

To get a sense of how the help files work, go to: Help  $\rightarrow$  Help

• Read the 'Welcome to DNA Master' and the 'Getting Started Tutorial' sections.

## 1.7 Checking for updates

DNA Master is regularly updated, and with an internet connection it is easy to make sure your copy is up-to-date. Go to:

#### Help → Update DNA Master

- If a new version is available, it will update the program, and a dialog box will appear when completed. Please note that you must have an active internet connection to do this!
- When the update is complete, close and restart the program.
- As of the time of writing (November 2014), the most up-to-date version of DNA Master is Version 5.22.19 Build 2448, dated 13 June 2014. You can find your current version by going to: Help → About

## 1.8 Event Manager

The Event Manager (Tools -> Event Manage) is a useful tool when you receive error messages in DNA Master. There is far more information recorded here than what is provided on screen.

# 2 Provisional Cluster assignment of your phage

## 2.1 Overview

All sequenced mycobacteriophage genomes have been compared to one another, and based on these comparisons they have been grouped into **clusters** of related phages. Some of these clusters are small (Cluster V currently has only two members), whereas others are quite large (Cluster A has 259 members). Some clusters are further divided into **subclusters**; for example, Cluster B's genomes are currently divided into five subclusters: B1, B2, B3, B4, and B5. There are also some phages (nine currently) who have no close relatives, and therefore are classified as **Singletons**. Up-to-date statistics are available at:

http://phagesdb.org/clusters/

Your phage's final cluster designation depends on a variety of analyses, as described in:

Hatfull *et al.*, (2010) Comparative genomic analysis of sixty mycobacteriophage genomes: Genome clustering, gene acquisition and gene size. *J Mol Biol.* **397**, 119-143.

In the meantime, however, it is helpful to make a provisional cluster assignment for your phage. This can be done using just a completed genome sequence, before any annotation has taken place because clustered phages usually share a span of 50% or more recognizable nucleotide similarity across their genomes.

Performing a nucleotide BLAST search of your phage sequence against a database of mycobacteriophage genomes (at phagesDB.org) provides a simple and quick approach to making a provisional cluster assignment. The assignments made at the completion of a genome sequence are based on this BLAST alignment. Further analysis made be necessary for some genomes.

Note: It is not reliable to denote cluster designations without a complete genomic sequence available. Inferences can be made about cluster membership through a variety of ways however these should be included in isolation notes and not used to label a phage as a member of a particular cluster.

## 2.2 BLASTing your sequence against the mycobacteriophage database

To BLAST your genome on phagesdb.org (nucleotide BLAST):

- Go to <a href="http://phagesdb.org/phages/">http://phagesdb.org/phages/</a>
- Locate your phage in the phage list, then click to open its detail page.
- Click on the green "Locally BLAST this genome" button.
- It will open a page that looks like the one in **Figure 2.1**.

## Local Nucleotide BLAST

#### Go to Protein BLAST

This tool will run a local BLAST search against our phage database. This will include some genomes that are not yet in GenBank and thus not accessible via NCBI BLAST.

#### Choose program to use and database to search

Program blastn 🗘 Database Mycobacteriophages as of Oct 14, 2013 🛟
Enter sequence below in FASTA format
<pre>&gt;Echild complete sequence, 53159 bp including 10bp overhang (CGGTCGGTTA), Cluster A2 TGCGGCCGCCCATCCTGTACGGGTTTCCAAGTCGAAGTCGGAGGTCCCGAGC CGGCGCAGGAGCGCCTCAGCCTCTGTGGCGCCCCCAGGACGCCAGGAT CCCCGCTCACGCGGGTAGTTGTATGGGCTAATCGGCAAACGGCCTCTGAG GCCGCGAGACCAATGTCACACCAGGTGGTGGGTGGTTATTGACGCACGC</pre>
Or load it from disk
Browse No file selected.
Set subsequence: From To
Clear sequence BLAST

#### Advanced Options

The query sequence is NOT filtered for low complexity regions by default.

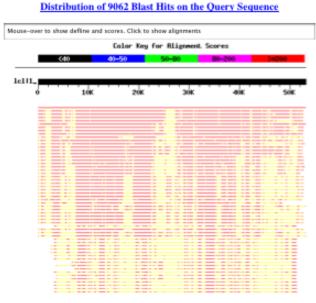
Filter 📃 Low complexity 📄 Mask for lookup table only	
Expect 10 🛟 Matrix BLOSUM62 🛟 🗆 Perform ungapped alig	Inment
Query Genetic Codes (blastx only) Bacterial (11)	
Database Genetic Codes (tblast[nx] only) Bacterial (11)	\$
Frame shift penalty for blastx No OOF	

Figure 2.1

- The defaults are set so that the program will run **blastn** (i.e. a nucleotide search against a nucleotide database) against a database of previously sequenced mycobacteriophage genomes (e.g., Mycophages as of 6.01.11).
- Click on the '**BLAST**!' button. It is just above the gray dividing bar at the center of **Figure 2.1** above.

A new page will open showing the results of the BLAST search, as shown in **Figure 2.2** below.

Your query is represented by a black bar underneath "Color Key for Alignment Scores". Subject sequences from the database that align well to your query sequence are represented by colored bars beneath the black bar. The length and location of the subject bars indicates the portion(s) of the query sequence the subject sequences match. The quality of each alignment is shown by color, with the best matches colored red.





To see which subject sequences your query has aligned to, simply mouseover any of the colored bars, and the subject's name will appear in the box above the "Color Key for Alignment Scores". Then, either scroll down or click on one of the lines to get the names of subject sequences that have the best alignments to your query sequence, listed in order from best match to worst match (see below). After each subject sequence name is the raw score of the alignment to your query sequence (higher is a better alignment), and the E value (lower is a better alignment).

	ore E its) Value	
Sequences producing significant alignments: (b: Echild complete sequence, 51159 bp including 10 bp comproximation (CGGT Jarm complete sequence, 53163 bp including 10 bp 1' overhang (CGG Mabigail7 complete sequence, 53167 bp including 10 bp 1' overhang (CGG Mabigail7 complete sequence, 53035 bp including 10 bp 1' overhang (CGG Explored the sequence, 53035 bp including 10 bp 1' overhang (CGG ChipMuck complete sequence, 53035 bp including 10 bp 3' overhang Evileenius complete sequence, 53335 bp including 10 bp 3' overhang Evileenius complete sequence, 53335 bp including 10 bp 3' overhang Evileenius complete sequence, 53335 bp including 10 bp 3' overhang Evileenius complete sequence, 53335 bp including 10 bp 3' overhang Evileenius complete sequence, 53335 bp including 10 bp 3' overhang Evileenius complete sequence, 53454 bp including 10 bp 3' overhang Evileenius complete sequence, 53454 bp including 10 bp 3' overhang Evileenius complete sequence, 53454 bp including 10 bp 3' overhang Evileenius complete sequence, 53454 bp including 10 bp 3' overhang Evileenius complete sequence, 53454 bp including 10 bp 3' overhang Serenity complete sequence, 53454 bp including 10 bp 3' overhang Addard Ddin complete sequence, 52467 bp including 10 bp 3' overhang ADGYY complete sequence, 52519 bp including 10 bp 3' overhang Harmar Cloudiang1 complete sequence, 52873 bp including 10 bp 3' overhan Blue7 Complete sequence, 52867 bp including 10 bp 3' overhan Blue7 Complete sequence, 52873 bp including 10 bp 3' overhang Elses a complete sequence, 52873 bp including 10 bp 3' overhan Elses Complete sequence, 52873 bp including 10 bp 3' overhan Elses Complete sequence, 52888 bp including 10 bp 3' overhans Elses complete sequence, 52128 bp including 10 bp 3' overhang (CG Chilaitor Final Sequence, 52132 bp including 10 bp 3' overhang Elses complete sequence, 52132 bp including 10 bp 3' overhang Elses complete sequence, 52484 bp i	its) Value      3.054m+05     0     1.104m+04     0     1.014m+04     0     2.0514m+04     0     2.0514m     0.0     2.05     2.05     2.05     2.05     2.05     3.22     0.0     3.22     0.0     3.2     0.0     0     3.2     0.0     0	.0
PackMan Complete Sequence, 51339 bp including 10 bp 3' overhang Hyzus Complete Sequence, 53425 bp including 10 bp 3' overhang (C Easan complete sequence, 52160 bp including 10 bp 3' overhang (C	714 D.D 714 D.D 476 D.D	
JewelBug complete sequence, 50341 including 10 bp 3' overhang (C Bysis complete sequence, 51256 bp including 10 bp 3' overhang (C BeFly complete sequence, 52502 bp including 10 bp 3' overhang (C DaVinci Final Sequence, 51547 bp, 10 bp 3' Overhang (CGG7CGGTTA) BricB Complete Sequence, 51702 bp including 10 bp 3' overhang (C	468 0.0 460 0.0 460 0.0 460 0.0 452 0.0	
Catalina complete sequence, 51715 p including 10 bp 3' overhaag (CG Alma Complete Sequence, 51317 bp including 10 bp 3' overhang (CG Morpher26 complete sequence, 51274 bp including 10 bp 3' overhan LittleGuy complete sequence, 51274 bp, including 10 bp 3' overhan Wile Complete Sequence, 51378 bp including 10 bp 3' overhang (CG	632 0.0 634 e-179 605 e-170 605 e-170 597 e-168	

Figure 2.3

Scroll down further (or click on the blue raw score number) to get the nucleotide alignment of your query sequence (top) to each subject sequence (bottom). The numbers on the sides of the sequences indicate the nucleotide coordinates within each sequence. Identical nucleotides are connected with vertical lines and smaller gaps in the alignment are shown by horizontal dashes.

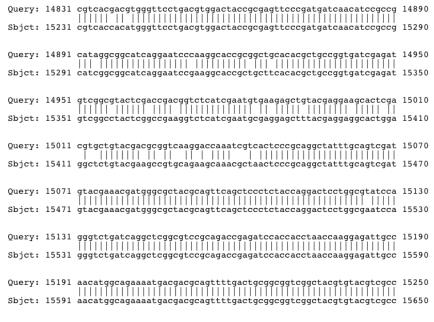


Figure 2.4

## 2.3 Cluster assignment

You should now be able to make a provisional cluster assignment. If one of your subject matches is a red line extending over at least 50% of the genome, then it is likely that your phage belongs in the same cluster as that subject. If the cluster is divided into subclusters, then a long but interrupted red line likely indicates that it is similar to a particular subcluster.

We'll use a case study—the phage Adephagia—to demonstrate how to assign a provisional cluster to your phage. **Figure 2.5** shows the output of a BLAST search with the Adephagia genome as the query.

## Distribution of 1211 Blast Hits on the Query Sequence

Mouse-over to s	how defline and	scores. Click to	show alignment	s			
			for Alignmen				
	<40	40-50	50-80	80-200	>=200		
lcl 1_							
0	10K	20K	30K	40K	50K		
						K1	
						K2, K3, K4	, K5
				1			
		=					
Sequences pro	ducing sign	nificant al:	ignments:			ore E its) Value	
Adephagia Fin BEEST Complet JAWS Complete BarrelRoll Co Angelica Fina Anaya Complet	e Sequence, Sequence, mplete Sequ l Sequence e Sequence,	59906 bp in 59749 bp in inence, 59672 bp, 60835 bp,	including 1 ncluding 11 2 bp includ including	1 bp 3' over bp 3' over ing 11 bp 3 11 bp 3' ov	erhang (C chang (CT 3' overha verhang (		0.0 0.0 0.0 0.0 0.0
CrimD Final S Pixie Final S						1.826e+04 1193 0.0	0.0
			Figure ?	5			

Figure 2.5

Adephagia's best hit is to itself. After that, there are six heavy red lines that indicate very similar genomes to Adephagia's. Scrolling down to the "Sequences producing significant alignments" section, we can see that these red lines correspond to the genomes of BEEST, JAWS, BarrelRoll, Angelica, Anaya, and CrimD. Using phagesdb.org, we can then look up the cluster assignments of these six phages. All six, it turns out, are members of Cluster K, and Subcluster K1.

There are four more genomes that appear to have significant similarity to Adephagia, though the matches are less solid and cover less of the query sequence. These more tattered-looking red lines correspond to Pixie, TM4, Larva, and Fionnbharth. Using phagesdb.org, we can see that these are all member of Cluster K, though they belong to Subclusters K2-K5, not K1.

Therefore, we can provisionally determine that Adephagia is a member of **Cluster K** and **Subcluster K1**.

**NOTE**: Though the example above may seem clear-cut, cluster assignment will not always be so simple. If it's not, don't be concerned. You may have found a new singleton phage, or a phage that will lead to a new subcluster being created. The main point of doing this now is so that you have an idea of which phages are most closely related to the one you are annotating. These closely related phages can be very useful guides as you go through the annotation process.

# 3 Importing your phage genome sequence into DNA Master

## 3.1 Overview

Now that you have a sense of your software and an overview of your phage genome, you are ready to move onto the really exciting stuff! The first thing you need to do is to download your phage's genome sequence, then import it into DNA Master.

## 3.2 Where do I get my phage genome sequence from?

Sequencing a phage genome involves two parts: Shotgun Sequencing and Finishing (aka Polishing). The second part, **Finishing**, involves generating targeted reads to fix weak areas, determining the type and/or sequence of genome ends, and orienting a genome to match convention. When performing annotations, you **must always use a Finished sequence file**, or your annotation work may have to be redone.

Fortunately, **phagesdb.org** only posts Finished sequence files, so be sure to get your sequence from phagesdb.org. Though you may have access to preliminary, un-Finished files from other sources, **the phagesdb.org site should be the only source for sequence when beginning annotation**.

## A NOTE ON FILE TYPES

DNA, RNA, and protein sequence files are often saved in **fasta** format. This is the standard format required by many bioinformatics programs, including BLAST. Fasta files are simply text files where:

- 1. The first line begins with ">" and contains information about the sequence
- 2. Subsequent lines contain the sequence itself

For example, the first few lines of a phage genome sequence fasta file may look like:

A few things to keep in mind:

- Fasta files can be opened with any text editor.
- A file does not need to have the extension **.fasta** to be in fasta format. For example, if you rename Giles.fasta to Giles.txt, the file will still be fasta-formatted.
- Sequence files from phagesdb.org are in fasta format and have a .fasta extension.

To download your genome sequence as a fasta file, go to:

- <u>http://phagesdb.org/phages/</u>
- Scroll down to find your phage and click its name to open its detail page.
- Scroll down to the section titled 'Sequencing Information'.
- Click on the '**Download fasta file** ' link, and save the file to a known location.

#### **IMPORTANT NOTES:**

- If you can't find the downloaded file, simply search your computer for a file named YourPhageName.fasta.
- If you are using a Windows emulator on a Mac (and use your internet browser on the Mac side to get the fasta file), then you should either copy the fasta file from the Mac side to the Windows side, or alternatively set up your emulator preferences so that it can directly read files from the Mac side from a shared folder.
- If for some reason you're using a sequence file from a location other than phagesdb.org, be mindful that there are two possible orientations for a genome, and that yours needs to conform to the standard convention (the virion structural genes on the left, transcribed rightwards). If you determine that a sequence needs to be reverse-complemented, instructions are provided at the end of this section for doing so.

## 3.3 Importing your DNA sequence into DNA Master

You are now ready to import your fasta file into DNA Master. Open DNA Master, then go to:

🕎 DNA	Master			
Ele Ioc	ols <u>W</u> indow	Help		
New		Ctrl+N	rit pristingt	Frighter 1
Open.		•	Archived DNA Master file	Ctrl+Alt+O
Import Close Close		Ctrl+W	GCG-Formatted File GenBank-Formatted File	Ctrl+O
Export Export	as DNAMS File : GCG file : Apollo XML File : Split Files	Ctrl+S Ctrl+E	Previously downloaded NCBI file Entrez ASN.1 Sequence File Entrez XML Sequence File Apollo XML File	Ctrl+Alt+A
✓ Autop	200	<	FastA Multiple Sequence File	Ctrl+Alt+F
Autop		Ctrl+P	Phylip Multiple Sequence File	
Utilitie		•	Sequence from Accession Number	
Quit		Ctrl+Q	and the second	All and

File  $\rightarrow$  Open  $\rightarrow$  FastA Multiple Sequence File

Figure 3.1

• Browse to the correct folder and select your fasta file.

• A window like the one shown in **Figure 3.2** appears.

FastA Sequence	s from Echild.fas	la			and.	The states
Sequences Pairwise /	Alignment   Multiple Al	gnment Site Analysis			and the second	- Andreast
Idx Description			Length	Insert Sequence	A States	-the star
<ol> <li>Echild complete se</li> </ol>	quence, 53159 bp ind	luding 10bp overhang (CGI	GTC(53159	Delete Sequence	-	and the second s
				Update Sequence		ittere
				Verify updates	and the	Provide to
				Save FastA File	Street L.	Profiler 1.
				Export Summary	the later	1 million
Sequence		Complement Hide	Annotation			and the second
TGCGGCCGCCCCATCC AGCCGGCGCGCAGGAGCG CAAGATCCCCGGCTCAC GCCTCTGAGGCCGCGGA	CCTCACCCAGCCTCT GCGGGTAGTTGTATG	GTGCGCCCCCAGGACG	2. Select			
TGACGCACGCGTCCGT AACTTAGATTCAAAAC				Create Seq	Jence from this e	ntry only
GCTCCTGGCGGCGCGG	GGCCAGGTCCACCAC	CGCAGGGAGGACCGCC			arate sequences f	
ATGAAGATCATCCGCT CGCCGCCCGAGGCGGC			1. Click		ate sequences foi l entries into a sir	
AGGACGAATACCTCCT					iltering for valid C	
TCAAGTGATACTCATG TAAGAGCCGCACAGGC				Create Seq	uence with entries	s as replicons
		GACCCGGCAACCGCCG 👱			語る	The state
1 Sequences 53,1	159 bp Creates a	sequence form		GC 🖹 Export ?	and an indian . Subtranet A	partition of
all the	and a	and a	and a	and the	and a	"mante

Figure 3.2

- Click on the Export button in the lower right hand corner (1).
- From the menu that opens, select 'Create Sequence from this entry only' (2).
- A new window titled 'Extracted from FastA library YourPhage.fasta' will open within DNA Master.
- We recommend that you now save the file with a new name, for example Sheen.dnam5. Once the file is in this format, you will always open an 'archived DNA Master file.

Let's take a moment to look at some of the new views that are available.

- There are five tabs in the new window: [Overview], [Features], [References], [Sequence], and [Documentation].
- Select the **[Overview]** tab if it's not already selected. Your window should look similar to the one in **Figure 3.3**.

👺 Extracted from FastA Library Echild.fasta		
Overview Features References Sequence Documenta	ition	
Genome Uggariam Genome Uggariam Genome U 0 Ugariam Genome U 0 Date Created: 10/17/2013 Repticon: 1 Genome Date: 10/17/2013 Cell Content: 0 0 Content: 0 0 Content: 0 0 Content:	Taxonomy Domain : Division : Family :	Notes Notes on the Genome Nates and Lawcomp use only available when sequences are loaded from the database
Features: 100         GL:           DRFs: 100         Status: Current           RNAs: 0         Deletion: Vulnerable           GCContext:: 0.00         Origin: 0           DRF GC: 63.4 ± 2.8         Terminus: 0           GC3: 83.5 ± 6.4         CAI: 0.6281 ± 0.0998           Translation Table: Unspecified: default to standard code	3	3
Image: Contract of the second seco	ition : 1 🔽 Controls >> N	Ap 🔽 Map >> Controls

- Check the sequence length (shown in the red circles in **Figure 3.3**) and verify that it matches the published sequence length on your phage's detail page on phagesdb.org. If there is a discrepancy, restart the program and try importing again, or re-download your sequence file from phagesdb.org.
- Select the **[Sequence]** tab. This tab displays the DNA sequence of your phage. You can click and drag to select part of the sequence, whereupon DNA Master will display the coordinates and length of the selected portion near the top of the window, as in **Figure 3.4**.

Overview	Features Re	ference	s Seq	uence	Docur	mentation										
Feature			-	F	矖	Position	: 866 - 12	234 (369	bases)	Ra	w 🕨	BLASTN	BLAST	PA	dd Feature.	1 62
1	TGCGGCCGCG	CCATO	CTGT	CGGGT	TTCCA	LAGTCGA	TCGGAG	TCCCGA	GCCGGCGI	CAGGAG	CGCC	TCACCC	AGCCTCT			
93	CGCAAGATCO	CCCGC1	CACGO	GGGTA	GTTGI	ATGGGC	TAATCG	GCAAAC	GCCTCT	GAGGCC	GCGA	GACCAA	TGTCAC	CCA	GGTGGTGG	ATG
185	TTATTGACGO	CACGCO	TCCGT	TAAGA	GGACA	TGGCCT.	AGGTAT	GGCTAC	CAAACT	TAGATT	CAAA	ACCAGT	CCCCTGG	ccci	COTCOTO	GGT
277	GTTGCCGCT	CTGGG	GGCGC	GGGGGC	CAGGI	CCACCA	CCGCAG	GGAGGA	CCGCCAT	GAAGAT	CATO	CGCTCG	CTCGCCC	GGGG	CGCTCGCA	CTC
369	ATTCGCATCO	GCCGCC	CGAGO	GCGGCT	GGGAT	GCGGAA	ATCTAC	GAGCCG	FEGEATE	AGGACG	AATA	CCTCCT	ATAGTG	TCT.	ACGCCACT	TGC
461	TCGGTGGGT	TCAA	TGAT	CTCAT	GTATO	TAGTTA	TTGAGG	GCCTAA	AGGCCCG	AATAAG	AGCO	GCACAG	GCGGCT	TCL	AAGAGCGC	CCA
553	CTAGGGCGCT	CGAAC	TAAT	CCGGC	CTTGA	GGGCCG	GTTATC	TGACCC	GCAACCI	GCCGGG	TCTI	CTGCCG	CGCCCAG	TGG	CGCGGCTC	ATA
645	GAAGGGGTGA	AGGCAA	CCGTO	TACGG	CACTO	GCTCGA	GTGCCT.	ACTGGG	CCTCGCA	GCCGGG	GAAG	TTCGAC	GTTCTGA	ACC'	FGCGGATG	GACG
737	TTCCCGAGC	ACGTCO	GCCGC#	CGACA	тсссс	GATCTT	CGGCCG	TCTGAC	TCGTTC	CGGCCA	ACCT	CGCAGC	GTGGAAG	ATG	CCACGCCA	TCG
829	CGAATACGCO	GCCAT	TTCGC	GCGGC	GCACT	CCACTT	CITCCT	TGACGA	TACCGT	TTCGAG	ACCO	TCTGGT	CGTCCCC	CGA	GCGTCTTC	TCC
921	CCCGCGTGC	AGCG	TTGGC	GCCGC	TCTGA	CGCCCG.	ATTTCA	GCGTGT	GCCGAGA	AATGCC	GCGG	GTAGCG	CAGGTCI	GGA.	ATACCTAT	rege
1013	AGCCGCTGG	recee	GCGTA	TTGGC.	AATCT	CAGGGG.	ATCGAG	GTAATT	CCACGG	TCGGTT	GGGG	TEGACE	GGACACA	TAC	GAATTCTC	CTT
1105	TGACGGCCT	GCCGG1	rcggco	GAAAT	GTCGC	CATTTC	CTGCCT.	AACCCT	CCGATCG	AAGCAA	GAGO	ATCGCG	AGCTGT	CAC	CCGAGGCC	TGC
1197	AAGAACTCG	ICTGG/	IGAACO	GCAGCC	GAAGA	CCCTTT	ICGTCT.	ACGGCC	GCTCAG	ATTCTG	CGAC	GAAATA	GACCTTO	CCG	AGGTCCGG	GAG
1289	TACCCGACCT															
1381	AGCCGCGCCC	GAAAO	GGCGC	TGGCG	GTGCC	ACAACT	CCCGGA	GGCGGT	GATCTG	GAGGCG	GTGG	AGGCGG	CGGAAAA	GGC.	AAAGGCTC	CGC
1473	AAGCGCCTCA	ACCG	TGGCC	TAAAG	AACCO	CGGAGG	CGGCGG.	ATCCGG	CTCCGGT	GGTGGC	GGTA	GCCAGA	TGCCGAG	CAA	GCCCAAAC	CGT
1565	GGCCGGATA	ACCTAC	TCCA	AAGTC	CGACA	AGCGCC.	TCGACA	CAGACC	AGCGTCG,	AGCCGA	CAAC	ATCGCC	GCGGTC	ACCI	CGCGGTGG	GAAC
1657	GAGCGCACGO	CCCGAC	GATCO	GAACA	CCGAT	CGTCAG	TGGAAC.	ATCAAC	FGCACCA	GGTGTG	CGGC	CACCGT	CGAGATO	AGG	GCCCGCGG	TTA
1749	CGACGTAACT	recect	reccca	AGCCC	GAGAA	CCTCAA	TGACAA	CTACAC	CACATG	TCGTCC	GCGA	AGTGGA	TCGATGA	GCA	CGGCAACO	CGG
1841	CTCAGTGGG	ACTACO	TCGAC	GGCTC	CCGCG	GACAGAA	TCATGG.	ACGAGA	FCAACGCI	CCGAGC	TGCG	TTGTGG	CCCGAGO	GAG	CCCGAGGG	стс
1933	ATCCGCACG	ACCTGO	GAGGO	CGGCG	GCGGT	CACATC	TTCAAC	TGGGAG	AGCGCG	ACGGCG	TGAT	CCGGTT	CATCGA	GGC	CAGCCCAA	CCA
2025	GTGGGACGCO	AGCTO	GTCGT	GGACC.	AACCA	GCCGTC	GGAGTG	GGGGCA	CAGCACCI	GCGGTT	GTCC	GCGTTG	ACCACCT	GAA	CCCGCACG	GAAG
2117	ACCTGTCGA	GTGG	TACGO	AACCG	CACGO	CAGAGG	AGATCA	ACGCTC	CCTGCG	CCGCGA	GGTT	ACAGTO	GAGATGO	AAC	GCGCGGG	TAC
2209	AAGCTCGGT															
2301	CCCTGAGTTO	CAAGGO	GATCT	ACGAA	GCGGC	AGTGAA	ATGGGC	AAGGAG	CCGATC	TGATGG	TGAL	GTTCGA	GCAGGCT	AGAI	GCATCOT	GTT
2393	CGACAACCG															
2485	GTCCCTACGO	GATGO	TCTA	CCGGC	ACGCA	ATGCCG	ATGACC	STOTOT	GATCTC'	TGACCA	GGAD	GGCCCG	TTEGTER	CGG	TEGACAAR	GCC
2577	ACAGGCGAG															
2669	ACCATGAGC	ngggcr	TCCTO	GAGAC	GCCGE	TATGAC	CTTCCC	CEGGAE	EGGGAGC'	TGAACT	ACAP	GCTCCC	GGTCCT	CGG	BATGCCAA	CTG
2761	GATCTGCGA															
	Accorton		Corres	201000	CA & CA	1.100.001										222
444			1 - 5000	10		Position	:1		Controls >>	Map 🔽	Map	>> Cont	rols			
													_			
D Features	Live														53159	2

Figure 3.4

• Until you run an automated annotation in the next section, the tabs for [Features], [References], and [Documentation] are largely empty.

**Congratulations!** You have now imported your phage sequence into DNA Master and are ready to run an Auto-Annotation.

## 3.4 Reverse-complementing your sequence

To re-emphasize, if you download your genome sequence from phagesdb.org, it will **NOT** need to be reverse-complemented. If you need to reverse-complement a sequence from a different source to match conventions, you can do so easily within DNA Master. To reverse-complement a sequence:

- Go to the [Sequence] tab.
- Make sure that no segment of the sequence is selected (otherwise you will only flip that part—a big mess). If in doubt, just click somewhere within the sequence, but without selecting anything.
- Select: DNA → Convert → Complement

- A dialog box will open that asks if you want to convert XXXXX bp to  $5' \rightarrow 3'$ . Click '**Yes**'. (XXXXX should be your full genome length)
- Select: **File** → **Save as** , then save your reverse-complemented file with a new name.

# 4 Performing and viewing a rapid automated annotation of your genome

## 4.1 Overview

DNA Master has an **Auto-Annotate** function that provides quick and simple identification of genes within your phage genome. It works by running Glimmer, GeneMark, and Aragorn, then combining the outputs from these programs to arrive at consensus gene calls. The consensus output is used to populate DNA Master's Documentation and Feature Table sections.

Generally, this auto-annotation will identify 90% or more of the genes accurately, but the careful refinement that you will perform in **Section 8** will be essential for obtaining the best possible annotation that will be ready for GenBank submission.

## 4.2 Running Auto-Annotate

• As shown in **Figure 4.1**, go to:

Genome $\rightarrow$ Annotation $\rightarrow$ Auto-Annota
---

	DNA M	laster					
File	Edit	Genome DNA Tools	Window Help	_			
13.	řř.	Add to Database		from FastA Li	hrary Echild	fasta	
28	1115	Analyze all gone start	s				
	1	Auto-Annotate		ures Reference	es Sequence		
892	NOT	BLAST All Genes		I ■ ■	Name !	5'End Length	Description
1		Coding Capacity		Direct SQL	Þ		Name
63		Codon Spacing		A	1		Type
6-53		Features	•	F			5'End
1.3.	rr.	Gene Orientation	Ctrl+Alt+G	L			
SE.	11/2	Karlin's Dinucleotides					3'End
20		Learn Features	Ctrl+Alt+L				Length
202	No.	Mutational Bias					Direction
14		Nucleotide Phasing	Ctrl+Alt+H				Translation Ta
83		Origin prediction Predict stop					EC Number
2.43		Profile	Ctrl+Alt+P				
17	PP-	Rearrange	COTTACT				Product
56	1	RNASeq	,				
22		Run Length					
	has the	Six frame translation	Ctrl+Alt+T				Function
10		Skewed sequences	Ctrl+Alt+K				
28		tRNA Profile					
200		at the Street	Tag like				Notes
17	P.P	ations of the second	-				
150	1	And the second second	Hide Igno	red Features			
22		lester -the	Select A	Il Features			
	1 The	in the second					
10		Str. Star					
28		1日					
2.27		and the				>	
14	Pre	Street & Paris			Insert Dele	ete Post Validate	1
236	1	and the second	जित्ती ब ि		1.50000	Position : 1	Contr
92		Line -	14 44 4 H		1 - 30000	Position. 1	j♥ conu
	A.T.						
19	1.00	apr and	0 Features	Live Who	e genome analyti	cal tools	

Figure 4.1

• An Auto-Annotate dialog box will open, with 4 sub-windows to configure. We recommend that you use the settings shown in **Figure 4.2**.

👺 Auto-Annotate 📃 🗖 🔀
Current Replicon  Export log file     Annotate
C All Replicons Gene Calling Tags & Comments BLAST Searches Automated Analyses
Document tRNAs found with Aragom
Document ORFs found with
C Glimmer 3.02 analysis
GeneMark HMM analysis     Both analyses, combining them as follows:
Favor Glimmer calls
C Favor GeneMark calls
Exclude genes called by only one method
Exclude genes called only by second method Maximum wait time 4 minutes
Insert template into comments for each gene
Examine and learn gene features
Status : Idle No BLAST ?
🚆 Auto-Annotate 📃 🗖 🔀
Current Replicon      Export log file     Annotate     Annotate
Gene Calling Tags & Comments BLAST Searches Automated Analyses
Locus Tags
Prefix DNAM
Start 1 🛨
Spacing 1
SSC: CP: SD: SCS: Gap: Blast: LD: ST: F: FS:
🕎 Auto-Annotate 📃 🗖 🔀
Current Replicon
C All Replicons
C All Replicons Gene Calling   Tags & Comments   BLAST Searches   Automated Analyses
C All Replaces Gene Calling   Tags & Comments   BLAST Searches   Automated Analyses   Perform BLAST search on NCBI nr database
AlRepicons     Gene Caling Tags & Commerts [BLAST Searches] Automated Analyses     Perform BLAST search on NCBI rr databate     Save hits with E-values less than 10E 3
All Replicons     Gene Calling   Tags & Comments   BLAST Searches  Automated Analyses     Perform BLAST search on NCBI rr databate     Save hits with E-values less then 10E   3      Pagardless of their E-values, save   1          hal(s)
C All Replicons Gene Calling   Tags & Comments   BLAST Searches  Automated Analyses   □ Perform BLAST search on NCBI rr databate Save hits with E-values less than 10E   3 □ Regardless of their E-values, save   1 □ Limit number of hits to be saved o   100 □ Limit number of hits to be saved o   100 □ hits
All Replicons     Gene Calling   Tags & Comments   BLAST Searches  Automated Analyses     Perform BLAST search on NCBI rr databate     Save hits with E-values less then 10E   3      Pagardless of their E-values, save   1          hal(s)
<ul> <li>C All Reploces</li> <li>C All Reploces</li> <li>Perform BLAST search on NCEI nr databare</li> <li>Save hits with E-values less than 10E</li> <li>I I I I I I I I I I I I I I I I I I I</li></ul>
C All Fieldcore       C All Fieldcore     BLAST Search Not REE Muture Analyses       Gene Caling     Tags & Commerks:       Perform BLAST search on NEE muture Adabate       Save hits with E values less than 10E       If Regardless of their E values, some       If Link number of hits to be saved to       If Up between multiple BLAST queries       Very between multiple BLAST queries       Send requests before retrieving for       Watro Annotate
All Replicons     Gene Calling Tags & Comments BLAST Searches     Automated Analyses     Perform BLAST search on NCBI'r database     Save hits with E-values less than 10E [ 3
<ul> <li>All Replicons</li> <li>Gene Caling Tags &amp; Comments BLAST Searches Automated Analyses</li> <li>Perform BLAST search on NCBI rr databate</li> <li>Save hits with E-values test than 10€</li> <li>If Regardless of their E-values, save</li> <li>If and their E-values is than 10€</li> <li>If and their E-values is the 10€</li> <li>If and the test is the top the saved to 100 the hits</li> <li>If and the test is the top the saved to 100 the hits</li> <li>If and test is the top the saved to 100 the hits</li> <li>If and test is the top the saved to 100 the hits</li> <li>If an test is the top the test is the top the saved to 100 the hits</li> <li>If a save hits to be saved to 100 the hits</li> <li>If a save hits to be saved to 100 the hits</li> <li>If a save hits to be saved to 100 the hits</li> <li>If a save hits to the test is top the test is the te</li></ul>
All Repictors     Gene Calling Tags & Comments BLAST Searches     Automated Analyses     Perform BLAST search on NCBI rr database     Verform BLAST search on NCBI rr database     Verform BLAST search on NCBI rr database     Verform Rula Search on NCBI re database
All Replicons     Gene Calling Tags & Comments BLAST Searches:     Automated Analyses     Fertom BLAST search on NCBI'r database     Save hits with E-values less than 10E [3
All Replicons     Gene Caling Tags & Comments     BLAST Searches     Automated Analyses     Fertom BLAST search on NCBI rr database     Save hits with Evalues less than IDE [3]     Flags & Comments     Flags & Comments     Flags & Comments     Flags     Flags & Comments     Flags
<ul> <li>All Replicons</li> <li>Perform BLAST searches</li> <li>Rutonated Analyses</li> <li>Perform BLAST searches</li> <li>Automated Analyses</li> <li>Perform BLAST searches</li> <li>Automated Analyses</li> <li>Perform BLAST searches</li> <li>Perform Searches</li> <li>Perform Search all Sheen Blasted Annotated Analyzes</li> <li>Perform all analyses of BLAST searches</li> <li>Perform all Sheen Blasted Annotated Analyzes</li> <li>Perform all analyses of BLAST hts</li> <li>Perform all Connerts</li> <li>Perform all Connerts</li> <li>Perform all Sheen Blasted Annotated Analyzes</li> <li>Perform all annotates</li> </ul>
<ul> <li>All performs</li> <li>Perform BLAST searches</li> <li>Rutonated Analyses</li> <li>Perform BLAST searches</li> <li>Rutonated Analyses</li> <li>Perform BLAST searches</li> <li>Automated Analyses</li> <li>Perform BLAST searches</li> <li>Rutonated Analyses</li> <li>Perform BLAST searches</li> <li>Rutonated Analyses</li> <li>Perform BLAST searches</li> <li>Perform Bealcon</li> <li>Perform Searches</li> <li>Perform Blast</li> <li>Perfo</li></ul>
<ul> <li>All Replicons</li> <li>Perform BLAST searches</li> <li>Rutonated Analyses</li> <li>Perform BLAST searches</li> <li>Automated Analyses</li> <li>Perform BLAST searches</li> <li>Automated Analyses</li> <li>Perform BLAST searches</li> <li>Perform Searches</li> <li>Perform Search all Sheen Blasted Annotated Analyzes</li> <li>Perform all analyses of BLAST searches</li> <li>Perform all Sheen Blasted Annotated Analyzes</li> <li>Perform all analyses of BLAST hts</li> <li>Perform all Connerts</li> <li>Perform all Connerts</li> <li>Perform all Sheen Blasted Annotated Analyzes</li> <li>Perform all annotates</li> </ul>
<ul> <li>All periods</li> <li>All periods</li> <li>Gene Calling, Tags &amp; Comments: BLAST Searches: Automated Analyses</li> <li>Perform BLAST search on NCBI'r database</li> <li>Save hits with E-values less than 10E [3 ]</li> <li>We have a search on NCBI'r database</li> <li>Vertice that the search on NCBI'r database</li> <li>Perform BLAST search on NCBI'r database</li> <li>Vertice that the search on NCBI'r database</li> <li>Save hits with E-values less than 10E [3 ]</li> <li>We have a search on NCBI'r database</li> <li>Vertice that the search on NCBI'r database</li> <li>Perform BLAST search on NCBI'r database</li> <li>Auto-Annotate</li> <li>Auto-Annotate</li> <li>Auto-Annotate</li> <li>Auto-Annotate</li> <li>Auto-Annotate</li> <li>Auto-Blast on the search on NCBI's callenge</li> <li>Auto-Annotate</li> <li>Auto-Annotate</li> <li>Auto-Annotate</li> <li>Auto-Annotate</li> <li>Auto-Annotate</li> <li>Auto-Annotate</li> <li>Search Sings &amp; Comments</li> <li>BLAST Searches</li> <li>Automated datase</li> <li>Export alternative start codons</li> <li>Export alternativ</li></ul>
C All Fieldons Gene Caling   Tags & Commerks   BLAST Searches   Automated Analyses   Perform BLAST search on NCBI Inr database Save hits with E-values less than 10E   3 ↓ Regardless of their E-values, sear   1 ↓ hits  ↓ Imit number of hits to be saved a 100 ↓ hits  ↓ Limit number of hits to be saved a 100 ↓ hits  ↓ Limit number of hits to be saved a 100 ↓ hits  ↓ Limit number of hits to be saved a 100 ↓ hits  ↓ seconds Send requests before retrieving for ↓ Limit number of hits to be saved a 100 ↓ seconds Send requests before retrieving for ↓ Limit number of hits to be saved a 100 ↓ seconds Matrix Annotate ↓ Autoave DMAMS file as   Steers, Blasted Annotate ↓ Autoave DMAMS file as   Steers, Blasted Annotate datama5 ↓ Export alemative stat codors ↓ Export A

Figure 4.2

- 1. In the Gene Calling window: Choose both analyses, favoring Glimmer.'
- Include the following Template in Tags & Comments window: 'SSC: CP: SD: SCS: Gap: Blast: LO: ST: F: FS': (See Section 9.6 for details) if you entered this in the Preferences Settings, this list will automatically appear.
- 3. Unclick Perform Blast searches in the **BLAST searches** window.
- 4. Select No analyses in the **Automated Analyses** window.
- 5. Click the '**Annotate**' button to launch the automated annotation. (Click '**Yes**' when prompted to "Erase features?")

The auto-annotation (without BLAST) takes only minutes. As the auto-annotation proceeds, the status of the process will be displayed at the lower left corner of the auto-annotate window. You will see Predict genes, predict tRNAs. Parsing, ...) When complete, an auto-annotation log will be generated. Review that what you requested was actually done. Then close that window and you are ready to review the genome.

## SOME NOTES ON AUTO-ANNOTATE OPTIONS

• One key Auto-Annotate option is the '**Perform BLAST searches on nr database**' checkbox. When checked, this option will BLASTP the protein product of each gene Auto-Annotate finds, then save the results for viewing later—a powerful tool, and recommended if you have the time. However, performing that many BLAST searches often takes more than 45 minutes, during which DNA Master will be inaccessible. If you'd like to move on to further steps quickly, uncheck this box and Auto-Annotate will run in fewer than five minutes.

See **Section 4.5** for how to BLAST genes at a later time.

• In the Gene Calling pane, we prefer to use the default option of using '**Both analyses**' (Glimmer and GeneMark), with '**Favor Glimmer Calls**' selected. Often, the two programs' gene calls differ only in the location of the start codon, so it doesn't really matter. If desired, you can try modifying options to see their effects on the resulting gene calls. Auto-Annotate runs quickly enough to experiment!

When there is a conflict between Glimmer and GeneMark calls, both calls will be reported in the gene's Notes. If the two programs agree on a gene, the Notes will list the favored program's call only.

• The checkbox to '**Export a Six-Frame map**' produces a translation of the sequence in all six frames, a useful asset for annotation. See Section 5 for generating maps and translations. at a later time.

## 4.3 Saving your file

As with any program, it is important to **save your file often** to protect changes you've made from being lost. This can be done by going to:

#### File → Save as DNAM5 file

Choose a new file name if you wish to keep both previous and current versions. This is a way to keep backups of work you've done. To avoid confusion about which file is the current version, it is helpful to establish systematic naming conventions when saving files. Remember to recreate your Documentation prior to saving.

If you are prompted to save the file before closing, always save it with a new name. Do not rewrite the last version. (When prompted to "Save before closing?" after you have saved it as a newly named file, your response is "No".)

The input file format was a fasta, but now it is a DNA Master file. The format of the file will now be [your phageName].dnam5. Therefore, when you save and re-open the file you will open as an "Archived DNA Master file".

## 4.4 Looking at the output of your automated annotation

Once the Auto-Annotate function has run, it will return you to your main phage window. Under the **[Overview]** tab, however, you will see some immediate differences.

Extracted from Fas	tA Library Echild.fasta			×
Overview Features Refe	erences Sequence Documenta	tion		
Genome Organism : GenomeID : 0 Length : 0 Repticons : 1 Features : 100 CAT Table : <u>Directory : DMTemp 41564</u> Repticon	PrototypeID : 0 Date Created : 10/17/2013 Genome Date : 10/17/2013 GC Content : 0.00	Taxonomy Domain : Division : Family :	Notes Notes on the Genome Notes and Taxonomy are only available when sequences are loaded from the database	
Replicon NCBI Date : 10/17/2013 Topology : Linear	NCBI Genome : Accession :			~
Length : 53159 Features : 100 ORFs : 100 RNAs : 0 GC Content : 0.00 ORF GC : 63.4 ± 2.8 GC3 : 83.6 ± 6.4 CA: 1 : 0.000 ± 0.0000 Translation Table : Unspeci	Version : GL: Status : Current Deletion : Vulnerable Origin : 0 Terminus : 0 fied: default to standard code		Notes on the Replicon	
Image: Section 1         Image: Section 1	1 - 50000         Posi           13         1         1         20         0	ion : 1 Controls >> 1 29 21 31 1 1 43	fap IZ <sup>-</sup> Map >> Controls IA M <so im="" imsin,="" imsin,<br="" msin,="">IS3159   5</so>	

Figure 4.3

For example, note that there is a map showing the predicted genes at the bottom of the window. Genes transcribed leftwards and rightwards are shown in different colors depending on how you have set your DNA Master preferences (**Section 1.6.2**; green and red in **Figure 4.3**).

This map is dynamic and can be manipulated as follows:

- Roll your mouse over the map. You will see the number changing in the box above it labeled '**Position**'. This reports the coordinate in the genome where your mouse is pointing.
- Click on the eq button to zoom in and the eq button to zoom out.
- Click on the left and right arrows to move a little each way, a lot each way, or to the extreme left or right ends.

## 4.4.1 Viewing the documentation

Auto-Annotate writes its output to the **Documentation**. Though you will generally work in the **[Features]** tab, it is useful to be familiar with this underlying Documentation. Click on the **[Documentation]** tab to take a look.

You will see that DNA Master has populated the Documentation with the consensus outputs from Glimmer, GeneMark, and Aragorn. In the example shown in **Figure 4.4**, the first line says "CDS (330-443)". This means the first feature is a protein-coding sequence (CDS) transcribed left to right and located at coordinates 330 – 443.

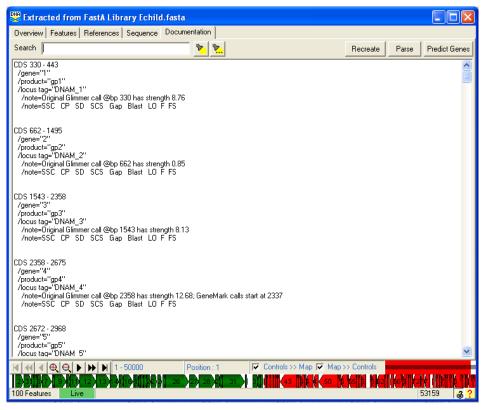


Figure 4.4

Additional data for each feature are shown in the indented lines that follow. For example, the first feature has a gene name of "1", a protein product named "**gp1**", a locus tag of "**DNAM1**", and a note about where Glimmer called the start position. The Notes also contain the template that you entered in Preferences (**Section 1.6**).

The data contained in the Documentation are also viewable in the Features Table (see below).

## 4.4.2 Viewing features in the Feature Table

The Documentation that you viewed above has been automatically **Parsed** by DNA Master into the **Feature Table**. Click on the **[Features]** tab to view the Features Table (**Figure 4.5**).

Note that the Documentation window is a text file. You can copy and paste its entire content. What you see is what you get. The same is not true for the Feature Window. When you are in the Feature Window, you are looking at the contents of a database that is being constructed about each feature of this project. You cannot Copy & Paste this window as you see it.

**Figure 4.5** demonstrates the first modification you will want to make to this window, namely "Widen the Feature Table". In the current form, the Feature Table displays the Name, 5' End, and Length of each feature. In order to also display the 3' End, right click on the heading of the table (1) and select "Widen Feature Table" (2). The list will now display each gene's **Tag** (Locus Tag), **Name, 5' End, 3' End**, and gene **Length**. You can select any gene by clicking on it. Gene "**1**" is selected in the example below, as indicated by the small black triangle next to it.

👺 Extracted from FastA Li	brary Echild.fasta	👺 Extracted from FastA Library Echild. fasta
Overview Features Reference	Sequence Documentation 1. Right click	Overview Features References Sequence Documentation
Sort By Index 👻 🤇	Name 5'End Length Description Sequence Product Regions Blast Context	Sort By Index 👻 📕 Tag Name 5'End 3'End Length 🔨 Description Sequence Product Re 🕨
Select Features Direct SQL	Name 1 <u>GenelD</u>	Select Features Direct Set
Type is All 💌		Type is DNAM_2 2 662 1495 834 Type CDS C 61
Name like	A ORF Analysis 330 Locus Tag DNAM 1	Name like DNAM 4 4 2358 2675 318 5 End 330 Locus Tag
GenelD =	5 BLAST (local) 443 Regions 1	GenelD = DNAM_5 5 2672 2968 297 3'End 443 Regions
Locus like	6 BLAST (nr) Tag	Locus like DNAM_6 6 3007 3477 471 Length 114 🕅 Tag
Start	7 Choose Start	Start DNAM_7 7 3559 4524 966 Direction Forward
Length		DNAM_8 8 4639 4797 159
Regions >	Sopy as rasking Underned     Copy Coordinate List to Clipboard	
	10 Save Coordinate List to File	
	12 Widen Feature List 2. Click	DNAM 12 12 2711 0220 1700 Broduct
	13 9995 1464 901	DNAM 13 13 9325 10788 1464 991
EC# like	14 10785 885	EC# like DNAM_14 14 10785 11669 885
Product like	15 11720 513 Function	Product like DNAM_15 15 11720 12232 513 Function
Function like	16 12262 957	Function like DNAM_16 16 12262 13218 957
FeatureID =	17 13288 183	FeatureID = DNAM_17 17 13288 13470 183
Tag like	18 13473 378 Notes 19 13947 195 Original Glimmer call @bp 330 has strength 8.76	Tag like DNAM_18 18 13473 13850 378 Notes DNAM_19 19 13847 14041 195 Original Gimmer call @bp 330 has
Hide Ignored Features	CONTRACTOR	Strend 10 10 1001 1001 100 shows the 270
-	20 14038 369 21 14406 336	SSU CP SD SCS Gap Blast LU F
Select All Features	22 14751 426	Select All Features DNAM_21 21 14406 14/41 336 F5
	23 15194 594	DNAM_23 23 15194 15787 594
	24 15899 408	DNAM_24 24 15899 16306 408
		DNAM_25 25 16420 16731 312
	Insert Delete Post Validate	Inset Delete Post Validate
	I - 12500 Position : 1 📝 Controls >> Map 🔽 Map >> Controls	H ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( ( (
2 3 4 5		
00 Features Live	53159 🙀 ?	100 Features Live 53159 3

Figure 4.5

If you look to the right, you will see six sub-tabs named **[[Description]]**, **[[Sequence]]**, **[[Product]]**, **[[Regions]]**, **[[Blast]]**, and **[[Context]]**.

The **[[Description]]** sub-tab is shown by default and contains basic information about the gene that you'll recognize from the documentation, including gene name, coordinates, product name, and notes.

DNA Master imports the data of the best Glimmer and GeneMark predictions. It reports it in the order in of your auto-annotation. If you ran both analyses and favored Glimmer, the notes will reflect that in the following ways.

The **Notes** for gene 1, shown above, indicate that Glimmer called the start at position 330. There is no mention of GeneMark in these notes, which means that GeneMark's gene call agreed with Glimmer's gene call. If the two programs do not agree, this will be mentioned in the Notes as shown in **Figure 4.6**. Check out genes 61 and 63.

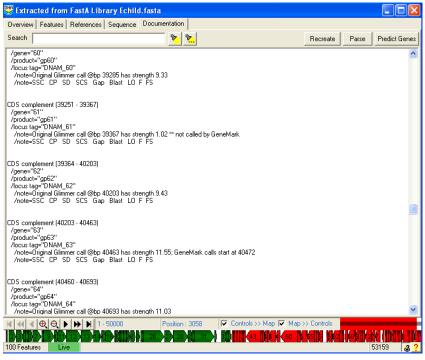


Figure 4.6

In the next example, gene 61 was predicted by Glimmer, but was "not called by GeneMark".

For gene 63, the assigned start is 40203 as called by Glimmer, but there is a note that "GeneMark calls start at 40472".

There is one more alternative to this notation. The notes can say GeneMark call @bp XXXX. This notation means that this gene was not called by Glimmer.

Your refinement of your annotation in **Section 8** will focus substantially on evaluating the predictions made by Glimmer and GeneMark. You will be resolving any ambiguities that have arisen and adding or deleting genes that were missed or errantly called by these programs.

You don't need them just yet, but you can see that there are also buttons (at the bottom of the central box middle) that will let you either '**Insert**' or '**Delete**' features. And eventually the '**Validate**' button will help you assess whether all your gene calls make sense.

#### 4.4.3 Viewing the sequence in the Sequence tab

Click on the [Sequence] tab.

You will see the sequence appear as before, but now you can use the '**Feature**' dropdown menu at the top left. When you click on this menu, a list appears that shows each gene and whether it is transcribed leftwards (R, for reverse) or rightwards (F for forward).

You can scroll down and select any of these and it will then select and highlight the corresponding part of the DNA sequence. This can be a very useful feature for examining specific parts of the genome.

)verview	Features Reference	s Sequence Documentation	
eature		💌 F 🏷 🌃 Position : 1	Raw ▶ BLASTN BLASTP Add Feature
	ORF 330 (1)	est Match ( at 3'End C Overall	OueralMateh
, P	ORF 662 (2)	orwards	Oligo
	ORF 1543 (3) ORF 2358 (4)	atchstring	
	ORF 2672 (5)	ackwards	
	ORF 3007 (6)	ACGGGTTTCCAAGTCGATCGGAGTCCC	GAGCCGGCGCAGGAGCGCCTCACCCAGCCTCTGTGCGCCCCCAGGA
	ORF 3559 (7)	CGGGTAGTTGTATGGGCTAATCGGCAA	ACGGCCTCTGAGGCCGCGAGACCAATGTCACACCAGGTGGTGGATG
	ORF 4639 (8)	TTAAGAGGACATGGCCTAGGTATGGCT	ACCCAAACTTAGATTCAAAACCAGTCCCCTGGCCCCCGTCGTCGG
	ORF 4797 (9)	CGGGGCCAGGTCCACCACCGCAGGGAG	GACCGCCATGAAGATCATCCGCTCGCTCGCCGGGGGCGCTCGCACTC
	ORF 6083 (10)	GCGGCTGGGATGCGGAAATCTACGAGC	CGTGGGATGAGGACGAATACCTCCTATAGTGATCTACGCCACTTGC
	ORF 6541 (11)	ACTCATGTATCTAGTTATTGAGGGCCT	AAAGGCCCGAATAAGAGCCGCACAGGCGGCTCTCTAAGAGCGCCC#
	ORF 7541 (12)	ACCGGCCTTGAGGGCCGGTTATCTGAC	CCGGCAACCGCCGGGTCTTCTGCCGCGCCCAGTGGCGCGGGCTCATA
F	ORF 9325 (13) ORF 10785 (14)	FTACGGCACTCGCTCGAGTGCCTACTG	GGCCTCGCAGCCGGGGAAGTTCGACGTTCTGAACCTGCGGATGACC
IF	ORF 10765 (14)		ACTTOGTTCOGGCCAACCTCGCAGCGTGGAACATGCCACGCCATCO
	ORF 12262 (16)		GATTACCGTTTCGAGACCGTCTGGTCGTCCCCCGAGCGTCTTCTCC
	ORF 13288 (17)		GTGGCGAGAAATGCCGCGGGTAGCGCAGGTCTGGAATACCTATCG
	ORF 13473 (18)		TTCCCACGGTCGGTTGGGGTCGACCGGACACATACGAATTCTGCT
120 F	ORF 13847 (19)		CTCCGATCGAAGCAAGAGGATCGCGAGCTGTTCACCCGAGGCGTGG
F	ORF 14038 (20)		
IP	ORF 14406 (21)		CCGGCTCAGATTCTGCGACGAAATAGACCTTCCCGAGGTCCGGGAC
317 F	ORF 14751 (22)		GGGCAAGCGCGGAGGGAGCGGAGGCGGTCCCGGAACAGGACGAGG
	ORF 15194 (23)		GTGGATCTGGAGGCGGTGGAGGCGGCGGAAAGGGCAAAGGCTCCGC
	ORF 15899 (24) ORF 16420 (25)		GGCTCCGGTGGTGGCGGTAGCCAGATGCCGAGCAAGCCCAAACCGT
			CCAGCGTCGAGCCGACAACATCGCCGCGGTCAACCCGCGGTGGAAC
593			ACTGCACCAGGTGTGCGGCCACCGTCGAGATGAGGGCCCGCGGTTA
787			ACCCACATGTCGTCCGCGAAGTGGATCGATGAGCACGGCAACCCGG
881	CTCAGTGGGACTAC	CTCGACGGCTCCCGCGACAGAATCATGGACGA	GATCAACGCCCGAGCTGCGTTGTGGCCCGAGGGAGCCCGAGGGCTC
975	ATCCGCACGACCTG	GGAGGGCGGCGGCGGTCACATCTTCAACTGGG	AGAAGCGCGACGGCGTGATCCGGTTCATCGACGGCCAGCCCAACC#
069	GTGGGACGCCAGCT	CGTCGTGGACCAACCAGGCGTCGGAGTGGGGG	CACAGCACCGCGGTTGTCCGCGTTGACCACCTGAACCCGCACGAAC
163	ACCTGTCGAAGTGG	GTACGCAACCGCACGCCAGAGGAGATCAACGC	TCCCCTGCGCCGCGAGGTTACAGTCGAGATGCAACGGCGCGGGTAC
257	AAGCTCGGTGAGCC	GATCCGCCAGGTGTTCCTCGAGGGCTGGGATG	ACATCCGATCCGGCCGGGGGTACGACCCGACGAAGTACGCCAACGA
351	CCCTGAGTTCAAGG	CGATCTACGAAGCGGCAGTGAAATGGGCAAGG	AGGCCGATCTGATGGTGACGTTCGAGCAGGCTAGAGGCATCGTGT
445	CGACAACCGAGCCC.	ACCTGTACCCGGCCGAGGCCGACTTCCAGGTA	GCCACCTGGGGCTGGGAGAACGACACGGCGTACCAGCTCGTCTGCC
	GTOCOTACGOGATG	GTCTACCCGGCACGCAATGCCGATGACCGTCT	CTGGATCTCTGACCAGGACGGCCCCTTCCTCACGGTCGACAAGGCC
		1 - 50000 Position : 3058	Controls >> Map 🔽 Map >> Controls
1 2 2 2		1 034011, 0000 1	

Figure 4.7

## 4.4.4 Viewing ORFs in the Frames window

The Frames window is an especially important one for determining and assessing start site choices. To open the Frames window (we use Angelica in the example below) select:

#### DNA → Frames

A window will open that has a graphical representation of the six possible translation reading frames, with each row representing one reading frame. Full-row-height vertical lines represent in-frame stop codons, and half-row-height vertical lines are possible start codons. At the lower left in the window is a box displaying the nucleotide coordinate corresponding to the position of your pointer as you mouse over the display. There are also buttons that allow you to scroll through your genome and zoom in and out.

At the lower right corner of the Frames window, there are seven additional buttons. Click on the button labeled '**ORFs**' (red circle in **Figure 4.8**).

This will highlight all the features currently in your feature table as shown in the screenshot below. Genes in forward reading frames are green, those in reverse reading frames are red, and tRNAs are blue. Note that the genes are numbered in this format also.

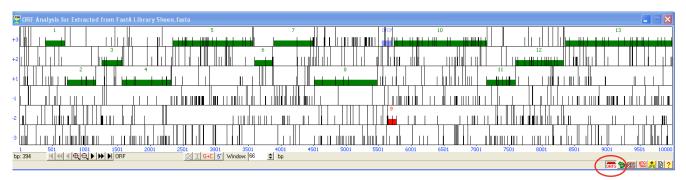


Figure 4.8

This next screen shot of the Frames window has been modified to expand the frames. This is done by clicking on the 0 button to zoom in. In order to a select a gene (gene 6 in this example) click in the ORF (box) that contains the highlighted gene (Figure 4.9).

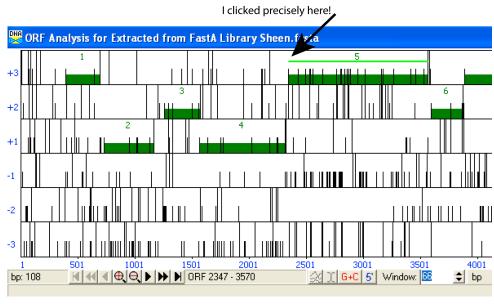


Figure 4.9

A thin, horizontal green line will appear that extends from the nearest upstream start codon to the next downstream stop codon. Now click on the '**RBS**' (ribosomal binding site) button in the bottom right corner of the Frames window (**Figure 4.10**).

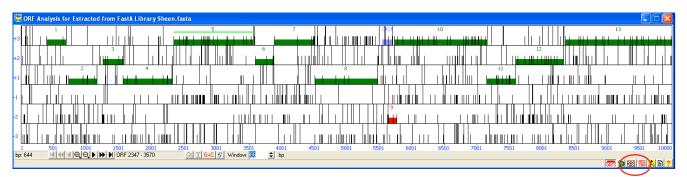


Figure 4.10

Another window titled "Choose ORF start" will appear, shown in Figure 4.11.

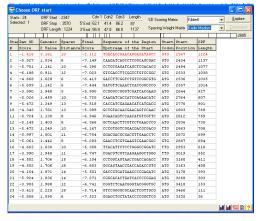


Figure 4.11

The information displayed here will depend on the setting you choose in the upper right corner of this window. (SD scoring Matrix & Spacing Weight Matrix.) For this display, Kibler 6 and Karlin Medium settings were chosen. Each row in this window represents all of the possible start codons **in the ORF you clicked on** in the Frames window, the corresponding upstream nucleotide sequence, the gene length resulting from that start, and three columns of scores: a Raw SD score, a Genomic Z value, and Final score. Also included is the spacer distance between the chosen SD site and the start. One line's text may be red, and this is a direct result of where you clicked in that ORF. You will find more information about choosing starts in Section 8.

When evaluating your gene calls and choosing between possible start sites, you may find it helpful to have all three windows open at once, as shown in **Figure 4.12** for the Sheen genome.

🚆 ORF Analysis for Extract	ed from FastA I	Library Sh	een.fasta																	
				5		_	7		I		0	10	10 11 I III						13	
						6			ЦЦ	ЦЩ	┶┲╨┷┰┲					12				
									.			և հայ հ				12		hallb. I		
								ЦШШ			<u> </u>	шңн тш						ЩЦЦ		
		, , , , , , , , , , , , , , , , , , ,	du i								h									
		╷╵┛╷╜╷╴╸					# I I I			╷╝╴┥╴┥╷└╴										
	. III - 1		- mahar	I	ա տես	l l l i	dalah dala	dan Lu							սիս է հ		hlin			իսի արելո
╵┞╶╷╝╶┦┙	╷╷╷║└──┬╴┥		_							ш	┯┯╋┲┯╋				штт	┉╆╴╴┾╀╴╀╷┥				
2 1 1 1 1 1 1 1	llein hid				1111	1		I				Li II			ս ս հ	. It ille of state bold	ևսևս		ш	
	╵╵┛╵║╴╶╿╶		1	тпт			╘╌╹╝╻╕				┶┰╁┶┟╨	┍╍╷╶└┤					պոսրո	-11-11-11		┶┯╗╧┲╬┱╼╄╇┯
3 11 1 1 1 1 1 1 1 1		տոհենտ						1.0			_ المعالم						l . lu			
1 501 1001	1501	2001	2501			501	4001	4501		DNA (	Choose OF	lF start								
	H ORF 234				5 Window		bp	1001			rts : 24	ORF Start			1 Cdn2 Cdr	SD S	coring Ma	trix Kibl	er6	
Position of internal starts										Sele	ected : 1			5'End 62.1	41.4 86.2	2 87	-	t Matrix Kar		Document
Contraction Contraction									28072635			ORF Lengt	th:1224	3'End 55.9	47.8 86.5	5 1137 Spac	ang weign	R Maduk Jusen	In orderan	2805
👺 Extracted from FastA Lit										St o	a Raw SD	Genomic	Spacer	Final	Semience	of the Region	Start	Start	ORF	2005
Overview Features References	s Sequence Do	cumentation								#	Score	Z Value	Distanc			of the Start	_	Position		
Sort By Index 👻 🔳	Tag	Name	5' End	3 End	Length 🔥	Descript	on Sequence	e Produ	act   Regio	1	-1.418	3.161	10	-2.112		AACAGGAGATAGCC	GTG	2347	1224	
Select Features Direct SQL	DNAM_1	1	397	693	297	Name	5	<u> </u>	GenelD	2	-5.927	1.034	8	-7.149	CAAGATCA	GCCTCGGCAGCGAC	GTG	2434	1137	
Type is All 💌	DNAM_2	2	732	1166	435	Туре	CDS		<u>61</u>	3	-5.701	1.141	10	-6.396		ATCAGCTCGACACG		2494	1077	
Name like	DNAM_3	3	1259	1576	318	5'End	1000		Locus Ta	4	-6.188	0.911	12	-7.023		TEGGTETGTEEGGE		2533	1038	
GenelD =	DNAM_4 DNAM_5	4	1566 2347	2318 3570	753	3'End			Regions	8 6	-4.668	1.628	6	-6.413		GTCTGTCCGGCGTG AACTCACCGGCCCG		2536 2557	1035	
	DNAM 6	6	3599	3877	279		1004		-	7	-3.990	1.948	5	-5.990		GGTCGATCACGAAG		2644	927	
	DNAM_7	7	3889	4512	624	Length			Tag	8	-5.026	1.459	10	-5.720		CCATCGAGAACATC		2764	807	
Start >	DNAM_8	8	4509	5477	969	Direction		2		9	-5.472	1.249	13	-6.518	CACCATCO	AGAACATCATGACG	ATG	2770	801	
Length >	DNAM_86		5551	5628	78		on Table Un	defined		10	-4.343	1.781	13	-5.389		ACGAACAGTCCAAC		2803	768	
Regions >	DNAM_87		5629	5704	76	EC Num	ber			11	-5.724	1.130	8	-6.946		CCAACATGTCGTTC		2812	759	
% GC <	DNAM_9 DNAM 10	9	5635 5731	5772 7155	138	Product				12 13	-5.145	1.403	8 10	-6.366		TCGTCCTGAACCCG CGGACGACGCGACG		2836 2863	735	
CAI >	DNAM_10 DNAM 11	11	7152	7595	444	gp5				14	-5.997	1.001	11	-6.754		CGACGTTGAACCTC		2863	699	
EC# like	DNAM 12	12	7592	8332	741					15	-5.061	1.442	8	-6.283		TGAAGTCGAACGGC		2887	684	
Product like	DNAM_13	13	8359	10059	1701	Function				16	-4.663	1.630	10	-5.358	TTACATTO	TCCTGGGCCGGATC	TTG	2953	618	
Function like	DNAM_14	14	10056	11552	1497					17	-3.990	1.948	11	-4.747	CGACGTCG	TCAAGAAGGCTGGG	TTG	3019	552	
FeatureID =	DNAM_15	15	11549	12562	1014					18	-4.502	1.706	15	-6.104		TAACCGACCAGACC		3160	411	
Tag like	DNAM_16	16	12621	13130	510	Notes Driginal	Glimmer call @	abo 2347	has streng	19	-4.502	1.706	18	-6.803		CCGACCAGACCGTG		3163	408	
Hide Ignored Features	DNAM_17	17	13160	14149	990	SSC: C	P: SD: SCS:	Gap: Bla	sol: LO: F:	20 21	-4.154	1.871	14	-5.501		TGAACCCCGAGATC TGATCACCCCGGAG	_	3178 3268	393	
Colori All Eastures	Insert Delete	Post Vali	date							22	-3.905	1.988	12	-4.741		ACGGGTACGGGTGC	_	3418	153	
	- 50000	Position :	1	Contro	ls⇒> Map 🔽	Map XX D	etrole			23	-3.413	2.220	18	-5.714		CGCAACTCGTTGCG		3460	111	
		Position :		Control				E.L.	ER K LI	24	-5.808	1.090	6	-7.553	GGAGCTGC	TATACCCCGGCTCG	ATG	3535	36	
87 Features		FF761 2	1 28	12/1		TUPIL 44	PUR R CS	- R R	PIR R R											
	412 28			22.478	(a)	20.20	-	(Jack)			_									<b>100 100 100 100</b>

Figure 4.12

# 4.5 Running the BLAST function

When determining the settings for the automated annotation above, we cautioned about the time it takes to run the BLAST function and you may have elected to skip BLASTing. Sooner or later, however, you will need to do this. When you can allow an hour or so for **continuous** Internet connectivity, you should run the BLAST function. To do so, go to:

#### Genome → BLAST All Genes

- In the dialog box, we recommend that you use the settings shown in Figure 4.13.
- You may want to modify these setting and run more than one BLAST. For example, you may want to run a BLAST excluding Mycobacteriophage and Mycobacterium phage. You will have to save each of these in separate .dnam5 files.

Genome BLAST of Extracted fr	om FastA	Library Sheen.fa	sta 📃 🗖 🔀
Retrieve			
Communicating with the Server			Ignore Definitions including the following terms
Number of hits to request from server	100 🚖		
Number of hits expected	10 🚖		
Delay between multiple BLAST queries	20 🚖	seconds	
Send requests before retrieving for	100 🜲	seconds	
Maximum number of outstanding requests	10 🔶	pending queries	
Saving Results Locally			
Save hits with E-Values smaller than 10E-	3 🚖		
I▼ Regardless of E-Value, save at least	1 🔹	Hits	
Limit number of hits saved to	100 🜲	Hits	
Execute BLAST Query			
C BLAST all protein-coding genes	🛛 Skip gene	s already analyzed	
C BLAST all RNA-encoding genes	Clear prev	ious BLAST results	
BLAST pre-selected set or genes     BLAST all genes	Export sur	nmary of results	
	hours for lar 9:00 pm Ea	rge batch jobs are astern time	~
Peak Hours			?

Figure 4.13

- Click on 'Blast All'.
- The setting are different for Peak and off-Peak hours. This is due to additional parameters put in place by NCBI to accommodate the load put on the NCBI servers as people around the world use BLAST.
- DNA Master will send the predicted protein sequences in your file in batches to the NCBI server, then retrieve the results and store them. Be patient during this process! Windows may briefly indicate that DNA Master is "Not Responding" during this period, but that's because it's processing! You won't be able to use DNA Master until the BLAST is complete. The 'hang time" is most likely a reflection of the extensive retrieval attempts as too many outstanding requests are still pending.

Even though you still only have a draft annotation that was generated automatically, it is very helpful to do the BLAST search **before** finalizing gene calls, because the data will be extremely helpful during the process of annotation refinement. However, as you finalize your file for the Quality Control process you must re-BLAST the genome to ensure that the BLAST data saved in the file is a result of the ORFs of any changes you have made.

When all BLAST searches are complete, DNA Master will report "**Genome BLAST has been completed**" as shown in **Figure 4.14**.

- Saving Results Locally	
Save hits with E-Values smaller than 10E- 3	
✓ Regardless of E-Value, save at least	
✓ Limit number of hits saved to     100      ✓ Hits	
Execute BLAST Query	
C BLAST all protein-coding genes	
C BLAST all RNA-encoding genes	
BLAST all genes	
Blast All Cancel	
	$\sim$
Sent 30 requests (90 total) to the QBlast server	^
Retrieving completed results Continuing to send.	
Sent 10 requests (100 total) to the QBlast server	
All requests have now been sent to the QBIast server Retrieving completed results	
Waiting for 0 results	
Genome BLAST has been completed	
	*
	?

Figure 4.14

- You may now close this BLAST window.
- You can now view BLAST results for any gene by returning to the [Feature] tab and selecting a gene, then clicking on the [[Blast]] sub-tab to the right. (Figure 4.15)
- Save the file with a new **Name**. (This will not auto-save.)

	र्ग 🔳	s   Sequence   Docu   Tag	Name	5' End	3'End	Length 🔼	Description Sequence Product Regions Blast Context
lect Features		DNAM 1	1	330	443	114	
		DNAM 2	2	662	1495	834	Score Target Description 808 hypothetical protein First 005 [Mycobacterium phage First]
oe is All	<b>_</b>	DNAM 3	3	1543	2358	816	799 gp6 [Mycobacterium phage Turbido]
me like		DNAM_4	4	2358	2675	318	675 hypothetical protein [Mycobacterium hassiacum]
nelD =		DNAM 5	5	2672	2968	297	666 gp5 [Mycobacterium phage PackMan]
cus like		DNAM_6	6	3007	3477	471	663 gp5 [Mycobacterium phage PackMan]
		DNAM_7	7	3559	4524	966	005 gp3 [Mycobacterium priage 025]
art <mark>&gt;</mark>		DNAM_8	8	4639	4797	159	BLAST Hit
ngth <mark>&gt;</mark>		DNAM_9	9	4797	6086	1290	Accession YP 007677414 Export
gions ゝ		DNAM_10	10	6083	6544	462	GI 472437721 Export A
GC 🔀		DNAM_11	11	6541	7518	978	Length 156 Delete
		DNAM_12	12	7541	9328	1788	Max Score 808 Date 10/18/2013 Delete A
		DNAM_13	13	9325	10788	1464	High-Scoring Pairs (HSP)
		DNAM_14	14	10785	11669	885	HSP Data Alignment
iduct like		DNAM_15	15	11720	12232	513	Bit Score 315.8 Identities 154
nction like		DNAM_16	16	12262	13218	957	Score 808 %Identity 98.72
aturelD =		DNAM_17	17	13288	13470	183	E-Value 0.0E0 Positives 156 Length 156 %Similarity 100.00
g like		DNAM_18	18	13473	13850	378	% Aligned 100.0 % Gaps 0
	_	DNAM_19	19	13847	14041	195	Query 1-156
Hide Ignored F	eatures	DNAM_20	20	14038	14406	369	Target 1 - 156
Select All Fea	itures	Insert Delete	Post Valid	ate		<b>⊻</b>	

Figure 4.15

In the example above we clicked on gene 6. Under the **[[Blast]]** sub-tab, you can see a window with the BLAST hits listed, with a score and a description. Below that is a pictorial report on the extent of the match (shown as a red bar depicting the part of the gene product – i.e. gp6 in this case – that matches the selected subject). Below that are the data for the hit (HSP Data), and if you click on the **[[[Alignment]]]** sub-sub-tab it will show the actual alignment.

In the example shown in **Figure 4.16**, we clicked on a BLAST hit further down on the list of matches, and then clicked on the [[[Alignment]]] sub-sub-tab. Note that you can now see the amino-acid matches in the bottom right pane.

Sort By In	dau	Ţ	1	Ē	Tan	Name	5'End	3'End	Length		Description Sequence Product Regions Blast Context
· )				Н	DNAM 1	1	330	443	114		
Select Feat	ures	Dire	ect SQL	н	DNAM 2	2	662	1495	834		Score Target Description
Type is	s	All	-	H	DNAM 3	3	1543	2358	816		569 hypothetical protein ODIN_4 [Mycobacteriophage Odin]
Name li	ike			E	DNAM 4	4	2358	2675	318	- =	561 hypothetical protein [Mycobacterium phage SWU1]
GenelD	= LÍ	_		F	DNAM 5	5	2672	2968	297		554 gp5 [Mycobacterium phage EricB]
	ike			Þ	DNAM_6	6	3007	3477	471		544 hypothetical protein [Mycobacterium hassiacum]  422 gp2 [Mycobacterium phage Benedict]
				É	DNAM 7	7	3559	4524	966		422 gp2 [Mycobacterium phage Benedict]
	>			F	DNAM 8	8	4639	4797	159		BLAST Hit
Length	>			F	DNAM 9	9	4797	6086	1290		Accession AEJ93398 Export
Regions	>			F	DNAM 10	10	6083	6544	462		GI 339753374 Export /
% GC	<			F	DNAM 11	11	6541	7518	978		Length 149 Delete
	,			E	DNAM 12	12	7541	9328	1788		Max Score 422 Date 10/18/2013 Delete /
	_			F	DNAM_13	13	9325	10788	1464		High-Scoring Pairs (HSP)
	ike				DNAM_14	14	10785	11669	885		HSP Date Alignment
Product li	ike			Г	DNAM_15	15	11720	12232	513		2 GTRODICKED BEEVERNAPE SPIETVOVIC PVEIPELCOV
Function li	ike				DNAM_16	16	12262	13218	957		1     ++  ++    +   +  +
FeatureID	= 1				DNAM_17	17	13288	13470	183		4 ATRGPVPERT DQTVRHS EPVDKVEVFG EVKVPDLGDV
	ike				DNAM_18	18	13473	13850	378		52 TEMYESIKNS AAVKYYEPID WIYAKLILYI LNOELIASRO
-					DNAM_19	19	13847	14041	195		51 ++++++++ + +++++++++++++++++++++++++
Hide Igr	norec	Feat	ures		DNAM_20	20	14038	14406	369		51 QDLYQAMQES GQSKFYEPSD WQVARLTLLA LNEELIAARH
Select	AILE	eatur	es		DNAM_21	21	14406	14741	336		102 TAINOMLSSL LLTEGDRRRV RLEIERNPGD PTAGKVVDHT
	_	_		L	DNAM_22	22	14751	15176	426		101   +   ++  +            + +
				μ	DNAM_23	23	15194	15787	594	~	101 TALNQHLTTL MLTEGDRRRV RIELERKPSQ P-EGVVINAA
					nsert Delete	Post Valid	ate				
[4] ]	€le		I III	1 -	50000	Position : 3	7143	Control	s >> Map		Map >> Controls
MARIE						22 2 2	Vary as	V.G. or		L KK	111 43 1161686 50 1163165616 6 6211 666111 27 8 11 1011965

Figure 4.16

• Save your file as described in **Section 4.3** to ensure your BLAST data are stored.

# 4.6 Re-opening an archived (saved) file

When you save files, Opening archived (saved) files is straightforward. Go to:

#### File → Open → Archived DNA Master file

• Browse to your saved .dnam5 file and select and open it.

# 5 Gathering additional information for refining your annotation

There are three additional pieces of data that we recommend gathering at this point. The first is a **six-frame translation** of your sequence labeled with your predicted genes. The second is a **provisional genome map**. The third is a **graphical output of the GeneMark-Smeg** analysis. Depending on your genome, you may also need the **tRNA predictions** from the web-based Aragorn and tRNAscan-SE algorithms. The output of these programs will be used in **Section 8**.

# 5.1 Generating a six-frame translation

With your genome open in DNA Master (we used Etude below), go to:

Genome → Six-frame translation

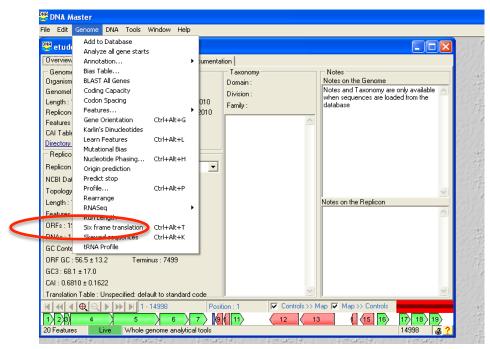


Figure 5.1

The six-frame translation window will open.

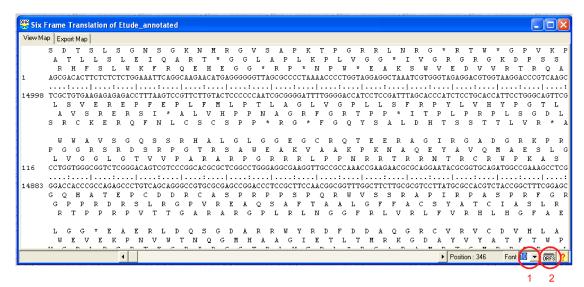


Figure 5.2

- Adjust the size of the font by entering '8' in red circle **#1** in **Figure 5.2**.
- Click on the ORFs button in the red circle **#**2 in **Figure 5.2**.

Note that the ORFs predicted in your auto-annotation are now highlighted. Also note that this window scrolls right and left rather than up and down. When you first click on the ORFs button you may not see highlighted text if there is no gene predicted in the extreme left end of your genome (which is what is shown by default). If you like, you can scroll to the right using the scroll bar at the bottom to see more sequence.

But you can also be assured that your selection has been chosen because the ORFs button at the bottom right is now shown in red (see **Figure 5.3**).

	Frame Translation of Etude_annotated
View №	1ap Export Map
1 14998	S D T S L S G N S G K N H R G V S A P K T P G R R L N R G * R T W * G P V K P W W A V S A T L L S L E I Q A R T * G G L A P L K P L V G G * I V G R G R G K D P S S P G G R S R R H F S L W K F R Q E H E G G * R P * N P W * E A K S W V E D V V R T R Q A L V G G L G ACCALCITICTCTCTGGAATCAGCGAACATGAGGGGGGTTAGCGCCCCCTAAAACCCTCGGAGGGTAAATCTGGGGAGGAGGGGGCGTGGTAAGAGCGCGGGTCTGGGGGGGG
133 14866	S R C K E R Q F N L C S C S P P * R G * F G Q Y S A L D H T S S T T L V R * A R T P P R G Q S S R H A L G L G G E G C R Q T E E R A G I R G A D G R K P R L G G * E A E R L D Q D S R P G T R S A W E A K V A A K P K N A Q E Y A V Q H A E S L G W E V E K P N V W T N T V V P A R A R P G R R R L P P N R R T R R N T R C R W P K A S V G R L R S R T F G P I GGACAGTCGTCCCGCTCGCGCTGGGGGGGGGGGGGGGGG
	P C D D R C A S P R P P S P Q R W V S S R A P I R P A S P R F G R N P P Q S A S R K S W S L R G P V R E A Q S A F T A A L G F F A C S Y A T C I A S L R P Q S T S F G F T Q V L P V T T G A R A R G P L R L N G G F R L V R L F V R H L H G F A E T P L N L L R V N P G S G D A R R W Y R D F D D A Q G R C V R V C D V H L A * W P H S Y C * R A G * R L S R G <b>G G H H A A G I R T L T M R K G D A Y V Y A T F T W P N G R I R T V D V R V N G F H E D</b>
265 14734	R G C T P L V S R L * R C A R A M R T C M R R S P G L M A A F V L L T C G L T A F T R T TCAGGGGAGCGACCCCCCCCTGCTATCGGAGCTTCACGGCGATGCGGAGCGGAGCGGAGCGGAGCGGAGCGGAGCGGAGCGGAGCGGAGCGGGGGG

Figure 5.3

Now click on the **[Export Map]** tab at the top left of this window. We recommend using the default settings as shown in **Figure 5.4** below.

Six Frame Translation of Etude_annotated	
View Map Export Map	
<ul> <li>Export Forward Translations</li> <li>Export Reverse Translations</li> <li>Export Forward DNA Sequence</li> <li>Export Reverse DNA Sequence</li> <li>Export Center Ruler</li> <li>Colorize Annotated ORFs on Translation</li> <li>Colorize Annotated RNAs on Sequence</li> <li>Prevent line breaks within blocks</li> </ul>	100       Nucelotides per Line         8       Point Text         0.5       Inch Margins         • Portrait Mode         • Landscape Mode         Export RTF File
•	▶ Position : 13 Font 🖲 💌 륝 ?

Figure 5.4

- Click the 'Export RTF File' button.
- In the dialog box that opens, choose a name and location for your translation file, then click '**Save**'.

To view your translation file, it's best to open it with Microsoft Word. Please note that your computer may not be set to open a file with an **.rtf** extension with Word. If not, open the file using the '**Open'** function from within Word. You may want to save as a .pdf in the future to preserve formatting more easily.



Figure 5.5

The formatting in Word has been set so that you have 100 base pairs per line. Both strands are shown with the coordinates of the bases on the left. There are vertical markers between the strands positioned every ten bases. All six possible translations are shown, with the predicted genes from your automated annotation highlighted.

**Note:** Because this document can be > 100 pages, we recommend you only print one or two copies to share among a class, or a single copy for an individual. Hard copies are particularly useful when annotating potential ribosomal frameshifts (**Section 8.4.3**).

# 5.2 Generating a provisional genome map in DNA Master

Another useful tool in DNA Master is the ability to make a genome map. This map is not comparative (though you will make a comparative map using Phamerator in the next section), but rather just a separate file of the map shown at the bottom of the sequence panel. Still, it is a useful way to see your gene calls in the context of the entire genome.

To make a genome map (we use mycobacteriophage Timshel below), go to:

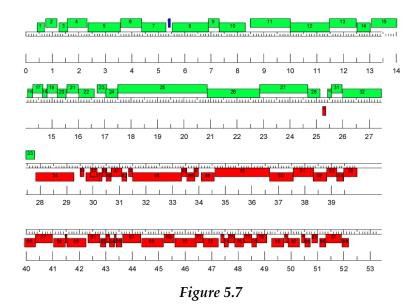
🙀 Export ORF Map		
Map Format	ORF Labels	Export Options
Scale 1/2" / kb 💌	Center Label Name : repA	Entire Sequence
Tiers 3 🗲	Size Adjust 🔿 None	C Region Selected in Sequence Form
AutoSize 25 🔹 kb per tier	C Trim to fit	Specified Region :
	Shrink to fit	From 1 🗲 🔳
Estimated	Top Label None-	
Dimensions	Size Adjust 🔘 None	To 53278 🗲 🔳
	<ul> <li>Trim to fit</li> <li>Shrink to fit</li> </ul>	Length 53278 🚖
Stagger ORFs	• Shrink to rit	
Divide By Direction	Text Size Default 💌	🔽 Draw ORFs 🛛 🗖 Draw Promoter
Central Marker 100 bp Ruler 💌		🔽 Draw RNAs 🔲 Draw Oligo
		🔽 Draw in Color
Bottom Ruler 500 bp Ruler		Show image when complete
Label Bottom Ruler		🍞 Draw Map
		?

#### DNA → Export Map

Figure 5.6

- In the dialog box that appears, many options are available. We recommend you use the settings shown in **Figure 5.6**, except that the '**Tiers**' field may need to be adjusted. Three or four tiers are acceptable for a genome of up to about 60 80 kb in length. If your genome is larger, increase the number of tiers accordingly.
- Click on 'Draw Map'.
- Choose a filename and location to save to, then click 'Save'.

The file will be saved as YourFileName.wmf (Windows metafile). This file can be opened by Preview (on a Mac), Paint, Canvas, or similar drawing programs. Depending on the program, you can manipulate this file in numerous ways. At the very least, you should see an illustration of your genome, similar to one shown in **Figure 5.7**.



# 5.3 Generating a graph of coding potential using GeneMark

As we noted above, GeneMark is a gene prediction program, and the version embedded in DNA Master runs heuristically, using parts of the genome you enter to train the program to identify coding potential. When using the version on the web, you can:

- 1. Use an existing coding model to predict the genes.
- 2. Generate a graph of the coding potential.

The host profile we recommend using is that of *Mycobacterium smegmatis* mc155, assuming that you used this host to isolate your phage. If you used a different host, you will obviously need to select a different bacterial profile for GeneMark. Even if you isolated a phage using *M. smegmatis* mc155 as your host, you may find different mycobacterial models (or even Actinomycetales models) yield higher coding potential outputs. As a learning opportunity, you can even use the programs at GeneMark's home page <a href="http://exon.gatech.edu/GeneMark/">http://exon.gatech.edu/GeneMark/</a> to obtain a graph of coding potential of the heuristic predictions (like what is imported into DNA Master). Use the version found here (<a href="http://exon.gatech.edu/GeneMark/genemarks.cgi">http://exon.gatech.edu/GeneMark/genemarks.cgi</a> ).

To run host trained, web-based model GeneMark (we use Mycobacteriophage Sheen below), go to:

- <u>http://exon.gatech.edu/GeneMark/genemarks.cgi</u> . Also found on the Links page of <u>http://phagesdb.org</u> as GeneMark (version 2.5).
- Once on the site, Select '**Browse**', then find and select your sequence file. This is the same YourPhage.fasta file that you imported into DNA Master.
- From the 'Select Species' dropdown box, select '*Mycobacterium\_smegmatis* mc<sup>155'</sup> (assuming you are annotating a mycobacteriophage genome).
- In the **'Output Format'** section, check **'LST**' under 'Output format for gene prediction and 'PDF' under output options. You can ignore the optional email option (See **Figure 5.8**).
- Click on the 'Start GeneMark' button at the middle left (See Figure 5.8).

GeneMark.hmm prokaryotic									
Prokaryotic GeneMark.hmm Alexander Lukashin and M GeneMark.hmm: new soluti Nucleic Acids Research (199	Mark Borod ons for gene	finding.							
Prokaryotic GeneMark.hmm version 2 John Besemer, Alexandre Lomsadze and Mark Borodovsky GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions, Walelick Acids Research (2001) 29, pp 2007-2618									
sets of pre-computed species species is available <u>here</u> . Th unsupervised training on eac <u>Browse GeneMark.hmm</u>	This webpage provides access to gene prediction program GeneMark.hmm prokaryotic (version 3.25) and to the sets of pre-computed species specific algorithm parameters (model parameters). The list of currently supported species is available here. These parameter sets were derived by application of the GeneMarkS that carried out unsupervised training on each genome. Browse GeneMark.hmm prokaryotic manual								
Input sequence and Selec	-	r multi FASTA format)							
Select species	or, upload file: Browse Sheen.fasta Select species Mycobacterium_smegmatis_MC2_155 :								
Action Start GeneMark.hmr	m Reset								
Options									
Outpu	it format prediction	Output options	Optional: results by E-mail						
Coding potential graph GFF     GFF     Coding potential graph (not for multi FASTA) g PDF     PostScript     Compress files     Compress files									
Advanced options									
Switch off gene start relations	ated motif(s)								
		Contact Us 1 Hama							
		Contact Us   Home							

Figure 5.8

Once GeneMark has run, a new window will appear as shown in Figure 5.9.

• Click on the link 'gmhmmp.out.pdf' just below.



Figure 5.9

- Save and open the .pdf file. We recommend changing the file name to something more useful, like Sheen\_smeg.pdf.
- The coversheet of the file provides you with the specifications of that particular run of the program. You will want to check that the specifications match your expectations (See **Figure 5.10**), i.e. check that the following items are identified: the correct sequence (and its length) and that *M. smegmatis* is listed as the Matrix (or whatever host that you picked.

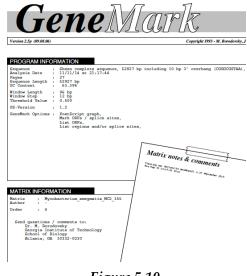


Figure 5.10

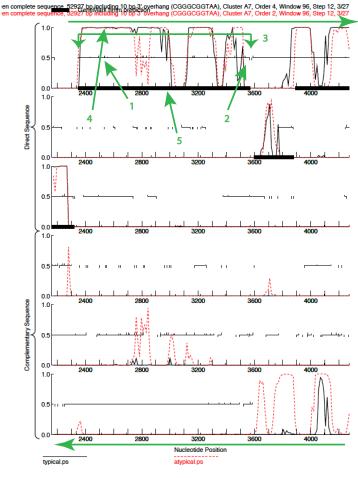


Figure 5.11

We recommend that you **print** this file because it is a good place to make notes as you refine your annotation. Several features of this output are described. (See **Figure 5.11**).

- All six frames are represented and are separated from one another by solid horizontal lines.
- The top three frames are in the forward orientation; the bottom three in the reverse orientation.
- In each frame, the start codons are shown as small upward facing ticks (#1 in figure). Note ATG start upticks are a bit longer than GTG starts while TTG are not shown.
- In each frame, the stop codons are shown as small downward facing ticks (#2).
- The horizontal lines in the middle of each row represent open reading frames (ORFs) (#3).
- A graphical representation of coding potential is shown (#4). Note that the Black lines are typical coding potential and the red lines are atypical.
- The black bars at the bottom of each ORF with coding potential (#5) signify regions that GeneMark predicts as likely coding regions, based on coding potential and positioning of stop codons, but for the most part is of limited utility in gene identification.

# 6 Phamerator & other Tools to assist with annotation

This section describes the basic applications of two valuable tools in our annotation suite. Phamerator was designed and implemented by Dr. Steve Cresawn at James Madison University. A new method for pham building was developed and implemented by Charlie Bowman at the University of Pittsburgh in 2014.

At present Section 6 contains information about using Phamerator (Section 6.1) and Starterator (Section 6.2) for genome annotation.

## 6.1 Phamerator

#### 6.1.1 Overview

**Phamerator** is a Linux-based program that compares phage genomes, their genes, and their gene products, and then displays the results of these comparisons in a variety of useful ways. Phamerator is comprised of two basic parts: an underlying database that contains the results of the comparisons, and a graphical interface to that database.

One of Phamerator's key features is that it groups gene products into "**Phamilies**" (generally referred to as "**Phams**") when the pairwise alignment scores (using BLASTP and ClustalW) are above a defined threshold.

In previous years, we created phamilies (phams) by performing pairwise amino acid sequence alignments between every pair of genes in the database using CLUSTAL and BLASTP. Proteins were assembled into phams if the pairwise alignment scores were above an empirically determined threshold value for each program. This was very computationally expensive---and therefore very time consuming. Incremental building of the 627-member Mycobacteriophage\_Draft database required about 6.5e7 pairwise alignments to be performed to build phamilies (or the approximate number of years since the proposed Cretaceous-Paleogene extinction!).

**New this Year:** In order to continue adding new phages to our database, we switched to a different program, kClust, to build phams. kClust is an alignment-free approach for clustering amino acid sequences that is based on the concept of k-mer profiles. A k-mer profile is built by dividing a protein sequence into small units (called k-mers; ours are about 4-6 aa long) and then creating a list of all k-mers found in the sequence, followed by determining the number of times each k-mer appears in the sequence. This is done for every protein sequence in the database. Then, instead of comparing protein sequences directly, these k-mer profiles are mathematically compared and scored.

The specific pipeline we use produces groupings very similar to the original groupings produced by BLAST and CLUSTALW, with low incidences of both false-positive and false-negative groupings. We accomplish this by using parameters and a workflow that we have optimized through testing and validation, and is loosely based on the kClust\_iter pipeline(CITE).

Our pipeline works as follows (See Figure 6.1):

First, protein sequence k-mer profiles are compared, and then proteins are grouped based on 75% predicted amino acid conservation with a 25% size cutoff. The size cutoff refers to the minimum size a gene must be to a larger gene to be compared. This grouping collates genes that are very similar, which cuts down on false groupings that were influenced by size differences.

Second, a multiple sequence alignment (MSA) is generated for each group from every sequence in the group and a consensus sequence is determined. Finally, the consensus sequences from the groups, as well as orphams, are then reclustered with kClust solely based on an e-value of 1e-4 with a 50% size cutoff; resulting in a pham profile in Mycobacteriophage\_Draft very similar to the one built using the BLASTP/CLUSTALW method; with fewer falsely inflated phams.

Phams are thus groups of proteins with a high degree of similarity to one another.

Phamerator is especially useful for generating and comparing genome maps of multiple phages through the visual interface that displays whole genome nucleotide and protein sequence relationships, as well as the conserved domains within genes.

See the Phamerator Help Menu for the User Manual.

For more on Phamerator and its mechanics, see the following paper.

Cresawn SG, Bogel M, Day N, Jacobs-Sera D, Hendrix RW, Hatfull GF. "Phamerator: a bioinformatic tool for comparative bacteriophage genomics." *BMC Bioinformatics*. 2011 Oct 12; 12(1):395.

Hauser M, Mayer CE, and Soding J. "kClust: fast and sensitive clustering of large protein sequence databases." *BMC Bioinformatics*. 2013 Aug 15; 14:248

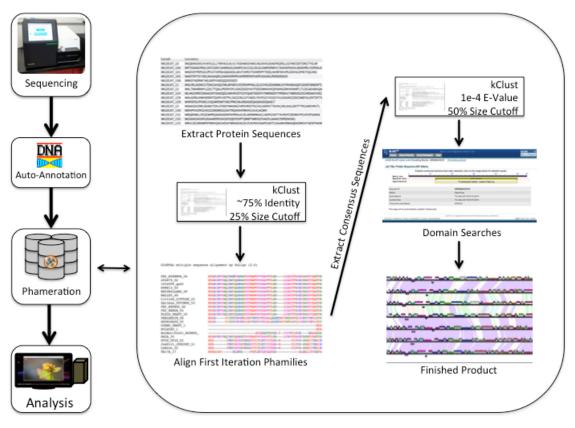


Figure 6.1

**Figure 6.1:** Figure New Phamerator workflow. After sequencing and autoannotation, genomes are run through the Phamerator pipeline on our servers. First, all protein sequences are extracted and subjected to a grouping using kClust. Then, Multiple Sequence Alignments are made from the resulting groupings. The consensus sequences from these groupings are

removed and added to a second iteration of kClust along with orphams. The results of the second iteration are stored in the database as phamilies. Domain database searches are performed on all of the genes, and then the data is ready to be viewed in the Phamerator client.

#### 6.1.2 Why Phamerator is useful to you at this stage of your annotation

Phamerator maps provide an easy-to-understand representation of how your genome compares to similar genomes. This is useful during annotation because it draws attention to places where your automated annotation diverges from the finalized annotation of a closely related (and often GenBank-published) genome. It also provides a genome-wide perspective and thus a context for the annotation refinement, functional analysis, and other explorations to follow. It is also the primary source of protein functions as denoted in the **Descriptions.** See **Section 10.3.3** for more details.

## 6.1.3 How did my genome get into Phamerator already?

In order to expedite your annotation workflow, we have taken each newly sequenced genome, generated an automated annotation (just as you did in **Section 4**), and entered all of these files into a Phamerator database that contains all sequenced mycobacteriophages. The database generated is called 'Mycobacteriophage\_Draft' because it contains auto-annotated draft genomes along with finalized and published annotations. The auto-annotated genome names are given the suffix "\_Draft," so as to distinguish them from the GenBank-quality files. At a later time, when you've refined your annotation and it is submitted to GenBank, your draft annotation may be replaced in Phamerator with your final annotation.

### 6.1.4 Making Phamerator maps

- Open the Phamerator program. (Allow up to a minute for the main window to appear, as Phamerator will check for new databases when it boots.)
- Click on '**Phages**' in the left 'Sources' pane.
- The name of the current database will be displayed at the top of the window (top red oval in **Figure 6.2**). Make sure the database is "Mycobacteriophage\_Draft". If not, go to **Edit** → **Preferences** and select Mycobacteriophage\_Draft from the Database dropdown menu.

Note: Some of the screen shots in this section may not directly match the current version of Phamerator, but the functionality is the same. Enjoy!

Figure 6.2

You can now choose genomes you want to compare to one another. We recommend:

- Your phage
- Some closely related phages (in the same cluster or subcluster)

You should decide carefully which genomes you want to compare. For example you may not want to compare all of the genomes from a particular cluster if there are a large number. If your phage belongs in a cluster with several different subclusters, you may want to use a representative of each subcluster.

A good rule of thumb is to shoot for no more than about six genomes to start with. You can always return to this and generate more maps as you need them.

- Scroll through the list—or use the search bar—to find your phage.
- Click on it to select it. It will be highlighted.
- To add additional genomes to your selection, scroll through to find the genome you want (if you used the search function, make sure you clear all search terms so that you can see all of the genomes).
- Use Ctrl-click (or equivalent if using an emulator—on Macs it is often Ctrl-Shiftclick) to add another genome to your selection. You can also select consecutive genomes in the list by using Shift-click.
- Repeat to select as many genomes as you want to include.
- The phages can also be sorted by simply clicking on the column headers—such as cluster, Length, GC%—to help find relevant genomes.

a 🖬						
		B				
Open Add	Map Pham Circle Goog	gle Maps				
Phages	Name	Length (bp)	GC %	Number of Genes	Cluster	
Phams	Thibaut-DRAFT	100327	00.793590		J	
Genes	Optimus	109270	60.790702	230	J	
Domains	Anaya	60835	66.394345	98	К1	
	CrimD	59798	66.876819	95	К1	
	JAWS	59749	66.617015	94	К1	
	BEEST_draft	59906	66.555938	91	К1	
	BarrelRoll-DRAFT	59672	66.615833	95	К1	
	Angelica	59598	66.388134	94		
	Adephagia	59646	66.606311	94	K1	
	TM4	52797	68.113719		K2	
	Pixie	61147	67.303384	100	K3	
	Fionnbharth-DRAFT	58076	68.007439	93	К4	
	Larva-DRAFT	62991	65.295042	95	K5	
	Ender-DRAFT	74731	58.796216	111	L1	
	JoeDirt	74914	58.780735	126	L1	
	LIDIE	70704	FO 703334	* 0.0	1.4	

In **Figure 6.3**, four genomes are currently selected, indicated by the orange highlight.



- Once you've finished selecting genomes, click on the button that says '**Map**' (red circle in **Figure 6.3**). Be patient, as it can take a minute (or more for a large number of genomes) to generate the map.
- When the map window appears, you will see something like this:



Figure 6.4

Congratulations! You've made a Phamerator map using your phage's draft annotation.

#### 6.1.5 Understanding and using the genome maps made by Phamerator

When the **Genome Map** window appears, you will probably only be able to see a small portion of the genomes. You can resize the window to see more, but you probably won't be able to see the entire picture unless you change the zoom factor. A sample is shown in **Figure 6.5**.

• To see a view of your entire genome, click the '**Zoom Out**' icon at the top left repeatedly until you can see the genome ends.

😣 🖨 🗉 Genome Map	
File View Color	
Q     Q     Image: State	
Manager and the second	
Winner	فالماسم والتجاري والمستجا المتعادية
<u> <u> Ser andre ser and</u></u>	
The second states in the secon	
1	
4	
click on a gene	click on a gene

Figure 6.5

Each genome is represented as a hash-marked horizontal bar. Forward-transcribed genes are shown as rectangles above the bar, and reverse-transcribed genes as rectangles below the bar. Each gene is colored according to the **Pham** to which it belongs, making it easy to see relatives in other genomes.

You may have noticed that some genes appear to have smaller yellow boxes within them. These represent matches to the NCBI Conserved Domain Database. These will be particularly useful later when attempting to determine gene functions, but they can be confusing at this stage. Fortunately, Phamerator makes it easy to toggle the display of these domains. Just go to:

**View** → **Show Domains,** then click to unselect this option.

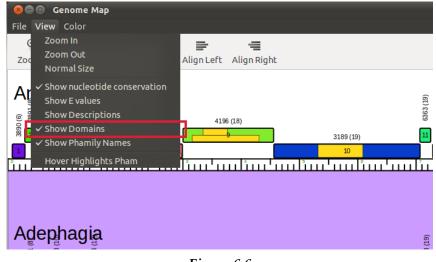


Figure 6.6

Lots of information is displayed on Phamerator maps.

• Click the '**Zoom In**' icon several times to get a closer look.

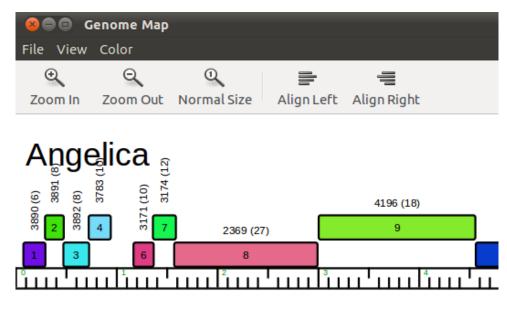
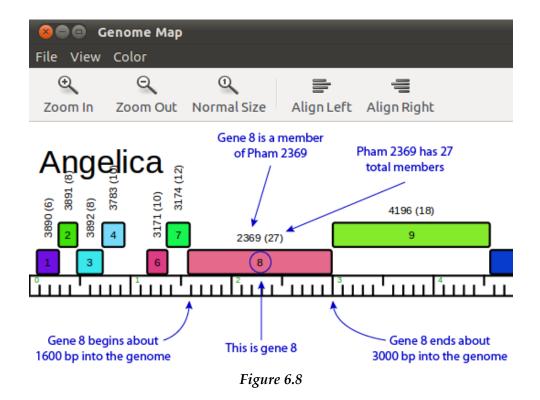


Figure 6.7

Again, the white bar at the bottom represents the genome sequence itself, and is marked with green numbers every 1,000 base pairs (bp). The small hash marks coming up from the bottom show 100 bp intervals, while the ones coming down from the top show 500 bp intervals.

Each gene's box has a number within it that represents that gene's number in this genome. There are also two numbers above each gene; the first is the number of the Pham this gene belongs to, and the second—in parentheses—is the total number of members of that Pham.

Putting all this together, we can determine that Angelica's gene 8 begins at ~1600 bp, ends at ~3000 bp, is a member of Pham 2369, and that there are 26 other members in that Pham:



#### 6.1.6 Viewing nucleotide sequence similarities in Phamerator

# A NOTE ON TWO DIFFERENT TYPES OF SIMILARITY

**Nucleotide** sequence similarity is a comparison of the **DNA sequence** (A, C, G, T) of two **genomes**. It is often determined by running BLASTN. On Phamerator maps, nucleotide similarity is shown by colored vertical boxes between genomes.

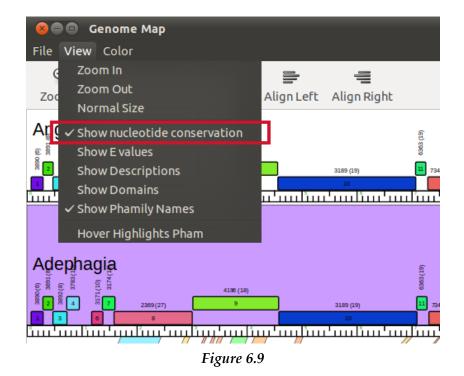
**Protein** similarity is a comparison of the **amino acid sequence** of two **proteins**. It is often determined by BLASTP or ClustalW. On Phamerator maps, protein similarity is shown by similarly colored gene boxes.

Phamilies, or **Phams**, are determined based on **protein similarity and NOT nucleotide similarity**.

Don't confuse these two types of similarity, or you may misinterpret the data that Phamerator is showing!

While Phamerator was conceived to compare protein sequences to other protein sequences, it can also show nucleotide sequence similarity between genomes. To enable this function:

View → Show nucleotide conservation should be checked (as in Figure 6.9).



Once you've turned on '**Show nucleotide conservation**', you may see colors between the genomes on your map, as shown in **Figure 6.10**.

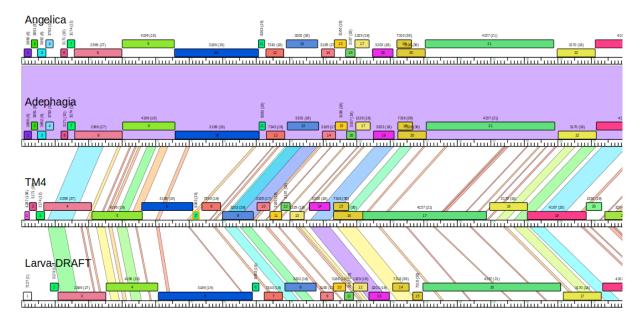


Figure 6.10

Nucleotide sequence similarity is shown by the (often slanted) shaded regions (boxes) **between genomes**. Each box represents one BLASTN alignment, and is colored based on its E value, with violet representing the best matches (lowest E values) and red the worst matches (highest E values). White areas indicate that there is **no** nucleotide similarity in those regions.

Looking at the screenshot above, it is apparent that the top two phages (Adephagia and Angelica) have widespread nucleotide similarity to one another, as indicated by the solid

purple between the two genome maps. The other two phages shown (TM4 and Larva) have multiple regions of nucleotide similarity, though these areas are interrupted by dissimilar (white) areas and have higher E values. This segmented similarity is a reflection of what you saw in the BLAST searches performed earlier. The top two genomes are members of Subcluster K1, while the bottom two are members of other subclusters within Cluster K.

Phamerator-generated maps can be extremely helpful when trying to evaluate a gene start codon in your novel genome that (for example) produces a bigger gene than in the compared genomes. A quick look at the Phamerator-generated map lets you know that the upstream sequence does or does not have sequence similarity.

#### 6.1.7 Other Phamerator features

There are many other functions in Phamerator. Several examples are below.

- 1. Click on the colored portion of any gene's box to select it, and the nucleotide and amino acid sequences of that gene are shown in the bottom panels.
- 2. You can move the order of genomes around in the display. This is important, because the nucleotide similarities are only displayed by comparing two adjacent genomes in the display. To do this, click and hold on the **NAME** of a phage you want to move (it is on the extreme left, and you may need to scroll over to it), then drag the genome either up or down to where you want it and release it.
- 3. You can move a genome to the left or right to better compare it to its neighbors. To do this, Ctrl-Click-hold on the **NAME** of the phage (on a Mac, this might be Ctrl-Shift-Click-hold), then drag to the left or right and release.
- 4. You can also align genes from multiple genomes, such as those within a particular Pham. For example, you may have noticed that gene 13 in Adephagia is in the same Pham as gene 9 in TM4. Select gene 13 from Adephagia, then Ctrl-click to select gene 9 from TM4, and verify that both genes are highlighted. Then press the "Align Left" or "Align Right" button at the top of the genome map.
- 5. You may want to also explore the '**Hover Highlights Pham**' function, available in the **View** menu.

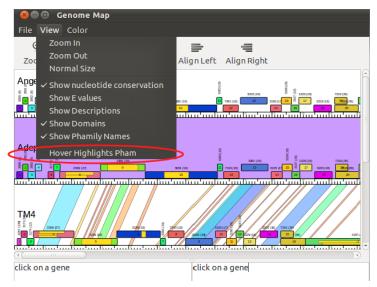


Figure 6.11

This function's use is that when your mouse hovers over any gene, only the gene members of that particular Pham are shown in color, while all others go white. This is a very useful function for easily seeing gene conservation or loss in different genomes.

#### 6.1.8 Saving Phamerator maps

Finally, if you would like to save the map as a file, from the Genome Map window go to:

- File  $\rightarrow$  Save As
- Enter a name and select your desired file type (pdf files are a good choice).
- Click 'Save'.

## 6.2 Starterator

#### 6.2.1 Overview

A companion program of Phamerator, Starterator uses comparative genomics to highlight conserved start codons for a given pham. Often, the scientific evidence for selecting a specific start codon for a gene is not in agreement (discussed in section 8.4). However, by analyzing that gene as part of a larger multiple sequence alignment, a start codon that is common to all genes in the alignment can sometimes be found.

Starterator aligns the longest possible ORF for each gene (from stop codon to stop codon) in a pham using ClustalW, and produces a graphic that shows all of the possible starts for all of those ORFs in the pham. It is possible to run Starterator on multiple genes or phams at a time, however, the more genes that are selected, the longer the program will take to finish. An entire genome can take several hours. During classroom annotation, it may be more advantageous to run Starterator on each gene individually as the need arises; or to generate a single .pdf of an entire genome outside of class-time to be shared in a future class.

A more detailed guide for the installation and operation of Starterator can be found here at the Software page of <u>PhagesDB</u> and <u>Seaphages</u>.

#### 6.2.2 Interpreting Starterator results:

The Starterator output of a pham is comprised of a graph followed by a text summary. In the graph below, each horizontal bar, or "track" in the output represents a single ORF--unless some of the phages in the alignment have identical sequences. Identical sequences are then pooled together in the same track. The length of all of the bars is determined by the longest ORF in the pham (in the example below, the ORFs in tracks 1, 6, 14, 16, and 18 are longer than the rest of the ORFs in the pham). The pink regions of the bars are included in the Clustal alignment, the white regions represent gaps in the alignment. The multi-colored vertical bars in each track represent all the possible start codons for each ORF. The starts are numbered from left to right in order of appearance. Starts that are aligned between tracks have the same number and the same color. Starts that are currently annotated in Phamerator are colored blue, the other start colors are selected randomly. The **Figure 6.12** was generated by running Starterator on Sisi gene 5: Pham 1523

Ì °	N 91 9	<u>`</u> 0	N 9	.*
Track 1	Track	Track 1	Track 1	Track 1
Track 2	Track 2	Track 2	Track 2	Tyack 2
Track 3	Track 8	Track 3	Track 3	Jack 3
Track 4	Track 4	Track 4	Track 4	Track 4
Track 5	Track 5	Track 5	Track 5	Track 5
Track 6	Tragk 6	Track 6	Track 6	Track 6
Track 7	Tragk	Track 7	Track 7	lýack 7
Track 8	Track 8	Track 8	Track 8	Track 8
Track 9	Track 9	Track 9	Track 9	gack 9
Track 10	Tragk 10	Track 10	Track 10	Track 10
Track 11	Tragk 11	Track 11	Track 11	Cack 11
Track 12	Track 12	Track 12	Track 12	Track 12
Track 13	Track 13	Track 13	Track 13	Track 13
Track 14	Tragk 14	Track 14	Track 14	P Track 14
Track 15	Track 15	Track 15	Track 15	Track 15
Track 16	Track 16	Track 16	Track 16	Track 16
Track 17	Track 17	Track 17	Track 17	Track 17
Track 18	Tragk 18	Track 18	Track 18	lyack 18
Track 19	Track 19	Track 19	Track 19	lýack 19
Track 20	Track 20	Track 20	$\frac{\text{Track 20}}{12}$	Jack 20

Figure 6.12

While the genes in the pham above are very similar to each other—as shown by the perfect alignment of starts from start "6" through the ends of the ORFs--- the region upstream of the start is not. All of the genes in this pham have start "6" in common. However, start "6" has not been selected as the start in the annotation in Phamerator in a number of genes. Many of these genes are the lengthier genes in the pham, and these starts may have been chosen by annotators who selected the longest ORF possible.

Following the graph is a report that lists all the genes that were included in the alignment and the track each is represented by:

#### Pham 1523 Report

 Track 1 : Ardmore\_5, Taj\_5, Tweety\_gp5, Shauna1\_5, Mutaforma13\_5, Wee\_gp5, SG4 5

Track 2 : Florinda 5

Track 3 : Ruby\_Draft\_4, MisterCuddles\_Draft\_4, Girr\_Draft\_4

- Track 4 : Brocalys Draft 6, Saal 5
- Track 5 : Cabrinians\_Draft\_5
   Track 6 : Spartacus\_5, Hades\_Draft\_5
- Track 7 : Che8\_5
   Track 8 : SuperGrey\_Draft\_5, Bipolar\_Draft\_5, Batiatus\_Draft\_5, Ovechkin\_Draft\_5
- Track 9 : GUmbie\_5, Llij\_5, Mantra\_Draft\_5, PMC\_5, Dante\_Draft\_5
- Track 10 : ShiLan\_5
- Track 11 : Dorothy\_5, Inventum\_Draft\_5, Daenerys\_5, Pacc40\_5
   Track 12 : Bubbles123\_Draft\_5
- Track 13 : Llama 5
- Track 14 : DotProduct\_5
   Track 15 : Empress\_Draft\_5
- Track 16 : SiSi 5
- Track 17 : OlympiaSaint Draft 6
- Track 18 : Hamulus 5
- Track 19 : Fruitloop\_5
   Track 20 : Ibhubesi\_5

#### Figure 6.13

Sisi gene 5 is Track 16 of the graph.

The final component of the report lists a "recommended start"; based on the most frequently annotated start(s) found in Phamerator, along with all the possible starts for the Starterated gene(s).

#### Suggested Starts:

SiSi\_5, (6, 4435) Gene Information: Gene: SiSi\_5 Start: 4435, Stop: 5028 Candidate Starts for SiSi\_5: [(1, 4292), (2, 4313), (4, 4379), (6, 4436), (7, 4448), (9, 4490), (10, 4583), (11, 4739), (12, 4796), (14, 4880)]

#### Figure 6.14

Starterator suggests that Sisi gene five should use start 6, at Sisi genome coordinate 4435.

It is important to remember that the Phamerator database contains draft annotations, and therefore a commonly selected start found by Starterator may be an artifact of unrefined autoannotations generated by the computer gene-calling programs.

# 7 Guiding Principles of Bacteriophage Genome Annotation

## 7.1 Overview

Genomes are best annotated when you understand their context. Their context can include how similar or different they are to other phages. What cluster are they a member? How similar are the phages of that cluster? How similar is your phage to the next closest phage? You may want to BLASTN your sequence on PhagesDB, align your sequence with its closest match at BLASTN at NCBI (use the align two sequences tool and format using display with "Pairwise with dots for identities"). You will want to use your Phamerator data for nucleotide and protein comparisons, and the DNA Master Genome Comparison tool (See Protocols -> Further Discovery -> Exploring Bacteriophage Biology). Once you have an overview of your phage genome, you are ready to start calling the genes.

Though the automated annotation you have created using DNA Master will usually identify more than 90% of genes correctly, some genes will need to be manually added, modified, or deleted. Therefore, all gene predictions must be reviewed to identify those that must be changed. In this section, we provide a set of principles that should guide you as you evaluate and improve upon your draft annotation.

It is helpful to think of the process of evaluating your draft annotation's gene calls as an application of these principles: together they will help you make the best possible gene predictions. It is essential to understand that any annotation consists of making a **prediction** as to how the genetic information is organized and used. In the absence of experimental evidence to support a given gene call, there is no right or wrong answer; there are, however, well-supported or ill-supported predictions.

As with any set of principles, the ones presented here will conflict with one another at times. It's your job to weigh one against another and make the best gene calls possible.

Because of the importance of these principles, this section is dedicated wholly to presenting them. Read them carefully before beginning an annotation, and keep them nearby as you work.

# 7.2 The Guiding Principles

The following two pages list the principles themselves. As mentioned above, we recommend that you print those two pages, read them carefully, and keep them close at hand as your refine your gene calls. Skip ahead to **Figure 8.1**. This is a diagrammatic representation of the work involved in annotating a genome for GenBank submission. The gene prediction analysis described in the Guiding Principles are part of the first two boxes of that diagram (Sections 8.4.1 and 8.4.2).

Because these are principles, and not unbreakable rules, you'll see words like "usually," "generally," and "typically" used quite frequently. Remember that phages are famous for finding exceptions to "rules", so very little is truly set in stone.

# **GUIDING PRINCIPLES** OF BACTERIOPHAGE GENOME ANNOTATION

- 1. In any segment of DNA, typically only one frame in one strand is used for a proteincoding gene. That is, each double-stranded segment of DNA is generally part of only one gene.
- 2. Genes do not often overlap by more than a few bp, although up to about 30 bp is legitimate.
- 3. The gene density in phage genomes is very high, so genes tend to be tightly packed. Thus, there are typically not large non-coding gaps between genes.
- 4. Protein-coding genes should have coding potential predicted by Glimmer, GeneMark, or GeneMark Smeg. Start sites are chosen to include all coding potential. These are, by far, the strongest pieces of data for predicting genes.
- 5. If there are two genes transcribed in opposite directions whose start sites are near one another, there typically has to be space between them for transcription promoters in both directions. This usually requires at least a 50 bp gap.
- 6. Protein-coding genes are generally at least 120 bp (40 codons) long. There are a small number of exceptions. Genes below about 200 bp require careful examination.
- 7. Switches in gene orientation (from forward to reverse, or vice versa) are relatively rare. In other words, it is common to find groups of genes transcribed in the same direction.
- 8. Each protein-coding gene ends with a stop codon (TAG, TGA, or TAA).
- 9. Each protein-coding gene starts with an initiation codon, ATG, GTG, or TTG. But note that TTG is used rarely (about 7% of all genes). ATG and GTG are used at almost equivalent frequencies.

CONTINUED...

# **GUIDING PRINCIPLES**

- 10. An important task is choosing between different possible translation initiation (i.e., start) codons. The best choice of start site is gene-specific, and gene function and synteny must be carefully considered. As phage genes are frequently co-transcribed and co-translated, less weight may be given to optimal ribosome binding site sequences in start site selection. Identifying the correct start site is not always easy and is predicated on the following sub-principles:
  - a. The relationship to the closest upstream gene is important. Usually, there is neither a large gap nor a large overlap (i.e., more than about 7 bp). If the genes are part of an operon, a 4bp overlap (ATGA), where a start codon overlaps the stop codon of the upstream gene, is preferred by the ribosome. Therefore RBS scores may have little bearing in this type of gene arrangement.
  - b. The position of the start site is often conserved among homologues of genes. Therefore, the start site of a gene in your phage is likely to be in the same position as those in related genes in other genomes. But be aware that one or more previously annotated and published genes could be suboptimal, and you may have the opportunity to help change it to a more optimal one. Homologues in more distantly related genomes (those of a different cluster) may prove more informative because alternate incorrect start sites are less likely to be conserved. Use Starterator!
  - c. The preferred start site usually has a favorable RBS score within all the potential start codons, but not necessarily the best. A notable exception is the integrase in many genomes, which has a very low RBS score. Our experimental data suggests that some genes do not have an SD sequence.
  - d. Manual inspection can be helpful to distinguish between possible start sites.
     The consensus is as follows: AAGGAGG 3-12 bp start codon.
  - e. Your final start-site selection will likely represent a compromise of these subprinciples.
- 11. tRNA genes are not called precisely in the program embedded in DNA Master, and require extra attention. (Please refer to **Section 9.5**.)

# 8 Gene by gene: evaluating and improving your draft annotation

#### 8.1 Overview

This section describes the heart of the matter: how to go through a draft annotation, one gene at a time, and decide whether or not the automated annotation called the gene correctly. You will spend most of your annotation time in this section, because you'll need to follow the steps here between 50 and 250 times per genome, once per gene!

If you've been following this guide step-by-step, you probably have all the items listed below ready to use. If you've jumped directly to this step, you may want to gather the items listed below to assist you as you go.

- 1. Your draft annotation file (from **Section 4**) open in DNA Master. (It is helpful to have DNA Master's Frames window open as well, with the windows arranged as shown as the last figure in **Section 4.4.4**.)
- 2. A printout of the Guiding Principles of Bacteriophage Annotation (Section 7.2).
- 3. Phamerator running, preferably with a map displaying your genome and related genomes (**Section 6**).
- 4. Your GeneMark outputs, most notably the one generated using a host (*M. smegmatis*) as the target. (**Section 5.3**).
- 5. (Optional) A printout of your DNA Master-generated map (Section 5.2).
- 6. (Optional) A printed six-frame translation of your sequence (Section 5.1).

One useful configuration is to have a pair of annotators work together on a genome, using two computers, one with DNA Master running, and the other with Phamerator.

## 8.2 Button-pushing mechanics reserved for Section 9

The goal of this section is to help you **decide** what modifications need to be made to your draft annotation. In order to keep this section manageable and streamlined, we've moved the detailed **mechanics** (button-pushing) of many of these operations to **Section 9** of this guide.

Section 9 should be used more as an à-la-carte reference than as a step-by-step guide. For example, you probably won't need to read **Section 9.4.1** about properly annotating a programmed translational frameshift until you come across one during your annotation review.

## 8.3 Decision Tree for evaluating the draft annotation

To help clarify how to use Sections 8 through 12 of this guide, a decision tree is shown in **Figure 8.1**. There are three beginning tracks depending on what feature of your genome you're currently investigating: one for **Protein-Coding Genes** (Section 8.4), one for **Gaps in the Annotation** (Section 8.5), and one for **Special Considerations** (Section 9.4).

**Blue** boxes are **decision** points, most of which are covered in the rest of **Section 8**. To answer the question in each decision box you'll need to keep in mind the Guiding Principles described in **Section 7** of this guide as well as the rest of the information in this section.

**Purple** boxes are **action** points where you implement the changes you've decided on. These actions are described in detail in parts of **Section 9**.

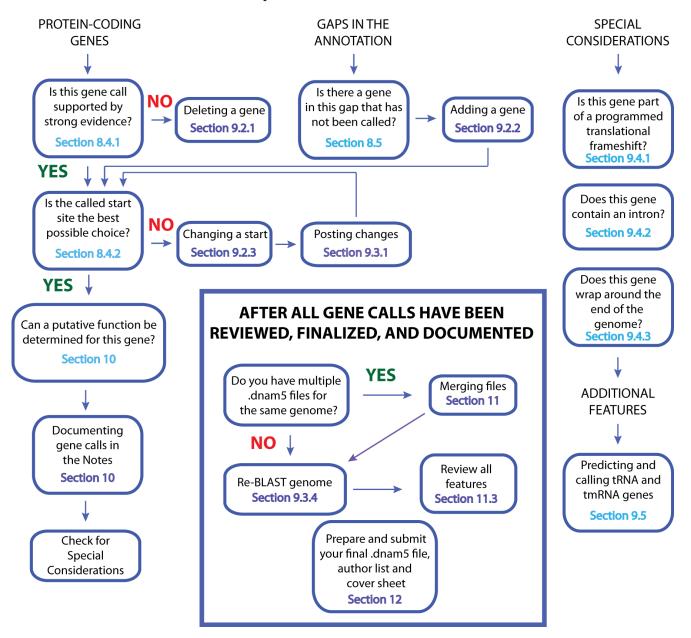


Figure 8.1

# 8.4 Evaluating protein-coding gene calls

The vast majority of features you will need to investigate are protein-coding genes, so you will use this section extensively. The first few genes you review will probably take some time as you become quite familiar with the process, but as you gain experience things will move faster.

It is best to start with your second open reading frame, which will typically be called gene '2' in the DNA Master feature table. We recommend skipping gene 1 until you have some practice. With no upstream sequence, some rules do not apply. Just remember to revisit this gene later!

In evaluating the veracity of the prediction of this gene that was performed automatically by Glimmer and GeneMark, there are several questions you should ask, described in the following sub-sections. We'll use a sample gene, but you can proceed with your genome from here on.

It is also recommended that—in accordance with good lab practice—you keep notes of your thoughts and decisions as you proceed. You'll use them to enter your final Notes (**Section 9.6**).

#### 8.4.1 Is the designation of this ORF as a gene well-supported?

If it's not already selected, click on the [Features] tab.

In the central column, click on the gene in question to select it. A small black triangle will appear to the left of that gene, indicating that it is active.

Look at the "Notes" field under the [[Description]] sub-tab.

Sheen	_Blas	ted								
Overview	Featu	ures Refere	nces	Sequence Docu	mentation					
Sort By	Index	-	•	Tag	Name	5' End	3' End	Length	^	Description Sequence Product Regions Blast Context
Select Fe	atures	Direct SQL	۱L	SHEEN_1	1	397	693	297	_	Name 5 GenelD
Туре	is	All •	i-	SHEEN_2	2	732	1166	435		
			ĽĽ	SHEEN_3	3	1259	1576	318		
Name	like		$ \downarrow $	SHEEN_4	4	1566	2318	753		5'End 2347 Locus Tag SHEEN_5
GenelD	=	(		SHEEN_5	5	2347	3570	1224		3'End 3570 Regions 1
Locus	like			SHEEN_6	6	3599	3877	279		Length 1224 🔳 Tag
Start			- L	SHEEN_7	7	3889	4512	624		Direction Forward
			- []	SHEEN_8	8	4509	5477	969		
Length	>		_ [	SHEEN_9	9	5551	5628	78		Translation Table Undefined
Regions	>			SHEEN_10	10	5629	5704	76		EC Number
% GC	<			SHEEN_11	11	5635	5772	138		
CAL			- [	SHEEN_12	12	5731	7155	1425		Product
			- [	SHEEN_13	13	7152	7595	444		gp5
EC#	like		_ [	SHEEN_14	14	7592	8332	741		
Product	like			SHEEN_15	15	8359	10059	1701		Function
Function	like			SHEEN_16	16	10056	11552	1497		
FeatureID			- [	SHEEN_17	17	11549	12562	1014		~
	like		- [	SHEEN_18	18	12621	13130	510		Notes
Tag	іке			SHEEN_19	19	13160	14149	990	6	Original Glimmer call @bp 2347 has strength 10.85
Hide Ignored Features SSC: CP: SD: SCS: Gap: Blast: LD: ST: F: FS:										
Cala	- All C	anturan I		Insert Delete	Post Valid	ate				
Image: Market Arrow Controls         Image: Market Arrow Controls         Image: Market Arrow Controls										
10 13 13 16 15 16 17 13 16 17 13 16 17 13 10 10 10 10 10 10 10 10 10 10 10 10 10										

Figure 8.2

The notes should report whether Glimmer and/or GeneMark made the prediction. In the example above, both Glimmer and GeneMark did predict the gene with the same start. (Remember that if both programs agree, only one program's output is reported.) The gene was called by both programs, which supports its legitimacy. Good so far.

Find this gene in your GeneMark-Smeg output, and check if there is coding potential that supports this gene call (Figure 8.3).

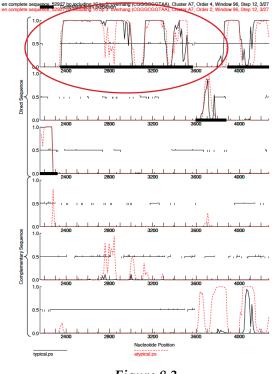


Figure 8.3

In **Figure 8.3**, the region of gene 5 is circled. You can find a gene by looking at its coordinates in the Feature Table, then finding those coordinates on the GeneMark output. Use the **stop coordinate** to identify the correct reading frame. For any given ORF, there may be many possible starts, but only one stop.

GeneMark (smeg) shows that this ORF has coding potential starting near position 2300+ and ending near 3600. You now have verified that all three coding prediction programs have predicted this gene. Strong evidence that you have a gene! Now you want to evaluate whether this is the best way to call the gene. Do you have the best start chosen?

How many other mycobacteriophages have this gene and do they all have the same start? Examine the BLAST data under the [[Blast]] sub-tab (**Figure 8.4**), and see if there are genes in the databases that are high-quality matches to this one.

Sort By Index 👻 .	▲ Tag	Name	5'End	3'End	Length /	Description Sequence Product Beninns Blast Context
elect Features Direct SQL	SHEEN_1	1	397	693	297	Score Target Description
1	SHEEN_2	2	732	1166	435	1033 putative structural protein [Mvcobacterium phage HINdeR]
	SHEEN_3	3	1259	1576	318	1066 gp5 [Mycobacterium phage Timshel]
ame like	SHEEN_4	4	1566	2318	753	571 gp27 [Mycobacterium phage Henry]
enelD =	SHEEN_5	5	2347	3570	1224	568 gp27 [Mycobacterium phage Toto]
ocus like	SHEEN_6	6	3599	3877	279	568 gp27 [Mycobacterium phage Bask21]
art >	SHEEN_7	7	3889	4512	624	
ength	SHEEN_8	8	4509	5477	969	BLAST Hit Exp
	SHEEN,- High-So				-	Accession YP_008051857
-		itā Alignmi	ant			Length 406
GC <	SHEEN, Bit Score		Identities		38	Max Score 1089 Date 11/11/2014 Delet
N >	SHEEN Score SHEEN E-Value	1089	%Identity Positives		425	Unit Company (UCD)
C# like	SHEEN Length		*Ositives %Similari		741	HaP Data Alignment
roduct like	SHEEN & Aligner		Gaps	3	701	
unction like		3 - 401	2.		197	3 LSLGTKLPRL IQWGDDKVIR KISLCSDVIW TAADPDLAQG LRDQ 1 ++         +  +  +  + ++      +++  +
	SHEEN Target	4 - 401		16.06	<b>K</b>	4 ITLGDKPIRR ISWGSQE-IK KVSLGNSLLW AAEDPDMSQA NKTA
eatureID =	- SHEEN 18	18	12621	13130	510	
ag like	SHEEN 19	19	13160	14149	990	53 TLESTFGLSG VMFKLTGPVG KYEAAFGSTA GRALALNDNF RAGS 51 +   +     ++
Hide Ignored Features	SHEEN 20	20	14229	14390	162	53 AFMTQFKIPG VSFMISGPAG KYAKAFGTTA GRPLALGDHF RCGS
Select All Features	SHEEN 21	21	14394	14768	375	
Select All Features	SHEEN 22	22	14765	14920	156	103 GQAALQQIAA GNIBYDATID TYVSGYPNGD KITIENINTM TSGL 101    +     +   ++ +    +++     +++
	SHEEN 23	23	14917	15300	384	103 SQAVWEQIDK GTLHLDDTVN QWVPGLPSGN LIKIPHLLQM TSGL
	SHEEN 24	24	15303	15647	345	N
	SHEEN_25	25	15660	16109	450	153 MSFMLNFVLN PMTPWSDDAT LNLVKSNGVY FQPGKGYKYA NSNY
	SHEEN_26	26	16124	16708	585	153 TSFMLNFVLN PQTAWSAANT LALAKANPLA HPPGTKYAYT MSQY
	SHEEN_27	27	16821	17186	366	
	Insert Delete	Post Valid	un l			203 LETURETIS DILINDVVKK AGLNNTTUPT AGGUOPGGIP TPAA
		Post Valid				
4 4 <b>0 0 1 1</b>	1 - 50000	Position : 4	16272 F			Map >> Controls
28(4) 5) 718) 12 104	15 16 17 19 2	EFE E B	29	30	31 8 3	36   e
Features Live						52927

Note to Figure 8.4: The inserted box is obtained by clicking on the sub-tab **[[HSP data]].** DNA Master can display either HSP data or Alignment data at any one time.

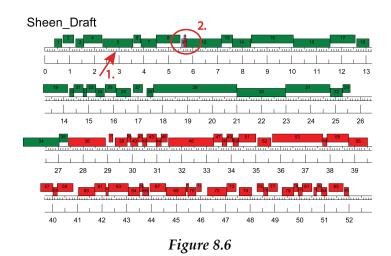
- 1. The first match to Sheen's 5th gene under the parameters of BLASTp in DNA Master is to a putative structural gene in mycobacteriophage HINdeR.
- 2. Evaluate the E-Value. The score is 0.0E0, which is not exactly identical but close. (The E-value on NCBI's BLASTp (default parameters) is 1e-122). Why is that?
- The HSP data describes it as 98.0% aligned, 50.75 % identity, and 68.59% similarity. See <u>BLAST glossary</u> (<u>http://www.ncbi.nlm.nih.gov/books/NBK62051/</u>) for definitions of terms.

Review gene length to make sure it meets the expected parameters. You will recall (see **Section 7.2**) that you should carefully examine genes less than 150 bp in length with an eye towards gauging their legitimacy, and genes below 120 bp should be viewed very skeptically. You can see gene length only when you are in the Widened Feature Table Mode (right click on **Name** at the top of the table. (see **Figure 8.5**), or you can select your gene and the length will be listed under the **[[Description]]** sub-tab to the right (see **Figure 8.5**). In this case, the gene length (1224 bp) is fine. Note that the amino acid length is 408.

verview	Featu	iles	References	Sequence	Documentation				
ort By	Index	•	_ <b>_</b>	Tag	Name	5'End	3' End	Length 📥	Description Sequence Product Regions Blast Context
elect Fe	atures	Dire	ect SQL	SHEEN_1	1	397	693	297	Name 5 GenelD
уре	is [	All		SHEEN_2	2	732	1166	435	Type CDS V GI
				SHEEN_3	3	1259	1576	318	
ame	like			SHEEN_4	4	1566	2318	753	5'End 2347 Locus Tag SHEEN_5
enelD	=			SHEEN_5	5	2347	3570	1224	3) ind 3570 Regions
ocus	like			SHEEN_6	6	3599	3877	279	Length 1224 🔳 Tag
tart	ЪÍ		L	SHEEN_7	1	3889	4512	624	Direction Server
ength	5		L	SHEEN_8	8	4509	5477	969	
-			L	SHEEN_9	9	5551	5628	78	
egions	>			SHEEN_10	10	5629	5704	76	EC Number
GC	<			SHEEN_11	11	5635	5772	138	
41	>		L	SHEEN_12	12	5731	7155	1425	Product gp5
C#	like	_	— L	SHEEN_13	13	7152	7595	444	gp5
			— L	SHEEN_14	14	7592	8332	741	
roduct	like			SHEEN_15	15	8359	10059	1701	Function
unction	like			SHEEN_16	16	10056	11552	1497	
eatureID	= [			SHEEN_17	17	11549	12562	1014	
ao	like		— L	SHEEN_18	18	12621	13130	510	Notes
-				SHEEN_19	19	13160	14149	990	Original Glimmer call @bp 2347 has strength 10.85 SSC: CP: SD: SCS: Gap: Blast: LO: F: FS:
Hide I	gnored	Feat	ures	1		1		<b>v</b>	336. CF. 3D. 363. diap. biast LO.F. F3.
Cala	AF ALLE.	huri	L	Insert Dele	te Post Valid	ate			
44	⊕ ∈		<b>H H</b> 1.	50000	Position : 4	7418 I		s>> Man 🔽	Map >> Controls

Figure 8.5

Verify that there is only one gene called in this region of DNA, as per Guiding Principle #1. The easiest way to do this is by viewing either the Phamerator map or DNA Master map you've generated to see if there are other genes called that substantially overlap this one on either strand.



- 1. In this example above, we can see from the DNA Master generated map that there are no other genes occupying the same portion of DNA at gp5.
- 2. Note: Genes 9, 10 and 11 will need to be reconciled at a future point.

### DECISION TIME: Is the designation of this ORF as a gene well-supported?

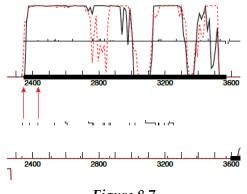
**GUIDANCE:** Most gene calls will pass this stage. Exceptions are genes that are called by only one program, have little or no coding potential, have very weak or no BLAST matches, are too short, and/or substantially overlap other genes.

YES	NO
ACTION: Continue to Section 8.4.2.	ACTION: You need to delete this gene. Go to
	Section 9.2.1 for instructions.

#### 8.4.2 Is the called start site for this gene the best possible choice?

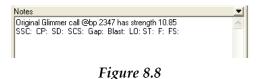
This can be a tricky, but the simplest way to answer is to address the following questions.

**Does the currently predicted start site include all of the coding potential in the GeneMark output?** The current start position for our example is in location **A** in **Figure 8.7** below, and captures all of the coding potential. A hypothetical start at position **B**, however, would be a poor choice because it excludes coding potential.





**Did Glimmer and GeneMark agree on the start for this gene?** Check the 'Notes' field under the **[Feature]** tab and the **[[Description]]** sub-tab to answer this question. In our example, shown in **Figure 8.8**, the two programs agree.



**Is the predicted start codon the longest possible for the ORF without causing excessive overlap?** The start codon, 2347, provides the longest gene possible for this ORF and does not overlap with the previous gene.

**Does the start site match other starts for similar genes in GenBank and the mycobacteriophage database?** To view the relevant information from NCBI BLAST, go to the **[[Blast]]** sub-tab, then the **[[[Alignment]]]** sub-sub-tab. You can select different BLAST alignments in the top pane to see how your start compares to those in a variety of other genomes. Refer to **Figure 8.7** for the alignment display. To review the relevant information from PhagesDB, select your protein and perform a BLASTp. Either and/or both will provide you with a global look at this protein and its homologues. Does it match proteins only found in the same subcluster or cluster of mycobacteriophages? Does it match phages from other clusters? Does it match phages of other species?

Remember to not be overly enthusiastic about alignment to other gene products, because you don't know *a priori* whether these were correctly identified. You just know that someone made that choice during a previous annotation. In addition, our example only has 68.59% similarity, 50.75% identity even though the 98% of the sequence aligns. It would be wise to look at the alignment across more than one homologue. The best approach to this is to use the two matches from the same subcluster, HINdeR and Timshel. The chosen start is the best alignment to those genes.

**Using Starterator to evaluate gene starts.** Sometimes the comparison with genes in a particular pham is what you will want to evaluate. You can use Starterator to evaluate that. Starterator provides you with alignment adta to evaluate the starts of a given Pham.

Refer to **Section 6.2** for how to run Starterator for your gene of interest, in this case Sheen, gene 37 (stop at 29071). From the Phamerator map, you can identify that this gene belongs to Pham 6183. The Starterator report has 3 components: the visual representation of each member of the Pham (represented as "Tracks"; a list of labels for each track; and the Recommend Start. As you start from the top and review coding potential, gaps, and BLASTP alignments, you will want to change the start of this gene (It does not capture all of the coding potential, doesn't fill the gap, or possess the best SD scoring.) You could run Starterator and evaluate its output because this output would show you if the start that you think is most appropriate has been called in other genomes. In this example (See **Figure 8.9**), the Starterator output confirms that the most upstream start is clearly a better choice.



Figure 8.9

Other Starterator notations are "NA" for "Not Applicable"—for orphams, or for genes in which the evidence overwhelmingly supports a single start choice; or "SS" for "Suggested Start".

# Does the predicted start have an associated ribosome binding site [RBS; Shine-Dalgarno (SD)] with a high score or recognizable sequence?

While ribosome binding site sequences can be found in phage genomes, the presence of a high scoring RBS is not the strongest determinant for what start to choose for any particular gene for a number of reasons. First, the most common arrangement of start and stop codons in phage genomes is a 4bp overlap between genes (for example, with the sequence ATGA: in which the –TGA is the stop of the upstream gene, and the ATG- is the start of the downstream gene). This arrangement allows for co-translation and does not require the ribosome to rebind to an RBS for the downstream gene. Second, leaderless transcripts (mRNAs that begin with the first base of the start codon) have been identified for some genes in phage genomes, suggesting that the ribosome is binding to the mRNA through some other mechanism that the SD sequence. Finally, ribosome binding site may be essential for limiting the levels of a protein made during infection. So while Shine-Dalgarno sites should be evaluated for every gene, the start with the best SD score may not be the best choice for a given gene's start.

The Shine-Dalgarno (SD) sequence in *E. coli* is AGGAGGA. It is located 7- 10 bases upstream of the start site. The Shine-Dalgarno sequence helps recruit the ribosome to the mRNA by aligning it with the start codon. Both the sequence and the spacing of the sequence are important in the evaluation. Dr. Lawrence recently rewrote the algorithms that underlie this evaluation using metrics described by Kibler (SD Scoring Matrix) and Karlin (Spacing Weight Matrix).

You will want to open the "Choose ORF start window to evaluate the ORF. (Check **Section 4.4.4** for details on how to open the "**Choose ORF start**" window.)

In the Choose ORF start window (See the oval of **Figure 8.10**), you must first select the SD scoring and Spacing Weight matrices. You will find multiple options (in a drop down menu) for both matrices. With Dr. Lawrence's recommendations and some trial and error, we suggest using Kibler 6 and Karlin Medium.

**Figure 8.10** shows that there are 24 possible start codons for our example, gene 5 (coordinates 2347 – 3570) of Mycobacteriophage Sheen. For each start, this window displays, a raw SD score, Genomic Z value, spacer distance, final score, upstream sequence to that start (can you find an AGGAGGA-like sequence?), the start codon, start position, and ORF length.

😌 DNA Master		
File Tools Window Help		
2		😤 Choose ORF start
		State: 24 DRF Stat : 2347 Cdn 1 Cdn 2 Cdn 3 Length SD Scoring Matrix Kaller6 V Chone
		Selected: 10 DRF Stop: 3570 5'End 0.0 33.3 100.0 9 S0 Scoring Matrix Delete
1 417 853 1249 1665 2031 2497 25 bx 765 Hillel 4 (4) 4 (4) (4) (4) (4) (4) (4) (4) (4)		DRF Length : 1224 3' End 57.3 49.0 87.0 789 Spacing Weight Matter Kalin Medium
bp: 765 H 44 4 ⊕ ⊕ ▶ ₩ ₩ 0RF 2347 - 3570 34 1 G+C 5 Winds Position of internal state	00 <u>-</u> 00	2236
Diverview Peakules References Sequence Documentation	Contraster	StaRaw SD Genomic Spacer Final Sequence of the Region Start Start OFF
	Duration in the state of the state of the	Score Z Value Distance Score Upstream of the Start Coden Fosition Length     -1.410 3.161 10 -2.112 TGALACCOAACADGAGATAGGC GTG 2347 1224
		2 -5.927 1.024 8 -7.149 CARGATCAGCCTCDDCAGCGAC GTG 2424 1127
CHEEN 2 2 722 1100 425	None 8 GenelD	2 -5.701 1.141 10 -6.296 CCTRCSAGATCASCTCSACACE ATG 2494 1077
Type is Al  SHEEN 3 3 1259 1576 318	Type CDS 💌 🖬	4 -6.199 0.911 12 -7.023 STCSACCTTCSSTCTSTCCGGC GTG 2533 1039
Nome Rie SHEEN_4 4 1566 2318 753	5'End 4509 Locus Tag SHEEN_8	5 -4.668 1.628 6 -6.412 GACCTTCGGTCTGTCCGGCGTG ATG 2536 1035
GenelD - SHEEN_5 5 2347 3570 1224	3'End 5477 Regions 1	6 -5.699 1.142 6 -7.444 GATGTTCAAACTCACCOGCCCG GTG 2557 1014
Locus Re SHEEN_6 6 3599 3877 279	Length 969 🖷 Tag	7 -3.990 1.948 S -5.990 CC000CCGGTCGATCACGAAG ATG 2644 927
Start > SHEEN_7 7 3889 4512 624	Direction Forward	8 -5.026 1.459 10 -5.720 CARGATCACCATCORGAACATC ATG 2764 807
HEEN_8 8 4509 5477 969		9 -5.472 1.249 13 -6.518 CACCATCGAGAACATCATGACG ATG 2770 801
Length > SHEEN_9 9 9551 9528 78		10 -4.343 1.781 13 -5.389 GCTGTACAACGAACAGTCCAAC ATG 2803 768
Region: > SHEEN_10 10 5623 5704 76	EC Number	11 -5.724 1.130 8 -6.946 CGRACADTCCAACATGTCGTTC ATG 2812 759
% GC < SHEEN_11 11 5635 5772 138	Product	12 -5.145 1.403 8 -6.366 GCTCAACTTCGTCCTGAACCCG ATG 2836 735
CAI > SHEEN_12 12 5731 7155 1425	gp8	13 -5.472 1.249 10 -6.167 CCCGTGGTCGGACGACGCGACG TTG 2863 708
ECH ike SHEEN_13 13 7152 7595 444 SHEEN_14 14 7592 8332 741		14 -5.997 1.001 11 -6.754 GGACGACGTCGACGTCGACCTC GTG 2872 699
Product B/e SHEEN_14 14 7592 8332 741 Product B/e SHEEN_15 15 8359 10059 1701	Function	15 -5.051 1.442 8 -6.253 GAACCTCGTGAACGGC GTG 2887 684 16 -4.663 1.630 10 -5.358 TTACATTCTCCTGGGCCGGATC TTG 2953 618
Function like SHEEN_16 16 10056 1152 1457	rucch All	16 -4.063 1.050 10 -5.350 IIECATICICCIONCOMIC IIG 2353 618 17 -3.990 1.948 11 -4.747 CGACGTCGTCAAGAAGGTGGG TTG 3019 552
	the second se	17 -3.590 1.940 11 -4.747 CONCICCICARGANDELEGE 116 3019 552 18 -4.502 1.706 15 -6.104 CCC00CAGTAACCGACCC GTG 3160 411 Genomic SD Z values
SHEEN 18 18 12621 12130 510	Notes -	18 -4.502 1.706 15 -6.003 GECAGTAACCGACCAGACCGTG ATG 3160 411
Tag like SHEEN 19 19 13160 14149 990	Driginal Gimmer call (9bp 4509 has strength 5.52	20 -4.154 1.071 14 -5.501 GACCGTGATGAACCCCGAGATC ATG 3170 393
Hide Ignored Features	SSC: CP: SD: SCS: Gap: Blast: LO: F. FS:	21 -5.924 1.036 14 -7.271 CGGCACACTGATCACCCCGGAG ATG 3268 303 Genomic SD scores
Culues All Customs   Insert Delete Post Validate	V	22 -2.905 1.989 12 -4.741 CGGTCTCAACGGGTACGGGTGC GTG 2418 153
		23 -2.412 2.220 18 -5.714 STC09555C5C4ACTC0TT6C5 ATG 2460 111
		24 -5.000 1.090 6 -7.552 GEAGCTGCTATACCCCSGCTCE ATG 2525 26
Execution (Content of the Content of the Conte	I (I DI KAS DI NA CEL KAKAN COKADIN CIN CIN CIN CIN CIN CIN CIN CIN CIN C	La La mar no el composición de la composi Composición de la composición de la comp

Figure 8.10

Use the following to best evaluate the information in this window:

- 1. You are provided the Raw Score (how well the SD sequence matches) in addition to the spacing and the final score. This allows one to assess the match independent of the spacing. The higher the score (the less negative), the better the sequence match. This Raw Score compares the SD of the start of interest with all of the SD scores across the genome (a distribution of scores). This scoring system is based on log probability scores. They are negative numbers and higher (less negative) are better. Moreover, magnitude is now meaningful. A score of -2 is ten times more likely than -3.
- 2. You can see that distribution by clicking on the Genomic SD score button in the bottom right of the choose start window. (See arrow in **Figure 8.10**.) To evaluate the distributions, change the bin counts to 20. If you see a normal distribution, proceed.
- 3. The normal distribution of the scores will be mostly random. These are ALL SD sequences, even in frames where no genes are predicted. So you are looking at the scores with the highest value (least negative number). However, the scores you see in your list are NOT random, but remember all but one are not correct!
- 4. The distribution of scores throughout the genome is used to provide a Z-value for each raw score. You can see the distribution of Z scores by clicking on the Genomic SD Z value button, and changing the bin count to 20. Z is the number of standard deviations from the mean. A Z-value higher than 2 is getting good.
- 5. The spacer distance is the number of bp this sequence is found upstream of the start. You will recall that you want the number to be between 7- 10.
- 6. The final score is determined by both the raw score and the spacing. It is evaluated in the same way as the raw score, where the higher the score (the less negative), the better the sequence match.

Note: The settings for this window are open for exploration. In general, when this particular gene is not controlled by a Shine Dalgarno sequence, this evaluation should not provide a clear answer. Further evaluation on this point is needed.

Applying this to our example, the start 2347 has the highest Raw SD Score and has a spacer distance of the predicted length (10). It is the highest final score (-2.112) and is at least 100 times more likely than the next lowest score (-4.714). The Z-value of 3.161 is the best in the list.

Note: The old DNA Master score is still available, but deprecated. It will eventually be removed. No Z-values, base distributions or probabilities are available in that format.

# You now need to put this information together to make the best choice and record your decision in the notes.

In this example, the initial information in the Notes window is:

Notes	•
Original Glimmer call @bp 2347 has strength 10.85	~
SSC: CP: SD: SCS: Gap: Blast: LO: ST: F: FS:	

Figure 8.11

To complete the notes for gp 2, you can enter something like this:

Notes

Original Glimmer call @bp 2347 has strength 10.85 SSC: 2347 - 3570 CP:only start to capture all coding potential SD: best scores SCS: Agrees with all Gap: a gap of 30 Blast: First match is to HINdeR and Timshell with 3:4 alignment LO:YES! F: structural protein FS: phagesDB w/ 68%pos. and 50% ID to HINdeR but still requires further investigation

### Figure 8.12

SSC: Record actual coordinates. 2347-3570. This may seem redundant, but this points to a common oversight. As you make decisions to change a start, you forget to actually change the start in the Feature Table (**Section 9.2.3**)

SD: Does this call have a relevant SD and is it the highest score?

SCS: This gene was called by all coding potential prediction programs with the same start.

Gap: Yes, there is 30 bp gap that cannot be filled with a coding potential prediction. This points in the direction that there may be a ribosomal binding site upstream of this gene.

Blast: The first hit is to Mycobacteriophage HINdeR. BP 3 of Sheen aligns with BP1 of HINdeR.

LO: Is this the Longest gene possible in this ORF? Yes

ST: Starterator data (where applicable). (Enter NA when it is not applicable---as in the case for an orpham with no close matches, or NI if Starterator was run but is Not Informative.)

F: and FS: Refer to Section 10 to complete this part of the gene evaluation. In this example, "structural protein" was noted in the BLASTp at PhagesDB.

For complete instruction on documenting gene calls, refer to Section 9.6.

### **DECISION TIME:** Is the currently called start site for the gene the best choice?

**GUIDANCE:** Ten percent or more of your genome's start sites will likely have to be changed, and in some cases NEITHER Glimmer nor GeneMark will call the correct start. For each gene, gather the information described in this sub-section, and try to weigh all possibilities to arrive at the best call.

YES	NO
ACTION: Continue to Section 8.5.	<b>ACTION:</b> You need to change this gene's start.
	Go to Section 9.2.3 for instructions.

### 8.5 Checking gaps in the draft annotation for uncalled genes

According to Guiding Principle #3, the genes in phage genomes are generally tightly packed, so any large gaps (>50 bp) in your annotation should be reviewed.

In circumstances where you have a series of genes in the same orientation that are likely to be expressed as an operon, these genes are typically nestled closely end-to-end. However, non-coding gaps are perfectly legitimate and to be expected, and filling gaps with poorly justified gene calls is not appropriate.

There are two basic things you should look for in gaps.

**Can the start site of the downstream gene be extended so that the gene covers more of the gap?** Carefully consider all possible start sites for the downstream gene. If a longer one is available, compare it to the current start site to see if it is a similar or better choice. All other things being equal, a longer call is usually preferable, but do not extend genes just to fill a gap.

### If YES, go to Section 9.2.3 to change the start site.

**Is there a protein-coding gene in this gap?** You have several resources to help answer this question. First, you can use Phamerator maps to see if any similar genomes have a gene called in this gap. Second, you can look at the GeneMark-Smeg output to see if any of the reading frames in this gap show some coding potential. Third, you can copy the DNA sequence from your gap and use it to run a BLASTX search on NCBI. The combination of these techniques may yield convincing evidence that the gap contains a protein-coding gene that was missed by both Glimmer and GeneMark.

#### If YES, go to Section 9.2.2 to add a gene.

Remember too that you should expect non-coding gaps between divergently transcribed genes as there is a strong prediction that promoters lie within these regions. For example, in **Figure 8.13**, we should expect some gap between gene 47 (transcribed leftwards) and gene 48 (transcribed rightwards).

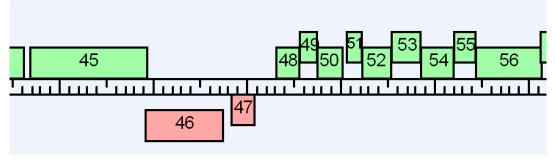


Figure 8.13

### 8.6 Finding and refining tRNA and tmRNA genes

DNA Master searches for tRNAs by default, but may miss some tRNAs that other approaches can find, or may miscall the precise boundaries of these genes. See **Section 9.5** for information on how to search for and call tRNAs and tmRNAs.

### 8.7 Completing your annotation refinement

Much of the work of annotation is following the steps above—for each gene and gap in your genome—until you've settled on the best calls for each with the information given.

As a double-check, you should scroll through the Feature table and the genome map (using buttons at the bottom of the [Feature] tab) to make sure that all the changes you've made have been committed to the file.

One suggestion to confirm that the changes you have made are actually in your file is to use the Genome Comparison Tool in DNA Master. (See Protocols -> Further Discovery -> Exploring Bacteriophage Biology). As you make changes in your file, the gene calls can become disparate with the auto-annotation found in Phamerator. The Genome Comparison Tool produces a quick, easily produced genome map compared to any genome(s) of your choice.

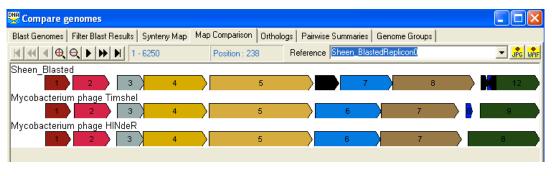


Figure 8.14

Several important steps remain.

1. **Documenting your gene calls**. You can use the Notes field (under [Feature] [[Description]]) to record notes about each gene as you go. Your final submitted file, however, should have each gene's Notes field filled in according to specific instructions so as to facilitate checking the annotation. These documenting instructions are described in Section 9.6.

- 2. **Determining putative functions**. You've figured out where the genes are (and aren't), so the next step is to see if you can make a well-supported guess as to what they do. This process is covered in **Section 10**.
- 3. Merging several different portions of the annotation into a single file. In a classroom setting, often you will choose to split the genome into sections and have different groups or students work on different sections. If you've split the genome up, now is the time to bring everyone's work back together or "Merge" the different annotations. This process is described in the first part of Section 11. We recommend that you make this a vibrant part of your classroom structure. It can be quite problematic to think that you will do all the 'checking' work out of class.
- 4. **Checking the final annotation**. Once you've produced a nearly final annotation, it still needs a (relatively) expert eye to double-check it, as described in **Section 11**.
- 5. **Submitting final files**. When you're confident in your annotation, have investigated every nook and cranny, and are ready to send it out the door, you'll need to generate and submit a final DNA Master file, as well as a list of those who have worked on the annotation and should be authors on the GenBank submission. This is described in **Section 12**.

## 9 The mechanics of making changes to your annotation

### 9.1 Overview

This section, unlike most sections of this guide, is not intended to be a sequential step-by-step description of any part of the annotation process. Rather, it is intended to be used as a reference section for how to make specific changes to your annotation. The actual decision-making steps were described in **Section 8**, and a graphical summary can be seen in the Decision Tree in **Section 8.3**.

The three most common operations you'll need are covered first. They are:

- Deleting a gene
- Adding a gene
- Changing the start site for a gene

The following sub-sections describe some common steps you should take after making any changes to your annotation. They are:

- Posting changes
- Validating your calls
- Renumbering your genes
- Re-BLASTing a gene you've changed

There are also some less common operations that you may need. They are:

- Annotating a programmed translational frameshift
- Annotating introns
- Annotating wrap-around genes

Next is a sub-section on RNA genes. It is:

• Predicting tRNA and tmRNA genes

Finally, there is a sub-section of how to document the annotation work you've done:

• Documenting your gene calls

### 9.2 Making common changes to your annotation

### 9.2.1 Deleting a gene

- Select the [Feature] tab of your main genome file.
- In the center column, click on the feature you would like to delete to select it. (The selection can be verified by the presence of a black arrow to the left of the gene name.)
- Click the 'Delete' button, found at the bottom of the center column.

• Click the '**Post**' button to commit your changes to the database.

### 9.2.2 Adding a gene

If it's not already open, open the Frames window by going to **DNA** → **Frames** 

- Locate the ORF that corresponds to the gene you would like to add.
- Click within that ORF, and a green or red line will appear, depending on its orientation.
- Click on the '**RBS**' button in the lower-right corner.
- Confirm that you have selected the correct frame by verifying the coordinate of the **STOP** codon. There can be many possible starts for each ORF, but there is only one possible stop!
- Choose the best start, as described in **Section 8.4.2**, then click anywhere in that start site's row in the "Choose ORF start" window to select it.
- Return to the [Feature] tab and click on the 'Insert' button at the bottom of the center column.
- A new window will appear that allows you to add the feature. Verify that the correct orientation (forward/reverse) is selected and that the coordinates are correct. Do not worry about adding the correct gene number or gene product (gp) number, as the genes will get renumbered using the Validation function when you are done.
- Check the boxes 'add to feature table' and 'add to documentation'.
- Click 'Add Feature'.
- Click the '**Post**' button to commit your changes to the database. This is also a good time to save your file.
- Your new gene will likely be placed at the end of your feature list, because the default sorting is by index number, rather than genome position. To sort by position, find the dropdown box at the top left of the [Feature] tab labeled 'Sort by', and change it from "Index" to "Start."
- You may want to collect BLAST data for your new gene. See **Section 9.3.4** for instructions.

### 9.2.3 Changing the start site for a gene

- Select the [Feature] tab of your main genome file.
- In the center column, click on the gene you want to change to select it.
- Click on the [[Description]] sub-tab to the right.
- In the box labeled "Start", third from the top under "Description", type in the new start coordinate you've selected.
- Click on the Calculator button (this is an icon of a calculator, found just to the right of the "Length" display) to recalculate the ORF length. The new length (in bp) will be shown and should reflect your change.

- Click the '**Post**' button at the bottom of the central column to ensure your changes are saved to the database. This is also a good time to save your file.
- Because you've changed the start site, you'll probably want to re-BLAST this gene so that the BLAST results reflect your change. See **Section 9.3.4** to do so.

### 9.3 Common steps to take after making changes

### 9.3.1 Posting changes

When making gene changes—including changing start codons, deleting genes, annotating programmed frameshifts, adding notes to the Notes field, etc.—you need to both **enter** and **post** the changes. Simply entering them is insufficient, and the changes may be lost. Once you've learned how to post, it doesn't hurt to **post often**!

Normally, a selected gene in the feature table will be indicated by a triangle, as shown below.

DNA Echil	d_10.1	18.13											
Overview	Featu	res Re	ferences	Sequence	Documentation								
Sort By	Index	•	- <b>-</b>	Tag	Name	5' End	3' End	Length	^	Descripti	on Sequence Proc	luct Regions	Blast Context
Select Fr	eatures	Direct 9		DNAM_1	1	330	443	114		Name	2	GenelD	
-	. r	A II	<b></b>	DNAM_2	2	662	1495	834					
Туре	is	All		DNAM_3	3	1543	2358	816		Туре	CDS 🗾	<u>GI</u>	
Name	like			DNAM_4	4	2358	2675	318		5' End	662	Locus Tag	DNAM_2
GenelD	= [			DNAM_5	5	2672	2968	297		3' End	1495	Regions	1
	···· ŕ		— F	DNAM C	c	2007	2477	471	-				

Figure 9.1

When you make a change to a feature listed in the Feature table (e.g., begin typing in the Notes field), the icon next to the feature changes to an Insert icon, as shown below.

Echild_1	0.18.13											
)verview Fe	atures Referen	ces	Sequence Docu	mentation								
Sort By Inde	× 💌	┛	Tag	Name	5' End	3'End	Length	^	Descripti	on Sequence Pr	duct Regions	Blast Context
Select Feature	S Direct SQL		DNAM_1	1	330	443	114		Name	2	GenelD	
<b>.</b> .		I	DIIAM_2	2	662	1495	834			-	1	
Type is	All		DNAM_3	3	1543	2358	816		Туре	CDS 💌	<u>61</u>	
Name like			DNAM_4	4	2358	2675	318		5' End	662	Locus Tag	DNAM_2
GenelD =			DNAM_5	5	2672	2968	297		3' End	1495	Regions	
Loous like			DNAM 6	6	3007	3477	471		1	004	al -	ŕ



For the most part, this change to Insert Mode happens automatically when you start typing in any of the fields under the Description tab. Your changes , however, **won't be posted to the database until you exit Insert Mode**.

The following are ways to make sure your edits get posted to the database.

- Click on the '**Post**' button at the bottom of the center column.
- Click on the **Calculator** icon, after changing a start or stop.
- Click on a different feature in the center column.

You will be able to tell that your changes have posted to the database because the Insert icon will change back to the right-pointing triangle.

**Important Note**: The following are ways that your changes will **not be posted** to the database, and **WILL BE LOST**.

- Saving your file while still in Insert Mode.
- <sup>2</sup> Clicking on a different tab or sub-tab while still in Insert Mode.

### 9.3.2 Validating your annotation

As you work through your genome, DNA Master has a handy **validate** feature that helps ensure your gene calls have valid start/stop codons and do not have any internal stop codons.

To perform a genome validation, follow the steps below.

• Click on the 'Validate' button, at the bottom of the central column in the [Features] tab (located in the red circle in Figure 9.3 below).

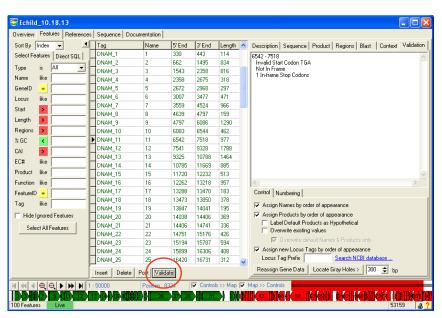


Figure 9.3

DNA Master will let you know when gene calls are not in frame or if they have incorrect start or stop codons. A genome is not complete unless validation returns as "All ORFs are valid".

	atures	Reference	s Sequence Name	e Docume 5'End			To the Interior Interior (1964)
Sort By Inde			1 1	330	Length 114	$\sim$	Description Sequence Product Regions Blast Context Validatio
elect Feature	s   Di	rect SQL	2	662	834		All ORFs appear valid
ype is	All	•	3	1543	816		
lame like			4	2358	318		
ienelD =	I T		5	2672	297		
.ocus like	<u> </u>		6	3007	471		
itart >	-		7	3559	966		
_	_		8	4639	159		
ength			9	4797	1290		
Regions 🔉			10	6083	462		
GC <			11	6541	978		
AI 🔉	<b>—</b>		12	7541	1788		
C# like	-		13	9325	1464		
			14	10785	885		
roduct like			15	11720	513		N
unction like			16	12262	957		< >
eatureID =			17	13288	183		Control Numbering
ag like			18	13473	378		Assign Names by order of appearance
- Hide Ignor			19	13847	195		Assign Products by order of appearance
Hide Ignor	earea	atures	20	14038	369		Label Default Products as Hypothetical
Select A	Featu	ures	21	14406	336		Overwrite existing values
			22	14751	426		Overwrite default Names & Products only
			23	15194	594 408		Assign new Locus Tags by order of appearance
			-	15899	408	~	Locus Tag Prefix Search NCBI database
			<		>		
			Insert [	elete Po	st [Valid	ate	Reassign Gene Data Locate Gray Holes > 300 🚖 bp
[44] 4 [ <b>⊕</b>		)	- 50000	F	Position : 1	_	Controls >> Map 🔽 Map >> Controls

Figure 9.4

If the validation generates failures, you should check the coordinates in those features to see what might have gone wrong and make necessary changes. You can then re-run the validation to ensure all ORFs are valid.

### 9.3.3 Renumbering & formatting annotated features

When you add or delete a gene, you may want to renumber the genes to reflect the change. Genes added manually after auto-annotation will appear at the bottom of the feature list when sorted by **Index**. Sorting by **Start** will place the gene in its correct order by start coordinate.

To renumber your features:

- In the **[Features]** tab, click the '**Validate**' button located at the bottom of the central column. This will open the **[[Validation]]** sub-tab on the right side.
- Check the boxes as shown in **Figure 9.5**.
- In the field marked "Locus Tag Prefix", type in your phage's name (it will be all capital letters). This is a necessary attribute of each feature for a GenBank submission. We are adding it here so that 'Tags' are renamed in the same manner as 'Names' and 'Products'. For returning users, we are changing the format this year.
- Click the '**Reassign Gene Data**' button.
- Click '**Yes**' to confirm in the window that pops up.
- Genes will now be re-numbered sequentially.

ort By Inde		Reference	÷	Sequence Docu Tag	Name	5'End	3' End	Length	~	Description	Sequence	Product	Begions	Blast	Context	Validation
elect Featur		irect SQL		ECHILD 1	1	330	443	114	-	All ORFs app		Troduct	riogionio	Didot	1 contone	
	10	INCC SQL		ECHILD_2	2	662	1495	834		All Unit's appl	ear valio					
vpe is	All	•		ECHILD 3	3	1543	2358	816								
ame like	8		Н	ECHILD_4	4	2358	2675	318								
enelD =			П	ECHILD_5	5	2672	2968	297								
cus like	e 🗖		П	ECHILD_6	6	3007	3477	471								
art 🔉			П	ECHILD_7	7	3559	4524	966								
	-			ECHILD_8	8	4639	4797	159								
ngth ゝ				ECHILD_9	9	4797	6086	1290								
gions ゝ				ECHILD_10	10	6083	6544	462								
GC <			Þ	ECHILD_11	11	6541	7518	978								
. 🔉				ECHILD_12	12	7541	9328	1788								
" # like	- C			ECHILD_13	13	9325	10788	1464								
	_			ECHILD_14	14	10785	11669	885								
oduct like				ECHILD_15	15	11720	12232	513								
nction like	•			ECHILD_16	16	12262	13218	957		<						>
atureID =				ECHILD_17	17	13288	13470	183		Control Nu	mbering					
g like	i –			ECHILD_18	18	13473	13850	378		Assign N						
-				ECHILD_19	19	13847	14041	195		$\sim$ $\sim$						
Hide Igno	red Fe	atures		ECHILD_20	20	14038	14406	369		Assign P	roducts by o I Default Pro					
Select A	II Feat	ures		ECHILD_21	21	14406	14741	336			i Derauit Pro write existing		ypumetica			
				ECHILD_22	22	14751	15176	426			Verwrite def		2 Deaders	in and		
				ECHILD_23	23	15194	15787	594		<u> </u>						
				ECHILD_24	24	15899	16306	408		Assign n						
			Ш	ECHILD_25	25	16420	16731	312	~	Locus Ta	ag Prefix El	JHILD	Sarch N	CBI data	abase	
			1	nsert Delete	Post Valid	ate				Reassign G	iene Data	Dcate G	ray Holes	> 300	) 🔹 bp	

Figure 9.5

**Note:** If you're annotating a portion of a genome as one part of a larger group, you may not want to renumber genes because this may cause confusion if some groups do so and others do not. Make your own decisions, but bear this in mind. You can re-number as often or as little as you like.

Additional Note: Remember that the auto-annotation is based on a random sample of the genome. This means that all auto-annotations will NOT be identical. Therefore, your auto-annotation may not exactly match what was loaded into Phamerator. As you re-number, the gene numbers may again be modified from what is in Phamerator. It is prudent to identify gene by their STOP coordinate. Until a genome is published using gene names (in our case, numbers) all other means of identifying a gene a can lead to confusion.

### 9.3.4 Re-BLASTing a gene

Once you have finished adding a gene, changing a gene's start site, or entering multiple regions for a gene, it can be useful to re-BLAST the gene. This is particularly helpful to check whether or not a gene's modified start site now matches those published in GenBank.

- From the [Features] tab, select the [[Blast]] sub-tab.
- Click the '**Delete All**' button, identified in **Figure 9.6**. Do not skip this step. DNA Master does not overwrite Blast hits, storing multiple copies that can be confusing.

Verview Features Reference	Name	Start	Stop	~	Description Sequence Product Regions Blast Contex
Sort By Index 💌 💆	Name		254	^	Description Sequence Product Regions Blast Contex
Select Features Direct SQL		117			Score Target Description
Type is All 💌	2	251	481		1447 putative terminase gp4 [Mycobacterium phage T .
Name like	3	478	645		1405 gp8 [Mycobacterium phage JAWS]
	4	645	914	_ [	1402 gp8 [Mycobacterium phage CrimD]
GenelD =	5	911	1273		1396 gp7 [Mycobacterium phage Pixie]
Locus like	6	1273	1959		1395 gp8 [Mycobacterium phage Adephagia]
Start >	1	1985	2224	ſ	
Length	8	2221	2493	Г	BLAST Hit Export
	▶9	2465	3910		Accession NP_363740
Regions >	10	3922	5508		GI 18496890 Export A Length 474 Delete
%GC <	11	5489	8248		Max Score 1447 Date 10/5/2011 Delete A
CAI >	12	8245	8442	L	High-Scoring Pairs (HSP)
EC# Ike	13	8534	9106		HSP Data Alignment
Product like	14	9165	10094		
	15	10234	10608		1 MGLGFDRUQD DLGKLICAKR PDGLYAADMF GMS
Function like	16	10611	10967		1
FeatureID =	17	10954	11214		SS HOLDFDLOQD DEGREICKER DDGEIKADEF KER
Tag like	18	11211	11636		51 LCVAAVDRPL TVIWTAHRTR TAAETFKSMQ GF3
	<			×	51   + +                +
Hide Ignored Features					86 LCIKTPNT TVIWTAHRTR TAAETFRSMQ GLJ
Select &II Features	Insert I	Delete Po	ost Valida	e	
	1 - 50000	F	Position : 1		Controls >> Map 🔽 Map >> Controls

Figure 9.6

• A dialog box will pop up and ask if you really want to delete all the BLAST hits for this gene. Click '**Yes**'. The BLAST tab will now be empty of hits, as shown below.

Fionn	bhart	th_c	hecked_	w	þ				
Overview	Feat	ures	Reference	s	Sequence	Documer	ntation		
Sort By	Index	-	- I		Name	Start	Stop	^	Description Sequence Product Regions Blast Context
Select Fe	atures	Dire	ect SQL		1	117	254		There are no BLAST results for this feature
T	is	All			2	251	481		Blast this gene   Blast ALL Genes   Clear All
Туре		AII			3	478	645		
Name	like			L	4	645	914		
GenelD	-			L	5	911	1273		
Locus	like			L	6	1273	1959		
Start	>	í –		L	7	1985	2224		
		<u> </u>		L	8	2221	2493		
Length	>			▶	9	2465	3910		
Regions	>			L	10	3922	5508		
% GC	<			L	11	5489	8248		
CAI	>				12	8245	8442		
EC#	like	<u> </u>		L	13	8534	9106		
		<u> </u>		L	14	9165	10094		
Product	like			L	15	10234	10608		
Function	like				16	10611	10967		
FeaturelD	) =				17	10954	11214		
Tag	like	<u> </u>		L	18	11211	11636		
-					a ( <i>a</i> aa)			~	
🔲 Hide	Ignored	d Feal	tures				>		
Sele	oct All P	eatur	ec		Insert De	lete Pos	t Valida	ate	
H 4 4		€   ▶	H H	1 -	50000	P	osition : 3	7354	4 🔽 Controls >> Map 🔽 Map >> Controls
6 9	10	11 )	h n 4) ]]hh	₿	22	23 24 2	26 28	29	31   B         36                    45  A
95 Feature	s	Live	-						58076 🛃 🤶
				-					

Figure 9.7

- Click the 'Blast this gene' button.
- A new window will appear, labeled "**BLAST search for** [your gene coordinates]". The status of the BLAST attempt will continually be updated in this window until the BLAST is done. When it is finished, the window will display the BLAST results as shown in **Figure 9.8**.
- If you have not BLASTed all genes, or want to re-BLAST all genes you can do it here also.

BLAST	search for 246	5 - 39	10 (9)				
Retrieve [	XML Results   Text	Results	Save to Datab	ase			
Score	Description				_	BLAST Hit	
1567	putative terminase g	jp4 (My	cobacterium pha	ge TM4]		Accession N	IP_569740
1474	gp8 [Mycobacteriun	n phage	JAWS]			GI 1	8496890
1471	gp8 [Mycobacteriun	n phage	e CrimD]			Length 4	.74
1466	gp8 [Mycobacteriun	n phage	Adephagia]			Max Score 1	567
1463	gp6 [Mycobacteriun	n phage	Earva]				
1456	gp7 [Mycobacteriun	n phage	e Pixie]				
1449	gp8 [Mycobacteriun	n phage	Angelica]			HSP Co	verage Map
1442	gp9 [Mycobacteriun	n phage	Anaya]			1101 00	rorago map
1151	terminase [Segnilipa	arus rotu	indus DSM 4498	5]		a	
1010	0.00 I I I I						
E-Value 0.0E0	Starts 1	Bit So Score E-Val	9 1567	Query Query Target Length % Aligned	1 - 453 1 - 446 453 94.1	Positive Identity Similarity Gaps	340
		1	MSNSTATLTD	VARHVVAPTG	IVSTGFSAV	R ATCRHMGLGF	DRWQDDLGKI 🔼
		1	++     ++ MNHSTATLSE	+     VARHVIAPQG	+ +  IVSTAWPSV	I III IIIII R ATCGAMGLGF	DIMODDICKI
		51	ICAKRPDGLY	AADMFGMSIP	RQTGKTYLL	G AIVFALCVAA	. VDRPLTVIWI
		51 51	IIIII IIII	11111 1111 55500005000075	DOTCRTVI	+     + G ALVFALCIKT	+       DMTTVIN1
		101			•		FTNGSRILX
							?

Figure 9.8

• To save your new BLAST hits to your genome file, select the [Save to Database] tab.

BLAST search for 2465 - 391	0 (9)	×
Retrieve XML Results Text Results	Save to Database	
Maximum E-Value of HSPs to save	Ignore Definitions including the following terms	
0.0E0 🗨		^
Save 16 Values		
		~
		?

Figure 9.9

- Click on the drop-down arrow next to the empty field under 'Maximum E-Value of HSPs to save'.
- Scroll through the listed E-values (these are from your new BLAST matches) and pick an appropriate value (greater than 10<sup>3</sup>) that also gives you a useful number of matches (at least 10 or so). If you only have E-values higher than 10<sup>3</sup>, just pick at least one match so you will know that you have BLASTed this gene, and it doesn't have any good matches in GenBank.

- Click the 'Save [n] Values' button. The "n" will be automatically filled in for you based on the number of matches you picked from the drop-down menu. It should then say "[n] saved" in this window under the button. Close the BLAST window.
- Now your new BLAST hits should be listed in your genome file (you may not see them until you select a different feature and then reselect the one you just BLASTed to refresh the view).

Overview F Sort By In	dex 👻	Reference:	s Sequenci Name	e Docume Start	Stop	~	Description Sequence Product Regions Blast Context
Select Feat		ect SQL	1	117	254	_	
Jelecci eau	1.0.0		2	251	481		Score Target Description
Type i:	s All		3	478	645		1567 putative terminase gp4 [Mycobacterium phage T     1474 gp8 [Mycobacterium phage JAWS]
Name li	ike		4	645	914		1474 gpo [Mycobacterium phage JAWS] 1471 gp8 [Mycobacterium phage CrimD]
GenelD	-		5	911	1273		1466 gp8 [Mycobacterium phage Adephagia]
Locus li	ike		6	1273	1959		1463 gp6 [Mycobacterium phage Adephagia]
	>		7	1985	2224		1403 gp0 [mycobacterium phage Larva]
			8	2221	2493		BLAST Hit
Length	<u>&gt;</u>		▶ 9	2465	3910		Accession NP_569740 Export
Regions	>		10	3922	5508		GI 18496890 Export All
% GC	<		11	5489	8248		Length 474 Delete Max Score 1567 Date 10/8/2011 Delete All
CAL	>		12	8245	8442		
	ike 🗌		13	8534	9106		High-Scoring Pairs (HSP)
			14	9165	10094		HSP Data Alignment
	ike		15	10234	10608		1 MSNSTATLTD VARHVVAPTG IVSTGFSAVR ATC
Function li	ike		16	10611	10967		1  ++     ++     +          + +       1 MNHSTATLSE VARHVIAPOG IVSTAWPSVR AT(
FeatureID	=		17	10954	11214		1 MNHSTATLSE VARHVIAPQG IVSTAWPSVR ATC
Tag li	ike 🗌		18	11211	11636		51 ICAKRPDGLY AADMFGMSIP RQTGKTYLLG AIL
☐ Hide Igr	nored Feat	ures	<		>	· .	51 IIIII IIII IIII IIII IIII IIII IIII
Select	∆ll Featur	et	Insert [	Delete Po	st Valid	ate	<u> </u>
	€lel⊾		- 50000	F	Position : 2	3111	🔽 Controls >> Map 🔽 Map >> Controls

Figure 9.10

### 9.4 Making less common changes to your annotation

### 9.4.1 Annotating programmed translational frameshifts

Assuming you have identified the two genes involved in the frameshift (see **Section 8.4.3**), the next critical piece of correctly annotating a frameshift is locating the precise position where the shift occurs. A printed six-frame translation of the region in question is helpful during this process (see **Section 5.1**).

Frameshifting occurs when the ribosome encounters a "slippery" sequence in the mRNA, such as GGAAAA, and loses track of how to count to three. In the most common shift, the -1 shift, the first "A" of the above sequence is "counted" twice; it is read as the third nucleotide in the last codon of the upstream region, AND the first nucleotide in the first codon of the downstream region. (There are also examples of +1 shifts, in which a nucleotide is skipped, or -2 shifts, in which two nucleotides are counted twice.)

For those unfamiliar with finding the slippery sequences and determining where and how the shift is occurring, it is probably easiest to examine a similar phage's genome in Phamerator that has a correctly annotated fusion gene, and compare it to the six-frame translation of your own phage's fusion gene. This will help to determine what the correct amino acid sequence should be, and therefore which nucleotide the shift must occur at.

To annotate a programmed translational frameshift within your phage, you should do the following (we use Fionnbharth below).

### Determine the precise location of the shift

- Using Phamerator or BLAST, find the most similar genome you can that has a correctly annotated frameshift. For Fionnbharth, we've selected Angelica.
- Make a Phamerator map using your genome plus the similar genome you've chosen (see **Section 6.4**).
- Click on the first gene in the correctly called frameshift in Phamerator to select it. Its border will change from black to orange to indicate that it's selected, and its nucleotide and amino acid sequences will be displayed in the panels at the bottom of the window, as shown in **Figure 9.11**.

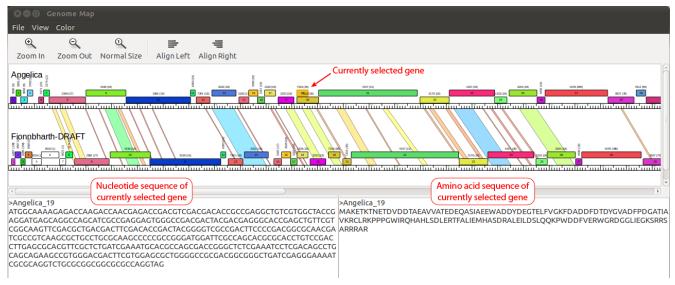


Figure 9.11

- Copy the amino acid sequence from the bottom-right panel and paste it into a new text file.
- Now select the second correctly called frameshift gene (just below the first), and copy and paste its amino acid sequence into a new text file as well.
- Locate the precise position where these two amino acid sequences diverge. (This can be done by manual inspection of the amino acid sequences, or by using BLASTP with the "Align two or more sequences" option checked.) In our example, the two Angelica sequences diverge after amino acid 135, as shown:
  - ... GGLIEGK**SRRSA...** in the first protein.
  - ... GGLIEGKIAQVC... in the second (fusion) protein.
- Now back to your genome. An examination of your six-frame translation shows the two genes as they were called by DNA Master's Auto-Annotate function.

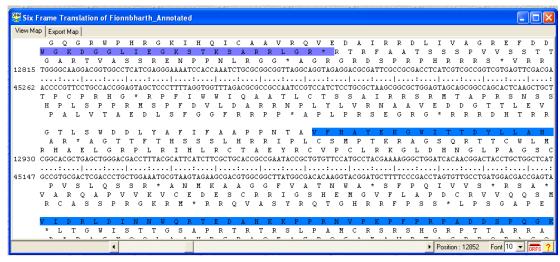


Figure 9.12

- In **Figure 9.12**, the purple bar shows the end of the first protein, and the blue bar shows the beginning of the auto-annotated version of the second protein. Note that the purple highlight is in reading frame 2 while the blue is in reading frame 1. This means that this phage likely has a -1 frameshift, and we need to identify a nucleotide somewhere in this region that should be "counted" twice by the ribosome.
- Near position 12841 there is an obvious slippery sequence, "GGGAAAA" (underlined in red below). If we count the first A (at position 12844) of this sequence twice, we shift frames as shown by the red box, and generate the amino acid sequence ...GGLIEGKIHQIC... in the fusion protein. This sequence is not identical to Angelica's fusion sequence, but it is very close. Counting carefully from the left, we can determine that the first "A" at position 12844 (underlined in green) is the coordinate of our frameshift.

🞇 Six Frame Translation of Fionnbharth_Annotated
View Map Export Map
G Q G R W P H R G K I H Q I C A A V R Q V E D A I R R D L I V W G K D G G L I E G K S T K S A R R L G R * R T R F A A T S S
G A R T V A S S R E N P P N L R G G * A G R G R D S P R P H R 12815 TGGGGCAAGGACGGTGGCCTCATCGAGGGAAAATCCACCAAATCTGCGCGGCGGTTAGGCAGGTAGAGGACGCGCATTCGCCGCGACCTCATCG
452.62 ACCCCGTTCCTGCCACCGGAGTAGCTCCCTTTTAGGTGGTTTAGACGCGCCGCCAATCCGTCCATCTCCTGCGCTAAGCGGCGCTGGAGTAGC
T P C P R H G * R P F I W W I Q A A T L C T S S A I R R S R M H P L S P P R M S P F D V L D A R R N P L Y L V R N A A V E D
PALVTAEDLSFGGFRRPP*APLPRSEGRG*R
GTLSWDDLYAFIFAAPPNTA <mark>VFHAYEKGWIT</mark> AR*AGTTFTHSSSLHRRIPLCSMPTKRAGSQ
R H À E L G R P L R I H L R C T À E Y R C V P C L R K G L D H 12930 CGGCACGCTGAGCTGGGACGACCTTTACGCATTCATCTTCGCTGCACCGCCGAATACCGCTGTGTTCCATGCCTACGAAAAGGGCTGGATCAC
45147 GCCGTGCGACCCCGCCCGGCTGCGGAAATGCGTAAGTAGAAGCGACGTGGCGGCCTTATGGCGACACAAGGTACGGATGCTTTTCCCCGACCTAGTG
PVSLQSSR*ANMKAAGGFVATNWA*SFPQIV VARQAPVVKVCEDESCRRIGSHEMGVFLAPD RCASSPRGKRM*RROVASYROTGHRRFPSS*

Figure 9.13

#### Annotate the frameshift in DNA Master

- Go to the **[Features]** tab and click on the **second** of the two genes involved in the frameshift. (We do not need to modify the first gene, only the second.)
- In the **[[Description]]** sub-tab in the right-hand section, locate the field labeled "Regions" (far right column, shown below). Change the number from "1" to "2", then click the '**Post**' button at the bottom of the central column to save this change.

Fionn	bhart	h_A	nnotate	d										
Overview	Featu	ues	Reference	es	Sequence	Docum	entation							
Sort By	Start	-	E	<u>ا</u>	Name	Start	Stop	^	Descripti	ion Seque	nce Prod	uct Regions	Blast	Context
Select Fe	atures	Dire	ect SQL		5	911	1273		Name	21		GenelD		
-	. 1	4.0			6	1273	1959							
Туре	is	All	<u> </u>		7	1985	2224		Туре	CDS	-	<u>61</u>		
Name	like				8	2221	2493		Start		12991	Locus Tag	DNAM20	1
GenelD	=				9	2570	3910		Stop		13311	Regions		2
Locus	like			1L	10	3922	5508		Length	321		Tag	í –	
Start					11	5489	8248		D	<b>F</b>		3		
					12	8245	8442			Forward	=			
Length	N				-	-	-	-	Translati	on Toble II	Indafinad			-

Figure 9.14

- Change from the **[[Description]]** sub-tab to the **[[Regions]]** sub-tab in the right-hand section of the Features tab.
- You will now enter the two regions that constitute the fusion protein. These must be entered in order, upstream first and downstream second.
- The **Start** coordinate for the **first region** is the start of the whole frameshift region (same as the start for the previous gene). The **Stop** coordinate for the first region is the position you've identified where the frameshift occurs; in our example it is 12844. For the **Length** field, just enter the number 1, because DNA Master will calculate this for us automatically in the following steps, but does require that some number be entered as a placeholder until then.

Extracted from	n FastA Lib	rary Fionn	bharth.1	fasta							
Overview Features	References	Sequence	Documer	ntation							
Sort By Index 🔹		Name	Start	Stop	^	Description	Sequence F	roduct	Regions	Blast	Context
Select Features D	irect SQL	18	11211	11636		Start	Stop	Length			[
		19	11717	12328		* 124			1		
	[	20	12431	12880		P		·			
Name like		21	12991	13311							
GenelD =		22	13314	17489							
Locus like		23	17508	18623							
Start >	<u> </u>	24	18623	20410	-						
	L	25	20411	20929							
Length >		26	20926	22023							
Regions >		27	22035	22280							
%GC <		28	22273	24582							
CAI >		29	24586	25590							
EC# like		30	25612	25989							
		31	26020	27630							
Product like		32	27630	28421							
Function like		33	28459	28857							
FeatureID =		34	28854	29189							
Tag like	<u> </u>	35	29186	29446							
					$\mathbf{\Sigma}$						
Hide Ignored Fe	atures	<		>			1	1.			
Select All Feat		Insert De	lete Pos	t Valida	ate	Insert	Delete	A	ssign Leng	ths	
<b>                   </b>	)	376 - 15625	P	osition : 1	5098	) 🔽 Cont	trols >> Map 🔽	Map >>	Controls		
14	15 16	17 18		9		20 21	X		22		
95 Features Li	ve									58076	a ?

Figure 9.15

- With the "Length" field selected (as shown in **Figure 9.15** by the blue highlight), press **Tab** to move to the second line. For the **second region** of the fusion protein, the **Start** coordinate is the position of our frameshift (again, in our example this is 12844). The **Stop** coordinate is the previously called stop for the second gene (the end of the entire frameshift region, in our example 13311). Again, the **Length** should be entered as "1" for now.
- Click the 'Assign Lengths' button at the bottom of the [[Regions]] sub-tab (see below). DNA Master will calculate the length of each region and display it in the "Length" column.

🞇 Extrac	cted f	rom	FastA L	ibı	ary Fionn	bharth.	fasta						
Overview	Featu	ires	Referenc	es	Sequence	Documer	ntation						
Sort By	Index	-		Ľ	Name	Start	Stop	^	Description S	Sequence Pr	oduct Re	gions Blast	Context
Select Fe	atures	Dire	ect SQL		18	11211	11636		Start	Stop	Length	-	~
T	is	All		L	19	11717	12328		12431			14	
Туре		AII	<u> </u>	L	20	12431	12880		12844		4	68	
Name	like				21	12991	13311		-				
GenelD	=			L	22	13314	17489					T I	
Locus	like			L	23	17508	18623						
Start				L	24	18623	20410				<b>C</b> 1		
				L	25	20411	20929				Calcu		
Length	>			L	26	20926	22023				leng	gths	
Regions	>			L	27	22035	22280						
% GC	<			L	28	22273	24582						
CAI				L	29	24586	25590						
EC#	like			L	30	25612	25989						
				L	31	26020	27630						
Product	like			L	32	27630	28421						
Function	like			L	33	28459	28857						
FeatureID	-			L	34	28854	29189						
Tag	like			L	35	29186	29446						
-				I.				<b>×</b>					
🔲 Hide I	gnored	l Feat	ures	4			>		1	1			
Sele	rct Δ∥ F	eatur	es		Insert De	lete Pos	st Valid	ate	Insert	Delete	Assig	n Lengths)	
H 4 4		2.	<b>H</b>   <b>H</b>	93	76 - 15625	P	osition : 1	3717	7 🔽 Contro	ols >> Map 🔽	Map >> Co	ontrols	
14	$\rightarrow$	-	15 16	X	17 18		19		20 21	×		22	
95 Feature:	s	Live	;									58	076 🛃 🤶

Figure 9.16

• Finally, change back to the **[[Description]]** sub-tab, and enter the correct start and stop coordinates for the entire gene (both regions). In our example, these coordinates are 12431 and 13311. Then click the **Calculator** icon to post changes and calculate the length of the entire gene.

🞇 Extra	cted fron	n FastA Li	brary I	Fionnbh	arth.	fasta								
Overview	Features	Reference	s   Sequ	ience 🛛 D	ocume	ntation								
Sort By	Index 👻		Nam	e Sta	art	Stop	^	Descripti	ion	Sequence	Prod	uct Regions	Blast	Context
Select Fe	atures Di	irect SQL	18	11	211	11636		Name	21			GenelD		
T	is All		19	11	717	12328					_	GI		
Туре			20	12	431	12880		Туре		-	-	_		
Name	like		▶ 21	·	431	13311		Start		1:	2431	Locus Tag	DNAM2	201
GenelD	=		22		314	17489		Stop		1:	3311	Regions		1
Locus	like		23		508	18623		Length	882	2		Tag		
Start			24		623	20410		Direction	n For	ward			·	
Length			25		411	20929				able Undel	fined			
-			26		926	22023				able jonde	Inneu			
Regions			27		035	22280		EC Numb	Der					
% GC	<		28		273	24582								*
CAI	>		29		586	25590		Product gp21						
EC#	like		30		612	25989		gpzi						-
Product	like		31		020	27630								<u> </u>
			32		630	28421		Function	1					
Function	like		33		459	28857								-
FeaturelD	) =		34		854	29189								×
Tag	like		35	29	186	29446		Notes	C.F.v.		- 1000	)4	0.00	<u> </u>
	Ignored Fea	atures	<			>		Uriginal	uim	imer call ଓଡ଼ା	p 1293	31 has strengt	n 9.38	
· · · · · · · · · · · · · · · · · · ·	-		Insert	Delete	Po	st Valida	ate							~
	ect All Featu	1 1 1	·					,	_				_	
<u>       </u>	<b>Q</b> Q		3747 - 15	5996	F	Position : 1	5905		Cont	rols >> Map	M	ap >> Control	s 💼	
14	15	16 17	-18	19	$\rightarrow$		21					22		
95 Feature	s Liv	ve											580	76 🛃 🙎

Figure 9.17

Now if you change back to the **[[Regions]]** sub-tab, you will see a graphic representation of your two frameshifted regions in black bars at the bottom of the tab, as shown in **Figure 9.18**. (You may need to select a different feature, then come back to this one to refresh the view.)

verview Features Reference	<ul> <li>■ Name</li> </ul>	Start		~	Descrin	tion   S	equence   F	roduct	Regions	Blast	Context		
elect Features Direct SQL	5	911	1273	1	Start		Stop	Leng		Didde	-   oomont		
	6	1273	1959	ŀ	Start	12431	1284		414				6
ype is All 💌	7	1985	2224	ŀ	Þ	12431	1284		468				
lame like	8	2221	2493	ŀ		12044	1001		400				
enelD =	9	2465	3910										
ocus like	10	3922	5508										
tart >	11	5489	8248										
ength	12	8245	8442										
	13	8534	9106										
legions	14	9165	10094										
GC <	15	10234	10608										
Al >	16	10611	10967										
C# like	17	10954	11214										
roduct like	19	11717	12328										
unction like	20	12431	12320										
	20	12431	13311										
eatureID =	22	13314	17489										
ag like	23	17490	18623										
Hide Ignored Features	24	18623	20410										
Select All Features	25	20411	20929										
				~						-			_
			>				1	1					
	Insert D	elete Po	st Valida	e _	li	isert	Delete		Assign Lei	ngths			
(4) 4 ( <b>Q</b> ( <b>b</b> ) <b>b</b> ) H	1 - 50000	F	Position : 25	028	~	Control	s>> Map 🔽	Map >	> Control:	s 🗾			

Figure 9.18

**Note:** The frameshift described here is a -1 programmed frameshift. Not all tail assembly chaperone frameshifts are -1. They can also be +1 (typical of Cluster F1 genomes) or -2. (Xu, J., Hendrix R.W., Duda, R.L. (2004) Conserved translational Frameshift in dsDNA Bacteriophage Tail Assembly Genes. Molecular Cell 16, 11-21.

### 9.4.2 Annotating introns

Genes with introns in them can be annotated as two regions by following the procedure above under the heading "**Annotate the Frameshift in DNA Master**." In this case, the two regions you enter will correspond to the exon portions of the gene. However, determining the precise boundaries of these regions is beyond the scope of this guide, and you need to refer to relevant literature or previous examples to figure this out. At this moment in time, we are not calling introns without experimental data.

### 9.4.3 Annotating wrap-around genes

Wrap-around genes are those that 'connect over the right and left ends of the phage genome. Wrap-around genes can be annotated by following the procedure above under the heading "Annotate the Frameshift in DNA Master", (Section 9.4.1). In this case, the first region will the portion of the gene at the right end of the genome, starting at your chosen start site and stopping at the end of the genome. The second region would be the portion of the gene at the left end of the genome, starting at position 1 and ending at the stop codon for the frame. For example, in a 60,000 bp genome, the two regions might be something like 58,734-60,0000; and 1-4.

There is a caveat associated with wrap-around genes. GenBank software cannot tolerate a wrap-around gene when annotated in a linear genome. Since all phage genomes are submitted as linear genomes (because in a phage, they are all linear), a gene that extends over the ends is not permitted as a CDS. However, it is tolerated if labeled as a Miscellaneous Feature instead.

### 9.5 Predicting tRNA and tmRNA genes

DNA Master's Auto-Annotate feature runs the tRNA search tool **Aragorn**, v1.1, which may identify some tRNA genes in your genome. However, the version of Aragorn that is within DNA Master does not call the tRNAs (and their ends) as well as it could. The newest, web-based version of Aragorn is the best of the tRNA programs at determining the correct ends of tRNAs. The other web-based program, **tRNAscan-SE**, is useful for finding non-canonical tRNAs as it is possible to relax its search parameters.

tRNAScan-SE and web-based Aragorn must be run on every sequence.

### 9.5.1 Running web-based Aragorn (version 1.2.36)

• Go to: <u>http://130.235.46.10/ARAGORN/</u>

### ARAGORN, tRNA (and tmRNA) detection

Dean Laslett, an Australian specialist in stable RNAs, is the developer of ARAGORN.

ARAGORN	Search online
Download	Upload (multi) fasta file:
Publication	Browse Echild.fasta or choose a genome: E. coli M. jannaschii Yeast Select options Full list of options
	Type tRNA tmRNA both
Other software	Allow introns, 0-3000 bases no yes
ARWEN	Sequence topology
BRUCE	Strand both single
tRNAscan-SE	Output format standard tab-delimited
	Submit Reset
Other links	
tRNAdb	

### Figure 9.19

- In the '**Upload (multi) fasta file**' section, click '**Browse**...' then select your phage's DNA sequence as a FASTA file.
- Choose the following settings:

Type: Both (tRNA & tmRNA)

Allow introns: no

Sequence topology: **circular** (because phage genomes circularize upon infection)

Strands: both

Output format: standard

- Click the '**Submit**' button.
- Your results will load in a new page. The output includes the secondary structure of the tRNAs found. An example is shown in **Figure 9.20**.

```
ARAGORN v1.2
                                        Dean Laslett
                          _____
Please reference the following paper if you use this program as part of any published research.
Laslett, D. and Canback, B. (2004) ARAGORN, a program for the detection of transfer RNA and transfer-messenger RNA genes in nucleotide sequences. Nucleic Acids Research, 32;11-16.
Searching for tRNA genes with no introns
Searching for tmRNA genes
Assuming circular topology, search wraps around ends
Searching both strands
Using standard genetic code
Bongo Complete Sequence, 80228 bp including 11 bp 3' overhang (ACCTCCTGCAA), Cluster M 80228 nucleotides in sequence
Mean G+C content = 61.6%
 1.
                              c-g
t.t
c-g
c-g
              g-c
t tgcc
a g ::!
agcg tgcg
l:!! c '
tggc
          αta
               tggc ggg-c
j a g c-g
ga a g-
                           g.g
                                               a+t
                                                 g-
                                                    t
         tRNA-Arg(ccg)
96 bases, %GC = 65.6
Sequence [32355,32450]
```

#### Figure 9.20

The principles underlying Aragorn are described in:

Laslett, D. & Canback, B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**;11-16. <u>PMID: 14704338</u>

### 9.5.2 Running tRNAscan-SE (version 1.23)

• Go to: http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::trnascan

Click the button marked Advanced options. Scroll down through the list, and change the following settings:

Improve detection of prokaryotic tRNAs to "yes"

Analyze sequences using COVE only to "yes"

Show both primary and secondary structure components..... to "yes"

Disable pseudogene checking to "yes"

Strict or relaxed tRNAscan-SE mode to "relaxed"

Strict or relaxed EufindtRNA mode to "relaxed"

Cove cutoff score for reporting tRNAs to "0" (you have to scroll down a bit to find this one)

Save secondary structure results file to "yes"

Anything you change from the default settings will automatically get highlighted in yellow.Now if you click the "only simple options" button, all of your yellow-highlighted selected options on the advanced page should be returned to the main window, and it should look like this:

Run Reset tRNAscan-SE 1.23 ? **Detection of transfer RNA genes** advanced options \* Sequence File (EDIT) CLEAR) paste db upload Browse... No file selected. \$ select >Rey Final Sequence, 83724 bp, 11 bp 3' overhang (ACCCCATGCAA), Singleton, GPCL454, 19 Primers 0 ATCGGGCCTTCTCTCTCCGGCACTTTTGGGCCGAGACCCTTCGATTTCAA ACTCCCTGGTCGAGACGAATGAGTGCTCGAAAACAGCTGGTGGCACCTCG GTTGAGGGGTTTGCTGTATGCAAAAAACCCGCCCCTCCCGATGCAGGAAG Search Mode options Improve detection of prokaryotic tRNAs (-P) ? Yes Select archeal-specific covariance model (-A) ? No 🛟 Analyze sequences using Cove only (-C) ? Yes 🛟 Show both primary and secondary structure components to covariance model bit score (-H) ? Yes Disable pseudogene checking (-D) ? Yes Special options Strict or relaxed tRNAscan mode (-t) ? Relaxed (R) EufindtRNA mode (-e) ? Relaxed (R) Specify Alternate Cutoffs / Data Files options Cove cutoff score for reporting tRNAs (-X) ? 0 Output options Save secondary structure results file (-f) ? Yes References : Fichant, G.A. and Burks, C. (1991) Identifying potential tRNA genes in genomic DNA sequences, J. Mol. Biol., 220, 659-671. Ide C. D. and Durkin D. (1004) DNA annual and build under annual an and the Nucl. Aside Day, 22, 2070 2000

Figure 9.21

- Next to the field labeled "Sequence file", click the 'Upload...' tab and select your phage's DNA sequence as a FASTA file.
- Now click "Run"

Note: This program can take some time to run with the relaxed parameters (~20 min), and so it may be worth pre-running it prior to class time.

When the job is finished, you will get a file emailed to you with your results that can be opened by any browser, or you can right-click on the finished job in the far left column (one is circled in the below figure) and select "Open link in new tab". This will open the results of your search.

#### set email | sign-in | sign-out refresh workspace

### Mobyle @Pasteur

tRNAscan Search or All [more]	Welcome Forms Data Bookmarks Jobs Tutorials	
Programs	Overview trnascan - 10/16/13 16:18:53 x trnascan - 10/16/13 16:34:04 x	
<ul> <li>sequence</li> <li>nucleic</li> <li>pattern</li> <li>trnascan</li> </ul>	trnascan - 10/16/13 16:49:24 x  thtp://mobyle.pasteur.fr/data/jobs/trnascan/Q32538305022955  update get help back to form (remove job)	
Data Bookmarks [overview]	parameters	
Sequence : Rey.fasta Jobs [overview] trnascan - 10/16/13 16:18:53 trnascan - 10/16/13 16:34:04 trnascan - 10/16/13 16:49:24	Sequence File (DNA Sequence) Rey.fasta (FASTA) save >Rey Final Sequence, 83724 bp, 11 bp 3' overhang (ACCCCATGCAA), Singleton, GPCL454, 19 Primers ATCGCGCCTTCTCTCCGGCACCTTTGGGCGGGGGGGGGG	

Figure 9.22

The outputs from this program looks like the sample in Figure 9.23

welkin@pitt.edu (guest) set email | sign-in | sign-out

### **Mobyle @Pasteur**

Search [more]	Welcome	Forms D	ata Bookma	rks Jo	bs	Tutorials						
Programs	Overview	trnascan - 1	0/16/13 16	:49:24	x							
alignment	http://mob	yle.pasteur.fr/d	ata/iobs/trna	scan/03253	830502	2955						
assembly		,,,.	, ,,		back to f		nove job	download	)			
<ul> <li>database</li> <li>display</li> </ul>				gerneip	Dack to I		love job	uowinoau	)			
genetics	results											
■ hmm	Poculte filos	Results files (Text)										
information		Revsta.ss     save										
nucleic nucleic	Reysla.s	W Reystalss Save										
phylogeny		24 (61775-61		Length:								
protein     sequence	Type: Un		codon: ???		0-0)	Score	: 19.77					
sequence structure	HAM SC	HMM Sc=-0.81 Sec struct Sc=20.58										
		Seq: GCCTGTTCGTCTAACGGCAGGATTCCTGGgagTTGGTCCCAGGGGATGGAGGTATCGAATCCTCCACGGGGCA Str: >>>>>>,										
Vorkflows	Str: >>>	>>>>>>>		»»»»	•••••		>>>>	•••••	~~~~~	•		
alignment	Rev.trna	25 (61855-61	927)	Length:	73 bp							
database phylogeny	Type: Al	a Anti	codon: TGC	at 34-36		8-61890)	Scor	e: 12.09				
a phylogeny o blast_to_multialign	HMM Sc=-	7.22 Sec	struct Sc=	19.31								Ψ.
• hmm_build_search	full screen	)			boo	kmark) as	Reysta.s	s or		\$	further a	nalysis)
mafft-cons-tree					_							
oprotein_distance_phylogeny	Standard ou	tput (Report)										
utorials	_□ trnascan		e									
data formats	Rey	18	61189	61259	Lys	CTT	0	0	54.65	34.43	20.22	6
BMPS_tutorial	Rey Rey	19 20	61261 61341	61336 61410	Ile Arg	GAT CCT	0	0	47.92 2.49	27.52	20.40	
registration	Rey	20	61425	61502	Glv	TCC	ŏ	ŏ	40.71	9.99	30.72	
setpbystep	Rey	22	61563	61634	Val	TAC	Ō	Ō	61.45	38.25	23.20	
ata Bookmarks [overview]		23	61696	61773	Thr	AGT	0	0	58.76	32.03 -0.81	26.73	
equence : Rey.fasta	Rey Rey	24 25	61775 61855	61848 61927	Undet Ala	??? TGC	0	0	19.77 12.09	-0.81	20.58	
obs Foverview	Rey	26	62051	62123	Asp	GTC	ŏ	ŏ	44.73	14.87	29.86	
	Rey	27	62184	62254	Glu	CTC	0	0	30.93	9.60	21.33	*
trnascan - 10/16/13 16:18:53								_				Ŧ
trnascan - 10/16/13 16:34:04	(full screen	)			boo	kmark) as	6 trnascan	out or				
trnascan - 10/16/13 16:49:24												

Figure 9.23

The top window displays the predicted tRNA sequence with ">>>" under the sequence to indicate portions that pair to make the tRNA stems and "..." underneath to make the tRNA loops.

The bottom window lists the number of found tRNAs, their start and stop coordinations, the amino acid, the anti-codon, the intron boundaries (almost always zeros in phages), the COVE score, the HMM score, and the secondary structure score. We are primarily interested in the COVE score as a measure of the quality of the putative tRNA.

The principles underlying the tRNAscan-SE program are described in:

Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, 25, 955-964.

### 9.5.3 tRNA secondary structure and end determination

Some manual checking is required to determine the precise 3' end of a tRNA gene.

In the tRNA schematic below, the 5' end of the tRNA is a 7 base-pair segment called the Acceptor Stem. The remainder of the tRNA is depicted in the diagram; it winds all the way through three additional stem-loops of variable lengths and then back to the matching base pairs of the acceptor stem. Conserved bases are labeled in nucleotide single-letter shorthand at the appropriate position. The tRNA algorithms score potential tRNAs based on their adherence to the conserved bases and stem-loop lengths.

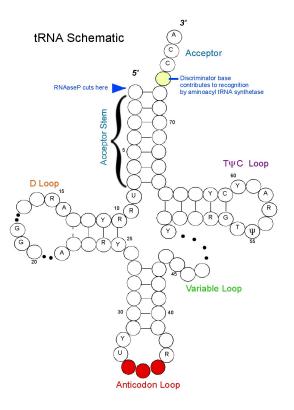


Figure 9.24

After the Acceptor Stem, the 3' end has up to four unpaired bases. The first is called the discriminator base, and it is part of the recognition system that the tRNA synthetase uses to charge the tRNA with the correct amino acid. The discriminator base is followed by the sequence CCA.

The ends of the tRNA must be carefully checked. The acceptor stem loop must be seven base pairs. The CCA sequence at the 3' end must be present on the final tRNA molecule for the tRNA to be charged. Sometimes in the tRNA gene within the DNA of the genome the CCA sequence is truncated, in which case the additional part of the CCA sequence is added after transcription. Therefore if the 3' end of the sequence is not CCA, it should be trimmed at the first deviation from the CCA sequence, and the remainder should not be included in the gene call. This is usually done by web-based Aragorn perfectly.

The tRNA Schematic shown in **Figure 9.24** is an adaptation of the schematic found on the Lowe website <u>http://lowelab.ucsc.edu/tRNAscan-SE/</u> with review and guidance from Dr. Craig L. Peebles.

#### In Summary:

The phages that contain more than 1 tRNA within their genomes tend to localize the tRNAs to certain regions of the genomes (also called "tRNA clusters" in the phage tRNA literature.) It is highly unusual that a phage will contain a sole tRNA distant genomically from all the others within its genome found by the programs, or encoded on the opposite strand as all the others, or encoded within a ORF called by GeneMark or Glimmer that has high coding potential. In general, violation of any of the three preceding conditions is sufficient for exclusion of a potential tRNA from an annotation. It is possible for a phage genome to have multiple tRNA clusters (for example, the Cluster C mycobacteriophages have three tRNA clusters).

The "best" tRNAs are those with a tRNAscan-SE COVE score higher than 20, and that are also found by web-based Aragorn. These criteria include almost all known bacterial tRNAs. Some phage tRNAs meet these standards, however, others don't. Until the phage tRNAs are more extensively tested for expression and functionality in the wet lab, we will err on the side of inclusion.

#### In our annotations, we will include:

- All tRNAs with COVE scores above 2
- All tRNAs found by web-based Aragorn, even if they are not found by tRNAscan-SE at all or have a COVE score lower than 2.

# The ends of the tRNA should be trimmed to match the web-based Aragorn start and stop coordinates.

### 9.5.4 Entering a tRNA in DNA Master

DNA Master may have already called some of your tRNA genes using the old stand-alone version of Aragorn. If so, go to the **[Feature]** tab and the **[[Description]]** sub-tab, and enter the following information. (See **Figure 9.25** for an example.)

- Type: tRNA (not CDS)
- 5' and 3': Exact coordinates as determined above
- Feature Product: "tRNA \_\_\_\_\_" (In the blank, write the amino acid 3-letter abbreviation, e.g. "Lys".)
- Feature Notes: "tRNA \_\_\_\_\_" (In the blank, write the amino acid 3-letter abbreviation followed by the anti-codon, e.g. "Lys (ttt)".) Include the COVE score from tRNAscan-SE and whether or not it was detected by Aragorn.

Descriptio	on Sequence Pro	duct Regions	Blast Context	
Name	128	<u>GenelD</u>		
Туре	tRNA 💌	<u>61</u>		
5' End	61696	Locus Tag	REY_128	_
3' End	61773	Regions		1
Length	78	Tag		
Direction	Forward 🗧 🔁			
Translatio	m Table Undefined			-
EC Numb	er			
Product				
tRNA Th	r			^
Function				<u> </u>
				~
				V
Notes				-
	r (agt), COVE = 58.76 ed Aragorn calls.	6, also called by	Aragorn. Ends trimmed to	^

Figure 9.25

If you are adding a brand new tRNA, click the '**Insert**' button at the bottom of the central column. Then enter in the above information in the window that opens and click '**Add Feature**'. (You can leave the name blank, and it will be automatically assigned when you renumber genes, as described in **Section 9.3.3**.)

### 9.5.5 Identifying and annotating tmRNA genes

Description from Wikipedia:

"Transfer-messenger RNA (tmRNA) is a bacterial RNA molecule with dual tRNA-like and messenger RNA-like properties. In *trans*-translation, tmRNA and its associated proteins bind to bacterial ribosomes which have stalled in the middle of protein biosynthesis, for example when reaching the end of a messenger RNA which has lost its stop codon. tmRNA can recycle the stalled ribosome, add a proteolysis-inducing tag to the unfinished polypeptide, and facilitate the degradation of the aberrant messenger RNA."

The coordinates for tmRNAs can be annotated as web-based Aragorn (or the algorithm BRUCE on the Aragorn web page) calls them. Entering tmRNAs into your DNA Master annotation can be done using the same procedure as for entering tRNAs (**Section 9.5.4**), only the "**Type**" of feature in the should then be "tmRNA" (not CDS or tRNA).

### 9.6 Documenting your gene calls

Just like in at the wet bench, it is important to takes notes and document your findings during genome annotation. While you may want to keep an additional notebook or word document for lengthier rationales or questions, there is a good place to put an abbreviated version of your rationale for each gene in the DNA Master file. In the **[Feature]** tab and **[[Description]]** sub-tab, there is a convenient box marked "Notes" that will allow you to do this.

**Every gene call should be documented in its Notes as described below.** These notes are extremely important for the annotation review process. This is the place where you will want to advocate for those difficult calls. Once checked, these notes will be removed from the GenBank submission file.

Iverview	Feat	1	_	Sequence			<b>I</b>	
Sort By		<u> </u>	-	Name	5'End	3'End	Length	Description Sequence Product Regions Blast Context
Select Fea	atures	Direct SQL	<u>I</u>		55	444	390	Name 1 GenelD
Туре	is	All 👻	1-	[2	413	2089	1677	Type CDS 💌 🖬
Name	like		╘┝	C 3 C 4	2106 3560	3569	1464	5'End 55 Locus Tag DONOVAN 1
GenelD	=		-	05	4896	4834	366	
			-  -	6	5299	6243	945	a num la sur sur la sur l
Locus	like		- 1-	17	6243	6590	348	Length 390 Tag
Start	>			68	6594	6956	363	Direction Forward
Length	>			69	6956	7306	351	Translation Table Undefined
Regions	>			010	7322	7645	324	EC Number
% GC	<		- 1-	011	7645	7914	270	
CAL			- 17	012	7914	8291	378	Product
	_		- 17	013	8288	8725	438	gp1
EC#	like		- 0	014	8819	9817	999	
Product	like			015	9933	10532	600	Function
Function	like			016	9933	10851	918	
FeatureID	=			C 17	10865	14632	3768	
Tag	like	í	-  _	018	14629	16374	1746	Notes
-				E 19	16374	18110	1737	Original Glimmer call @bp 55 has strength 12.44; GeneMark calls sta 151
Hide I	gnore	d Features		020	18133	19026	894	SSC: 55-444 FWD. CP: all covered. SD: 483, highest. SCS: GL calle
Sele	t All F	eatures		E 21	19027	21288	2262	GM called shorter but I went with GL due to higher SD and BLAST matches. Gap: 1st gene. Blast: Q1:S1 gp1 BigNuz. L0: longest. F:
				C 22	21288	22325	1038	Terminase small. FS: Blast P on phagesdb.
			H	E 23	22340	22489	150 363	Y
			H	C 24	22506	22868	363	
				Insert De	elete Post	Validat	•	
	<b>A</b>		11.	47162	D.	sition : 1		✓ Map >> Controls

Figure 9.26

To edit the Notes field, simply click within the field and type. Make sure you Post changes (**Section 9.3.1**) when done so that you don't lose your work. The following information should be recorded for every gene, in order if possible.

- The original auto-annotation information. So do not delete!
- **SSC:** Start/stop coordinates. (This may seem redundant because there are "Start" and "Stop" fields that already contain this information, but it serves as a double-check that all changes you made are actually contained in the final file.)
- **CP**: Whether or not your start includes all the coding potential identified by GeneMark.
- SD: Whether or not the start has the best SD score of all this ORF's possible starts.
- **SCS** (Start choice source): Whether or not the gene was called by Glimmer and GeneMark, and if the start was called by same.
- **Gap** (or overlap): Any significant gap or overlap with preceding gene (in basepairs).
- **Blast:** The best BLAST match, and the alignment of the gene start with that BLAST match. (For example, "Matches KBG gp32, Query 1 to Subject 1", or "Aligns with Thibault gp45 q3:s45".)
- LO (Longest ORF): Whether or not the coordinates you have chosen yield the longest reasonable possible gene for that ORF. A start that overlaps the upstream stop codon by 4bp can reasonably be called the "longest ORF" if the only other start choice would cause a 700bp overlap with the preceding gene.
- **ST:** Starterator data is entered here. If it is not applicable, record NA. If it was run and was not informative, write NI. Other notations are "suggested start" or "conserved start" (a start found throughout the analyzed pham that isn't the suggested start).
- **F** (Function): Gene Function
- **FS** (Function source): source for the function (see **Section 10**), and supporting evidence (BLAST e value? HHPred probability? Phamerator?). Please mention if BLASTP assignments come from phagesdb.org or from GenBank. *Include the phage name and gene number of the phage used as the basis for your functional assignment*. Only enter the putative function in the Notes, **do not write anything into the function field**.
- Anything else you think is important. In particular if you made a different choice than previous annotators have made in published genomes, and feel very strongly about your choice, this is the place to let us know. **Example:** If your gene start does not match the published starts of similar genes in GenBank, an explanation of why not. ("Published Thibault gp45 start not present in my sequence" or "Thibault start caused a 200 bp overlap with upstream gene")

#### An example of good Notes:

**SSC:** Start: 2435 Stop: 2650 (FWD). **CP:** Agrees with both Glimmer and GeneMark predictions. **SD**: 310, best score. **SCS**: ORF includes all coding potential shown on GeneMark-Smeg output. **LO:** 213 bp; longest possible ORF. **Gap:** 84 bp gap with Previous Gene. **BLAST:** gp3 of Oline; Oline aa 1 aligns with query aa 1. **ST:** Not informative; **F** (Function): NKF (No Known Function).

### 10 Assigning gene functions

### 10.1 Overview

Before the age of bioinformatics, the only way to determine a gene function was to perform wet bench experiments: cloning and expressing a gene, or knocking a gene out, and then characterizing the resulting mutants. These kinds of studies are still the gold standard for determining gene function.

Because of recent advances in sequencing technology, however, we are identifying potential genes far more rapidly than we can perform the supporting wet bench experiments for functional determination. Bioinformatic tools can make some strong predictions through comparative approaches, especially by comparing the sequence of any particular gene to the sequences of genes with known functions (i.e., those that have been characterized experimentally).

Even with the new tools that are available, we are unable to assign functions to the majority of the genes that we annotate in bacteriophage genomes. A common occurrence is that students assign poorly-supported gene functions. All reference to functions need strong evidence to be considered.

There are several categories in which genes can be assigned functions with some confidence.

- 1. Virion structural and assembly genes, i.e. those encoding proteins that are either components of virion particles or assist in their formation. These include genes encoding the terminase, portal, capsid maturation protease, scaffolding proteins, major capsid protein, major tail subunit, tail assembly chaperones, tape measure protein, and minor tail proteins.
- 2. **Genes involved in phage DNA replication**. These include DNA polymerase, DNA primase, DNA helicase, nucleotide metabolism genes, and ssDNA binding proteins.
- 3. **Genes involved in life cycle regulation**. These include various regulators such as repressors and activators, integrases, recombination directionality factors, etc.
- 4. **Genes involved in lysis**, including endolysins (referred to as Lysin A in the mycobacteriophages), Lysin B, and Holins.
- 5. **Other well-characterized genes**, including transcription factors, toxin/anti-toxin systems, peptidases, phosphatases, host gene homologues, methylases, nucleases, and DNA binding proteins, among others.

Not all phages contain all of the above genes—or at least genes that can be recognized as having these functions (e.g., we still are not sure where the tail assembly chaperones are in the cluster B phages). Even with a substantial body of knowledge about the mycobacteriophages, we can still only assign functions to 10-20% of the genes in a given genome. Remember that it is okay to write "No Known Function" or "NKF" for a gene.

For more information on the specific function of some of the above phage genes as they relate to mycobacteriophages, see:

http://phagesdb.org/glossary/

Remember assigning functions is an area where the amount of data that is available for comparison is greater than ever before. There is a lot of history to how you can attribute function to the gene product you are investigating. As with everything else you have learned in the annotation process, this requires that you contextualize the evidence you are evaluating. In addition, we are getting better at this, so comparisons to more recent annotations is a better reflection of our understanding than older annotations. If it looks too good to be true, it probably isn't! You can always write a note that there was a weak 'hit' to some obscure function, but DO NOT CALL gene function without good evidence.

### 10.2 Using bioinformatic tools to assign gene function

There are four main tools that are useful for predicting potential gene functions. These are:

- 1. BLASTp (at PhagesDB or NCBI)
- 2. HHpred
- 3. Phamerator
- 4. Conserved Domain Identification (either through NCBI or Phamerator)

### **10.2.1 BLASTP**

**BLASTP** [BLAST (Basic Local Alignment Search Tool) P (Protein)] is a program that searches your query protein sequence against all known predicted protein sequences. You have already come across this in the context of using BLAST to refine your annotations, but it is very useful for predicting potential gene functions.

There are three basic ways of doing BLASTP searches. They can be done:

- 1. within the DNA Master environment (Sections 4.5 and 9.3.4)
- 2. on phagesdb.org (Figure 10.1)
- 3. at BLASTP on the NCBI BLAST server. (Figure 10.2) http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE\_TYPE=BlastSearch&L INK\_LOC=blasthome

When you BLASTP your protein sequence, you are comparing it to all the other protein sequences in your database. If you are using BLASTP through DNA Master or on NCBI's website, your database is GenBank. So even though the results are derived from the same the same database, they display is different. This BLASTP is important to do because this is where your protein sequence is compared to the latest information across all of the protein sequences in NCBI's non-redundant protein databases. This is where you can find 'new' information.

If you are using BLASTP on PhagesDB, your database is comprised solely of mycobacteriophages QC'd by the University of Pittsburgh or mycobacterial proteins. This is valuable because it gives you information that we know about. Remember when you use this that you will see lots of DRAFT annotations. You will want to pay attention to the QC'd entries (those whose name does not contain the word "Draft").

Some Pointers:

- BLASTP at PhagesDB is fast! Actually, it is very fast!
- BLASTP at NCBI and PhagesDB is pointing to information that was input by people, so human error exists. In addition, some folks use an automated process without review, so the perpetuation of those errors is rampant. At all times, you will need to evaluate the data you accumulate.

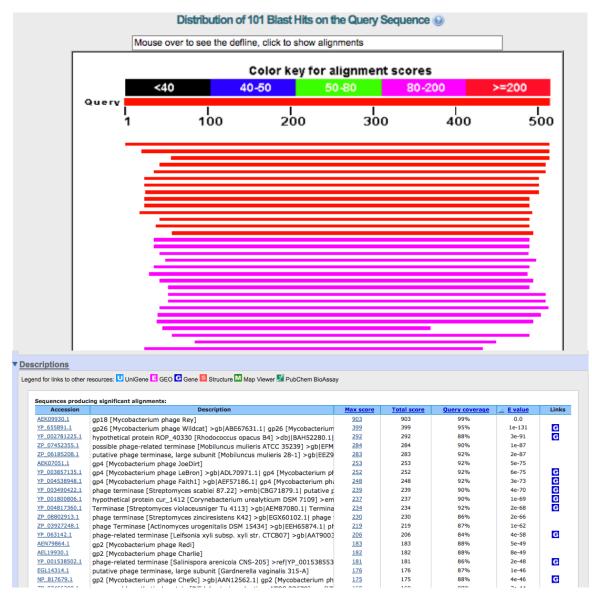
When assigning functions using BLASTP you should consider the following points.

**E value.** E values are a measure of the likelihood that this alignment would appear at random. Therefore, lower E values are better (less likely to be random) matches. For any potential functional match, the E value should be  $10^{\circ}$  or less. The E value should be evaluated in the context of the length of the alignment. (See next point!)

**The length of the alignment**. Does the alignment extend the entire length of your protein? If it only matches a portion, you should interpret this cautiously. For example, if you find a relatively small segment of a protein that matches others at a statistically significant level, you may want to consider annotating this as a domain rather than a full protein function. For example, if a small segment of your protein matches other proteases, you might want to consider writing "peptidase domain", rather than "peptidase" in your Notes.

**Likelihood of the proposed match**. Even if you have an exact match to a piece of a protein in *Vinis vinifera*, it is pretty unlikely that a protein from grapes has the same sequence and function as a protein in a mycobacteriophage. Most of the time when BLASTP aligns bacteriophage proteins with eukaryotic proteins, the alignment is occurring between repetitive sequences, rather than the functional domains of the protein.

**Figure 10.1** is an example of a good BLASTP match, generated using NCBI's web-based BLASTP, where a putative function can be assigned. In the graphical portion of the results, there are many matches in red (the color for the highest match scores) that extend over the entire length of our query sequence. In the list of matches, we can see that all of the E values are well below 10<sup>4</sup>. Many of the hits have a Description that involves terminases. We can now say, with some confidence, that the protein we BLASTEd is a terminase.





### BLASTing a protein on phagesdb.org:

From the top banner, choose BLAST, and select BLASTP from the dropdown menu. Paste your protein sequence in fasta format into the search window, and select the database you would like to search against. The remainder of the settings are similar or identical to the setting choices you see with the program on the NCBI website.

Other than the database, the primary difference between searching GenBank and searching the PhagesDB is in how the results are displayed. We have included both the gene number and the function on the result on PhagesDB to aid functional assignments and highlight mosaicism of the genomes. Gene number and functional assignment are in two different fields in GenBank files and the BLASTP output on NCBI only returns what is written in the product field in a GenBank file.

#### **Distribution of 107 Blast Hits on the Query Sequence**

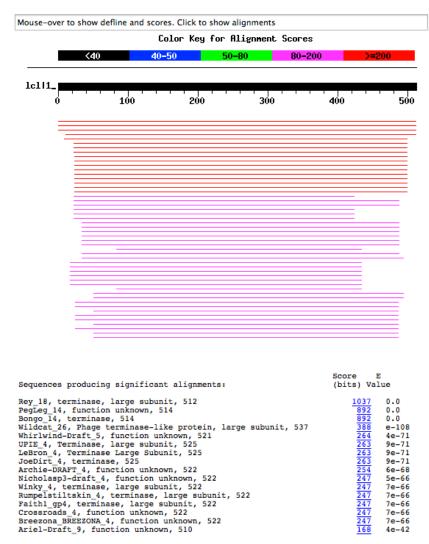


Figure 10.2

#### 10.2.2 Conserved Domain Database (CDD)

When you run your protein sequence through BLASTP on the NCBI webpage, one of the default settings is to examine your protein sequence for conserved domains. Conserved domains are smaller shorter amino acid sequences that are usually affiliated with a specific part (or domain) of a protein. These conserved domains also appear on Phamerator maps as yellow boxes *within* a gene's colored box if they have been enabled through the View menu on the map. You will want to evaluate a CDD match just like you do a BLASTP: look at the E value, the length of the alignment, the likelihood, AND the usefulness of the CDD assignment.

If you have a conserved domain detected within your protein, the function assigned to the domain will be frequently—but not always—be similar to ones found in BLASTP matches. Useful domains to indicate in your annotation are things like peptidases or phosphoesterases, but there is a wide variety that may appear.

Not all conserved domains will be useful. Some contain little information, such as "Conserved domain of unknown function, found in bacteriophages". Others are false positives such as the "Structural maintenance of chromosomes" domain that often appears in structural proteins.

Unfortunately, it is not clear *a priori* which are false fits and which are reliable. Consideration of the genomic context as well as the HHpred search described below are perhaps the most reliable indicators.

An example of a reliable Conserved Domain hit reported by BLASTP on the NCBI server might look like: If you hover your cursor over these boxes with the mouse, a pop-up window will appear that tells you about the conserved domain. When you click on the pink Terminase\_1 box in the provided example, you would detect an E value of 2e10<sup>45</sup>.

Graphic Summa		
	Query seq.	Putative conserved domains have been detected, click on the image below for detailed results.
	Superfamilies	Terminase_1 superfamily

Figure 10.3

The same gene in a Phamerator map might look like:

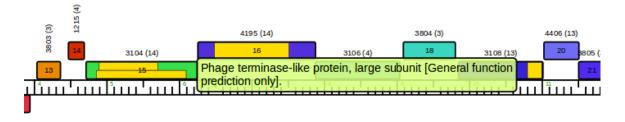


Figure 10.4

In this case, we moused over gene **15** in **Figure 10.4**, and the green box describing the domain appeared. A less informative match on NCBI might look like:

r !	Graphic Summary									
	Show Conserved Domains									
	Putative conserved domains have been detected, click on the image below for detailed results.									
	Query seq.	1 75 150 225 300 375 450	540							
	Multi-domains	Phage_prot_Gp6								

Figure 10.5

We already know that this is a phage protein, so this is not particularly useful information. And the same gene in Phamerator:

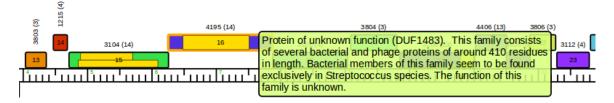


Figure 10.6

In this case, we moused over gene *16* in the above map, which is the well-characterized portal protein (shown in BLASTP hits). Based on the notes in the green box, we see that the Conserved Domain Database does not know that this is the portal protein. This is an example of the dependence of GenBank on its authors, who may not be as informed as they should be.

## 10.2.3 HHpred

HHpred is essentially a more sensitive way of searching for functions than BLASTP. In detail:

HHpred performs an iterated multiple sequence alignment using your query amino acid sequence and its best GenBank matches, using either PSI-BLAST or HHblits (Homology detection by iterative HMM-HMM comparison). It then builds a Hidden Markov Model (HMM) based on the alignment, and compares this model to HMMs based on the Protein DataBank (PDB) (which contains crystal structure coordinates for crystallized proteins). By comparing conserved residues to a 3-D coordinate map, we can sometimes detect and assign gene functions to genes that have very few informative matches using BLAST.

For more information about the design, abilities, and bioinformatics of HHpred, see:

http://toolkit.tuebingen.mpg.de/hhpred/help\_ov

Like BLAST, some matches in HHpred are very useful while others are more likely to be false positives.

To run HHPred, go to <u>http://toolkit.tuebingen.mpg.de/hhpred.</u>

You can register on this site, but it is not necessary. (Anonymous job results are stored **two weeks**, whereas jobs of logged-in users are stored for at least **two months**.) We recommend that you store this data in some fashion (it should save completely in html format. The interpretation that is written in the notes may not be enough information to prove a case to assign a function.

To run a protein sequence at HHPred refer to **Figure 10.7**:

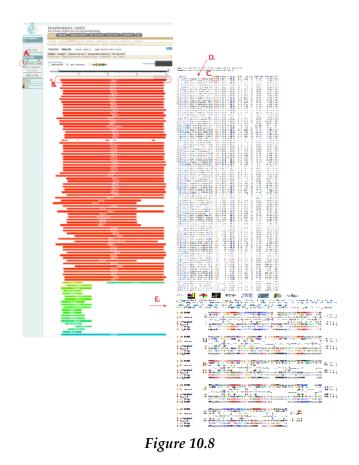
- Copy your sequence and paste it in the appropriate window (A & B)
- Choose 3 databases available in the list (they can be simultaneously used): pdb70\_[currentdate], PfamA\_[currentdate],, and tgrfam\_[currentdate], (C.) See <u>http://toolkit.tuebingen.mpg.de/hhpred/help\_params#databases</u> for more information.
  - Pdb70 is the Protein Data Base that contains all publicly available 3D crystal structures of proteins.
  - PfamA is a database of curated alignments of genetically mobile domains found in signaling, extracellular and chromatin-associated proteins
  - TGRFAMs is a collection of curated protein families that provides a tool for identifying functionally related proteins based on sequence homology.
- Name the job. We suggest phage name\_gene number OR phage name\_STOP coordinate (D.)
- Click the Submit button (E.)
- Check the status of the job by its color. (F.)

HEME		Lopped-in se d	ta@pht.edu looput User Account PtRolect Perce
	Bioinformatics Tool Max.Planck Institute for Devel		
Show results of	A	ralysia 2ary Brushare 2	
			Ch ProcBLAST PSI-BLAST SINSNITCB
Show results			ediction by HMM-HMM comparison
lecert jober	Input B.		
of the second se	Peste protein sequence or multiple alignment	NUSTOR PETERSION INTO	ANY MINIMUM AGENCED INVERSING SOMETICS OF METAWAR
Casr set Jahr			4
Undere set. John			of protein FIR49. Parts example acquirice.
cyevet cutning	er spikat a local file	Brawse No file selected.	-
alone .	Select input format	ASTA 💌	E
			Rosot form Submit job
	Search Dations		
		Standard	Proteomes
	Saind HMM databases (Tail Orte-salest assess)	pdb90_1394rv14	Autidapela, thailona
	Chocae i databasea: pb70 [current date] ptan4 [current date]	pdb_on_Nell_12Hov54 StDhot5_2.04 StDhot5_2.04 StDhot5_3.04	Cseno-hosofitis_skepana Drosophila_msterogeator Horno_saprens Misis_muccluus
	tortans_(current date) MSA Committee Method	🖲 HHolita 🕕 Pablaat	
	Max. HEA Germative: Residence	k <u>w</u>	
	Score secondary structure	🖲 yes 🕞 ne 💮 predicted	на ревосава отту
	Algoment mode	🖲 local 🕕 şidani	
	Dealgo alth BAC		
	More options Show		
	Job Optiona	D.	
	300-10	inter gat (there empty for	Automatic creation)
	Send wolfization to (botienal)		

Figure 10.7

The results can look like Figure 10.9. Note the following

- Green here indicates the job is finished. (A.)
- Each bar represents an alignment that is linked to more information below. It follows the rainbow color scheme, with red showing the highest probabilities (95-100 %). This demonstrates the length of the alignment. (B.)
- Each line in (B.) has a probability (C.) and a definition (D.).
- Once you get to where the description and alignment are (E.), you can evaluate if the structure has the same function in your phage genome. Remember functions have to be appropriate for phages!



**Figure 10.8**: (This representation has been compressed.)

Like BLAST, HHpred provides a graphical view where the best matches are shown in red and lower-quality matches go to orange -> yellow -> green -> blue -> black. Also like BLAST, below the graphical representation is a list with useful information, including the score each hit gets. In the above screenshot, the best hit, "D-alanyl-D-alanine carboxypeptidase" (this is the PDB designator, if you want to see the crystal structure), matches your protein with 100 % probability and an E value of 3.1e<sup>st</sup>.

Good HHpred matches have high probabilities (90 or above), and low E values (the lower the better). The scientists who wrote HHpred claim that matches with probabilities above 30% might be real matches. However, if you are going to claim a function found in HHpred with a probability between 30 and 90%, supporting data (such as a conservation of a domain, a function found in other mycobacteriophages, or synteny) is necessary.

For more on determining if your HHpred hit is a real match, see:

http://toolkit.tuebingen.mpg.de/hhpred/help\_faq#correct%20match

When we scroll down to the look at the specifics of the alignments, we see:

Another example of an informative report from HHpred is below.



The top hit, to 2k5k\_A, has a Probability score over 90, and it is an uncharacterized protein. The rest of the matches have low probabilities (80 or below), and high E values. So even though the other matches are to phosphatases, and one might be tempted to write "phosphatase", this would not be a supportable functional prediction for this protein.

## 10.3 Other ways to assign gene function

#### 10.3.1 Synteny

Many of the genes in bacteriophage genomes—but especially in the virion structure and assembly genes—appear in the same order (synteny). Therefore, sometimes functions can be inferred from gene order. The typical order is:

Terminase  $\rightarrow$  Portal  $\rightarrow$  Capsid Maturation Protease  $\rightarrow$  Scaffolding  $\rightarrow$  Major Capsid Subunit  $\rightarrow$  Major Tail Subunit  $\rightarrow$  Tail Assembly Chaperones  $\rightarrow$  Tape measure  $\rightarrow$  Minor Tail Proteins

Sometimes other smaller genes of unknown functions are interspersed within the structural genes, but in general the overall order remains conserved. While we may see conservation of gene order in some other areas of phage genomes, these other areas are far more mosaic than the structural genes are, and so the use of a synteny argument applies primarily when assigning gene function to the virion structure and assembly genes.

The longest gene in the genome of a phage with a flexible tail is almost always the tape measure protein gene. This gene is directly proportional to the length of the tail in the flexible-tailed phages.

### 10.3.2 Prior functional assignments

Many of the genes within the previously sequenced mycobacteriophages have already been assigned functions based on experiments, BLAST and/or HHpred matches, or synteny. Do not assume that all functions are known and recorded in our database. There is new data available all the time and should be reviewed. Even Dr. Hatfull periodically reviews the mycobacteriophage genomes and updates gene functions to genomes that were annotated long ago. If you are trying to assign a function to a gene that has a BLAST match to or is in the same pham as one of the genes with an assigned function in our published literature, you may assign your gene the same function.

### 10.3.3 Phamerator

Many of the genomes in Phamerator have already been published according to the most recent functional assignments, but not all. We are constantly in the process of improving our gene calls, and so Phamerator functional assignments reflect our best effort at assigning gene functions **at the time the QC'd form of the genome was entered** into Phamerator. This means that many of the more recent genomes might have better functional assignments than some of the older ones. If you're using comparisons in Phamerator to already-published genomes to determine function, your best source of gene function are available in the Phamerator Map, with **Descriptions** enabled through the View menu of the most recently published Mycobacteriophage genomes. **These same Phamerator descriptions also appear in the results line in PhagesDB BLASTP searches**.

# 11 Merging and checking annotations

## 11.1 Merging overview

In a classroom setting, different portions of a genome are often assigned to different students or groups of students to annotate. Once all portions have been annotated, they must be combined into a single file, and the "**Merge**" function in DNA Master performs this action. It takes multiple files from a single phage genome and creates a single master file that contains all of the gene calls from each individual file.

**Note**: merging will **only work on files that contain identical sequences**. If you are going to split a genome among different annotation groups, make sure that you keep the entire sequence intact, and simply work on a region identified by gene coordinates (e.g. between 20,000 and 30,000).

Typically, you'll merge all of a given genome's partial annotations together into a single file that can then be proofed and edited to become the final complete annotation. However, it is also possible to do several iterations of merging. For example, if two groups are working on the region from 10,000 to 20,000, you may want to merge their files first, come to a consensus on that region, then merge the newly checked version with the other final files from other sections of the genome. Merging is flexible enough to meet your pedagogical goals.

## 11.2 Merging multiple annotations into a single file

- Collect the files you'd like to merge into a single directory. Remember that these must all be from an identical DNA sequence (i.e., the same phage genome).
- You may want to include a newly made auto-annotated only genome. Go back to phagesDB.org to obtain the FASTA file and quickly auto-annotate. Use this as your reference file in the merged data. If the sequence has been corrupted in any of the student files, the merge will not work. This is an excellent quality control measure!
- Open DNA Master.
- Go to File → Merge

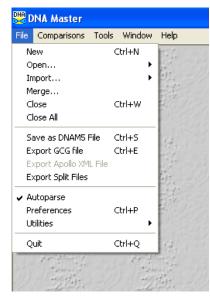


Figure 11.1

• A new window will open, as shown below.

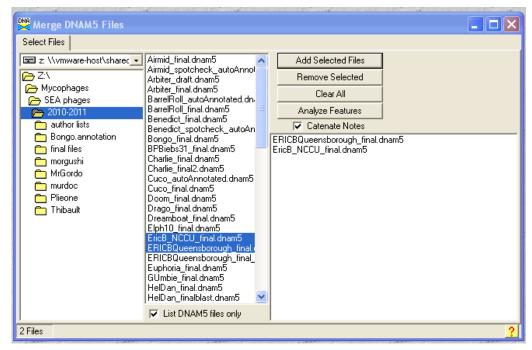


Figure 11.2

- In the left column, browse to the directory on your computer that contains the DNA Master (.dnam5) files that you want to merge.
- In the center column, click on files that you want to add to your merged file.
- Click the 'Add Selected Files' button. The files will then appear in the empty white box on the right. You can browse to additional directories (if necessary) to add additional files.

- Once all of the files that you would like to merge added, they will be listed in the white, check the box marked "**Catenate Notes**".
- Click the 'Analyze Features' button.
- The window will open a new tab, [Merge Files].

Select Files Merge F		
	sent in at least 1 🜩 files	Merge Files
	ear Selections	 Export Summary
Feature	1 2	
488 - 1177		
1333 - 1926		
1090 - 1926		
1926 - 2147		
2483 - 2953		
3038 - 4000		
4000 - 4665		
4662 - 5360		
5403 - 5474		
5480 - 5552		
5595 - 5666		
5697 - 5855		
5855 - 7066		
7066 - 7500		
7524 - 9278		

Figure 11.3

In the picture above, Features (or gene calls) are listed according to genome coordinates. Each file you selected is represented by a numbered column, displayed in the order that they were selected in the previous tab.

In each row, a black box is present if that file contains that feature, and a white box is present if the file does not contain that feature. The first feature, 488-1177, is present in both of the files that were merged. The next feature, from 1333-1926, was present only in the first file. The third feature, from 1090-1926, was present only in the second file. Because both of these features have the same stop codon, what we are looking at is a disagreement in the two files about where the start for this gene should be. File 1 calls it at 1333, while file 2 calls it at 1090.

• To export a spreadsheet that contains the above information (which can be useful to identify areas of disagreement that require further attention), click the 'Export Summary' button in the top right of this window.

To create a .dnam5 file with all of the gene calls from the files to be merged:

• Click the 'Select Features' button. (Selected features will turn red, as shown below.)

Merge DNAM		
Select Files Mer		
Select Features	present in at least 1 🚖 files	Merge Files
Toggle Selected	Clear Selections	Export Summary
Feature	1 2	2
488 · 1177		
1333 - 1926		
1090 - 1926		
1926 - 2147		
2483 - 2953		
3038 - 4000		
4000 - 4665		
4662 - 5360		
5403 - 5474		
5480 - 5552		
5595 - 5666		
5697 - 5855		
5855 - 7066		
7066 - 7500		
7524 - 9278		<u>.</u>
2 Files		

Figure 11.4

• You can tailor your selection by modifying the number in the dropdown box next to "present in at least \_\_\_\_\_ files". After changing the number, click the 'Clear Selections' button to erase previously selected genes, then click the 'Select Features' button again to make your new selection. In the picture below, now only the features present in at least two (both) files are selected and shown in red.

Merge DNAM5	Files	
Select Files Merge F	iles	
[Select Features] pre	sent in at least 2 🜩 files	Merge Files
Toggle Selected C	lear Selections	Export Summary
Feature	1 2	
488 - 1177		
1333 · 1926		
1090 · 1926		
1926 - 2147		
2483 - 2953		
3038 - 4000		
4000 - 4665		
4662 - 5360		
5403 - 5474		
5480 - 5552		
5595 - 5666		
5697 - 5855		
5855 - 7066		
7066 - 7500		
7524 - 9278		<b>~</b>
2 Files		?

Figure 11.5

- Once you have selected the features you would like in your merged file (picking all of them is a good choice, disagreeing features can always be deleted from the merged file after review), click the '**Merge Files**' button at the upper right corner.
- A new window titled '**Merged Sequence**' will appear, as shown below.

curry curry	and the same and the state	and the share	Carlos Carlos
Merged Sequence			
Overview Features Reference	s Sequence Documer	ntation	
Sort By Index 👻 🔳	Name Start	Stop 🔥	Description Sequence Product Regions Blast Context
Select Features Direct SQL	▶ 1 488	1177	Name 1 GenelD
Type is All 🔻	2 1333	1926	Type CDS V GI
	3 1926	2147 -	
Name like	4 2483	2953	Start 488 Locus Tag ERICB_1
GenelD =	5 3038	4000	Stop 1177 Regions 1
Locus like	6 4000	4665	Length 690 🔳 Tag
Start	7 4662	5360	Direction Forward
	8 5403	5474	
Length	9 5480	5552	Translation Table Undefined
Regions <mark>&gt;</mark>	10 5595	5666	EC Number
%GC <	11 5697	5855	
CAI >	12 5855	7066	Product
	13 7066	7500	gp1
EC# like	14 7524	9278	
Product like	15 9275	10738	Function
Function like	16 10813	11619	·
FeatureID =	17 11668	12198	~
	18 12231	13202	Notes
Tag like		~	Feature identified in 2 of 2 files
Hide Ignored Features		>	Notes from EricB :
Select All Features	Insert Delete Pos	t Validate	
	1 - 50000 P	osition : 17878	3 🔽 Controls >> Map 🔽 Map >> Controls
h214567 120 14 120	130 00 00 27 08 2		11/11// 43. DAARK 167/51. THE K BELIEVE AT K THER 20/20/20/10/01/67/11/06 V
112 Features Live		- <u>778</u> 2780 P.	51702 🔐 ?
			31702 🗑 🤙

Figure 11.6

- Save your file immediately by going to: **File** → **Save as DNAM5 File**
- Select a meaningful name for the merged file, such as "YourPhageName\_Merged.dnam5".

In the above picture, we are looking at feature 1. In the "Notes" field on the lower right, the top line indicates that this feature was called in 2 of 2 files. Further down in the Notes box, both sets of notes have been concatenated.

#### How features and notes are reconciled when there is disagreement:

While all the genes from the unmerged files will be present within the features of the merged file, DNA Master will not treat all these genes equally. Features that share the same stop codon but have different start codons will be listed as separate features in the merged feature list. Features that were selected by the majority of the files in the merge will be given preference in the merged file, and will be listed first in the feature table if it is sorted by Index.

The most popular features will have concatenated notes. That is, all the notes from the unmerged files will be listed in the Notes field of the merged feature. Less popular features will be in the merged file, but will be listed at the end of the feature list when sorted by Index. Less popular features will have their original notes, not merged notes.

• To clearly see discrepant calls, go to the "Sort By" drop-down menu at the top left of the [Feature] tab, and select "Start" rather than "Index".

🕎 EricBMerged Sequence														
Overview	Feat	ures	Reference	s	Sequence	Documer	ntation							
Sort By	Start	-	I I		Name	Start	Stop	^	Descriptio	on Seque	ence Prod	luct Regions	Blast	Con 4 🕨
Select Fe	atures	Dir	ect SQL	L	1	488	1177		Name	4		GenelD		
Turne	is	All		L	2	1090	1926			CDS		GI		
Туре			<u> </u>	L	2	1333	1926		Туре	lubs				
Name	like			L	3	1926	2147		Start		2483	Locus Tag	ERICB_4	
GenelD	=			Þ	4	2483	2953		Stop		2953	Regions		1
Locus	like			L	5	3038	4000		Length	471		Tag		
Start	>	<u> </u>		L	6	4000	4665		Direction	Forward		-	·	
	-			L	7	4662	5360				💻 📃			
Length	>			L	8	5403	5474			on Table	Underined			
Regions	>			L	9	5480	5552		EC Numb	er				
% GC	<			L	10	5595	5666							
CAI	>	í –		L	11	5697	5855		Product					
EC#	like	<u> </u>			12	5855	7066		gp4					<u>^</u>
					13	7066	7500							$\sim$
Product	like				14	7524	9278		Function					
Function	like				15	9275	10738							~
Featurel	=	í –			12	9296	10738							$\sim$
Tag	like	<u> </u>			16	10813	11619		Notes					•
Tay	like						_	~	Feature i	identified ir	n 2 of 2 files			^
🔲 Hide	Ignored	d Fea	tures	4			>		Notes fro	om EricB :				-
Sele	ont All P	eatu			Insert De	lete Pos	t Valida	ate	1.0003 m	Sin Enco .				×
सित्ति	<b> Q</b>   <b>(</b>	Ð		1 -	6250	P	osition : 2	5590		Controls >>	Мар 🔽 М	ap >> Control	:	
		1	2		3)		4		5	X	6	7	D 89 N	11) 12
112 Featur	es	Liv	e		1 - 7					4			5170	2

Figure 11.7

You can see that there are two versions of gene 2, one from each file, that share the same stop codon but differ in their choice of start codon. Now it's up to you to determine which is correct!

Note: if students are only working on a portion of the genome, it may be advantageous to delete the auto-annotation calls from the portions of the genome they are not working on prior to merging. That way, only the reviewed features will be in the merged file instead of many unreviewed computer-called features. Remember, only delete the features you don't want to merge, not the underlying sequence.

## 11.3 Checking an annotation

Once you've merged all files, made final decisions on each gene, and believe you've finished your annotation, there are a few final steps to take before submitting your genome for review and then GenBank submission. The steps below reflect what we typically do at the University of Pittsburgh to quality-control submitted annotations, so you can stay one step ahead and try to identify any remaining issues first.

- Review your cover sheet and check that you have correctly formatted and re-Blasted your phage DNA Master file. Annotations will be returned if all components are not completed.
- Click the '**Validate**' button bottom of the central column in the **[Feature]** tab. The response should be "All ORFs appear valid." If you get a different message here, check the gene(s) identified for errors.
- Zoom in on the interactive map along the bottom of the sequence, and carefully scroll along the whole length of the genome. Do all the genes seem to be tightly packed? Look for large overlaps, gaps, or duplications. You can also do this by generating an ORF Map (Section 5.2)

- Open an interactive Phamerator map of your phage along with two or three closely related cluster members that are already in GenBank. (Remember that it is still your auto-annotated genome in Phamerator.) Are there any areas where your phage has orphams (white boxes) or otherwise diverges from similar phages that you have **not** addressed during your refinement? Or you can address these same questions using The Genome Comparison Tool in DNA Master. See Exploring Bacteriophage Biology at Protocols at phagesDB for instructions.
- Re-BLAST your genome. The BLAST data must match the final calls of your phage genome final file. Don't guess or presume that the Blast data is up-to-date. ReBLAST for final review.
- Create a "Genome profile". This is a spreadsheet (.csv format) of all the information in the Features table. While this won't give you any new information compared to simply scrolling through your features, it may help you make sure you don't miss anything or that you have recorded in formation in an incorrect field.

#### Go to: Genome → Profile

In the window that opens, there are a number of settings. The default settings should be fine, but consider checking the "Export Notes" box if you'd like Notes included in your spreadsheet, and consider unchecking the "Load into Excel" box if you don't have Excel or would like to open the file later. You may also export "Product", "Function", and "Best BLAST hit" for your final review. You can quickly evaluate if your file is in tip-top shape.

🔛 Genome Profile : Airmid	
Basics       Codon Bias       Dinucleotide       Profile Merge         Image: Start and stop codons and their flanking bases       Start and stop codons and their flanking bases         Image: Start and stop codons and their flanking bases         Image: Start and stop codons and their flanking bases         Image: Start and stop codons and their flanking bases         Image: Start and stop codons and their flanking bases         Image: Start and stop codons and their flanking bases         Image: Start and stop codons and their flanking bases         Image: Start skew         Image: Start and string tags         Image: Custom integer and string tags         Image: Protein Features (pl, Hydrophobicity, etc)         Image: Arrive Acid Courts         Image: Window for Coiled-Coil         Image: Start and Acid Courts         Image: Start and acid Courts <t< th=""><th>Genes All genes Selected genes All ORFs All RNAs Replicons Current Replicon All Replicons <u>QK</u> <u>Load into Excel</u></th></t<>	Genes All genes Selected genes All ORFs All RNAs Replicons Current Replicon All Replicons <u>QK</u> <u>Load into Excel</u>
	?

Figure 11.8

• Now check each gene individually.

**Read the comments, and consider**: Do the start and stop coordinates listed match the coordinates in the file? Does the gene have Glimmer/GeneMark support? A good RBS/Shine-Dalgarno score? Include all the GeneMark-Smeg coding potential? Is the gene as long as possible without overlapping the previous gene too much? Match its best BLAST hit 1-to-1? If the phage has close relatives in GenBank (you can tell pretty quickly by using Phamerator), our frequent default position is to make a newly annotated gene match the

annotated genes already in GenBank. If it doesn't, use your best judgment based on the other metrics.

**Check the gene functions, and consider**: Do they make sense? Are reported E values low (below 10<sup>2</sup>)? Do they match the Hatfull-approved data (where appropriate)? Is there a source listed for a function (HHpred, BLASTP, CDD, GFHmap, other)? If there is no known function, is "NKF" written?

**When checking tRNAs, consider**: Is the tRNA amino acid and anti-codon written in the Notes box? Does the tRNA end with "CCA", and if not is it trimmed correctly? Did you report the COVE score and if it is found by Aragorn?

**For gaps in your gene calls, consider**: Is there an ORF with coding potential that was missed? Are there any BLASTX hits with good GenBank matches?

Keep track of any potential issues you encounter during checking, and revisit those areas of the genome to ensure the best call has been made. Make a note of all issues, and include them in a cover sheet to be submitted with your final annotation.

# 12 Submitting final files for review and GenBank submission

You've made it. Plowed through gene after gene, pored over BLAST results and coding potential diagrams, perhaps argued over some start sites, and have merged all calls and come up with a final annotation.. Congratulations!

But you are not quite done. Science is always in the details and there are details to be done!

To submit your competed project for QC, you still need to:

- Make a duplicate copy of your file, erase all of the notes, and add functions into the notes field. This is merely a formatting issue, but will help in the QC process.
- Upload both versions to PhagesDB
- Upload an author list to PhagesDB
- Upload a cover letter to PhagesDB

The submit annotation button is located on PhagesDB under the Data button.

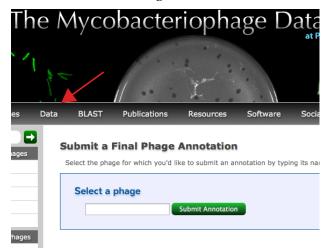


Figure 12.1

After expert review, your annotation will be either accepted or returned. If accepted we will provide a GenBank flat file for your inspection. Once accepted, the file is then sent to GenBank. If not accepted, your file will be returned with an explanation and request for revisions.

# 12.1 Details of your final DNA Master (.dnam5) file

A final .dnam5 file is one that has the following properties.

- 1. It must be named "YourPhageName\_CompleteNotes.dnam5", which will help distinguish it from other versions you may have been working on.
- 2. Do not delete the original Auto-Annotation remarks.
- 3. It must contain one entry and set of notes (and only one) per feature. That means that if you have merged multiple files, you need to have evaluated the data from each source, come to a decision, and deleted erroneous versions of each feature. There should also be only **one set of notes** for each feature, and it should contain

**everything** listed in **Section 9.6** about proper documentation of your gene calls. You may have to delete some notes, or even rewrite some notes from scratch to meet this criterion.

- 4. All features must be validated (Section 9.3.2).
- 5. All features must be re-numbered if necessary (Section 9.3.3).
- 6. Recreate the Documentation (**Section 1.4**).
- 7. All features must be re-BLASTed (Section 9.3.4).
- 8. Any functions are noted in the Notes fields, along with their source. (Section 9.3.3.)
- 9. The correct format will look like **Figure 12.2**.

Sort By	Index	T I	•[	Tag	Name	5'End	3'End	Length	~	Descriptio	n   Sea	ience   Proc	luct   Re 4
Select Fea	5	Direct SQL	┢	SHEEN 1	1	397	693	297	-		1		
				SHEEN 2	2	732	1166	435			5		<u>GenelD</u>
Туре	is A	u <u>-</u>	1	SHEEN_3	3	1259	1576	318		Туре	CDS	-	<u>61</u>
Name	like			SHEEN_4	4	1566	2318	753		5' End		2347	Locus Tag
GenelD	- [		1	SHEEN_5	5	2347	3570	1224		3'End		3570	Regions
Locus	like [		1	SHEEN_6	6	3599	3877	279		Length	1224		Tag
Start			- 10	SHEEN_7	7	3889	4512	624		Direction			iug
			- []	SHEEN_8	8	4509	5477	969				💻	
Length				SHEEN_9	9	5551	5628	78				Undefined	
Regions	>			SHEEN_10	10	5629	5704	76		EC Numb	er		
% GC	<			SHEEN_11	11	5635	5772	138					
CAI				SHEEN_12	12	5731	7155	1425		Product			
EC#	like			SHEEN_13	13	7152	7595	444		gp5			1
	like [		-  -	SHEEN_14	14	7592	8332	741					1
Product				SHEEN_15	15	8359	10059	1701		Function			
Function	like			SHEEN_16	16	10056	11552	1497					1
FeatureID	-			SHEEN_17	17	11549	12562	1014					1
Tag	like 🗌			SHEEN_18	18	12621	13130	510		Notes	ve.	1.01.000	21
- Hide le			Ŀ	SHEEN_19	19	13160	14149	990		strength		all @bp 2343	7 has 🧹
Hide I	gnorea r	-eatures		SHEEN_20	20	14229	14390	162		SSC: 23	47 - 3570	CP:only sta	
Sele	ct All Fe	atures		SHEEN_21	21	14394	14768	375				potential SE es with all G	
				SHEEN_22	22	14765	14920	156		of 30 Bla	st: First m	natch is to HI	NdeR and
			L	SHEEN_23	23	14917	15300	384	~			alignment LO	
				Insert Delete F	Post Validat	e				suructural	protein r	S: phagesD	D W/

Figure 12.2

## 12.2 The second "minimalistic" Final .dnam5 file

This file will contain only the formatting for GenBank submission. Now that you have arrived at you're your best annotation, we are ready to present it as complete and polished. To do so, we remove all notes except REPORTABLE functions. This year, with the increase in numbers of genomes, we are asking that you do this step. This step may also help you to separate out the functions you wish you could call and the ones that you are willing to stand behind and truly defend. There are a couple of ways to do this, but here is our advice for a simple approach:

- Create a Profile of your genome (**Section 11.3**). This will put your features in an Excel spreadsheet.
- Erase all notes (**Figure 12.3**).

Notes				$\frown$					
Add Private	Clear Private	Privatize	Publicize	Clear					
	All	All	All	All					
Original Glimmer call @bp 2347 has strength 10.85 SSC: 2347-3570 CP: captures all coding potential SD: best scores SCS: agrees with all Gap: 30 bp gap Blast: first match is to HINdeR and Timshel with 3:4 alignment L0:Yes ST: not informative F: D-ala-D-ala decarboxylase FS: HHPred probability 100%									
Figure 12.3									

• Add reportable functions into the empty notes fields, using the profile as your guide (**Figure 12.4**).

Notes	
D-ala-D-ala decarboxylase	~
Figure 12.4	

- Recreate documentation (Section 1.4).
- Save file as "YourPhageName\_Final.dnam5".

## 12.3 Details of your author list

Please create a list (.csv formatted file) of the authors from your school who are to be included in this GenBank submission. Your author list should meet the following criteria.

- It contains **ONLY** authors from your school who deserve to be listed on the GenBank file. **Do not** include names from Pitt, HHMI, sequencing centers, or any other source.
- It is a .csv file. A .csv formatted file can be created in Excel, using the '**Save as...**' function, and selecting .csv as the file type.
- It contains exactly three columns, with **NO HEADERS** at the top of each column.
- The first column contains the last name, the second column contains first name, and the third column contains a middle initial. If no middle initial is needed, type a period in that column instead. All three columns should contain some information for each author. See below for an example.

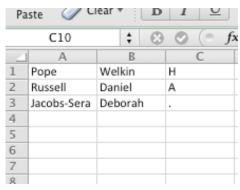


Figure 12.5

# 12.4 Details of your cover sheet

For submission along with your annotation, **please create a document with a brief list of genes in the genome that you feel warrant extra attention by the annotation quality control team prior to submission and why**, including but not limited to: start choice, functional assignment, or gene inclusion/exclusion; and/or areas that you have extensively investigated and you feel should remain genome gaps. Do not send this information in the text of the email, but rather as an attached document. This list should not be more than a page. If you feel confident in all areas of your annotation, please state so. Upload it along with your 2 .dnam5 files when complete.

#### Acknowledgements

DNA Master was designed and developed by Dr. Jeffrey G. Lawrence at the University of Pittsburgh. The program has gone through a multitude of advances, some of which were implemented by Dr. Adam Retchless when he was a graduate student with Jeffrey. Dr. Lawrence continues to provide support, updates and new functionalities to DNA Master.

DNA Master is much more than a genome annotation tool, although this is its main role in this guide. DNA Master has been developed for assisting in bioinformatic dissection of genomes – primarily microbial – with a view to understanding how they have evolved and how they are related. As you become familiar with the program and develop your own interests in genome evolution, we hope these utilities will be of use to you.

We are deeply grateful to Dr. Lawrence for making DNA Master available to us and for his constant willingness to listen to our suggestions and our particular needs. Over many years we have found DNA Master to be an incredibly effective platform for genome annotation and analysis, and Jeffrey's contributions cannot be overestimated.

We would also like to thank the literally hundreds of students and faculty who have used DNA Master and provided feedback that has helped us to develop and refine this annotation platform.

We thank our colleagues in the Science Education Program at HHMI, especially David Asai, Kevin Bradley, Lucia Barker, Razi Khaja, and Melvina Lewis, for their tremendous insights and feedback. Melvina F. Lewis provided the terrific cover design.

The electronic version of the guide is available on the wiki and PhagesDB. Additional helpful documents at PhagesDB include:

- System requirements and Installation of DNA Master
- DNA Master Quick Start Guide
- Gene Function with Bench Support and References
- Case Study: The Annotation of Etude
- Exploring Bacteriophage Biology