

Table of Contents

How	to use this guide	4
1 In	ntroduction to DNA Master	6
1.1	DNA Master overview	6
1.2	Installation	6
1.3	Quick Start Guide	6
1.4	DNA Master program structure	6
1.5	Analysis programs running within DNA Master	7
1	.5.1 Glimmer	7
1	.5.2 GeneMark	8
1	.5.3 Aragorn	
1.6	Setting Preferences	9
1	.6.1 Set Default Translation Table	9
1	.6.2 Set color preferences	9
1	6.4 Set default values for BLAST searches	10 11
1	.6.5 Choose a default location for saving files	
1	.6.6 Finishing up your Preference settings	
1.7	Getting help	
1.8	Checking for updates	
<u>о</u> .	revisional Cluster againment of your phage	1 6
	rovisional cluster assignment of your phage	
2.1		15
2.2	BLASTing your sequence against the mycobacteriophage database	15
2.3	Cluster assignment	18
3 In	nporting your phage genome sequence into DNA Master	21
3.1	Overview	21
3.2	Where do I get my phage genome sequence from?	21
3.3	Importing your DNA sequence into DNA Master	22
3.4	Reverse-complementing your sequence	24
4 Pe	erforming and viewing a rapid automated annotation of your g	enome .25
4.1	Overview	
4.2	Running Auto-Annotate	
4.3	Saving vour file	
4.4	Looking at the output of your automated annotation	
4	.4.1 Viewing the documentation	
4	.4.2 Viewing features in the Feature Table	
4	.4.3 Viewing the sequence in the Sequence tab	
4	.4.4 Viewing ORFs in the Frames window	
4.5	Running the BLAST function	35
4.6	Re-opening an archived (saved) file	37

5	Gathering additional information for refining your annotation	
Į	5.1 Generating a six-frame translation	
Į	5.2 Generating a provisional genome map in DNA Master	42
ļ	5.3 Generating a graphical output from GeneMark	43
6	Using Phamerator to assist with annotation	47
(6.1 Overview	47
(6.2 Why Phamerator is useful to you at this stage of your annotation	47
(6.3 How did my genome get into Phamerator already?	
(6.4 Making Phamerator maps	
(6.5 Understanding and using the genome maps made by Phamerator	
(6.6 Viewing nucleotide sequence similarities in Phamerator	52
é	6.7 Other Phamerator features	54
(6.8 Saving Phamerator maps	55
7	Guiding Principles of Bacteriophage Genome Annotation	
•	7.1 Overview	
,	72 The Guiding Principles	57
8	Gene by gene: evaluating and improving your draft annotation	61
8	8.1 Overview	61
8	8.2 Button-pushing mechanics reserved for Section 9	61
8	8.3 Decision Tree for evaluating the draft annotation	61
8	8.4 Evaluating protein-coding gene calls	63
	8.4.1 Is the designation of this ORF as a gene well-supported?	63
	8.4.2 Is the called start site for this gene the best possible choice?	
	8.4.3 Is this gene part of a programmed translational frameshift?	
	8.4.4 Does this gene contain an intron?	72 72
	8.4.5 Does this gene wrap around the ends of the genome?	
(6.5 Checking gaps in the drait annotation for uncalled genes	כ / קר
) (8.6 Finding and renning tRNA and tmRNA genes	
č	8.7 Completing your annotation relinement	
9	The mechanics of making changes to your annotation	77
(9.1 Overview	77
(9.2 Making common changes to your annotation	77
	9.2.1 Deleting a gene	77
	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 valuating your annotation	
	9.5.5 Renumbering annotated features	18 כס
(9.5.4 Re-DLAST IIIg a gene	۵۲ ۵۳
	9.4 Making less common changes to your annotational frameshifts	
	9.7.1 Annotating programmed dalistational dallestilles	05 01
	943 Annotating wran-around genes	91 11
	7.1.5 minotating wrap around genes	

9.5 Predicting tRNA and tmRNA genes	
9.5.1 Running web-based Aragorn (version 1.2.28)	
9.5.2 Running tRNAscan-SE (version 1.21)	
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	
10.1 Overview	
10.2 Using bioinformatic tools to assign gene function	
10.2.1 BLASTP	
10.2.2 Conserved Domain Database	
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	
11.1 Merging overview	
11.2 Merging multiple approximations into a single file	111
11.2 Checking an annotation	116
12 Submitting final files for review and GenBank submission	119
12.1 Details of your final DNA Master (.dnam5) file	
12.2 Details of your author list	

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-phase 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.

👷 DNA Master		
File Manager Tools	Window	Help
New	Ctrl+N	rit kolphaist kolphais
Open		Archived DNA Master file Ctrl+Alt+O
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+Alt+A
Export Split Files		FastA Multiple Sequence File Ctrl+Alt+F
 Autoparse Preferences 	Chrl±D	Phylip Multiple Sequence File
Utilities	Cultr	Dequence from Accession Number

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

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





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, and BLAST. 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 provided in **Appendix I**, and are also available at http://phagesdb.org/DNAMaster/.

1.3 Quick Start Guide

Appendix II is the **DNA Master Quick Start Guide**, which you may find useful if 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. 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.

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**.

1.5 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.

1.5.1 Glimmer

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

1.5.2 GeneMark

GeneMark (version 2.5) provides a similar functionality to Glimmer and is used to predict genes. 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, GeneMark is heuristic, in that it learns from the 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.

In addition, a second GeneMark prediction is helpful for accurately identifying the genes in your phage genome. 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, we use *Mycobacterium smegmatis*.

GeneMark online is available at:

http://opal.biology.gatech.edu/GeneMark/genemark_prok_gms_plus.cgi

The 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" output.
- It provides a graphical output (as .pdf) of the gene predictions and coding potential. This is especially useful when you are determining gene starts.

1.5.3 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 fine-tune the tRNA calls. Please refer to **Section 9.5** when you evaluate your tRNAs in your genome.

1.6 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.

1.6.1 Set Default Translation Table

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.
- Click 'Apply'. Note that the dialog box remains open.



Figure 1.2

1.6.2 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

1.6.3 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'

	ок
Internet Local Settings Automation Phylogeny Timed Events Miscellaneous	<u> </u>
Directories GCG Export New Features Colors Translation Validation	Bevert
Default Start Codons	<u></u>
☑ Use ATG start codons	<u>C</u> ancel
🔽 Use GTG start codons	
☑ Use TTG start codons	Apply
✓ Translate Cued Start Codons as Methionine	
Default Stop Codons	
☑ Use TAA stop codons	
✓ Use TGA stop codons	
✓ Use TAG stop codons	

Figure 1.4

1.6.4 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.

DNA Master Preferences	
Internet Local Settings Automation Phylogeny Timed Events Miscellaneous	<u>o</u> k
Email REBase NCBI Gene Prediction Blast	<u>R</u> evert
✓ Include "ignore self" terms when initiating BLAST form	<u>C</u> ancel
Number of hits to request from server 100 🚖	Apply
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 hits, regardless of E-Value	
Minimum number of hits to save 1 🚖	
☑ Limit number of hits saved	
Maximum number of hits to save 100 🚖	

Figure 1.5

1.6.5 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	Settings Automation Phylogeny Timed Events Miscellaneous		<u>o</u> k
Directories GC0	GExport New Features Colors Translation Validation		Beuert
			<u>Heven</u>
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	
Download to	C:\Program Files\DNA Master\Downloads\	Browse	
Manuals	C:\Program Files\DNA Master\Doc\	Browse	
BLAST Libraries		Browse	

Figure 1.6

1.6.6 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.

1.7 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 Reference	s Seque	nce Docume	entation			
Sort By	Index 💌 🔳	Name	Start	Stop		Description Sequence Product Regions Blast Context	
Select Fe	atures Direct SQL	▶ 1	408	704		Name 1 GenelD	
-		2	743	1177			
Туре		3	1254	1577		Type CDS • GI	
Name	like	4	1567	2319		Start 408 Locus Tag DNAM1	
GenelD	=	5	2345	3565		Stop 704 Regions 1	
Locus	like	6	3592	4380		Length 297 🔲 Tag	
Start		7	4377	5294		Direction Forward	
Juan			5374	5449			
Length		8	5547	6899		Translation Table Undefined	
Regions	>	9	6896	7309		EC Number	
% GC	<	10	7306	8283		× v	
CAL		11	8479	10005		Product	
500		12	10002	11459		gp1	
EU#		13	11456	12469			
Product	like	14	12520	13020		Function	
Function	like	15	13055	14014		A	
FeatureID		16	14080	14268		v	
Tag		17	14276	14650		Notes	
⊤ag ∏ Hide I	Ignored Features			Þ	-	Original Glimmer call @bp 408 has strength 2.35; GeneMark	
Sele	Select áll Features Insett Delete Post Validate						
89 Features	Id Id						

Figure 1.7

To get a sense of how the help files work, go to:

Help → Help

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

1.8 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 (October 2011), the most up-to-date version of DNA Master is Version 5.22.5 Build 2338, dated 17 Oct 2011. You can find your current version by going to:

Help → About

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 M currently has only two members), whereas others are quite large (Cluster A has over 90 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 (ten currently) who have no close relatives, and therefore are classified as **Singletons**. Up-to-date cluster assignments 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 BLAST search of your phage sequence against a database of mycobacteriophage genomes provides a simple and quick approach to making a provisional cluster assignment.

2.2 BLASTing your sequence against the mycobacteriophage database

To BLAST your genome on phagesdb.org:

- Go to http://phagesdb.org/phages/
- 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 Phage 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 Mycophages as of 6.01.11 +
Enter sequence below in FASTA format >Etude accGACACTTCTCTCTGGAAATTCAGGCAAGAACATGAGGGGGGGTTAGCGCCCCTAAA ACCCCTGGTAGGAGGCTAAATCCTGGGTAGAGGACGTGGTAAGGACCCGGCCCGCAAGCCCTGG TGGGCGGTCTCGGGACGTCGCCGCCCGGCCTGGGCCGAAGGTTGCCGCCCAAGCCCTCG CAAACCGAAGACGCGCAGGAATACGCGGGCGCAGAGGCCCACGCCGCTGGTACGGAGGTTGCGGCGAGGTCCCGC CAAACCGAAGACGTTTGGACCAATCAGGGGATCCACGCCGCTGGTATCGGGAGCTTTGGAC CATGCGCAAGGCCGATGCGTACGGGAGCCCACGCCGCTGGCCTAATGGCCGCATTCGG
Or load it from disk Browse
Set subsequence: From To Clear sequence BLAST!
The query sequence is filtered for low complexity regions by default. Filter Low complexity Mask for lookup table only
Expect 10 ; Matrix BLOSUM62 ; Perform ungapped alignment
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.





Figure 2.2

To see which subjects 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).

		Score	в	
Sequen	es producing significant alignments:	(bits)	Value	
UPIE Co LeBron	mplete Sequence, 73784 bp including 10 bp 3' overhang (TC.	··· 1.3	14e+04 78e+04	0.0
JoeDir	Final Sequence, 74914 bp including 10 bp 3' overhang (TC.	1.1	69e+04	0.0
Microw	olf Final Sequence, 50864 bp including 10 bp 3' overhang,	402	2 0.0	
Vix Co	mplete Sequence, 50963 bp including 10 bp 3' overhang (CGG,	399	8 0.0	
Methus	alah Complete Sequence, 50891 bp including 10 bp 3' overha.	399	8 0.0	
JHC117	Final Sequence, 50877 bp including 10 bp 3' overhang. Clu	399	8 0.0	
Byz2	rindr bequence, soon op increating to op 5 overhang, ere	399	8 0.0	
Faith1	Complete Sequence, 75960 bp including 10 bp 3' overhang (138	8 0.0	
Rumpel	tiltskin Complete Sequence, 69279 bp including 10 bp 3' o	136	4 0.0	
HelDan	Complete Sequence, 50364 hp including 10 hp 3' overhang (35	3 1e-95	
Rockst	ar Complete Sequence, 47780 bp including 10 bp 3' overhang	23	2 30-59	
Peacher	n complete begaence, 47700 bp including 10 bp 5 overhang.	21	2 26-53	
TiroTh	, eta9 Complete Sequence, 51367 bp including 10 bp 3' overha.	20	4 6e=51	
MeeZee	Complete Sequence, 51368 bp including 10 bp 3' overhang (.	20	4 6e-51	
Eagle	compress bequence, since by including to by 5 oferhang (20	4 6e-51	
LHTSCC	Complete Sequence (51813bp, including 10bp 3' overhang; C.	19	6 1e-48	
Shaka (Complete Sequence, 51369 bp including 10 bp 3' overhang (C.	18	8 4e-46	
Twister	Complete Sequence, 51094 bp including 10 bp 3' overhang	14	9 3e-34	
George	Final Sequence, 51578 bp including 10 bp 3' overhang, Clu.	13	7 1e-30	
Benedi	t Complete Sequence, 51083 bp including 10 bp 3' overhang.	13	7 1e-30	
Airmid	Complete Sequence, 51241 bp including 10 bp 3' overhang (.	13	7 1e-30	
Theia (Complete Sequence, 51543 bp including 10 bp 3' overhang (C.	12	9 3e-28	
Cuco Co	omplete Sequence, 50965 bp including 10 bp 3' overhang (CG.	12	9 3e-28	
Bxb1		12	3 2e-26	
Violet	Complete Sequence, 52481 bp including 10 bp 3' overhang (.	12	1 7e-26	
Switzer		12	1 7e-26	
Pari Co	mplete Sequence, 50614 bp including 10 bp 3' overhang (CG.	12	1 7e-26	
KBG		12	1 7e-26	
Doom F:	inal Sequence, 51421 bp including 10 bp 3' overhang (CGGAT.	12	1 7e-26	
Dreambo	oat Complete Sequence, 51083 bp including 10 bp 3' overhan.	· · · <u>11</u>	7 1e-24	
Jasper		11	<u>5</u> 4e-24	
Wile Co	omplete Sequence, 51308 bp including 10 bp 3' overhang (CG.	· · · <u>11</u>	<u>3</u> 2e-23	
U2		11	<u>3</u> 2e-23	
Solon		11	<u>3</u> 2e-23	
SkiPol		11	3 2e-23	
RidgeCl	3 Complete Sequence, 50844 bp including 10 bp 3' overhang .	11	3 2e-23	
Perseu	Complete Sequence, 53142 bp including 10 bp 3' overhang	11	<u>3</u> 2e-23	
MrGord	Complete Sequence, 50988 bp including 10 bp 3'overhang (.	11	<u>3</u> 2e-23	
Lockle	1	11	<u>3</u> 2e-23	
Lesedi		11	<u>3</u> 2e-23	
KSSJEB		11	3 2e-23	
JC27		11	3 2e-23	

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 si	now defline and	scores. Click	to show alignm	ents			
		Color Ke	y for Align	ent Scores			
	<40	40-50	50-80	80-200	>=200		
le111							
0	10K	20K	30K	40K	50K	,	
						K1	
						K2 K3 K4	1. K5
_			÷ :	-			.,
		= :	<u>.</u>				
		-					
			-				
equences pro	ducing sign	ificant a	lignments:		Sc (b	ore E oits) Value	
dephagia Fin EEST Complete	al Sequence, Sequence,	, 59646 b 59906 bp 59749 bp	p including including	g 11 bp 3' ov 11 bp 3' ove 11 bp 3' over	erhang, rhang (C hang (CT	4.766e+04 3.606e+04 3.010e+04	0. 0. 0.
arrelRoll Con ngelica Fina naya Complete	mplete Sequ L Sequence e Sequence,	ence, 596 60835 bp	72 bp including, including	ding 11 bp 3 g 11 bp 3' ov	erhang (3.008e+04 2.906e+04 2.080e+04	0. 0. 0.
imD Final Se ixie Final Se	equence, 59 equence, 61	798 bp, 1 147 bp in	1 bp 3' ov cluding 11	erhang (CTCGT bp 3' overha	AGGCAT),	1.826e+04 1193 0.0	0.

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:

- http://phagesdb.org/phages/
- 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			
File Tools Window H	<u>t</u> elp		
New	Ctrl+N	rit prophecit	Forgenery 1
Open	•	Archived DNA Master file	Ctrl+Alt+O
Import Close Close All	► Ctrl+W	GCG-Formatted File GenBank-Formatted File	Ctrl+0
Save as DNAMS File Export GCG file Export Apollo XML File Export 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
 Autoparse Preferences 	Ctrl+P	FastA Multiple Sequence File Phylip Multiple Sequence File	Ctrl+Alt+F
Utilities	•	Sequence from Accession Number	r
Quit	Ctrl+Q	and the second second	1 Hand

File → Open → 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.



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.

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**.



Figure 3.3

• 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**.

Extract	ed from FastA Library Bongo.fasta	_	
Overview	Features References Sequence Documentation		
Feature	F > Raw ▶ BLASTN BLASTP A	dd Feature	73.5 63.6 57.6
1	ATCAGCCCCTCTCTCCCCGGCAAGTTTGAGTGCCAAGTATCCAATTCCAATTTGCTGGTATGATGGAGGGGGGTT	TGGGGGGGA	3G 🔺
86	CCTTTTGAAACCCCAACTGCTGGTCGAAGCCTTGAGAAGCTTCCCATCTGCGGTTCGGCTCCTGGTTGAGGGGTT	TGCTGTATO	JT 🔲
171	AAAAAAACCCCAGGTCGAGTGCGACAAAACTGCAGGTCGCACCTCCCTGGGGTTGGCCGGGACGAGGTCAGGCTC	GCGTGAGCO	JT
256	CAGGTAGGTCGTCTTCACGTACTTGGTGTTGGGCGTGACGGTCATGTAGCCGGTGGCCCCTTCGCCCTCGATGAT	CTCGTGGT	3A 🛛
341	ATTGGCATGCCGCAGTTGGCGTCTGCGCATAGCGCGTAGGCCCCTTCCTCCCAGCCGACCACTTTGTACATGTAC	TCGCCGGGG	3G
426	CGATGTCGGTGCCGACGCGGTACATCGCGGACCTGGCCTCCGCTGCGGGGGCGAGTGCGATGGCGGCGCAGATTC	CTGCGATGA	NG
511	CCCCCTCACTCGCTCCCCCCCCCCCCCCCCCCCCCCCCC	GCGGCCGT	3G
596	TGTGACGGCGCGTGGCTGGTGTAGTACGCACTATCGGTGGCGTCAGGGTAAGCCGATGGATCGGCCGGTGCG	GACGTATC	AG
681	CGATATGAAGAGGGGCCGGATGGCAGCTGTGTGATTGCCGCTATCCGACCCCGTCGAAGGCTCGGGCGGG	TTCGCTAA	3G
766	AGCTGTCTCCCCCAGGAGCTTGGTCTCCGGAAGGGGGGATGATCCACTCCGAGCGGACTGCGTTCATCCGCTC	GTAGAGAT	3G 🛛
851	AGGACGGCGTCGAGGACTGGGTACGAGTCGCCTGCTTCGTCGGCGGGCCCAGTGACGATGACGGTGCCGTGGAG	GAGTCTCG	3C
936	CTTCCATCTCGGGCTGGAGTTTCCACCAGAGGTAGGTTGCCATGGGGTTCGTTGGCAAACCGAGGAGCTTGCCTT	CTTCGTTG	CA
1021	CCAGAGGGTGGCGTGCTCGGTGGTGACCGCCTCGATGTGACCACCGACGAGTTTGCGCATGGCGGGTAAGGCCTC	GTCGATCA	3G 🛛
1106	CGCACTTCGCGCGTGCCTTCTGGCGTGATGATTAACGCGCGCG	GTCGTGGT	CA
1191	TGAAATCCTCCTTCGTCGTCGTGGGTTGCGCCCAGTCACTTCTATAAGCCTTAGTTCACTCCTCTCGGTTGGAG7	TAAGGCCCC	CG
1276	CCACCTTTGAAAAGGTAGCCGTCAGCGTAGCTGTATCGACTGCCTTTTGTCTCCCCCGTGCGGGAGTGAATCCGG	AGGAATCTO	ст 📗
1361	CAGTTGAACACCTAATACAGATGAGGCCCTGGATGTAGGCCACGCCCTCCTTCTTAACTGTGGTGTAGTCCCAGG	CGTGACCG	AA
1446	AACACGGCACATCCGGTGTTGCGGCTTAACAACTCGGTCTTCTGTGTCGCTCATTTATGTGCGTTCTTCCAAGCT	TCTATGACO	3G
1531	CCGGTTTGATGTACCCACGCTTGGAGACTGTGAACCCGTTTTTGCCCGCCC	CCTTTGTG	AG
1616	GGCCGTGGTGGGATGGGCCGGTGGCTGCTCGCTTACAGCCGGTGTAACGCTCACCTCTTTCTGCGTCCTCTTTGG	CCACTTCT	гс
1701	TTCTCGTGAGCAACGCTTAGCCACTTATCGAGTGTGCGCTCTAGCTCATCTGCGTGTTCTTGCGTTAGGTCGATC	TCATACAAG	JA
1786	AGCCTCGGTACCCGATGACTTCGGTTACCGAGGCTTCTTTGGATGGA	CTTGCCGTO	3C
1871	CACTTCTTTCCCTTTCCCCTGCGAGGAGAACACTATCGAAGTGGGGCTAATTTTCGATCACATCAGGTGGGGCGA	GTTGCCTAG	3T
1956	CGGCGTGGTCAATTATCGCCATGAACACGAAGGCGATGAGGGTGATGACCGCCCACATGCCCGCTCCTACTGCCA	GATTGATG	3C
	TTCTGGTGAGACCGTAGAGGCGAGTATGGATCCGAGGAGGGCCGTAACCCAGTAGGTCGCGCTCAGCCAGC	GACTTGGGT	rg 🔳
H H I	⊕ ⊖ ▶ ▶ 1 - 50000 Position : 25073 I Controls >> Map I Map >> Controls	-	
0 Features	Live	80228	32

Figure 3.4

• Until you run an automated annotation in the next section, the tabs for **[Features]**, **[References]**, and **[Documentation]** are largely empty. We'll revisit these later.

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 \rightarrow Convert \rightarrow Complement
- A dialog box will open that asks if you want to convert XXXXX bp to $5' \rightarrow 3'$. Click 'Yes'.
- 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 80% 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:

DNA [DNA N	laster													
File	Edit	Genome	DNA	Tools	Window	Help	_								
12	ir.	Add to	Datab	ase			P.C.	DHO		1. 3×3×3	nd i s			1000.00 A	Anna an An
	12	Analyz	e all ge	ene starl	ts		180	🔀 Extrac	cted	from	FastA	Lib	rary	Etude	.fas
252	1.50	Annota	ation			•	Au	to-Annotate	•	s	Referen	nces	Sequ	uence	Do
192	N.C.	BIAST	able All Cor				ке	JOIL DY	icion IIIIUCA			┛	Nam	e	Star
den.	1 log	Coding	1 Capac	itv			1 de la	Select Fe	atures		et SQL	ı D			
292	14	Codon	Spacin	g			117	T	:.		_				
1911	fin-	Featur	es			•	20	туре	15						
134	11	Gene (Drienta	tion	Ctrl+A	lt+G	с <u>с</u> ,	Name	іке			-1			
13	1	Karlin's	; Dinucl	eotides	Challe Al	IF 1 1	3	GenelD	=			-1			
	192	Mutati	onal Biz	35 95	CUITA	IL+L	170	Locus	like						
	22	Nuclea	itide Ph	asing	Ctrl+A	lt+H	12	Start	>						
2.07		Origin	predict	ion -			1740	Length	>						
24	- Part	Predict	: stop				1	Regions	>						
	1 mars	Profile			Ctrl+A	lt+P	20	% GC	<						
	1.	Rearra DNASe	ange					CAI	>	í—		1			
62		Run Le	enath			,	2	EC#	like	i—		-1			
	1	Six fra	me trai	nslation	Ctrl+A	lt+T	1	Product	lika			-1			
1	-	Skewe	d sequ	ences	Ctrl+A	lt+K	377	Function	liko	-		-1			
8.83	15	trna f	Profile				损益	Function		-		-1			
2.2.3	The.	al est		They.	R. Ca.		12m	FeatureID	' <mark>=</mark>	<u> </u>		-1			
123	in	Store 1		PP 3	the state		in	Tag	like						
		and a contract		12	test a		12	🔲 Hide I	Ignore	ed Feat	ures		<		
	1.50	er an		sit	and the		-30	Sele	nt All	Feature	₽«		Insert	De	elete
12	NE.	and the		1	· · ·		2	M		۹Þ		1	- 1499	8	
and the	1 Section	34-		1 Areas	4-	all and	1 State								
22	120	4		1.	4		14	0 Features		Live					

Genome → Annotation → Auto-Annotate

Figure 4.1

• An Auto-Annotate dialog box will open. We recommend that you use the settings shown in **Figure 4.2**.

Auto-Annotate						
Gene Calling	BLAST Searches					
Document tRNAs found with Aragorn	Perform BLAST search on nr database					
Document ORFs found with	Save hits with E-values less than 10E- 3 🚖					
C Glimmer 3.02 analysis	Regardless of their E-values save 1					
C GeneMark HMM analysis						
 Both analyses, combining them as follows: 	Limit number of hits to be saved to 100 ÷					
Favor Glimmer calls	Remember these settings					
C Favor GeneMark calls						
Exclude genes called by only one method						
Exclude genes called only by second method						
Maximum wait time 1 minute 💌						
Examine and learn gene features	No BLAST scheduled					
Autosave						
✓ Autosave DNAM5 file as Extracted from FastA Libra	ry Bongo.fasta_Annotated.dnam5					
🔲 Export alternative start codons 👘 Expo	ort a Profile of features					
🔲 Export a summary of BLAST hits 👘 Expo	ort a list of Gray Holes and gene overlaps					
🔲 Export a GC-Content map 👘 Expo	ort a Six-Frame map					
🔲 Export Frames map	ort an ORF map					
Height 200 🜩 Width 3209 🜩 Sca	le 1/2" / kb 💌					
Tiers 4 🗲 Size	e to 25 🚔 kb per tier					
Export Directory Browse						
Analueie						
Current Replicon All Analyses Appot	ate					
C All Replicons No Analyses	ale					
Status : Idle						

Figure 4.2

Click the '**Annotate**' button to launch the automated annotation. (Click '**Yes**' when prompted to "Erase features?")

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, and since Glimmer recognizes TTG as a start codon, we prefer to favor its calls. 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, the Notes will contain only one program's call.

• 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.

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 FastA	Library Bongo.fasta		
Overview Features Refer	ences Sequence Document	ation	
Genome Organism: GenomeID:0 Length:0 Replicons:1 Features:149	PrototypeID : 0 Date Created : 10/6/2011 Genome Date : 10/6/2011 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
CATTable : <u>Directory : DMTemp_40822.6</u> Replicon Replicon	35846565\	- -	
NCBI Date : 10/6/2011 Topology : Linear Length : 80228	<u>NCBI Genome :</u> <u>Accession :</u> <u>Version :</u>		Notes on the Replicon
Features : 149 ORFs : 133 RNAs : 16	<u>GI :</u> Status : Current Deletion : Vulnerable		<u></u>
GC Content : 0.00 ORF GC : 55.4 ± 19.4 GC3 : 71.7 ± 25.8 CAL : 0.4954 + 0.2042	Origin : O Terminus : O		
Translation Table : Unspecifi	ed: default to standard code	sition : 36337 🔽 Controls >> N	lap 🔽 Map>> Controls
149 Features Live	25	26 27 2 29 30 33 34 87	80228 🛃 ?

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 P button to zoom in and the P 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 complement (238-450)". This means the first feature is a protein-coding sequence (CDS) transcribed right to left and located at coordinates 238 – 450.



Figure 4.4

This "complement" orientation is worth thinking about for a moment. It means that the first base of the first codon of this predicted gene is at position 450, while the last base of the termination codon is at position 238.

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 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**).

DNA Extrac	ted from	FastA Librar	y Bongo.fas	ita								<u>_ ×</u>
Overview	Features	References	Sequence	Documer	ntation							
Sort By	Index		Name	Start	Stop		Descriptio	on Sequer	nce Prod	uct Regions	Blast 0	Context
Select Fe	atures D	irect SQL	1	238	450		Name	1		GenelD		
Turne		i	2	760	1110		Turne	lene		CI.		
Type		<u>`</u>	3	1496	1873		Type	lus	<u> </u>	<u>ui</u>		
Name	like		4	1951	2136		Start		238	Locus Tag	DNAM1	
GenelD	=		5	2133	2366		Stop		450	Regions		1
Locus	like		6	2363	2575		Length	213		Tag		
Start		L	7	2572	2838		Direction	Reverse		2		
L U		L	8	2826	3134		- Lucion		<u> </u>			
Length			9	3131	3418		I ranslatio	on Lable U	ndefined			
Regions	>		10	3415	3645		EC Numb)er				
% GC	<		11	3645	3944		J					
CAL			12	4298	4693		Product					
		I	13	4719	6263		gp1					
EU#	пке		14	6260	7882							-
Product	like		15	7882	9042		Function					
Function	like		16	9093	9809							<u> </u>
Featurel			17	9854	11017							-
Taa	like [18	11031	11513		Notes					<u> </u>
ray	ince					-	Original (GeneMark o	all @bp 45:	50		<u> </u>
🗌 Hide	Ignored Fe	atures	•		<u> </u>							
Sel	ect All Feat		Insert De	elete Pos	t Valida	ite						Y
4 4 4		▶ ▶ ▶ 50	01 - 30000	P	osition : 6	272		Controls >> N	Map 🔽 M	ap >> Controls		
13 1	4 15		2 21 22	224			25	N	26 27	2 29	30	8 8 33 8
149 Featu	res Li	ve	a of of 2121					A		257	80228	a ?

Figure 4.5

The central box shows each gene's **Name**, **Start** and **Stop** coordinates, and—if you use the scroll bar to move to the right—the gene **Length**. You can select any gene by clicking on it. Gene "1" is selected in the example above, as indicated by the small black triangle next to it.

CRITICAL NOTE ABOUT A POTENTIALLY CONFUSING PROGRAM FEATURE

DNA Master **ALWAYS** lists the leftmost genomic coordinate as the **Start** position and the rightmost as the **Stop** position, regardless of a gene's direction of transcription. This means that genes that are transcribed from right to left, and thus have start codons at their rightmost coordinate, will still have their rightmost coordinate in the "Stop" column. Don't let this confuse you!

It is helpful to think of those column headings as "**Left**" and "**Right**" rather than Start and Stop.

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.

The **Notes** for gene 1, shown above, indicate that Glimmer called the start at position 450. 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 below.

🚟 Extracted from FastA Library Bongo.fasta	_ 🗆 🗙
Overview Features References Sequence Documentation	
Search Recreate Parse	Predict Genes
CDS 4042 - 4365 /gene="12" /product="gp12" /locus tag="DNAM111" /note="'Original Glimmer call @bp 4042 has strength 1.04 ** not called by GeneMark''	1
CDS 4475 - 4693 /gene="13" /product="gp13" /locus tag="DNAM121" /note="Driginal Glimmer call @bp 4475 has strength 8.33; GeneMark calls start at 4298"	
CDS 4719 - 6263 /gene="14" /product="gp14" /locus tag="DNAM131" /note="'Driginal Glimmer call @bp 4719 has strength 13.82"	
CDS 6260 - 7882 /gene="15" /product="gp15" /locus tag="DNAM141" /note="0riginal Glimmer call @bp 6260 has strength 18.22"	
CDS 7882 - 9042 /gene="16" /product="gp16"	-
Image: Weight of the second secon	

Figure 4.6

In the next example, gene 12 was predicted by Glimmer, but was "not called by GeneMark".

For gene 13, the assigned start is 4475 as called by Glimmer, but there is a note that "GeneMark calls start at 4298".

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.

📴 Extracted from FastA Library	r Bongo.fasta	X
Overview Features References	Sequence Documentation	
Feature	F № Raw ► BLASTN BLASTP Add Feature 2000	.0 .0
Move R ORF 238 (1) R ORF 760 (2) R ORF 1496 (3) R OBE 1951 (4)	est Match at 3'End Overall OverallMatch Oligo	
R ORF 2133 (5) 1 R ORF 2363 (6) 88 R ORF 2572 (7) 137 D ORF 2572 (7)	ackwaras GCAAGTTTGAGTGCCAAGTATCCCAATTCCCAATTTGCTGGTATGATGGAGGGGGGTTTGGGGGGGG	-
262 R ORF 3131 (9) 349 R ORF 3415 (10) 436 R ORF 3645 (11)	TACTTGGTGTGTGGCGTGACGTCATGTAGCCGGTGGCCCCTTCGCCCTCGATGATCGTCGTGGCG TACTTGGTGTGGCGTGACGGTCATGTAGCCGGTGGCCCCTTCGCCCTCGATGATCGTCGTGGGC CGTCTGCGCATAGCCGCGTGGCGCCCCTTCCCCCGACCGCCACCTTGTACATGTAGCCGCGGGG TTACATCGCGGACCTGGCCTCCGCCGCGGCGGACTGCGGATGGCGCGCGATTGGTGCCGCGGGGG TTACATCGCGGACCTGGCCCCCCCCCC	
500 F ORF 4042 (12) 523 F ORF 4475 (13) 610 F ORF 4719 (14) 697 F ORF 6260 (15)	ATGAGCCTTCACAGTAGAGGAGAGGGCCGTACATGGCCGTTGCAAGGTTGGAGGTTGCGGCGTCG FGTAGTACCCACTATCGGTGGTGGCGTCAGGGGAAGCCGATCGGCCGGTCGCGGACGTATCAG ATGGCAGCTGTGTGATTGCCGCTATCCGACCCCGTCGAAGGCTCGGGCGGG	
784 F ORF 7882 (16) 871 F ORF 9093 (17) 958 F ORF 9854 (18) F ORF 11021 (19)	GCTTGGTCTCCGGAAGGGGGGATGATCCACTCCGAGCGGACTGCGTTCATCCGCTCGTAGAGATCG GGTACGAGTCGCCTGCTTCGTCGGCGGGCCCCAGTGACGATGACGGTGCCGTGGAGTGAGT	
1045 F ORF 11510 (20) 1132 F ORF 11871 (21) 1219 F ORF 12285 (22)	FTGGTGACCGCCTCGATGTGACCACCGACGAGTTTGCGCATGGCGGGTAAGGCCTGGTCGATCACG CTGGCGTGATGATTAACGGGCGTATTGACGGTGCTGCCACGGTTGCACTGTAGCCCTGTCGTGGGC CGTGGGTTGCGCCCAGTCACTTCTATAAGCCTTTGTTCACTCCTCTCGGTTGGAGTTAAGGCCCTG CGTGGGTTGCGCCCAGTCACTTCGTTGATGAGCCCCGTTGGAGTTGAGGCCCGGTGGAGCACGCCGGTGGAGCACGCCGTGGGCCGGTGGAGCACGCCGGTGGAGCACGCCGGTGGGGCCGGTGGGGCCGGTGGGGCCGGGCGGG	
1306 F ORF 12816 [23] 1393 F ORF 13804 (24) 1480 F ORF 14136 (25) 1567 CCCCTTTCATCTACCC	ATGAGGCCCTGGATGTAGCGCCCCCCCCCTCCTTCTTAACTGTGGTGTGCCCAGGCGCGCCGCGGCGGCCCG GCGGCCTTAACAACTCGGTCTTCTGTGTGTGCCGCCCATTTATGTGCGTCCTCCAGGCTGACGGCGCGCCCCGCCCTTTTCTCTCAGGCGCGCCCCCCCC	
1654 GGCGTGGTGGGGATGG 1741 TTCTCGTGAGGAACGC	ACCENTION ACCENTION ACCESSION TO THE CONCENT A CONSIGNATION OF A	•
Image: Second	50000 Position : 20785 Controls >> Map V Map >> Controls	

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 reading frames, with each row representing one reading frame. Full-row-height vertical lines represent inframe 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 six 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 above. Genes in forward reading frames are green, those in reverse reading frames are red, and tRNAs are blue.

Next click on the frames window within the box that contains a highlighted gene.



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

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

DNA C	hoose ORI	⁼ start	and the second		
Start: Selec	s : 38 cted : 1	ORF Start : ORF Stop ORF Length	20846 Cdn 1 Cdn2 C : 23278 5' End 81.8 78.8 44 :: 2433 3' End 51.2 80.7 61	dn3 Length 3.5 99 - 3.4 2332	<u>D</u> ocument
	Shine D	algarno	Sequence of the Region	Start Start	ORF 🔷
#	Score	Space	Upstream of the Start	Codon Position	Length
1	567	8	CACGCAGGCGGGGGGGGCGGTAA	GTG 20846	2433
2	294	7	TCTGCCCAAGCTCGACCCGGCG	ATG 20945	2334
з	195	9	TCAGATCCCGCTCGGCTCGATC	GTG 21206	2073
4	465	9	CGTGCAGGAGTCGCCAAACCTG	TTG 21227	2052
5	378	7	TTTCGTCCGCTGGGCGGGCCTG	GTG 21434	1845
6	420	8	CGGGTCGATCGGTATCGGGCTG	ATG 21467	1812
7	441	8	CGGCGACGCTGGCGAGCAGCGT	GTG 21497	1782
8	462	7	GAAGGTCGGCGGCCAGTACGTC	GTG 21569	1710
9	315	7	GCCCGACACCGGCGTCGACAGT	GTG 21593	1686
10	294	7	CCTGCAAAACCAATGGCAGGCG	ATG 21791	1488
11	462	8	CGTCGGCGGCGCAATCGAGGAC	TTG 21836	1443
12	450	9	AATCGAGGACTTGTGGAATCGG	TTG 21848	1431
13	441	7	CTTGTGGAATCGGTTGCTGCAC	TTG 21857	1422
					🌃 🖹 ?

Figure 4.11

This window lists 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 a score for the Shine-Dalgarno sequence (higher is better). One line's text may be red, and this is because that row corresponds to the start site immediately upstream of where you clicked in the Frames window.

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 Etude genome.



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 DNA Master to run uninterrupted, you should run the BLAST function. To do so (we use phage Bongo below), go to:

Genome → BLAST All Genes

• In the dialog box, we recommend that you use the settings shown in **Figure 4.13**.

Genome BLAST of Extracted from FastA Library Bongo.fasta	
Retrieve	
Number of hits to request from server	Ignore Definitions including the following terms
	A
Number of hits expected	
Saving Results Locally	
Save hits with E-Values smaller than 10E- 3	
Regardless of E-Value, save at least 1 📑 Hits	
✓ Limit number of hits saved to	
Execute BLAST Query	
BLAST all protein-coding genes Skip genes already analyzed Skip genes already analyzed	
Clear previous BLAST results	
BLAST all genes	
Blast All Cancel	
	<u> </u>
	<u>?</u>

Figure 4.13

- Click on 'Blast All'.
- 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!

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.

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

Retrieve Ignore Definitions including the following terms Number of hits to request from server 10 • Number of hits expected 10 • Saving Results Locally • Save hits with E-Values smaller than 10E- • Image: Regardless of E-Value, save at least 1 • Image: Regardless of E-Value, save at least 1 • Image: Regardless of E-Value, save at least 1 • Image: Regardless of E-Value, save at least 1 • Image: Regardless of E-Value, save at least 1 • Image: Regardless of E-Value, save at least 1 • Image: Regardless of E-Value, save at least 1 • Image: Regardless of E-Value, save at least 1 • Image: Regardless of E-Value, save at least 1 • Image: Regardless of E-Value, save at least 1 • Image: Regardless of E-Value, save at least 1 • Image: Regardless of E-Value, save at least 100 • Results Locally Skip genes already analyzed BLAST all protein-coding genes Image: Regardless are coding genes Blast All Cencel Sent 30 requests (150 total) to the QBlast server Retriving completed results	🔀 Genome BLAST of Extracted from FastA Library Bongo.fasta	
Communicating with the Server Number of hits to request from server Number of hits expected Saving Results Locally Save hits with E-Values smaller than 10E- Regardless of E-Value, save at least Regardless of E-Value, save at least Recute BLAST all protein-coding genes BLAST all protein-coding genes BLAST all protein-coding genes BLAST all genes BLAST all genes BLAST all genes BLAST all genes Sent 30 requests (150 total) to the QBlast server Retrieving completed results Sent 2 requests (150 total) to the QBlast server All requests have now been sent to the QBlast server All requests have now been sent to the QBlast server All requests have now been sent to the QBlast server All requests have now been sent to the QBlast server All requests have now been sent to the QBlast server All requests have now been sent to the QBlast server All requests have now been sent to the QBlast server Catinum for 0 results Genome BLAST has been completed	Retrieve	
Number of hits to request from server 10 Image: Contract of the server of the ser	Communicating with the Server	Japone Definitions including the following terms
Number of hits expected 10 Saving Results Locally Save hits with E-Values smaller than 10E- 3 Image: Regardless of E-Value, save at least 1 Image: Regardless of E-Value, save at least Image: Regardless of E-Value, save at least genes Image: Regardless of E-Value, save at least genes <tr< td=""><td>Number of hits to request from server 100 🚖</td><td>Ignore Deminions including the following terms</td></tr<>	Number of hits to request from server 100 🚖	Ignore Deminions including the following terms
Saving Results Locally Save hits with E-Values smaller than 10E- Image: Regardless of E-Value, save at least Image: Regardless of E-Value, save at least of genes Image: Regardless of E-Value, save at least of genes Image: Regardless of E-Value, save at least of genes Image: Regardless of E-Value, save at least of genes Image: Regardless of E-Value, save at least of genes Image: Regardless of E-Value, save at least server Retrieving completed results Continuing to send Sent 2 requests (152 total) to the QBlast server Retrieving completed results <	Number of hits expected	<u> </u>
Save hits with E-Values smaller than 10E-3 Image: Regardless of E-Value, save at least Image: Regardless of E-Value, save at least of genes Image: Regardless of E-Value, save at least server Retrieving completed results Continuing to send Sent 2 requests (152 total) to the QBlast server Retrieving completed results Waiting for 0 results Genome BLAST has been completed	Saving Results Locally	
▼ Regardless of E-Value, save at least 1 Hits ▼ Limit number of hits saved to 100 Hits ▼ BLAST dll protein-coding genes ○ Skip genes already analyzed ○ BLAST all protein-coding genes ○ Skip genes already analyzed ○ BLAST all genes ○ Clear previous BLAST results ○ BLAST all genes ○ Clear previous BLAST results ○ BLAST all genes ○ Clear previous BLAST results ○ Blast All Cancel ○ Sent 30 requests (150 total) to the QBlast server Retrieving completed results ○ Continuing to send Sent 21 cequests (152 total) to the QBlast server Retrieving completed results ○ Vaiting for 0 results Genome BLAST has been completed ▼ ▼	Save hits with E-Values smaller than 10E- 3	
Limit number of hits saved to 100 Hits Execute BLAST Query BLAST all protein-coding genes BLAST all RNA-encoding genes BLAST all RNA-encoding genes BLAST all genes Sent 30 requests (150 total) to the QBlast server Retrieving completed results Continuing to send Sent 2 requests (152 total) to the QBlast server Retrieving completed results Continuing to send Sent 2 requests (152 total) to the QBlast server Retrieving completed results Waiting for 0 results Genome BLAST has been completed	✓ Regardless of E-Value, save at least 1	
Execute BLAST query BLAST all protein-coding genes BLAST all protein-coding genes BLAST all RNA-encoding genes Clear previous BLAST results Clear previous BLAST results BLAST all genes Sent 30 requests (150 total) to the QBlast server Retrieving completed results Sent 2 requests (152 total) to the QBlast server All requests have now been sent to the QBlast server Retrieving completed results Waiting for 0 results Genome BLAST has been completed	✓ Limit number of hits saved to 100	
BLAST all protein-coding genes BLAST all RNA-encoding genes BLAST all RNA-encoding genes BLAST all RNA-encoding genes Clear previous BLAST results Clear previous BLAST results BLAST all genes Sent 30 requests (150 total) to the QBlast server Retrieving completed results Sent 2 requests have now been sent to the QBlast server All requests have now been sent to the QBlast server Retrieving completed results Waiting for 0 results Genome BLAST has been completed	Execute BLAST Query	
C BLAST all RNA-encoding genes C BLAST pre-selected set of genes BLAST all genes BLAST all genes BLAST all genes Blast All Cancel Sent 30 requests (150 total) to the QBlast server Retrieving completed results Sent 2 requests (152 total) to the QBlast server All requests have now been sent to the QBlast server All requests have now been sent to the QBlast server Betrieving completed results Wating for 0 results Genome BLAST has been completed	○ BLAST all protein-coding genes ✓ Skip genes already analyzed	
BLAST all genes Sent 30 requests (150 total) to the QBlast server Retrieving completed results Sent 2 requests (152 total) to the QBlast server All requests have now been sent to the QBlast server Retrieving completed results Wating for 0 results. Genome BLAST has been completed T	C BLAST all RNA-encoding genes Clear previous BLAST results	
Blast All Cancel	 BLAST pre-selected set or genes BLAST all genes 	
Blast All Cancel		
Sent 30 requests (150 total) to the QBlast server Retrieving completed results Sent 2 requests (152 total) to the QBlast server All requests (152 total) to the QBlast server All requests have now been sent to the QBlast server Retrieving completed results Waiting for 0 results Genome BLAST has been completed	Blast All Cancel	•
Hetreving completed results Continuing to send Sent 2 requests (152 total) to the QBlast server All requests have now been sent to the QBlast server Retrieving completed results Waiting for 0 results Genome BLAST has been completed	Sent 30 requests (150 total) to the QBlast server	_
Sent 2 requests (152 total) to the QBlast server All requests have now been sent to the QBlast server Retrieving completed results Waiting for 0 results Genome BLAST has been completed	Retrieving completed results	
All requests have now been sent to the QBlast server Retrieving completed results Waiting for 0 results Genome BLAST has been completed	Sent 2 requests (152 total) to the QBlast server	
Retrieving completed results Waiting for 0 results Genome BLAST has been completed	All requests have now been sent to the QBIast server	
Genome BLAST has been completed	Hetrieving completed results	
	Genome BLAST has been completed	
		-
	1	

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.

DNA Extrac	ted fror	n FastA Libra	ary Bong	o.fasta				×
Overview	Feature	es Reference	s Seque	ence Doc	umentation	1		
Sort By	Index	• •	Name	Start	Stop	Le	Description Sequence Product Regions Blast Contex	t
Select Fe	eatures	Direct SQL	1	238	450	21	Score Target Description	
Turne			<u>▶</u> 2	760	1110	35	▶ 355 gp2 [Mycobacterium phage Rev]	
туре		<u> </u>	3	1496	1873	37	214 gp121 [Mycobacterium phage LeBron]	
Name	like		4	1951	2136	18	209 gp124 [Mycobacterium phage JoeDirt]	
GenelD	=		5	2133	2366	23	177 hypothetical protein MAB 1818 [Mycobacterium	
Locus	like		6	2363	2482	12	164 op128 [Mycobacterium phage Eaith1]	
Start			7	2572	2838	26		-
Julia			8	2826	3134	30	⊢ BLAST Hit	50
Length			9	3131	3418	28	Accession AEK09914 Export	
Regions	>		10	3415	3645	23	GI 339784206Export A	
% GC	<		11	3645	3944	30	Length 140 Delete	
CAL			12	4042	4365	32	Max Score 355 Date 10/7/2011 Delete A	
			13	4475	4693	21	High-Scoring Pairs (HSP)	
EC#	like		14	4719	6263	15	HSP Data Alignment	
Product	like		15	6260	7882	16	Bit Score 141.4 Identities 69	
Function	like		16	7882	9042	11	Score 355 %Identity 59.48	
Festural			17	9093	9809	71	E-Value 6.0E-42 Positives 85	
			18	9854	11017	11	Length 116 %Similarity /5.22	
lag	like		F			-	Aligned ou 7 % Graps 5	
🔲 Hide	Ignored F	Features	•				Target 28-140	
Sele	ect All Fe.	atures	Insert	Delete	Post Va	alidate	<u> </u>	
Reale			1 - 50000		Position	: 4143	🔽 Controls >> Map 🔽 Map >> Controls	
152 Featur	14 1 es	5 16 1 18 1	\$ 23 }	26	27	28 2 3		3 <mark>?</mark>

Figure 4.15

In the example above we clicked on gene 2. 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. gp2 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 the second BLAST hit and then clicked on the [[[**Alignment**]]] sub-sub-tab. Note that you can now see the amino-acid matches in the bottom right pane.

DNA Extract	ed fro	m Fa	stA Lib	ary	/ Bongo.	.fasta				x
Overview	Featur	es	Referenc	es	Sequen	ce Doc	umentati	on		
Sort By	Index	-	Ŀ	4	Name	Start	Stop	Le	Description Sequence Product Regions Blast Context	1
Select Fe	atures	Dire	et SQL		1	238	450	21	Score Target Description	
Turne	ъ IZ	MI			2	760	1110	35	355 gp2 [Mycobacterium phage Rev]	70
Type	- 18 pr	50	<u> </u>	L	3	1496	1873	37	▶ 214 gp121 [Mycobacterium phage LeBron]	
Name	like				4	1951	2136	18	209 gp124 [Mycobacterium phage JoeDirt]	
GenelD	=				5	2133	2366	23	177 hypothetical protein MAB 1818 [Mycobacterium	
Locus	like [6	2363	2482	12	164 gp128 [Mycobacterium phage Faith1]	
Start	IN É				7	2572	2838	26		•1
Start					8	2826	3134	30	BLAST Hit	
Length					9	3131	3418	28	Accession YP_003857243 Export	
Regions	>				10	3415	3645	23	GI 304361056 Export All	
% GC	<				11	3645	3944	30	Length 126 Delete	
CAL	Σİ.				12	4042	4365	32	Max Score 214 Date 10/7/2011 Delete All	
5.04					13	4475	4693	21	High-Scoring Pairs (HSP)	
EL#	like				14	4719	6263	15	HSP Data Alignment	
Product	like				15	6260	7882	16	6 QALPAMRKLV GGHIEAV TTEHA TLWC	111
Function	like 🛛				16	7882	9042	11	1 + 1 ++ 11 + ++ ++	111
FeatureID	L I				17	9093	9809	71	29 RELAKLQGLV GGYIEAVYGY TSDRAEQPDV TLFC	Ш
T					18	9854	11013	7 11	48 YLWIKLOPEN EGEDSLIGTV TUTGPADEAG DSYL	Ш
Tag	іке			E				-		111
🔲 Hide I	gnored	Featu	ires					F	79 TLWWSLNRYA TGVDHLHGVV VVTGGADRNG DTLF	.11
Sele	ct All Fe	ahure		Е	Insert	Delete	Post 1	Validate	<u> </u>	Ш
		1			50000		Devision		Controls 33 Mars III Mars 33 Controls	
				1.	50000		Positio	on:4143		un n
	14 1	5/16	州神	PP	38	26	- 27	28 2 3	80_33131334335313391	
152 Feature	es	Live							80228 😸	?

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.

Six	٤F	rai	m	•	T	re	m	ile	ti	or	1 9	f	Et	uq	le	a	n	no	ta	te	d																																														l				×
View M	Мар		E>	кр	orl	ħ	lap																																																																
1 14998	8	S AG TC L	R G A	G.C.R.	Г Н А.() : Г() 5 \		L G V	2 F F F F F F F F	L C .G E	I S TC AC R E	S T A R	2 C1 GJ E F	L G E	U TC AC	; E ;G ;C P	N K AJ TT S F	I A T F	2 F T A J I N		0 R A0 TC *		K CA GT	R AC TC L I	N GA CT	T A(: T(F	M H CA GT Z Z	* T(A(M F	R GA J CT H S	G G C C L F	G G G C C P	G G G G C C C C C C C C C C C C C C C C	V IG CI	L * TT AA T N	S 'A(.T(1 *	A GC I CG L A	A P GC CG A R	P G G G	L CC G	K * TA AT R *	K AA • TT L F	T J LAC TTC F	P CC GG V G	P I G G G	U CT GA R	G G G C C P Q	T T T T	G G L G L C C L L	R (GA(CT() S	GG GG CC L P	L (T) (GA S	N AA TT *	I 2 .A7	R TCC IGC IGC	V STO AC T I	G G G G G G G G G G G G G G G G G G G	7 GT.	R E AG I TC Y L	R G AG TC S	T D GA CT P S	R .CC 'GC V I	W 7 GT CA CA	G GG CC H P	* F T & F A T T T	R IA(TT(L I	J D JG P	P T AC TG S V	P CC GG G	V 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	S Q TC I AG T D	K S AA TT L	P .GC .CG L A
116 14883	3	P CC GG G G	W T A R	G . C Q :	ז 20 20 20 20		G G C I P	A F GC 100 P	G G C R	7 91 02 1 9	L C G	S TC AC F	G G G	G G C F P	T T T T T S		UG C L	S TC AC I	V G G R	S F TC AG T	P P G G G	G G G G G G G G G G G G G G G G G G G	H GC CC	H T CA GT C	R C G V H	R GC CG R	I A GC R	ST SA	R CC GC E F	G G P	P CC GG	U (1 (1 (1) (1) (1) (1) (1) (1) (1) (1) (G GG CC P	E GJ CT P	G R AG I. TC 3 L	A GC CG P A	E GJ C R	K RAA FT F	G G G G C C C C C C C C C C C C C C C C	C L TT AA T N	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	P CC GG	A IGC R I		K NAA TT W L	T I AC TC TC) R (C) G G (G)	E K J J J S A R	E R RG TC F L	N AA TT S F	R CO GC F	R R G G R G G G G G G G G G G G G G G G	Q IG IG IG IG I I I I I I I I I I I I I	G I I I I I I I I I I	N N SAJ	TAT	R T AC IG TG V	A GC CG R A	G G C C R	C T AC	Q GC. GG	D R AG TC C L	M A7 1. TJ S 1	G I I I I I I I I I I I I	P GC G A	R E CG GC GC G	K AA TT S F	S JAC IAC ITC F I	P GC GC CG L A	L ST GA G R	R G CG GC R E
		L	ម	~	3 1	-	G	7.	*	E _	E	ĸ	A 	P T	E	N	R	v,	L	U	D	<u>ر</u> ۲	Q N	4 -	s (5	G	; ~	D	د	A H	1 (R		RA	W.	G Ţ	ү ,	I ~	RE		т	, I ,	L	D +	T,	D	4	R	Q ,	ĸ	G	, I ,	D,	с ,	, , ,	Y Y P(R 	V on	V 1	7 46	C A	I	T Fo	v nt	F	н	ı Y		A	P ?

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**).

DNA Six	x Frame Translation of Etude_annotated
View N	Map Export Map
	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
	ATLLSLEIUART*GGLAPLKPLVGG*I <mark>VGRGRGRDPSSPGGRSR</mark>
Ι,	
1	
14998	3 TCCCTGTGAAGAGAGAGAGAGACCTTTAAGTCCGTTCTTGTACTCCCCCCAATCGCGGGGATTTTGGGGACCATCCTCGGATTAGCACCCATCCCGCACCATTCCGGGACCACCCCCCCAGAG
	L S V B R B P F B P L F M L P T L A G L V G P L L S F R P Y L V H Y P G T L G O H A T B
	AVSRERSI*ALVHPPNAGRFGRTPP*ITPLPRPLSGDLGPPRDR
	SRCKERQFNLCSCSPP*RG*FGQYSALDHTSSTTLVR*ARTPPR
	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 M 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
133	GACATTEGTEEGGACGECGTEGGECTEGGACGTTGEGGACGTTGECGECGAAGGTTGECGGEGGGGGGGGGG
14066	
14000	
	S L R G P V R B A O S A F T A A L G F F A C S Y A T C I A S L R P O S T S F G F T O V L
	PVTTGARARGPLRLNGGFRLVRLFVRHLHGFAETPLNLLRVNPG
	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
	Q G M H A A G I E T L T H 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
	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
265	TCAGGGGATGCACGCCGCTGGTATCGAGACTTTGACGATGCGCAAGGGGGATGCGTATGGGGGCATGCGCCTAATGGCCGCGATTGGCGGGTTAACGGGTTTCACGAGG
14704	
14734	Position : 25 Font 8 Postion : 25 Font 8

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	
 Export Forward Translations Export Reverse Translations Export Forward DNA Sequence Export Reverse DNA Sequence Export Center Ruler Colorize Annotated ORFs on Translation Colorize Annotated RNAs on Sequence Prevent line breaks within blocks 	100 Nucelotides per Line 8 Point Text 0.5 Inch Margins • Portrait Mode C 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.



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 ht	C Specified Region :
		From 1 🚖 🔳
Estimated		To 53278 🔺 🔳
Dimensions	Size Adjust 🗢 None	
	C Trim to fit	Length 53278 🜩
Stagger ORFs	 Shrink to fit 	,
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
	J	?

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 a graphical illustration of your genome, similar to one shown in **Figure 5.7**.



5.3 Generating a graphical output from 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 stand-alone version on the web, you can:

1. Use an existing coding model to predict the genes.

2. Generate a graphical output.

The host profile we recommend using is that of *Mycobacterium smegmatis*, 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.

To run web-based GeneMark (we use mycobacteriophage Bongo below), go to:

- <u>http://opal.biology.gatech.edu/GeneMark/genemark_prok_gms_plus.cgi</u> Also found on the Links page of <u>http://phagesdb.org</u>
- Select '**Browse**', then find and select your sequence file. This is the same YourPhage.fasta file that you imported into DNA Master.
- Enter your phage's name in the 'Title' box.
- From the '**Species**' dropdown box, select '*Mycobacterium_smegmatis*' (assuming you are annotating a mycobacteriophage genome).
- Maintain the default option of *E. coli* as the RBS model (there is no other).
- Maintain the default options for Window size, Step size, and Threshold.
- In the 'Graphical output options' section, check each box in the first column except 'Generate PostScript graphics (email)' and 'Mark putative exon splice sites'. You do not need to enter an email address.
- Uncheck all boxes under the 'Text output options' heading.
- Click on the 'Start GeneMark' button at the bottom left.

Reference: Borodovsky M. and McIninch J. <u>GeneMark: parallel gene recognition for both DNA strands</u>, Computers & Chemistry, 1993, Vol. 17, No. 19, pp. 123-133. [Download PDF]

Prediction models ready for a total of 265 completely sequenced prokaryotic genomes in NCBI R	efSeq database. Pre-calculated prediction database for these genomes
Input Sequence	
Title (optional):	
Bongo	
Sequence: •	
Sequence File unload e	
/Users/welkin/Documents/Mycophages/SEA phages/2010-2011/ Browse	
Running Options	
Species: Mycobacterium_tuberculosis_H37Rv	Window size: 96 5 5 5 5 5 5 5 5 5 5 5 5 5
RBS model: • Ecoli :	Step size: • 12 : bp
Use alternate genetic code:• Eukaryote (e.g. Yeast, ATG = only start) Mycoplasma (TGA = Tryptophan)	Threshold: 🖲
Output Options	
Graphical output optionse Graphical output optionse Generate PDF graphics (screen) Generate PotScript graphics (email) Mark orfs on graph Mark regions on graph Mark pot codons on graph Mark start codons on graph Mark frameshifts on graph Mark frameshifts on splice sites Print graph in landscape format Email address (required for PostScript email output)	Text output optionse List open reading frames (ORFs) predicted as coding sequences (CDSs) List regions of interest List putative eukaryotic splice sites Write protein translations of ORFs Write protein transcripts of ORFs Write nucleotide transcripts of regions Write nucleotide transcripts of regions Write nucleotide transcripts of putative exons Write nucleotide transcripts of putative exons
Kull	
Start Genemark Default	
Please send any suggestions for improvements or problems to the web page maintainer.	

Figure 5.8

Once GeneMark has run, a new window will appear and in the middle it will have a heading "Result of last submittal", as shown in Figure 5.9.

• Click on the link 'View PDF Graphical Output' just below.

Result of last submittal

GeneMark Results View PDF Graphical Output Sequence: Bongo Sequence length: 80228 GC Content: 61.62% Window length: 96 Window step: 12 Threshold value: 0.500 Matrix: Mycobacterium tuberculosis H37Rv, Thu Oct 27 16:10:50 2005 Matrix author: Dr. Borodovsky Laboratory, School of Biology, Georgia Tech Matrix order: 5

Figure 5.9

• Save and open the pdf.



Figure 5.10

We recommend that you **print** this file because it is a good place to make notes as you refine your annotation. Below, several features of this output are described.

- 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).
- 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).
- The shaded areas (#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 Using Phamerator to assist with annotation

6.1 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.

Phams are thus groups of proteins with a high degree of similarity to one another, though there is one caveat to be aware of. If protein A is similar to protein B and protein B is similar to protein C, all three will be grouped into the same Pham, even if proteins A and C are not above the threshold scores when compared directly. This can be very useful in identifying proteins with multiple domains that may be fused in one phage genome and split in 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.

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.

6.2 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.

6.3 How did my genome get into Phamerator already?

In order 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 'Mycobacteriohage_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.

6.4 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 (red oval in diagram below). Make sure the database is "Mycobacteriophage_Draft". If not, go to Edit → Preferences and select Mycobacteriophage_Draft from the Database dropdown menu.

	8 🗢 🖲 CPha	merator :: Mycobacter	iophage_Draft	>			
	File Edit To	ols Help					
	🗎 🔮	S	E				
	Open Add	Map Pham Circle Go	oogle Maps				
	Sources						1
<	Phages	me	Length (bp)	GC %	Number of Genes	Cluster	6
	Phams	Mitkau-DRAFT	68493	66.490006	95		
	Genes	Patience_draft	70506	50.303520	105		
	Domains	Plieone-DRAFT	155585	64.734390	219		
		Ruby-DRAFT	57726	61.367495	95		
		Turj99-DRAFT	51161	63.741913	77		
		DS6A	60588	68.437644	97		
		Shanov-DRAFT	51140	63 840438	91		

Figure 6.1

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.

In Figure 6.2, four genomes are currently selected, indicated by the orange highlight.

• *(S) () (IJ				
rces	(X)	gte maps				
ages	Name	Length (bp)	GC %	Number of Genes	Cluster	
ams	Optimus	109270	60 790702	230	1	
enes	Anava	60835	66 304345	08	K1	
omains	CrimD	59798	66 876819	95	K1	
	IAWS	59749	66 617015	94	K1	
	BEEST draft	59906	66 555938	91	K1	
	BarrelRoll-DRAFT	59672	66.615833	95	K1	
	Angelica	5959R	66 388134	94	K1	
	Adephagia	59646	66,606311		K1	
	TM4	52797	68 113719	91	K2	
	Pixie	61147	67.303384	100	K3	
	Fionnbharth-DRAFT	58076	68.007439	93	K4	
	Larva-DRAFT	62991	65.295042	95	K5	
	Ender-DRAFT	74731	58,796216	111	L1	
	JoeDirt	74914	58,780735	126	L1	
	LIDIE	73704	FO 703334	400		

Figure 6.2

- Once you've finished selecting genomes, click on the button that says '**Map**' (red circle in **Figure 6.2**). 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 6.3

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

6.5 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 6.4**.

• 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: Second State Zoom In Zoom Out Normal Size	nt
	<u> </u>
	الالالاسماد الاستان الميام المستحادة والمتحادة والمتحاد
	and the liter and the second and the
	1 _ = 11
	()))
cuck on a gene	cuck on a gene

Figure 6.4

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:

😣 🖨 💷 Genome Map	
File View Color	
C Zoom In Zoom Out Normal Size	📑 🚍 Align Left Align Right
Ar Show nucleotide conservation Show E values Show Descriptions	(1) (1) (1) (18)
Show Phamily Names	3189 (19)
• Hover Highlights Pham	
Adephagia	60

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

Figure 6.5

Lots of information is displayed on Phamerator maps.

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



Figure 6.6

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:



6.6 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 6.8).



Figure 6.8

Once you've turned on '**Show nucleotide conservation**', you may see colors between the genomes on your map, as shown in **Figure 6.9**.



Figure 6.9

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.

6.7 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 6.10

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.

6.8 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'.

7 Guiding Principles of Bacteriophage Genome Annotation

7.1 Overview

Though the automated annotation you have created using DNA Master will usually identify more than 80% of genes correctly, some genes will need to be manually added, modified, or deleted. Therefore, all gene calls 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.

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. 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.
- 5. 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.
- 6. Protein-coding genes should have coding potential predicted by *either* Glimmer, GeneMark, or GeneMark TB. Start sites are chosen to include areas of strong coding potential.
- 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 OF BACTERIOPHAGE GENOME ANNOTATION

...CONTINUED

- 10. An important task is choosing between different possible translation initiation (i.e., start) codons. The correct start site can often be distinguished by association with a credible ribosome-binding site (RBS; Shine-Dalgarno (SD) sequence). Identifying the correct start site, however, is not always easy and is predicated on the following sub-principles:
 - a. The preferred start site usually has one of the higher SD scores of all the potential start codons, but not necessarily the highest.
 - b. Manual inspection can be helpful to distinguish between possible start sites. The consensus is as follows: **AAGGAGG – 3-12 bp – start codon**.
 - c. 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 4 bp). A short overlap of 1-4 bp—where the start codon overlaps the stop codon of the upstream gene—is very common.
 - d. 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.
 - e. Your final start-site selection will likely represent a compromise of these subprinciples. For example:
 - i. A start codon that overlaps the stop codon of a previous gene trumps a somewhat lower score.
 - ii. A higher SD score or canonical RBS trumps a more extended gene overlap.
 - iii. If choosing between several starts with similar SD scores, it is usually best to choose the one that gives the longest open reading frame.
- 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. A printout of your GeneMark-Smeg output (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 **tRNA Genes** (Section 8.6).

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 first open reading frame, which will typically be called gene '1' in the DNA Master feature table.

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.

Overview Features References Sequence Documentation Sort By Index Image: State Stop Description Sequence Product Regions Blast Context Select Features Direct SQL 1 408 704 Name 1 GeneID Type is All 1 408 704 Name 1 GeneID GeneID - 5 2345 3555 Stop 704 Regions 1 Locus like - 6 3592 4380 Length 297 Tag Tag Start 2 - 4377 5294 Direction Forward Translation Table Undefined ▼ Regions 2 9 6896 7309 EC Number EC Number EC Number EC Number Int 8 111 8479 10005 Igp1 Int Int 122 1002 Int Int Int Int<	Extracted from FastA Library Timshel.fasta										
Sort By Index Image Stat Stop Description Sequence Product Regions Blast Context Select Features Direct SQL 1 408 704 Name 1 GeneID Type is All 2 743 1177 Type CDS Git Name ike 4 1567 2313 Statt 408 Locus Tag DNAM1 GeneID 5 2345 3565 Stop 704 Regions 1 Locus ike 6 3592 4380 Stop 704 Regions 1 Locus ike 7 4377 5294 Direction Forward Image	Overview Feature	8 References	Sequence	Documer	ntation						
Select Features Direct SQL 1 408 704 Name 1 GenelD Type is All 2 74.3 1177 Type CDS Gi Name like 4 1567 2319 Statt 408 Locus Tag DNAM1 GenelD 5 2345 3565 Statt 408 Locus Tag DNAM1 Locus like 6 3592 4380 Length 297 Tag Tag Start 2 7 4377 5234 Direction Forward Tag Tag 7 Regions 2 7 4377 5234 Translation Table Undefined EN EN EN Viet GC 10 7306 8283 Translation Table Undefined EN Interviet 12 Interviet 12 Interviet 145	Sort By Index		Name	Start	Stop	🔺 Desci	iption	Sequence F	Product Region	s Blast Con	text
Type is All 2 743 1177 Name like 4 1567 2319 Start 408 Locus Tag DNAM1 GenelD = 5 2345 3565 Stop 704 Regions 1 Locus like 6 3592 4380 Length 297 Tag Tag Start > 5374 5449 Direction Forward > Tag Tag > Regions > 9 6896 7309 Product > > > Product > </td <td>Select Features</td> <td>Direct SQL ⊄</td> <td>1</td> <td>408</td> <td>704</td> <td></td> <td>1</td> <td></td> <td>GenelD</td> <td>· ·</td> <td>· 11</td>	Select Features	Direct SQL ⊄	1	408	704		1		GenelD	· ·	· 11
Name ike 3 1254 1577 1705 103 104 Name ike 4 1567 2319 Start 408 Locus Tag DNAM1 GenelD 5 2345 3565 Stop 704 Regions 1 Locus ike 6 3592 4380 1 Length 297 Tag Tag Start 5 5374 5249 Translation Table Undefined Image: Start Start Start 10	Tupe is Al	i i l	2	743	1177	Тира		ne	- GI		
Name like 4 1567 2319 Start 408 Locus Tag DNAM1 GenelD = 5 2345 3565 Stop 704 Regions 1 Locus like 6 3592 4380 Length 29 Tag 1 Length 2 7 4377 5294 Direction Forward 2 1 Length 2 7 4377 5294 Direction Forward 2 1 Length 2 5374 5449 Start 20 Direction Forward 2 Regions 2 9 6896 7309 2 CNumber 2 CAI 3 11 8479 10002 11459 2 10 7 14268 7 Function 16 14080 14268 17 14276 14650 1	Type is in	L	3	1254	1577				· · ·		III
GeneID = 5 2345 3565 Stop 704 Regions 1 Locus like 6 3592 4380 Length 297 Tag Tag Start > 5374 5449 Translation Table Undefined ▼ Begions > 9 6896 7309 Translation Table Undefined ▼ K GC C 10 7306 8283 Translation Table Undefined ▼ Product like 11 8479 10005 Product ♥ ♥ Product ♥ ♥ Image: Start 1 1459 13020 Product ♥ <	Name like		4	1567	2319	Start		4	08 Locus Tag	DNAM1	
Locus like 6 3592 4380 Start 7 4377 5294 Tag Length 8 5547 6899 Direction Forward Image: Construct of the second	GenelD =		5	2345	3565	Stop		7	04 Regions		1
Start 7 4377 5294 Length 5374 5449 Regions 8 5547 6899 9 6896 7309 % GC 10 7306 8283 CAI 11 8479 10005 12 10002 11459 13 11456 12469 Function 15 13055 Function 15 13055 16 14080 14268 17 14276 14650 Original Glimmer call @bp 408 has strength 2.35; GeneMark Calls start at 297 Select All Features Insert Delete Post Validate V 4 <td>Locus like</td> <td></td> <td>6</td> <td>3592</td> <td>4380</td> <td>Lenat</td> <td>, 297</td> <td>7</td> <td>Tag</td> <td></td> <td></td>	Locus like		6	3592	4380	Lenat	, 297	7	Tag		
State 5374 5449 Length 8 5547 6899 Regions 9 6896 7309 % GC 10 7306 8283 CAI 11 8479 10005 % GC 12 10002 11459 EC# 1ke 13 11456 12469 Product 1ke 14 12520 13020 Function 1ke 15 13055 14014 FeatureID = 17 14276 14650 Valuet Tag ike 17 14276 14650 Veleet Valuet Valuet Valuet Valuet Select All Features Insert Delete Post Validate Valuet Valuet K4 Color101 11 12 1000 25 26 27 32 44 K4 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44	Start N		7	4377	5294	Direct	ion Eor	word		1	
Length 8 5547 6899 Regions 9 6896 7309 % GC 10 7306 8283 CAI 11 8479 10005 12 10002 11459 13 11456 12469 Product ike 14 12520 Function ike 15 13055 14 12520 13020 Function ike 15 17 14276 14650 Veter 0riginal Gimmer call @bp 408 has strength 2.35; GeneMark Calls start at 297 Controls >> Map < Map >> Controls Select All Features 11 124344544 14 56673 9 400 11 12 0000 14258 0riginal Gimmer call @bp 408 has strength 2.35; GeneMark calls start at 297 Veter Veter Veter Select All Features 11 12434545454 13 125 26 27 32,454 Veter 53278 4 53278 2				5374	5449	Direct	IUNI FUI	iwaiu j	<u></u>		III
Regions 9 6896 7309 % GC 10 7306 8283 CAI 11 8479 10005 12 10002 11459 13 11456 12469 Product 1ke 14 12520 Function 1ke 15 13055 14014 FeatureID 16 14080 14268 Tag 1ke 17 14276 14650 Vetex Insert Delete Post Validate Vetex Vetex Insert Delete Post Validate Vetex Vetex Vetex Insert Delete Post Validate Vetex Vetex Vetex Vetex Insert Delete Post Validate Vetex Vetex Vetex Vetex Vetex 89 Features Ivet 25 26 27 32 44 Vetex 44 Vetex 44 89 Features Ivet 53278 4 2 22 26 27 32 44 Vetex	Length		8	5547	6899	Trans	ation T	Table Undefin	ed		
% GC I 10 7306 8283 CAI > 11 8479 10005 III 8479 10005 12 10002 11459 III 11 8479 10005 12 10002 11459 III 11 11450 12469 14 12520 13020 Function like 15 13055 14014 14 14 12520 13020 Function like 15 13055 14014 14 14 14 1250 1000 Image: State of the state o	Regions >		9	6896	7309	EC N	ımber				
CAI > 11 8479 10005 EC# ike 12 10002 11459 Product ike 13 11456 12469 Function ike 15 13055 14014 FeatureID = 16 14080 14268 Tag ike 17 14276 14650 Hide Ignored Features Insert Delete Post Validate V Callect All Features Insert Delete Post Validate V Controls >> Map Map >> Controls Map >> Controls Map >> Controls 89 Features Live 25 26 27 32 44 Map >> Controls 89 Features Live 53278 4 </td <td>% GC 🔽</td> <td></td> <td>10</td> <td>7306</td> <td>8283</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>× 1</td>	% GC 🔽		10	7306	8283						× 1
ECH 12 10002 11459 9p1 Forduct ike 13 11456 12469 Product ike 14 12520 13020 Function ike 15 13055 14014 FeatureID = 16 14080 14268 Tag ike 17 14276 14650 Original Glimmer call @bp 408 has strength 2.35; GeneMark Calls start at 297 Gelect & II Features Insert Delete Post Validate I 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4		F	11	8479	10005	Produ	ct				
EC# like 13 11456 12469 Product like 14 12520 13020 Function like 15 13055 14014 FeatureID = 16 14080 14268 Tag like 17 14276 14650 Hide Ignored Features Insert Delete Post Validate Celect & II Features Insert Delete Post Validate V Controls >> Map >> Controls >> Map >> V Controls >> Map >> Controls >> Map >> V Controls >> Map >> Controls >> Map >> V Controls >> Map >> Controls >> Map >> V Controls >> Map >> Controls >> Map >> Controls >> V Controls >> Map >> Controls >> Map >> Controls >> Controls >> V Controls >> Map >> Controls >> Map >> Controls >> Controls >> V Controls >> Controls >>		F	12	10002	11459	gp1					<u> </u>
Product like 14 12520 13020 Function Function like 15 13055 14014 FeatureID = 16 14080 14268 Tag like 17 14276 14650 Original Glimmer call @bp 408 has strength 2.35; GeneMark Image: Call Features Image: Call Features Gelect & II Features Insert Delete Post Validate V Image: Call Features Image: Call Features Image: Call Features Image: Call Features Image: Call Features Image: Call Features Image: Call Features V Image: Call Features Image: Call Features Image: Call Features Image: Call Features V Image: Call Features Image: Call Features Image: Call Features Image: Call Features V Image: Call Features 89 Features Image: Call Features 89 Features Image: Call Features Image: Call Features Image:	EC# like	F	13	11456	12469						-
Function like 15 13055 14014 FeatureID = 16 14080 14268 Tag like 17 14276 14650 Hide Ignored Features Insert Delete Post Validate Select All Features Insert Delete Post Validate I 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4	Product like		14	12520	13020	Funct	on				
FeatureID = 16 14080 14268 Tag like 17 14276 14650 Hide Ignored Features 17 14276 14650 Select ΔII Features Insert Delete Post Validate I 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4	Function like		15	13055	14014						<u> </u>
Tag like 17 14276 14650 Notes Tag like 17 14276 14650 Original Glimmer call @bp 408 has strength 2.35; GeneMark Hide Ignored Features Insert Delete Post Validate Select & Il Features Insert Delete Post Validate V 4 4 4 4 9 Sob 51 55 5 <t< td=""><td>FeaturelD -</td><td>—_ F</td><td>16</td><td>14080</td><td>14268</td><td></td><td></td><td></td><td></td><td></td><td></td></t<>	FeaturelD -	—_ F	16	14080	14268						
Tag like Original Glimmer call @bp 408 has strength 2.35; GeneMark Hide Ignored Features Insert Delete Post Calls start at 297 Select All Features Insert Delete Post Validate Image: All Features Insert Delete Post Validate Image: All Features Insert Delete Post Validate Image: All Features Image: All Features Image: All Features Image: All Features 89 Features Live 53278 4 Features 53278		F	17	14276	14650	Notes	_				- <u>-</u>
Hide Ignored Features Insert Delete Post Validate Validate Validate Validate Validate Validate Validate Validate Validate	lag like				[y Origin	al Glim	nmer call @bp 4	108 has strength	2.35; GeneMark	
Celect ∆II Features Insert Delete Post Validate Image: Market All Features Image: Market All Feat	🔲 Hide Ignored F	eatures 📘			►	C calls	start at	: 297			
Select all Features Image: Select all Features Image: Select all Features Image: Select all Features Image: Select all Features Image: Select all Features Image: Select all Features Image: Select all Features Image: Select all Features Image: Select all Features Image: Select all Features Image: Select all Features Image: Select all Features Image: Select all Features Image: Select all Features Image: Select all Features Image: Select all Features Image: Select all Features	0.1.1.485		Insert De	lete Pos	st 🛛 Validal	te					
Image: A state of the sta											
2 4 5 6 7 8 6 0 1 12 3 6 1 2 2 2 2 2 2 2 2 2 2 2 2		b b b 1-	50000	P	osition : 49	1927	Con	trols >> Map 🖡	Map >> Contro	ns	
89 Features Live 53278 4 ?	2 4 5 6 7 8 8	10 11 12 131	15/11/22/2	25	26 27	2 32	4	4 44 4 4 49	9 605 5 565 5	6 6 6 6 1 7 1 7	
	89 Features	ive								53278	3?

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

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, although the predicted starts sites are different. (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

In **Figure 8.3**, the region of gene 1 is circled. You can find a gene by looking at its coordinates in the Feature Table, then finding those coordinates on the GeneMark output. It appears that this ORF has coding potential starting near position 400 and ending near 650. This is further evidence that there is a real protein-coding gene here.

• Examine the BLAST data under the **[[Blast]]** sub-tab, and see if there are genes in the databases that are high-quality matches to this one.

Extrac	ted from	FastA Librar	y Timshel.f	asta			×
Overview	Features	References	Sequence	Docume	ntation		
Sort By	Index 💌		Name	Start	Stop		Description Sequence Product Regions Blast Context
Select Fe	atures Di	rect SQL	▶ 1	408	704		Score Target Description
Tuno	io All	<u> </u>	2	743	1177		516 gp1 [Mycobacterium phage Timshel]
Type		<u> </u>	3	1254	1577		367 gp1 [Mycobacterium phage HelDan]
Name	пке		4	1567	2319		318 gp4 [Mycobacterium phage Trixie]
GenelD	=		5	2345	3565		317 gp4 [Mycobacterium phage RedRock]
Locus	like		6	3592	4380		295 gp1 [Mycobacterium phage Bxz2]
Start			7	4377	5294		
Length	The second se			5374	5449		BLAST Hit
Eongan			8	5547	6899		Accession AEJ92316 Export All
Hegions			9	6896	7309		Length 98 Delete
% GC	<		10	7306	8283		Max Score 516 Date 10/21/2011 Delete All
CAI	>		11	8479	10005		Uigh Cooring Pairs (UCP)
EC#	like 🗌		12	10002	11459		High-scoling Fails (HSF)
Deadurat	Bun [13	11456	12469		HSP Data Alignment
			14	12520	13020		1 MTWNTSDRAS RLPADWD NY RQPVLRDADY RCQI
Function	like		15	13055	14014		1 MTHNTSDAS DLDADHDINY DODULDDADY DOOL
FeatureID	=	-	16	14080	14268		
Tag	like		17	14276	14650		51 PCDDHSRDNL OAACSRCHCK KSSREGNDKK ARMF
- Hidel	approd Eq.	aturaa	•		F	Ľ	51 IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
1 maer	gnoreu r ea			u la	1.4.1	. 1	TI FOUNDROWN GARCONCHON NOORDANN AND
Sele	rct ∆ll Feah	ires .	Insert De	elete Po	st Valid	ate	<u>p. </u>
N 4 4		H H 1	- 50000	F	osition : 4	9927	Controls >> Map 🔽 Map >> Controls
12141516	7 8 81	0 11 12 13	NISHITED B	2 25	26 27	N.	321 34 MT81 44 MAX 49 505 51 555 55 6 16 5 1 7 10 11 1
89 Feature:	s Liv	/e /1 7 7 7			<u> </u>	210	
						_	

Figure 8.4

In **Figure 8.4** above, you can see that there are several good matches, which further supports this gene call's legitimacy.

- 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 200 bp in length with an eye towards gauging their legitimacy, and genes below 120 bp should be viewed very skeptically.
- You can check gene length by using the scroll bar to move to the right in the central column of the Feature 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**).

🚟 Extracted from FastA Library Timshel.fasta														
Overview	Featu	ires	Reference	s	Sequence	e Docun	nentation							
Sort By	Index	-	•		Clart	Stop	Length		Descripti	ion Seque	ence Prod	luct Regions	Blast C	Context
Select Fe	atures	Dire		▶	408	704	297	\supset	Name	1		GenelD		· 1
T	. [743	1177	435		T	CDC		<u></u>		
Type	- 18 	~"	<u> </u>		1254	1577	324		туре	lus	<u> </u>	<u>ui</u>		
Name	like				1567	2319	753		Start		408	Locus Tag	DNAM1	
GenelD	=				2345	3565	1221		Ston		704	Regions		1
Locus	like				3592	4380	789	(Lenath	297		Tag	í —	
Chart					4377	5294	918		Constant of the second se			, ug	·	
Start					5374	5449	76		Direction	rrorward	<u></u>			
Length	>				5547	6899	1353		Translati	on Table	Indefined			
Regions	>			Π	6896	7309	414		EC Numb	ber				
% GC	<			Π	7306	8283	978							*
CAL	Ti			Π	8479	10005	1527		Product					
				Π	10002	11459	1458		gp1					<u> </u>
EC#	like				11456	12469	1014							_
Product	like				12520	13020	501		Function					
Function	like [13055	14014	960							A
FeaturelD	, di				14080	14268	189							-
				H	14276	14650	375		Notes					•
Tag	like						1	-	Original	Glimmer ca	ll @bp 408	has strength 2	.35; GeneMa	ark 🔺
🔲 Hide I	gnored	Feat	ures 🤇	4				<u>ار</u>	calls sta	rt at 297				
Sele	et áll Fi	eature	~		Insert 1	Delete F	ost Vali	date						~
Reference	le le	• •		1	1562		Position :	14680		Controls >>	Map 🔽 M	lap >> Control	s	
14 44 4						4			1.0					
00.5														
89 Features	\$	Live											53278	- I 🗟 <u>/</u>

Figure 8.5

In this case, the gene length (297 bp) is not a concern nor requires special attention.

• 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.



Figure 8.6

In this example above, we can see from the Phamerator map that there are no other genes occupying the same portion of DNA. Good.

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 ACTION: Continue to Section 8.4.2.	NO 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 about half of the area with 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 disagree; Glimmer has called the start at position 408, and GeneMark at 297.



Figure 8.8

• Does the predicted start have an associated ribosome binding site [RBS; Shine-Dalgarno (SD)] with a high score or recognizable sequence? You may recall that you can use the Frames window (DNA → Frames) to review the SD scores for all start options in a given ORF. (Check Section 4.4.4 for details on how to open the "Choose ORF start" window and select a particular ORF.) Figure 8.9 shows that there are four possible start codons for our example. Their SD scores are shown in the red box. There is also a snippet of the upstream sequence so it can be inspected manually.

Chamaster		_ 8 ×
File Tools Window Help		
CHE DRF Analysis for Extracted from FastA Library Timshel.fast	ta	
+3		
+2		
-2		
-3		
1 63 125 187 249 311 3	73 435 497 559 621 683 745	807 869 931 993 1055 1117 1179 125
bp: 917	<u> </u>	
Analyze potential Hibosome Binding Sites		iners 🖼 RBS 1944 inc.
GenelD = 5 2345 3565	Stop 704 Regions 1	
GenelD = 5 2345 3565 Locus ike 6 3592 4380	Stop 704 Regions 1	start
GenelD - 5 2345 3565 Locus ike - 6 3992 4380 Stat 7 4377 5294	Stop 704 Regions 1 Length 297 Tag Starts : 4 Direction Forward Starts : 4	startX DRF Start : 408 Cdn 1 Cdn2 Cdn3 LengthX DRF Start : 408 Cdn 1 Cdn2 Cdn3 LengthCournent
GenelD = 5 2345 3565 Locus ike 6 3932 4380 Statt 2 7 4377 5294 Locus ike 5374 5449 Length 2 8 5547 6599	Stop 704 Regions 1 Length 297 Tag Choose ORI Direction Forward Stats: 4 Selected: 1 Translation Table Undefined Selected: 1	start X DRF Start : 408 Cdn 1 Cdn2 Cdn3 Length
GenelD = 5 2345 3565 Locus ike 6 3592 4380 Stat 2 7 4377 524 Length 2 5374 5449 Length 2 8 5547 6899 Regions 2 9 6896 7309	Stop 704 Regions 1 Length 297 Tag Cmit Choose ORI Direction Forward Starts: 4 Selected: 1 Translation Table Undefined EC Number	start Image: Cdn 1 Cdn 2 Cdn 3 Length DRF Stat : 408 Cdn 1 Cdn 2 Cdn 3 Length Document DRF Stop : 704 5 End 73.0 59.5 56.8 111 Document DRF Length : 297 37 End 53.5 79.6 65.3 295
GenelD - 5 2345 3565 Locus ike 6 3592 4380 Stat - - 7 4377 5234 Length - - - 5374 5449 Length - - - - - - % GC - - 9 6896 7309 - % GC 10 7306 6283 - - -	Stop 704 Regions 1 Length 297 Tag State: 4 Direction Forward Selected : 1 Translation Table Undefined EC Number Shine D	start X DRF Stat : 408 Cdn 1 Cdn 2 Cdn 3 Length X DRF Stop : 704 5 End 73.0 59.5 56.8 111 DRF Length: 297 3 End 53.5 79.6 65.3 295 algarno Sequence of the Region Start Start
GenelD - 5 2345 3565 Locus ike - 6 3392 4380 Stat > - 7 4377 5234 Lengtion > - 6 5374 5449 Regions > 9 6896 7309 % GC 10 7705 8283 CAJ > - 11 8479 10005	Stop 704 Regions 1 Length 297 Tag Start Actorse URL Direction Forward States: 4 Selected : 1 Fronslation Table Undefined EC Number Selected : 1 Product g Score	start LORF Start : 408 Cdn 1 Cdn2 Cdn3 Length Comment DRF Start : 408 Cdn 1 Cdn2 Cdn3 Length Comment DRF Stop : 704 5 End 73.0 58.5 56.8 111 DRF Length : 297 3 End 53.5 73.6 65.3 295 algarno Sequence of the Pegion Start Start ORF Space Upstream of the Start Codon Position Length
GenelD - 5 2345 3565 Locus ike - 6 3592 4380 Start > - 5374 5449 Length > - - 5374 5449 Regions > - 9 6896 7309 % GC 10 7306 8283 CAI > - 12 10002 11459	Stop 704 Regions 1 Length 297 Tag Stats: 4 Direction Forward Stats: 4 Selected :1 Translation Table Undefined Selected :1 EC Number Shine D Stats: 4 Product # Secre gp1 1 140	start DRF Start : 408 Cdn 1 Cdn2 Cdn3 Length DRF Start : 408 Cdn 1 Cdn2 Cdn3 Length DRF Start : 408 Cdn 1 Start Start DRF Length : 297 3 End 535 796 653 295 algarno Sequence of the Pegion Start Start ORF algarno Sequence of the Start Codom Position Length 12 CAAGGGCCTCCTTGGCCTCCCT TTG 297 408
GenelD - 5 2345 3565 Locus ike - 6 3592 4380 Start 2 - - 5574 5449 Length 2 - - 5574 5449 Regions 2 9 6995 7309 % GC C 10 7306 8283 CAI 2 - 112 10005 EC# ike - 13 11456 12469	Stop 704 Regions 1 Length 237 Tag Starts: 4 Direction Forward Starts: 4 Selected : 1 Translation Table Undefined Selected : 1 EC Number Shine Shine Product # Secre gp1 1 140 2 143	start DRF Stat: 408 Cdn1 Cdn2 Cdn3 Length DRF Stat: 408 Cdn1 Cdn2 Cdn3 Length DRF Stat: 408 Cdn1 Cdn2 Cdn3 Length DRF Length: 27 3 End 535 736 653 295 algarno Sequence of the Region Start Start OFF Space Upstream of the Start Codom Position Length 12 GAAGGGCCTCCTGCGCCCCCCT TTG 409 297 6 CTACCACCCCCCGGAGGATTCC TTG 409 297
GenelD - 5 2345 3965 Locus ike 6 3982 4380 Stat 2 7 4377 5234 Length 2 6 9582 4380 Regions 2 9 6996 7309 % GC 4 10 7306 6283 CAI 3 11 8479 10002 EC# ike 13 11456 12469 Fraduct ike 14 12501 13020	Stop 704 Regions 1 Length 297 Tag Status Choose ORI Direction Forward Image: Status Choose ORI Status Choose ORI Translation Table Undefined Status Choose ORI EC Number Image: Status Choose ORI Status Choose ORI Product Image: Status Choose ORI Status Choose ORI Image: ORI Image: Status Choose ORI Status Choose ORI Product Image: Status Choose ORI Status Choose ORI Image: ORI Image: Status Choose ORI Status Choose ORI Product Image: Status Choose ORI Status Choose ORI Image: ORI Image: ORI Image: Status Choose ORI Product Image: Status Choose ORI Status Choose ORI Image: ORI Image: ORI Image: Status Choose ORI Image: ORI Image: ORI Image: ORI	start DRF Start : 408 Cdn 1 Cdn2 Cdn3 Length DRF Start : 408 Cdn 1 Cdn2 Cdn3 Length DRF Start : 408 Cdn 1 Cdn2 Cdn3 Length DRF Length : 27 3 End 535 736 65 3 295 algarno Sequence of the Region Start Start ORF Space Upstream of the Start Codom Position Length 12 GAAGGCCTCTTGGCCTCCCT TIG 297 408 6 CTACCACCCCCGGAGGATTCC TIG 408 297 9 GGGAAACGACAACAAGAGCCCCGT ATG 654 51 0 CONTINCEDCOCCCCCT ATG 654 51
GenelD - 5 2345 3565 Locus ike 6 3392 4380 Start 3 7 4377 5294 Length 3 6 5374 5449 Regions 3 9 6896 7309 % GC 4 10 7306 6283 CAI 3 11 8479 10005 ECH ike 13 11456 12469 Function ike 15 13025 14014	Stop 704 Regions T Length 297 Tag Status 4 Direction Forward Status 4 Selected 1 Translation Table Undefined EC Number Product # Score gp1 1 140 2 143 Statis 4	start ↓08 Cdn 1 Cdn2 Cdn3 Length DRF Stap : 704 5 End 73.0 585 56.8 111 DRF Length: 297 3 End 53.5 73.6 65.3 295 algarno Sequence of the Region Start Start DRF Space Upstream of the Start Codon Position Length 12 GAAGGGCCTCCTTGGCCTCCCT TTG 297 408 6 CTACCACCCCCCGAGGATTCC TTG 408 297 9 GGGAAACGACAAGAAGGCCCGGTT ATG 654 51 8 CCGTATGCGAGCGGCCGCGTT ATG 672 33
GenelD - 5 2345 3565 Locus ike - 6 3392 4380 Stat > - 7 4377 5234 Lengtin > - 6374 5449 Regions > 9 6896 7309 % GC 10 7705 8283 CAI > - 11 8479 10005 EC# ike - 11 1479 10005 Function ike - 115 13055 14014 FeatureID - - 16 14080 14268	Stop 704 Regions T Length 297 Tag States: 4 Direction Forward States: 4 Selected: 1 Translation Table Undefined Selected: 1 EC Number Image: Secret gamma States: 4 Product g Score gp1 1 140 Function 3 345 4 441	start JORF Start : 408 Cdn 1 Cdn2 Cdn3 Length DRF Start : 408 Cdn 1 Cdn2 Cdn3 Length DRF Stop : 704 5 End [73.0 555 56.8 11] DRF Length: 297 3 End 53.5 73.6 65.3 295 algarno Sequence of the Begion Start Start DRF Space Dpstream of the Start Codon Position Length 12 GAAGGGCCTCCTTGGCCTCCCT TIG 297 408 6 CTACCACCCCCGGAGGATTCC TIG 408 297 9 GGGAACGAAAGAAGAAGGCCCGT AIG 654 51 8 CCGTATGCGAGCGGCCGGGTC AIG 672 33
GenelD - 5 2345 3565 Locus ike - 5 2345 3565 Stat S - 6 3592 4380 Freight S - 5374 5449 Length S 8 5547 6899 Regions S 6365 7303 % GC - 10 7305 82833 CAI S - 12 10002 11459 EC# ike - 14 12520 13020 Function ike - 15 13055 14014 16 14090 14289 17 14276 142650 Tag ike - - 17 14276 14550	Stop 704 Regions 1 Length 297 Tag Stats: 4 Direction Forward Stats: 4 Stats: 4 Translation Table Undefined Stats: 4 EC Number # Score Product # Score gp1 1 140 Function 3 344 Notes Driginal Gimmer call @bp 408 has strength 2.35; Gene Score	start DRF Star : 408 Cdn 1 Cdn2 Cdn3 Length DRF Length : 297 3 End 535 796 653 295 algarno Sequence of the Pegion Start Start ORF Space Upstream of the Start Codom Position Length 12 GAAGGGCCTCCTGGCCTCCCT TIG 297 408 6 CTACCACCCCCCGGACGATTCC TIG 408 297 9 GGGAAACGAACGACAAGAAGGCCCCCT AIG 654 S1 8 CCCTATGCGAGCGCGCGGGTTC AIG 672 33 CM D27
GenelD	Stop 704 Regions 1 Length 297 Tag Stell Choose ORI Direction Forward Image: Stell Selected : 1 Selected : 1 Translation Table Undefined E E EC Number Image: Stell Stell Stell Product Image: Stell Image: Stell Stell Function Image: Stell Image: Stell Image: Stell Original Gimmer call @bp 408 has strength 2.35; Gene Image: Stell Image: Stell	start ORF Star: 408 Cdn1 Cdn2 Cdn3 Length DRF Stop: 704 5 End 730 595 568 111 DRF Length: 297 3 End 535 736 653 295 algarno Sequence of the Region Start Start OFF Space Upstream of the Start Codom Position Length 12 GAAGGGCCTCCTGGCCTCCCT TIG 297 408 6 CTACCACCCCCCGAAGAATCC TIG 408 297 9 GGGAAACGAACAAGAAGGCCCGT ATG 654 51 8 CCGTATGCGAGCGGCGCGGCTC ATG 6572 33 8 CCGTATGCGAGCGGCGCGGCTC ATG 672 33
GenelD	Stop 704 Regions 1 Length 297 Tag Status Choose ORI Direction Forward Image Status Status Status Translation Table Undefined EC Status Status EC Number Image Status Status Status Product Image Image Status Status gp1 1 140 2 143 Function Image Image Image Driginal Gimmer call @bp 408 has strength 2.35; Gene Image Image	start DRF Star: 408 Cdn 1 Cdn2 Cdn3 Length DRF Star: 2408 Cdn 1 Cdn2 Cdn3 Length DRF Star: 2408 Cdn 1 Cdn2 Cdn3 Length DRF Length: 227 3 End 535 736 653 295 algarno Sequence of the Region Start Start DRF Space Upstream of the Start Codom Position Length 12 GAAGGGCCTCCTTGCCCTCT TG 297 408 6 CTCCACCCCCCCCGAGGATCC TTG 408 297 9 GGGAAACGACAAGAAGGCCCGT ATG 654 51 8 CCGTATGCGAGCGGGCGGGTTC ATG 672 33 W 12 C

Figure 8.9

- Is the predicted start codon the longest possible for the ORF without causing excessive overlap? According to Guiding Principle #3, bacteriophage genomes tend to have very small gaps between genes. Therefore, you may want to consider changing to a further upstream start codon if it provides better gene packing. For example, the start codon indicated by a red arrow in the figure above is worth investigating, because it would provide a longer gene without any overlap.
- Does the start site match other starts for similar genes in GenBank? To view the relevant information, 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.

In **Figure 8.10**, we're looking at our gene compared to two others. On the left is the alignment to gp1 of Bxz2, and on the right is the alignment to HelDan gp1.



Figure 8.10

The numbers highlighted with red circles are critical. The left pane shows a "1-to-1" match, meaning both proteins start at the same position—supporting evidence for this start. The right pane, however, shows a amino acid 1 of our protein matches amino acid 42 of HelDan gp1, meaning that gene's published start is further upstream. Because these differ, it's wise to check several BLAST alignments before making any decision. In this case, HelDan is the exception and most of the BLAST alignments are 1-to-1.

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.

You now need to put this information together to make the best choice. Occasionally, the answer to all five questions above will be "yes," and the auto-annotated start will clearly be the best choice. More commonly, the answers will conflict, and human judgment becomes necessary. This is the meat of the annotation process.

For our example:

By looking at the "Choose ORF start" window (shown in **Figure 8.9**), we see that there are 4 possible start sites for this gene. Numbers 3 and 4 can be easily dismissed because they both cut off some coding potential (counter to Guiding Principle #6) AND lead to very short genes (51 bp and 33 bp, respectively, counter to Guiding Principle #5).

That leaves us with two possible start sites to consider, one at 408 (called by Glimmer) and one at 297 (called by GeneMark). The SD scores for these are similar, though the one at 408 is a bit higher. Also, note that both utilize a TTG start codon.

In this instance, because the starts are rather similar in their features, it is also helpful to take a manual inspection of the upstream sequences. In this case, the GGAGGA, located about 4 bp upstream of the TTG start site at position 408 is a good match to the consensus sequence (remember Guiding Principle #10b).

Though the start at 297 would give a longer gene and was called by GeneMark, the best choice here is the start at 408. It was called by Glimmer, the RBS has a higher score and a better consensus sequence, the GeneMark coding potential begins near this start, and the majority of homologous genes have start sites at the same position.

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 8.4.3.	ACTION: You need to change this gene's start.
	Go to Section 9.2.3 for instructions.

8.4.3 Is this gene part of a programmed translational frameshift?

This is the first of three less common features you may come across during your annotation (followed by Introns and Wrap-around Genes).

Many dsDNA tailed phages encode a pair of genes that are expressed via a programmed translational frameshift, meaning the ribosome changes reading frame the in middle of translation. The resulting gene products are usually involved in assembly of phage tails, and the two genes are often just upstream of the tape measure protein gene.

The prototypical example for this is in phage lambda where the two genes are called 'G' and 'T'. The proteins expressed are, however, gpG (the product of the first gene), and gpGT, which starts at the beginning of 'G' (with the same start as gpG) but ends at the end of the second open reading frame. It accomplishes this by shifting translational reading frames about 8-9 codons upstream of the stop codon of 'G', and into the frame of 'T' (**Figure 8.11**).





There are a few quick ways to determine whether your gene might be part of a programmed translational frameshift.
- It is one of the two genes immediately upstream of the tape measure protein gene. The tape measure protein gene is almost always the longest gene in mycobacteriophage genomes, and is thus relatively easy to locate. Not all tape measure protein genes are preceded by frameshifts, however.
- A Phamerator map shows that homologues in finalized annotations of similar genomes are parts of a frameshift. In Figure 8.12, Angelica's frameshift is shown in the blue box, and appears as two genes that start at the same position but have different stops. Fionnbharth's annotation is just a draft, however, and so the equivalent region (shown in the red box) does not yet have the frameshift called properly.



Figure 8.12

• The GeneMark-Smeg output shows strong coding potential, but no way to call the second gene without significant overlap or excluding strong coding potential.

In **Figure 8.13**, we see that both genes have strong coding potential, but that to include all the coding potential for the gene that's the equivalent of lambda's 'T', we would have to choose a start site that would lead to a nearly 100 bp overlap between these genes (overlapping region shaded blue). Guiding Principle #2 tells us that this size overlap is very uncommon. The auto-annotated start, on the other hand, is too far downstream and misses strong coding potential. Seeing a situation like this on the GeneMark output should draw your attention and make you think of a possible frameshift.



Figure 8.13

For more on frameshifts in phages see the following paper.

'Conserved translational frameshift in dsDNA bacteriophage tail assembly genes' Xu J, Hendrix RW, Duda RL Mol Cell. (2004)16:11-21

DECISION TIME: Is the gene part of a programm	DECISION TIME: Is the gene part of a programmed translational frameshift?									
GUIDANCE: Each mycobacteriophage genome us common location is just upstream of the tape meas you suspect that you may have found a frameshift,	ually has only one or zero frameshifts. The most sure protein gene, but this is not a guarantee. If investigate it carefully.									
YES ACTION: You need to properly annotate this frameshift. Go to Section 9.4.1.	NO ACTION: Continue to Section 8.4.4.									

8.4.4 Does this gene contain an intron?

Even rarer than frameshifts in mycobacteriophage genomes are introns. We are only aware of perhaps two instances. Identification of potential introns is easiest using Phamerator maps that compare your genome to similar genomes. In **Figure 8.14**, the major capsid protein in Omega is labeled. In LittleE's genome, this protein appears to have been split into two pieces with a reverse-transcribed gene between them.



Figure 8.14

Of course, any tailed phage needs a functioning major capsid protein to be viable, so this situation in LittleE drew our attention. Subsequent experimental evidence has since verified that this is, in fact, an intron spliced out at the RNA level.

DECISION TIME: Does this gene contain an intron?

GUIDANCE: As mentioned above, introns are quite rare in mycobacteriophage genomes, so most of the time the answer is an easy "No." If you're lucky enough to locate a potential intron, you should use a variety of bioinformatic approaches to see if the putative intron shares any similarity to known introns. Experimental evidence may ultimately be needed for verification.

YES	NO
ACTION: You need to properly annotate this	ACTION: Continue to Section 8.4.5.
intron. Go to Section 9.4.2.	

8.4.5 Does this gene wrap around the ends of the genome?

In genomes that do not have defined ends, the left end of the genome is defined arbitrarily, and occasionally a gene may extend from the right end of the genome into the left end. We refer to these as "Wrap-around genes."

To even possibly be a wrap-around gene, an ORF must meet the following two criteria:

- 1. It is positioned at the right end of the genome.
- 2. It is transcribed in the forward direction.

If it does not meet these criteria, it is not a wrap-around gene, and you may proceed to the next steps.

If it does meet these criteria, you should check the GeneMark-Smeg output (**Section 5.3**) for strong coding potential near the right end of a genome with no stop codon present (see **Figure 8.15** below). The presence of strong coding potential implies that a wrap-around gene may be present.



Figure 8.15

DECISION TIME: Does this gene wrap around the ends of the genome?										
GUIDANCE: This question only applies at the extr most of the time it is not a concern. The GeneMar when they do exist.	reme right end of circularly permuted genomes, so k-Smeg output is critical for locating these genes									
YES ACTION: You need to properly annotate this wrap-around gene. Go to Section 9.4.3.	NO ACTION: You need to determine if your gene has a known function. Go to Section 10 .									

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 8.16**, we should expect some gap between gene 47 (transcribed leftwards) and gene 48 (transcribed rightwards).



Figure 8.16

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.

Several important steps remain, however.

- 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.
- 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** \rightarrow **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.

1	📴 Fionnbharth_/	Annotated								
2	Overview Features	References	equence Doc	umentation						
1	Sort By Start 👻] s	tart Stop	Length	^	Descripti	on Sequence	Prod	uct Regions	Blast Context
8	Select Features Dir	rect SQL 🜔	17 254	138		Name	1	_	GenelD	
1	T	2	51 481	231		T		_		
1	Type is All	4	78 645	168		Type	lcbs	_	<u>u</u>	
A	Name like	64	45 914	270		Start		117	Locus Tag	DNAM1

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.

Fionnbharth_Annotated	i				
Overview Features Reference	s∫Sequenc	ce Docu	mentation		
Sort By Start 👻 🔳	Start	Stop	Length	^	Description Sequence Product Regions Blast Context
Select Features Direct SQL	117	254	138		Name 1 GenelD
Turne in All	251	481	231		
Type is All	478	645	168		
Name like	645	914	270		Start 117 Locus Tag DNAM1
GenelD =	911	1273	363		Stop 254 Regions 1
	1070	1050	607		

Figure 9.2

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 follow are ways that your changes will **not be posted** to the database, and **WILL BE LOST**.

- 8 Saving your file while still in Insert Mode.
- ³ 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".

DNA etude									
Overview	Feat	ures	Reference	es	Sequence	Docume	ntation		
Sort By	Index	•	- I		Name	Start	Stop	^	Description Sequence Product Regions Blast Context Validation
Select Fe	atures	Dire	ect SQL		1	83	466		All OBEs appear valid
T		All			2	475	849		
туре	IS				3	850	1053		
Name	like	<u> </u>			4	1059	2636		
GenelD	=				5	2626	4296		
Locus	like			Ш	6	4318	5373		
Start	>	í –		Ц	7	5400	6038		
Length		<u> </u>		Ŀ	8	6239	6317		
Cengar		-		Ц	9	6344	6607		
Hegions				Ц	10	6647	6841		
% GC	<			Н	11	6902	7360		
CAI	>			Н	12	8315	9361		Control Numbering
EC#	like			Н	13	9411	10655		Assign Names builder of appearance
Product	like	i—		Н	14	11265	11486		Assign Products by order of appearance
Emotion	EL	-		Н	15	11588	12100		Assign Houses
Function	іке	<u> </u>		Н	15	12233	12637		Overwrite existing values Fig. Overwrite de fault Nersen & Des duste auto
FeatureID	=	<u> </u>		Н	17	10/00	13430		Uverwrite derault Names & Products only
Tag	like			۲	10	13403	14005	v	Assign new Locus Tags by order of appearance
□ Hide I	anore	d Feal	tures	<			>		Locus Tag Prenx ETUDE Search NUBI database
			1	1	nsert De	lete Po	st Valid	iete	Reassign Gene Data Locate Gray Holes > 300 🜩 bp
Sele		-eatur	PC	<u> </u>					
	Q	ર ▶		1-1	14998	F	osition : 1	619)6 🔽 Controls >> Map 🔽 Map >> Controls
1 2 3		4	X	5		6)	7> 🍳		
20 Feature:	s	Live	э.						14998 😸 ?

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 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** below.

DNR etude								
Overview	Featu	res Re	ferences	Sequence	Docum	entation		
Sort By	Index	-		Name	Start	Stop	^	Description Sequence Product Begions Blast Context Validation
Select Fe	ahiras	Direct		1	83	466	-	
00000110		Directs		2	475	849		All UNP's appear valid
Туре	is /	41	-	3	850	1053		
Name	like			4	1059	2636	-	
GenelD	- [— F	5	2626	4296		
Locus	like		-1	6	4318	5373		
Start			-[7	5400	6038		
Stat			<u> </u>	8	6239	6317		
Length			_[9	6344	6607		
Regions	>			10	6647	6841		×
% GC	<			11	6902	7360		3
CAJ			— L	12	8315	9361		Control Numbering
EC#	Re I		— I.	13	9411	10655		
Destant	n I		— I.	14	11265	11486		Assign Names by order or appearance
Product	ike			15	11588	12100		Assign Products by order of appearance
Function	like		ŀ	16	12239	12637		Overwrite existing values
FeatureID	- [17	13031	13438		Verwrite default Names & Products only
Tag	like		-	18	13469	14065	~	Ssign new Locus Tags by order of appearance
	Innored	Feelung		6		>	-	Locue Tag Prefix ETUDE Search NCBI database
1 Hidei	gnoreu	reatures		Jacob Da		and INTEREST	101	Reassign Gene Data 🕽 ocate Gray Holes > 300 🜩 bp
Sala	nt ôll Fr	a a hi ma e		insert De	iete P	Use Valua	<u>(6)</u>	
H 44 4			E 🕅 1	14998		Position : 16	\$19	36 I Controls >> Map I Map >> Controls
1 2 3	- 4		5		6	7) (9)	6	
20 Features	9	Live					-	14998 🚜

Figure 9.5

- In the field marked "Locus Tag Prefix", type your phage's name. (GenBank assigns a unique locus tag to every gene in GenBank, preferably constructed from the phage's name and gene number.)
- Click the 'Reassign Gene Data' button.
- Click '**Yes**' to confirm in the window that pops up.

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.

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.

👺 Fionnbharth_checked_	WP				
Overview Features Reference	s Sequence	Docume	ntation		
Sort By Index 👻 🔳	Name	Start	Stop	^	Description Sequence Product Regions Blast Context
Select Features Direct SQL	1	117	254		Score Target Description
Turn in All III	2	251	481		1447 putative terminase gp4 [Mucobacterium phage T
Type is pair 💽	3	478	645	Ξ	1405 gp8 [Mycobacterium phage JAWS]
Name like	4	645	914		1402 gp8 [Mvcobacterium phage CrimD]
GenelD =	5	911	1273		1396 gp7 [Mvcobacterium phage Pixie]
Locus like	6	1273	1959		1395 gp8 [Mvcobacterium phage Adephagia]
Start N	7	1985	2224		
	8	2221	2493		BLAST Hit
Length	▶ 9	2465	3910		Accession NP_569740 Export
Regions >	10	3922	5508		GI 18496890 Export All
% GC <	11	5489	8248		Length 474 Dete 10/5/2011
	12	8245	8442		Max Score 1447 Date 10/5/2011 Delete All
ECH Bu	13	8534	9106		High-Scoring Pairs (HSP)
EC# IKe	14	9165	10094		HSP Data Alignment
Product like	15	10234	10608		1 MGLGFDRUQD DLGKLICAKR PDGLYAADMF GMS
Function like	16	10611	10967		1
FeatureID =	17	10954	11214		36 MGLGFDLWQD DLGKLICAKR DDGLYAADMF AMS
Tag Re	18	11211	11636		51 LCVAAVDRPL TVIWTAHRTR TAAETFKSMQ GF3
rag into j				~	51 + + +
Hide Ignored Features			>		86 LCIKTPNT TVIWTAHRTR TAAETFRSMQ GLJ
Select OII Features	Insert De	elete Po	st Valida	te	
⋈⋈⋬⋐⋐⋟⋟⋟	1 - 50000	P	osition : 1		Controls >> Map 🔽 Map >> Controls
THEIR 9 10 11 VERALING	8 22	23 24 6	28	5 38	
95 Features live			-71 × 7	171	
uvo uvo					00010 10 1

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	bharth	_checked_\	NP						(X
Overview	Feature	s Reference	s Seque	nce Docu	umentation						
Sort By	Index		Name	Start	Stop	^	Description Sequence Product Re	gions	Blast	Conte	st
Select Fe	atures	Direct SQL	1	117	254		There are no BLAST results for this featu	ure			.
-	·		2	251	481		Blast this game Blast ALL Gam	ae	Cle	ar áll	
туре	IS A	<u> </u>	3	478	645						
Name	like		4	645	914						
GenelD	=		5	911	1273						
Locus	like		6	1273	1959						
Start			7	1985	2224						
i a			8	2221	2493						
Length) 9	2465	3910						
Regions	>		10	3922	5508						
% GC	<		11	5489	8248						
CAL			12	8245	8442						
EC#	like -		13	8534	9106						
EU#	like		14	9165	10094						
Product	like		15	10234	4 10608						
Function	like		16	10611	1 10967						
FeaturelD			17	10954	4 11214						
Tag	like [18	11211	1 11636						
rug	ince		(100)			~					
Hide	Ignored F	eatures				J.					
Sele	oct All Fea	ahures	Insert	Delete	Post Valid	ate					
H 44 4	€ Q	H H	- 50000		Position : 3	37354	4 🔽 Controls >> Map 🔽 Map >> C	ontrols			
6 9	10 11	30 4 4 10 1	22	23 24	4 2 26 28	29	8 31 8 136 1444 454 1115 8	Б Б4	is (r Kal	74	11
95 Feature	s l	Live							5807	6	3 ?

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**.

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	VP_569740
1474	gp8 [Mycobacteriun	n phage	JAWS]			GI 1	8496890
1471	gp8 [Mycobacteriun	n phage	: CrimD]			Length 4	174
1466	gp8 [Mycobacteriun	n phage	Adephagia]			Max Score 1	567
1463	gp6 [Mycobacteriun	n phage	Larva]				
1456	gp7 [Mycobacteriun	n phage	Pixie]				
1449	gp8 [Mycobacteriun	n phage	Angelica]			HSP Co	verage Map
1442	gp9 [Mycobacteriun	n phage	Anaya]			1101 00	rorago map
1151	terminase [Segnilipa	rus rotu	indus DSM 4498	5]			
1040	0.04 1 1 1	1	×			<u> </u>	
E-Value 0.0E0	Starts 1	Bit Sc Score E-Vali	-sconing Pair (HS :ore 608.221 : 1567 ue 0.0E0	Query Target Length % Aligned	1 - 453 1 - 446 453 94.1	Positive Identity Similarity Gaps	rs 383 340 y 85.9 7
		1	MSNSTATLTD	VARHVVAPTG	IVSTGFSAV	R ATCRHMGLGF	7 DRWQDDLGKI 🔨
		1	++ ++ MNHSTATLSE	+ VARHVIAPQG	+ + IVSTAWPSV	 R ATCGAMGLGF	7 DLWQDDLGKI
		51	ICAKRPDGLY	AADMFGMSIP	RQTGKTYLL	G AIVFALCVAA	VDRPLTVIWI
		51		11111 1111	111111111	+ +	+
		51	ICARRODGLY	AADMFAMSIP	RUIGKTYLL	G ALVFALUIKI	PNIIVIWJ
J		101	AHRTRTAAET	FKSMQGFARM	PRIEPYIEK	V SLARGEEAVI	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 10⁻³) that also gives you a useful number of matches (at least 10 or so). If you only have E-values higher than 10⁻³, 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).

Fionn	bhart	h_c	hecked_\	NP											
Overview	Featu	ures	Reference	s Seq	uence D	ocume	ntation								
Sort By	Index	-	•	Nam	ie S	art	Stop	^	D	escription	Sequence	Product	Regions	Blast	Context
Select Fe	atures	Dire	ect SQL	1	1	7	254		Г	Score	Target Des	cription			
Тире	ie	611		2	2!	51	481		Þ	1567	putative ter	minase gp	4 [Mycoba	cterium	phage T
туре	10		<u> </u>	3	4	78	645		F	1474	gp8 [Mycob	acterium p	hage JAW	(S]	
Name	like			4	6	45	914		F	1471	gp8 [Mycob	, acterium p	- ohage Crim	D]	
GenelD	=			5	9	1	1273		1466 gp8 [Mycobacterium phage Adephagia]						
Locus	like			6	1:	273	1959		F	1463	gp6 [Mycob	acterium p	hage Larv	a]	
Start				7	1	985	2224						_	-	
Length	Ξi			8	2	221	2493		Г	BLAST Hit					Event
Lengar				<u>▶</u> 9	2	465	3910		A	ccession N	NP_569740				Export All
Regions	>			10	3	922	5508			il 1 onath (18496890				Delete
% GC	<			11	5	189	8248		L N	tax Score 1	+74 1567	Date	10/8/2011		Delete All
CAI	>			12	8	245	8442			Link Coorin	n Daire (UCD				Delete All
EC#	like			13	8	534	9106			High-Scollin	Alianment	1			
Product	liko			14	9	65	10094		١.,	HSPData	Alignment				
Tioduct				15	11	J234	10608			1 MSN	NSTATLTD	VARHVVA	PTG IVS	TGFSA	VR ATC 🔨
Function	like			16	11	J611	10967			1 1111	+ ++ HSTATLSR '	+ VADHUTA	POG TVS	+ + TAMDS'	11 111
FeaturelD) =			17	11	J954	11214								
Tag	like			18	1	211	11636	-		51 ICA	AKRPDGLY .	AADMFGM	SIP RQT	GKTYL	LG AIL
- Hide I	, Ianored	Foot	TITOS	< 1			>			51 51 TC	NADDCTA	11111 885 M R 8 M	 פדם סחד	CRIVI.	+ LC ALL
j mae	ignored	neat			1	. n.	يتا يد			01 101	anoponi .		oir ngi	011111	
Sele	nt All F	eatur	P0	Inser	Delec			ate	Ľ					_	
H 4 4	€	₹. ►	H H	- 5000	0	P	osition : 2	3111		Cont	trols >> Map	🔽 Map >	> Controls		
95 Feature	10) 1 s	11 Live	₩)44)]]])))	ß	22 23	24	26 28	¥9	3	1)8) 36)	45		5 8 1 6 64	1 16) / / 580	74 74

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.

DNA Six I	ix Frame Translation of Fionnbharth_Annotated	
View Ma	V 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	DAIRRDLIV
	G Å R T V Å S S R F N P P N L R G G * Å G R G	TRFAATSS ; RDSPRPHR
12815	15 TGGGGCAAGGACGGTGGCCTCATCGAGGGAAAATCCCACCAAATCTGCGCGGCGGTTAGGCAGGTAGAG	GACGCGATTCGCCGCGACCTCATCG
450.00	·····	
45262	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
	P & L V T & E D L S F G G F R R P P * A P L	PRSEGRG*R
	GTLSWDDLYAFIFAAPPNTA <mark>VFH</mark>	IAYEKGWIT
	A R * A G T T F T H S S S L H R R I P L C S	M P T K R A G S Q
12930	R H A E L G R P L R I H L R C I A E I R C V P 30 CGGCACGCTGAGCTGGGACGACGTTTACGCATTCATCTTCGCTGCACCGCCGAATACCGCTGTGTTCC	C L R R G L D H
		. : :
45147	47 GCCGTGCGACTCGACCCTGCTGGAAATGCGTAAGTAGAAGCGACGTGGCGGCTTATGGCGACACAAGG	TACGGATGCTTTTCCCGACCTAGTG
	V A R Q A P V V K V C E D E S C R R I G S H E	MGVFLAPD
	RCASSPRGKRM*RRQVASYRQTG	HRRFPSS*

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	bharth_A	nnotated											
Overview	Features	References	Sequence	Documer	ntation								
Sort By	Start 🖉 👻] •	Name	Start	Stop	^	Descriptio	on	Sequence	Prod	uct Regions	Blast Cont	ext
Select Fe	atures Din	ect SQL	5	911	1273		Name	21		_	GenelD		
T	:. All		6	1273	1959		T				<u></u>		
Type	IS AII	<u> </u>	7	1985	2224		туре	Ju	15	-	<u>u</u>		
Name	like		8	2221	2493		Start		12	2991	Locus Tag	DNAM201	
GenelD	=		9	2570	3910		Stop		13	3311	Regions		2
Locus	like		10	3922	5508		Length	321			Тад		_
Chart			11	5489	8248		Disastina	-			3	,	
Stalt			12	8245	8442		Direction	FOR	waru	<u> </u>			
Length			-				Translatic	nn T	abla Undef	ined			-11



- 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.

Extrac	👺 Extracted from FastA Library Fionnbharth.fasta														
Overview	Featu	ures	References	Sequenc	e Do	cumer	ntation								
Sort By	Index	-		Name	Sta	rt	Stop	^	Desci	iption	Sequence	Product	Regions	Blast	Context
Select Fe	atures	Dire	ect SQL	18	112	11	11636		Sta	rt	Stop	Lend	ath		
Tuna	:	All		19	117	17	12328		*	124	31 12	844	1		
туре	15	AII	<u> </u>	20	124	31	12880		<u> </u>			i			
Name	like			21	129	91	13311								
GenelD	=			_ 22	133	14	17489								
Locus	like			23	175	08	18623								
Start				24	186	23	20410	-							
l				25	204	11	20929								
Length	2			26	209	26	22023								
Regions	>			_ 27	220	35	22280								
% GC	<			28	222	73	24582								
CAL	>			29	245	86	25590								
EC#	liko			30	256	12	25989								
EC#	iike			31	260	20	27630								
Product	like			32	276	30	28421								
Function	like			33	284	59	28857								
FeatureID	=			34	288	54	29189								
Tag	like			35	291	86	29446								
				(a)				<u> </u>							
🔲 Hide I	gnored	Feat	tures				>		-		1	. 1			
Sele	∼t ∆ll F	eatur	-P-1	Insert	Delete	Pos	t Valida	ate		Insert	De	lete	Assign Len	igths	
444	€ e			376 - 15625	;	P	osition : 1!	5098	3	Con	trols >> Map	🔽 Мар	>> Controls		
14	5		15 16	17 1		1	9		20	21	X		22		N
95 Features	;	Live	э / · · · ,	1 . 7							4			580	176 🛃 ?

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.

🔀 Extra	ted f	rom	FastA Lil	orary Fig	nnb	hartl	h.fasta										
Overview	Featu	res	References	Sequer	ice 🛛 I	Docun	nentation										
Sort By	Index	-	•	Name	S	tart	Stop	^	De	scription	Se	equence F	Product	Region	ns Blas	st C	ontext
Select Fe	atures	Dire	ect SQL	18	1	1211	11636		П	Start		Stop	Lena	th [~
T	. [All		19	1	1717	12328		H	124	131	1284	4	414			
туре	15	All	<u> </u>	_ 20	1	2431	12880		F	128	344	1331	1	468			
Name	like			21	1	2991	13311		۲								
GenelD	=			_ 22	1	3314	17489							î _			
Locus	like			23	1	7508	18623										
Start				24	1	8623	20410						~				
Lawath				25	2	0411	20929						Ca	icula	ted		
Length	2			26	2	0926	22023							engtl	าร		
Regions	>			27	2	2035	22280										
% GC	<			28	2	2273	24582										
CAI				29	2	4586	25590										
EC#	like		——I	30	2	5612	25989										=
EC#				31	2	6020	27630										
Product	like			32	2	7630	28421										
Function	like			33	2	8459	28857										
FeatureID	[34	2	8854	29189										
Tag	like			35	2	9186	29446										
T USA		E.e.sk		2			>	×									~
	gnored	геа	ules			1.	• •	. 1	-	Incort		Delete		Xaaign I	ongtheil		
Sele	ot All F	eatur	e?	Insert	Dele	te F	'ost Valida	ate	_	Insert	_	Delete		Masiyin L	enguis;)	
HHI	 €	٤ 🕨	H H S	376 - 1562	25		Position : 1	3717		Con	trols	s>> Map 🔽	Map :	>> Contr	ols		
14	\rightarrow	-	15 16	17	18)		19		20	21	$\left \right\rangle$			22			
95 Feature:	s	Live	•												5	8076	₿ ?

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 f	rom	FastA L	ibı	ary Fio	nnbharth.	fasta							
Overview	Featu	ures	Referenc	es	Sequence	e Docume	entation							
Sort By	Index	-	Ŀ	1	Name	Start	Stop	^	Descript	tion Seque	ence Prod	luct Regions	Blast Co	ontext
Select Fe	atures	Dire	ect SQL		18	11211	11636		Name	21		GenelD		
Tupe	in	ΔII	-		19	11717	12328		Tupe	ICDS	_	GL		
News	15	~"		Ŀ	20	12431	12880		Туре	LD3	10101	<u>.</u> -	DALANGON	
Name	like			₽	21	12431	13311		Start		12431	Locus Lag	DNAM201	
GenelD	=				22	13314	17489		Stop		13311	Regions		1
Locus	like			L	23	17508	18623		Length	882		Tag		
Start	>			Ŀ	24	18623	20410		Direction	n Forward			, 	
Length	πi			Ŀ	25	20411	20929		Translat	ion Toble []				
				Ŀ	26	20926	22023		EC Norma		Dridenned			
Regions	>			Ŀ	27	22035	22280		EUNUM	iber				
% GC	<			L	28	22273	24582							
CAI	>			Ŀ	29	24586	25590		Product					
EC#	like			Ŀ	30	25612	25989		gpzi					-
Deadlast				Ŀ	31	26020	27630							\leq
Product	пке			L	32	27630	28421		Function	1				
Function	like			L	33	28459	28857							
FeatureID) =			L	34	28854	29189							\sim
Tag	like			Ŀ	35	29186	29446		Notes	or			0.00	
- 10a.1				l.	e (mm)		>		Uriginal	Glimmer ca	II @DD 129:	11 has strengtr	19.38	\simeq
Hide I	Ignored	1 Feat	ures	Р				1						
Sele	of All F	eatur	es		Insert	Delete Po	st Valid	ate						
H 4 4		2	H H	97	47 - 15996	S F	Position : 1	5905		Controls >>	Map 🔽 M	lap >> Control:	s en la	
14	15	10	6 17		18	19		21	K			22		
95 Feature	s	Live	•										58076	32

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.)

👯 Fionn	bhart	ih_ci	necked_	W	þ															
Overview	Feat	ures	Referenci	es	Sequence	e Doo	umenta	tion												
Sort By	Index	-	•		Name	Star	SI	op	^	Descripti	on Se	equence	Product	Regions	Blas	t Context				
Select Fe	atures	Dire	et SQL]		5	911	12	273		Start		Stop	Lend	1th						
-		0.0		Ω	6	1273	3 19	959			12431	128	44	414						
Type	IS	All	-		7	1985	5 22	224			12844	133	11	468						
Name	like				8	222	24	193	_	۳		100		100						
GenelD	-				9	246	5 39	910												
Locus	like				10	3922	2 55	508	Ξ											
Start		<u> </u>			11	5489	9 82	248												
Longth				L	12	8245	5 84	142	-											
Lengin					13	8534	1 91	06												
Regions	>				14	916	5 10	0094												
% GC	<			L	15	1023	34 10	0608												
CAI	>			L	16	106	1 10	0967												
EC#	like	<u> </u>		H	17	109	04 11	214												
Product	lika	<u> </u>		H	18	112	1 11	636												=
T IOUUCI		<u> </u>		H	19	117	7 12	2328												
Function	like			Ŀ	20	124.	51 12 21 12	2880												
FeatureID	-			Ľ	21	124	81 13 14 15	2400												
Tag	like			H	22	1740	4 1/ 00 10	403												
□ Hide I	lanored	d Featr	res	H	23	1001	20 10	023 0410												
			1	H	25	204	1 20	1929												~
Sele	ect All F	eature	es	H	20	204		1020	~											
				<				>												
					Insert [Delete	Post	Valida	ite	Ins	ert	Dele	e	Assign Lei	ngths					
। स स्थित			H	1 -	50000		Posi	ion : 25	5028		Control	s>> Map	🗸 Map	>> Control	s 🗖					
	- 10	V.		14	RATE AS	Ŋ	22	123	24	Mac V	28	298 31	ASAR .	36	NA II	45 46 4 1 6 6		4116917	1 74	KRAA
95 Feature:	s	Live	216-7	17	n na an	UI -		<u>a-a</u>		2101-211		1 75/	(1700)	2000.03			A COTATATA CAL	20000710	58076	a ?

Figure 9.18

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.

9.4.3 Annotating wrap-around genes

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.

9.5 Predicting tRNA and tmRNA genes

DNA Master's Auto-Annotate feature runs the tRNA search tool **Aragorn**, 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. There is a newer, 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.

9.5.1 Running web-based Aragorn (version 1.2.28)

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

Online serv	vices dRXA
ARAGORN	ARAGORN, tRNA (and tmRNA) detection in nucleotide sequences
ARWEN	
BRUCE	compilation with a C-compiler (for example with gcc -O3 -ffast-math -finline-functions -o aragorn aragorn1.2.33.c; ignore any warnings
optalign	about trigraphs).
RAMI	Input sequence (both strands will be searched, max. 15 MB)
External links tRNAscan-SE Sprinzl compilation	Upload a fasta file with one or several sequences: Browse
	Select options (see here for all options in the standalone version) Search for (default tRNA): Search allowing introns, 0-3000 bases (default no): Sequence topology (default linear): Strand(s) (default both): Dutput format (default standard): Standard
	Notes ARAGORN detects tRNA and tmRNA in nucleotide sequences. The tmRNA detection algorithm is an development of the earlier published BRUCE algorithm. If your primary interest is in tmRNA detection you should also try BRUCE (see left panel). BRUCE is slightly more selective than ARAGORN, but on the other hand it may in rare cases miss a tmRNA. Note that with a short sequence and a selection of circular topology two tmRNA hits may be reported. If you want to search online for IRNAs in metazoan mitochondrial sequences, please use ARWEN. ARWEN is incorporated in the standalone version of ARAGORN. Please use linear topology when using short sequences (< 500 bp). These are not circular anyway.

Figure 9.19

- In the '**Input Sequence**' section, click '**Browse**...' then select your phage's DNA sequence as a FASTA file.
- Choose the following settings:

Search For: tRNA & tmRNA

Search allowing 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
a-t
                         g-c
                                      tc
             t tgcc a
a g :: 1 g
agcg tgcg c
1:11 c tt
        gta
             tggc ggg-c
g a g c-g
gaa g-
                       q.q
                                    q-c
                                      q+t
                                          g-
                       ccg
       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 1.21)

- Go to: <u>http://lowelab.ucsc.edu/tRNAscan-SE/</u>
- Next to the field labeled "or submit a file", click the 'Browse...' button and select your phage's DNA sequence as FASTA file.
- Choose the following settings:

Search mode: **Default**

Source: Bacterial

A note about settings: For most genomes, these default settings generate reliable results. However, you can always relax the parameters if you come across a suspicious area.

It is recommended that you run tRNAscan-SE using most of the website's default parameters, with source set to "Bacterial" (as described above) **UNLESS** your phage is a member of Cluster C. To date, only Cluster C mycobacteriophage genomes have been shown to include these non-canonical (or pseudo-) tRNA sequences.

The relaxed settings include changing the Search Mode to "Cove only", and setting the "Cove score cut-off to "2", as shown in **Figure 9.21**.

Search Mode: Cove only (very slow)	Source: (Bacterial +)
Format:	
Raw Sequence Sequence name (optional): (no spaces)	
O Other (FASTA, GenBank, EMBL, GCG, IG)	
Paste your query sequence(s) here:	
(Queries are limited to a total of less than 5 million nucleotides at any one time) Or submit a file: Clear Sequence ③ Show results in this browser. Browse ③ Receive results by e-mail instead: @ Run tRNAscan-SE (Clear Form) Clear Form	
Extended Options:	
 Disable pseudo gene checking Display results in ACeDB format Show false positives from tRNAscan/EufindtRNA Show primary and secondary structure components to Cove scores Genetic Code for tRNA lsotype Prediction: <u>Universal</u> Default cut-off values should only be changed for exceptional conditions Cove score cutoff: 2 	 Show origin of first-pass hits Show codons instead of tRNA anticodons
EdindRNA search parameters: Relaxed	Intermediate score cutoff:

Figure 9.21

The output from this program looks like the sample in **Figure 9.22**, and when you click on the '**View tRNA**' button, you will view tRNAscan-SE's interpretation of the secondary structure (**Figure 9.23**).

Results

Sequence Name	tRNA #	tRNA Begin	Bounds End 	tRNA Type	Anti Codon	Intron 1 Begin	Bounds End 	Cove Score
Bongo	1	54782	54852	Trp	CCA	0	0	54.55
View tRNA								

Figure 9.22



Figure 9.23

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.

It is recommended that both tRNAScan-SE and Aragorn be run on every sequence.

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.

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.

9.5.4 Entering a tRNA in DNA Master

DNA Master may have already called some of your tRNA genes. 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)
- Start & Stop: Exact coordinates as determined above

- Feature Product: "tRNA _____" (In the blank, write the amino acid 3-letter abbreviation, e.g. "Lys".)
- Feature Notes and Function: "tRNA _____" (In the blank, write the amino acid 3-letter abbreviation followed by the anti-codon, e.g. "Lys (ttt)".)

Fionn	bharth	.fasta											
Overview	Feature	es Referenc	es	Sequence	Docume	ntation							
Sort By	Index	•		Name	Start	Stop	^	Descriptio	on Sec	quence Pro	duct Regions	Blast C	ion 💶 🕨
Select Fe	atures	Direct SQL		32	27630	28421		Name	38		GenelD		
Tupo	in A		L	33	28459	28857		Tupe			GL		
Type		··	L	34	28854	29189		iype	UTINA	-	<u>.</u>		
Name	іке			35	29186	29446		Start		30843	Locus Lag	DNAM5_38	;
GenelD	=		L	36	29433	30623		Stop		30918	Regions		1
Locus	like		L	37	30626	30907	-	Length	76		Tag		
Start			Ĩ	38	30843	30918		Direction	Bevers	e 🚊	-		
Length	τ,		L	39	31057	31383		Tranalatio	vo Tiobla				
Lengun			L	40	31415	31714		Fent	in Labie	Torridenned			
Regions	>		L	41	31711	32013		EC NUMB	er				
% GC	<		L	42	32006	32395							
CAI	>		L	43	32392	32823		Product					
EC#	like		L	44	32896	33627		INNA LY	2				-
Destation			L	45	32922	33443							\sim
Product	like		L	46	33681	34946		Function	(111)				
Function	like		L	47	34927	35745		tHNA Ly:	s (ttt)				\sim
FeatureID) = [L	48	35830	36093							\sim
Tag	like [L	49	36320	36565		Notes					
⊢ Hide I	lanored F	eatures	<			>		tHNA Ly:	s (ttt)				\sim
Sele	ort All Fe	ahures		Insert De	elete Po:	st Valid	ate						\sim
H 4 4			30	078 - 31175	P	osition : 2	9640		Controls	>> Map 🔽	Map >> Control	s	
			36				У		37	3	8		39
96 Feature	s	Live					6					58076	3?

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.

DNA Extract	ed fro	m Fas	tA Libra	ry	Timshel.				
Overview	Featu	res R	eference	s	Sequence	Docu	mentation		
Sort By	Index	-	4		Name	Start	Stop		Description Sequence Product Regions Blast Context
Select Fe	atures	Direct	SQL		1	408	704		Name 2 GenelD
-		A II		Þ	2	743	1177		
Type	IS P	All			3	1254	1577		
Name	like				4	1567	2319		Start 743 Locus Tag DNAM11
GenelD	= [5	2345	3565		Stop 1177 Regions 1
Locus	like				6	3592	4380		Length 435 🗐 Tag
Start					7	4377	5294		Direction Forward
Start						5374	5449		
Length					8	5547	6899		Translation Table Undefined
Regions	>				9	6896	7309		EC Number
% GC	<				10	7306	8283		
CAL	ъİ				11	8479	10005		Product
с					12	10002	11459		gp2
EU#	ike j				13	11456	12469		V
Product	like				14	12520	13020		Function
Function	like				15	13055	i 14014		
FeatureID					16	14080	14268		V
Teg	like [17	14276	14650	/	Notes
rag	me l		_					ᆂ	Original Glimmer call @bp 743 has strength 6.35
Hide I	gnored	Feature	es	1	LI,			1	
Sele	ot All Fe	eatures	1		Insert D	elete	Post Valid	ete	· · · · · · · · · · · · · · · · · · ·
HAA			H H	•	50000		Position : 1		Controls >> Map Map >> Controls
2 4 5 6	7 8		1)12/13	N	5	2 25	26 27	N.	132 434 MTB1 44 MAX 49 605 511555 56 67 6 5 17 16 1111 6
89 Features		Live			20.001.0.20			2 1 2 1	53278 🛃 ?

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.

- 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.)
- Any significant gap or overlap with preceding gene (in basepairs).
- Whether or not the gene was called by Glimmer and GeneMark, and if the start was called by same.
- Whether or not the coordinates you have chosen yield the longest possible gene for that ORF.
- Whether or not your start includes all the coding potential identified by GeneMark.
- Whether or not the start has the best SD score of all this ORF's possible starts.
- 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".)

- 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")
- Gene Function, and source for the function (see **Section 10**). If the function assignment comes from a Hatfull-approved map in the Appendix, please also enter it into the field labeled "Function" directly above the "Notes" field. Otherwise, only enter the putative functional assignment in the Notes.
- 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.

An example of good Notes:

Start: 2435 Stop: 2650 (FWD). ORF Length: 213 bp; longest possible ORF. SD Score: 310, best score. Gap or Overlap with Previous Gene: 84 bp gap. Gene Predictions: Agrees with both Glimmer and GeneMark predictions. Coding Potential Support: ORF includes all coding potential shown on GeneMark-Smeg output. Best BLAST match: gp3 of Oline; Oline aa 1 aligns with query aa 1. Predicted Function(s): 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.

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/

10.2 Using bioinformatic tools to assign gene function

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

- 1. BLASTP
- 2. Conserved Domain Identification (either through NCBI or Phamerator)
- 3. HHpred

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 two basic ways of doing BLASTP searches. They can be done within the DNA Master environment, or they can be done using the NCBI BLAST server. The web address for this program is:

http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins

When you BLASTP your protein sequence, you are comparing it to all the other protein sequences in GenBank. One important thing to remember is that anyone can submit information to GenBank—whether it is correct and high-quality or not—so any GenBank hits that provide putative gene functions must be carefully considered.

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^{-4} or less. This is the perhaps the most important factor to consider, and if this condition is not met, you should not assign a function regardless of what kinds of functions appear in the results list.

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.



Figure 10.1

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^{-4} . And 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.

10.2.2 Conserved Domain Database

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 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 are 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.

Caraphic Summary	
V Show Conserved Domains	
Putative conserved domains have been detected, click on the image below for detailed results.	
75 1 ⁴ 225 340 375 4 ⁶ 54	
superfamilies Terminase_1 superfamily	

Figure 10.2

The same gene in a Phamerator map might look like:



Figure 10.3

In this case, we moused over gene **15** in **Figure 10.3**, and the green box describing the domain appeared.

A less informative match on NCBI might look like:

•	Graphic Summary
	Show Conserved Domains
	Putative conserved domains have been detected, click on the image below for detailed results.
	Hulti-domains Phage_prot_6p6

Figure 10.4

We already know that this is a phage protein, so this is not particularly useful information.

And the same gene in Phamerator:



Figure 10.5

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

HHpred is accessible at:

http://toolkit.tuebingen.mpg.de/hhpred

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

An example of an informative HHpred match:



Figure 10.6

Like BLAST, HHpred provides a graphical view where the best matches are shown in red and lower-quality matches are dark blue or 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, "2kng_A" (this is the PDB designator, if you want to see the crystal structure), matches your protein with 99.8 % probability and an E value of 2.1e⁻¹⁹.

Good HHpred matches have high probabilities (80 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 80%, supporting data (such as a conservation of a domain, or a function found in other mycobacteriophages) 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:
No 1	
>2kng_A Protein I Probab=99.77 E-	SR2; DNA-binding domain, immune response, DNA binding protein; NMR {Mycobacterium tuberculosis} value=2.1e-19 Score=117.31 Aligned_cols=47 Identities=38% Similarity=0.639 Sum_probs=0.0
Q ss_pred Q Sun_Sep_04_13: Q Consensus T Consensus T 2kng_A T ss_dssp T ss_pred	cccCCCCCCCcccchHHHHHHHHHCCcCCCCCCCCCCHHHHHHHH
No 2 >2zqe_A MUTS2 pro {Thermus thermophi Probab=83.45 E-va	Notein; alpha/beta, ATP-binding, DNA-binding, nucleotide-binding, DNA binding protein; 1.70A lus HB8} lue=1.2 Score=29.26 Aligned_cols=55 Identities=22% Similarity=0.060 Sum_probs=0.0
Q ss_pred Q Sun_Sep_04_13: Q Consensus T Consensus T 2zqe_A T ss_dssp T ss_pred	EEECCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH

Figure 10.7

The first line of text describes what "2kng_A" is: the LSR2 protein from *M. tuberculosis*. This is encouraging in two ways: first, it matches a mycobacterial protein; and second, similar mycobacterial proteins have already been found in other mycobacteriophages.

Under the line of text is the alignment. The top line is the secondary structure prediction for your query sequence, the second line is your query sequence, and the third line is the consensus sequence that was built from the iterative search. The bottom part of the alignment refers to the subject sequences (in this case the crystal structure data), their consensus sequence, and secondary structure prediction or determination from two algorithms: PSI-PRED, and DSSP.

For more on interpreting HHpred results, see:

http://toolkit.tuebingen.mpg.de/hhpred/help_results

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. Dr. Hatfull periodically reviews the mycobacteriophage genomes and assigns gene functions to the best of his knowledge. 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 on one of his maps, you may assign your gene the same function. The most recent version of these maps, one per cluster, are included in the Appendices of this guide.

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 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, try to compare to recently-published genomes.

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



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.)



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 Files		
[Select Features] present in	at least 2 🚔 files	Merge Files
Toggle Selected Clear Se	lections	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		~
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.

Carlo Carlo	and the star	CARLES	1.557.85.0	200	ALC: NO.	17. 80 J. 200	ALL CAR	1	Carl Carl Carl Carl Carl	
Merged Sequence										
Overview Features Reference	es Sequence	Docume	entation							
Sort By Index 👻 🗹	Name	Start	Stop	^	Description	Sequenc	e Prod	uct Regions	Blast Context	
Select Features Direct SQL	▶1	488	1177		Name 1			GenelD		
7	2	1333	1926		T T					
Type is All	3	1926	2147	_	Type U	DS		<u>61</u>		
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 69	n		Tag		
Shart N	7	4662	5360		Direction For	-		, ag	'	
	8	5403	5474		Direction Fo	Iwalu	<u> </u>			
Length	9	5480	5552		Translation 1	Table Unc	lefined			
Regions >	10	5595	5666		EC Number					
% GC <	11	5697	5855							
	12	5855	7066		Product					
	13	7066	7500		gp1				<u>~</u>	
	14	7524	9278						~	
Product like	15	9275	10738		Function					
Function like	16	10813	11619						~	
FeatureID =	17	11668	12198						~	
	18	12231	13202		Notes					
				Y	Feature ider	ntified in 2 (of 2 files		<u>~</u>	
Hide Ignored Features			>		Notes from	- EricB:				
Select All Features	Insert D	elete Po	st Valida	te		'			>	
	1 - 50000	F	Position : 17	878	Con	trols >> Ma	р 🔽 М	ap >> Control		
1214567 121 14 124	ETH MANY RANGER AND A VER BUTFARE OF AN AD ATTACK BUT AND A ATTACK AND									
112 Features Live	1.7150.057.055	- 75%	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	1		KIK IK VOL	IK K KI	K UKUU YU K UU	51702	

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".

EricB/	Merge	ed Se	equence										ſ.	
Overview	Feat	ures	Reference	s 9	Sequence	Documer	ntation							
Sort By	Start	-] 🔳	1	Name	Start	Stop	^	Descripti	on Sequ	ence Prod	luct Regions	Blast	Con + +
Select Fe	atures	Din	ect SQL		1	488	1177		Name	4		GenelD		
Tune	is	All		- 1	2	1090	1926		Tune	ICDS	-	GL		
Nama	liko		<u> </u>	-	2	1333	1926		Charl		2402	Leave Tee		
Name	IIKe	<u> </u>		_	3	1926	2147		Start		2403	Locus Lag	Enico_4	
GenelD	=			<u></u>	4	2483	2953		Stop		2953	Regions		1
Locus	like			_!	5	3038	4000		Length	471		Tag		
Start	>			_	6	4000	4665		Direction	Forward	<u> </u>		·	
Length		<u> </u>			7	4662	5360		Translatio	n Tabla	LIndefined			
Lenger		<u> </u>			8	5403	5474		EC Numb	un rable j	ondenned			
Regions	>				9	5480	5552		EUNUME	ber				
%GC	<				10	5595	5666							
CAI	>				11	5697	5855		Product					
EC#	liko	<u> </u>			12	5855	7066		gp4					\simeq
		<u> </u>			13	7066	7500							\sim
Product	like				14	7524	9278		Function					
Function	like				15	9275	10738							<u>~</u>
FeaturelD) =	í –			12	9296	10738							\sim
Taa	like	<u> </u>	<u> </u>		16	10813	11619		Notes					-
ray	like							~	Feature	identified i	n 2 of 2 files			~
🔲 Hide I	Ignored	d Fea	tures	<			>		Notes fro	 om EricB ·				
Sele	ont All P	eatu		lr	nsert De	lete Pos	t Valida	ite	11000010	Sin Endo .				
R			• H H 1	- 6	250	Po	sition : 2	5590		Controls >>	Мар 🔽 М	ap >> Control:	s e	
		1	2		3		4		5	K	6	7) <mark>8</mark> 9 ()	11 12
112 Featur	es	Live	e										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!

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. This 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.

- 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.
- 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?
- 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.

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.

🔤 Genome Profile : Airmid	
Basics Codon Bias Dinucleotide Profile Merge 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 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 Image: Start and stop codon and GI tag Image: Start and stag Image: Start an	Genes All genes Selected genes All ORFs All RNAs Replicons Current Replicon All Replicons <u>OK</u> <u>Cancel</u> Load into Excel
	?

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 10⁻⁴)? Do they match the Hatfull-approved maps (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 and in the function boxes? Does the tRNA end with "CCA", and if not is it trimmed correctly?

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.

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!

The next step is to submit your files for expert QC and GenBank submission. Read below to make sure that your files are ready for submission, then submit a final DNA Master (.dnam5) file and a final Author List via e-mail to:

phage.submission@gmail.com.

After expert review, your annotation will be either accepted or returned. If accepted we will provide a GenBank flat file for your inspection. 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_Final.dnam5", which will help distinguish it from other versions you may have been working on.
- 2. It must contain one entry and set of notes 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.
- 3. All features must be validated (Section 9.3.2).
- 4. All features must be re-numbered if necessary (Section 9.3.3).
- 5. All features must be re-BLASTed (Section 9.3.4).
- 6. Any functions are noted in the "Function" field as well as the "Notes" field.

12.2 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 12.1

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, Lu Barker, Razi Khaja, and Melvina Lewis, for their tremendous insights and feedback. Melvina F. Lewis provided the terrific cover design.

See <u>http://phagesdb.org/DNAMaster for .pdfs</u> of all Appendices

Appendix I

System requirements and Installation of DNA Master

Appendix II

DNA Master Quick Start Guide

Appendix III

Gene Function with Bench Support and References

Appendix IV

Hatfull Genome Maps

Appendix V

Case Study: The Annotation of Etudee

Appendix I

System requirements and DNA Master installation

DNA Master Installation Guide

1. Minimum System Requirements for Installation of DNA Master

PC Minimum Requirements

OS: Windows XP/Vista/7 32-bit or 64-bit CPU: Dual-core Processor 1.8GHz Memory (RAM): • XP: 1GB • Vista/7: 2GB Video Memory: 128MB Free Disk Space: 5GB DVD Drive INTERNET CONNECTION FULL ADMINISTRATOR RIGHTS

Mac Minimum Requirements

OS: Mac OS X 10.5 or Higher CPU: Dual-core Intel Processor 1.8GHz • Non-Intel Macs are NOT supported. Memory (RAM): • 2GB Video Memory: 128MB Free Disk Space: 25GB DVD Drive INTERNET CONNECTION FULL ADMINISTRATOR RIGHTS

2. Installing DNAMaster on a Windows Computer

IMPORTANT: For Vista/Windows 7 users, this program must have **full administrative rights**. It is **not** sufficient to install this program on a User account with administrative-level rights, you must specific that the program has these rights too. During installation or when starting the program once installed, press "**Yes**" or "**OK**" if you are prompted to allow the program to continue, failing to do so WILL NOT allow the program to run properly and WILL cause errors. One easy way to make sure the program always has admin rights is to create a short-cut for the program as outlined below. If the program is then **ALWAYS** run from this short-cut, odd error messages should not occur. Otherwise, when starting the program, right-click on the program icon, and select "Run as administrator" every time.

DNA Master Installation

-DNA Master can be downloaded at the following link: http://cobamide2.bio.pitt.edu/computer.htm

-Double click the installer, and follow the instructions to install the program.

Shortcut Creation

Navigate to the DNA Master directory: 32-bit OS: My Computer → C:/Program Files/DNA Master 64-bit OS: My Computer → C:/Program Files(x86)/DNA Master
-Right-click on "DNAMas.exe" and select "Create Shortcut" Windows XP: Drag the shortcut to the desktop. Windows Vista/7: Click "Yes" on the dialogue box.
-Windows Vista/7 Users Only—MUST confer admin rights to the program: Right Click on the newly created desktop shortcut: -Click properties -Click on the "Compatibility" tab -Check the box next to "Run as Administrator" under "Privilege Level" at the bottom -Click "OK"

Updating – INTERNET CONNECTION REQUIRED

-Double click your DNA Master shortcut.

-Go to Help → Update DNA Master

-Allow the update to run and restart the program.

-DNA Master is now updated and ready to use.

DNA Master is now installed and up-to-date. You can run it through the desktop shortcut.

3. Installing DNA Master on a Mac

To install DNAMaster on your Mac, you will need to install Windows as a second operating system on an emulator such as VirtualBox (<u>http://www.virtualbox.org/wiki/Downloads</u>). If you already have access to Windows on your Mac, skip down to "DNA Master Installation". Once you have downloaded and installed VirtualBox, as per the VirtualBox website instructions, install Windows as below:

Windows Image Installation

-Obtain a copy of Windows XP, Vista, or 7 32-bit edition.

Windows XP is the cheapest option, and requires the least resources.

-Open VirtualBox.

-Click "New" to create a new virtual machine.

-Click "Continue"

-Name the machine "Windows"

-Under the "Operating System" menu choose "Microsoft Windows"

-Under the "Version" menu choose the version you are installing

-Click "Continue"

-For Windows XP, allocate at least 512MB memory, Vista/7 requires >1GB.

-Click "Continue"

-Check the box for "Boot Hard Disk"

-Click the "Create new hard disk" option.

-Click "Continue"

-Click "Continue"

-Click "Dynamically expanding storage"

-Click "Continue"

-Set the starting size of the virtual storage using the slider to a minimum of 20GB.

-Click "Continue"

-Click "Finish"

-Click "Finish"

-Now start the machine by selecting it from the list to the left and clicking "Start"

-Click "Continue" when the first run wizard starts

-Insert your Windows Installation Disk. Ignore any autorun prompts.

-The dialogue in the middle should read "Host Drive" followed by a letter.

-Click "Continue"

-Click "Finish"

-Follow the instructions on-screen to install your copy of Windows.

-After the installation completes, close the virtual machine.

Click "Machine"

Click "Shut Down"

-In the main window, click the settings button.

-Click on the display tab.-Set the "Video Memory" to 64MB.-Save the changes.-Select the new machine in the main window and click "Start" to run it!

*Note: It is recommended that you update your copy of Windows before continuing

DNA Master Installation (Within your Windows Virtual Machine)

-DNA Master can be downloaded at the following link:

http://cobamide2.bio.pitt.edu/computer.htm

-Double click the installer, and follow the instructions to install the program.

IMPORTANT: For Vista/Windows 7 users, this program must have **full administrative rights**. It is **not** sufficient to install this program on a User account with administrative-level rights, you must specific that the program has these rights too. During installation or when starting the program once installed, press "**Yes**" or "**OK**" if you are prompted to allow the program to continue, failing to do so WILL NOT allow the program to run properly and WILL cause errors. One easy way to make sure the program always has admin rights is to create a short-cut for the program as outlined below. If the program is then **ALWAYS** run from this short-cut, odd error messages should not occur. Otherwise, when starting the program, rightclick on the program icon, and select "Run as administrator" every time.

Shortcut Creation

-Navigate to the DNA Master directory:

32-bit OS: My Computer -> C:/Program Files/DNA Master

64-bit OS: My Computer -> C:/Program Files(x86)/DNA Master

-Right click on "DNAMas.exe" and click "Create Shortcut":

Windows XP: Drag the shortcut to the desktop.

Windows Vista/7: Click "Yes" on the dialogue box.

-Windows Vista/7 Users Only—MUST confer admin rights to the program:

Right Click on the newly created desktop shortcut:

-Click properties

-Click on the "Compatibility" tab

-Check the box next to "Run as Administrator" under "Privilege Level" at the bottom

-Click "OK"

Updating – INTERNET CONNECTION REQUIRED

-Double click your DNA Master shortcut.

-Go to Help \rightarrow Update DNA Master

-Allow the update to run and restart the program.

-DNA Master is now updated and ready to use.

DNA Master is now installed and up-to-date. You can run it through your desktop shortcut.

Appendix II

DNA Master Quick Start Guide

DNA Master Quick Start Guide

Part 1: Creating a Draft Annotation

Setting Key Preferences

- 1. From the File menu, select Preferences to open the preference window.
- 2. Click on the Local Settings tab, then on the Colors sub-tab.
- 3. In the LEFT box, change the display colors for tRNAs, tmRNAs, and ORFs by clicking on the appropriate boxes. The default is black, but we strongly recommend the colors shown in **Figure 1**.
- 4. Still under the Local Settings tab, click on the Codons sub-tab.
- 5. If the box labeled "Use TTG start codons" is not checked, then check it.
- 6. Click the Apply button on the right to save changes, then OK to exit.

Importing a DNA Sequence

1. Verify that your DNA sequence, in fasta format, is saved in a known location (or you may download a fasta file from phagesdb.org).

NOTE: If using Virtual Box or another emulator to run Windows, you should copy the fasta file to the virtual machine desktop before proceeding.

- 2. From the File menu, select Open \rightarrow FastA Multiple Sequence File as shown in Figure 2.
- 3. Browse to your genome's fasta-formatted file, then click Open.
- 4. A window like the one shown in **Figure 3** should appear. Click on the button in the lower right hand corner that looks like a piece of paper with the upper right corner folded over, then select "Create Sequence from this entry only" and a new window titled "Extracted from YourPhage.fasta" will open within DNA Master.



Figure 3: Importing the Sequence into DNA Master

Auto-Annotating a Genome

1. Make sure the "Extracted from YourPhage.fasta" window is open and selected, then go to the **Genome** menu at the top of the program, and select **Annotation** → **Auto-Annotate**.

NOTE: Many options are available in the Auto-Annotate window that opens. Our standard choices for each are shown in **Figure 4**.

NOTE: Be prepared to wait **30-90 minutes** if you choose to "Perform BLAST search on nr database" because BLAST searches take time. If you'd rather move on quickly, then un-check that BLAST box. You will still be able to BLAST all genes—and store the results in DNA Master—later.

- 2. After selecting your desired options, click Annotate.
- 3. When DNA Master asks if you want to "Erase features and annotate genome?", click **Yes**.
- 4. Wait while DNA Master annotates your genome. This should be a fairly quick process (<5 minutes without BLASTing), and you can check the status in the lower-left corner of the Auto-Annotate window.



Figure 4: Auto-Annotation Options; BLAST Search Unchecked

5. When Auto-Annotation is complete, the Auto-Annotate window will close, and you'll be returned to the main window for your genome. Congratulations, you've created a draft genome annotation!



Figure 1: Setting Display Colors

Part 2: Refining Your Positional Annotation Features (Genes)

- 1. Click on the **Features** tab at the top of your genome's window (the red box in **Figure 5**). Here you can see a list of all features (genes) in your current annotation with their left and right coordinates. If you've set up your color preferences properly, forward genes should be in green and reverse genes in red. At the bottom of this window a map of your features is displayed, also color-coded, along with the number of features and total sequence length.
- 2. The currently selected feature will be identified by a black triangle (the blue box in **Figure 5**). Data for this selected feature will be shown in the right pane. The right pane has several tabs of its own, the default being **Description** (see the green box in **Figure 5**). By changing tabs, you can see this gene's DNA sequence, amino acid product, any saved BLAST information, and many other statistics. Please explore!

Frames View

3. From the main DNA Master menu, select DNA → Frames. A new window titled "ORF Analysis..." will open where the top three rows represent the forwards reading frames, the bottom three rows the reverse reading frames. To see your current genes displayed in this view, click the ORFs button in the lower right corner of this Frames

Elph10	fina	al											
Overview Fr	eatu	res F	Referenci	es Sequ	ence Docum	nentation							
Sort By Sta	art	-	1	Nam	e Start	Stop	^	Description	n Seq	uence Proc	luct Regions	Blast	Context
Select Featu	res	Direc	t SQL	1	34	270		Name	1	-	GenelD	-	
Town in	Ē	All	-	2	267	560		Tunn	COC	-	Low Ter	EL DUT	0.1
Type is	ľ	MI	-	3	557	742		Type	LUS		LOCUS I ag	ELENI	<u></u>
Name lik	e	1		4	739	1032		Start		34	Regions		1
GenelD =				5	1010	1159		Stop		270	Tag		
Locus lik	e		_	6	1152	1613		Length	237	8	Composition	56.54 %	GC
Cind N	i í	-		7	1685	1927		Disation	C	-	CAL	0 2055	1 1 50
Start		_		8	1927	2235		Direction	r unyaru	-	CHI	0.2333	1 -1.30
Length		_		9	2280	3950		Translabo	in Table	Undefined			-
Regions >				10	3962	4183		EC Numb	er				
%GC <	ſ		-	11	4170	5726							- 3
CAL	i i	-		12	5729	6730		Product					
ECH II				13	6818	8311		gp1					
EL# IK	e			1.4	8434	8613							100
Product lik	e			15	8613	9179		Function	-				
Function like	e		_	16	9194	9529							
FeatureID =	i f	-		17	9526	9906							100
Tag E		-		18	9884	10288		Notes					•
Fag lik I⊽ Hide Igno Select (ored	Featu	res	Insert	Delete F	Post Valid	late						1
		2 D	H	1 - 12500		Position : E	60763		Controls (ap >> Control	10	
147 Features	1	Live			77 9 11	1	16			ATA ISA	Sara Sa	746	75 3 ?

Figure 5: Looking at Auto-Annotated Features

window (enlarged portion of **Figure 6**). You should now see green and/or red regions highlighted that represent currently called genes. (You may want to zoom in for better resolution by using the buttons near the bottom left of this window.) All potential start codons are shown as vertical lines that are half the height of a given frame, while stop codons are shown as vertical lines that are the full height of a given frame.



Figure 6: The Frames Window

Comparing Potential Start Codons

- 4. From the Frames window, select an ORF by clicking within it (DNA Master will draw a line showing your selection).
- 5. Click the button labeled **5'-3'** in the bottom right corner (enlarged portion of **Figure 6**), and a new window called "Choose ORF Start" will open. In this window will be a list of **each possible start codon** for the selected ORF, along with each Shine-Dalgarno score, upstream sequence, start position, and resulting ORF length.

Part 3: Functional Annotation

BLASTing Predicted Proteins

- 6. A powerful feature of DNA Master is the ability to BLAST all gene products from an annotation, then store the results in the archived file so that they can later be accessed as needed, even without an internet connection. If a feature has stored BLAST results, you can view them by going to the **Blast** tab for that feature (green box, **Figure 5**).
- 7. If you did not BLAST during auto-annotation, you can do so at any time by going to the **Blast** tab for any feature (green box, **Figure 5**), selecting **Blast ALL genes**, modifying settings if desired, then clicking **Blast All**. BLASTing a complete phage genome annotation takes 30+ minutes. Be patient.
- 8. Once finished, "Genome BLAST Complete" should display, and BLAST data for each feature are in the Blast tab.

Part 4: Important Features

Suggested Annotation Layout

The figure below shows how you might arrange windows in DNA Master to work on an annotation. The **Frames Window** (on top of the figure) has all 6 reading frames clearly delineated, and allows you to see currently called genes, other choices for start codons, as well as all other potential ORFs in the genome. The **Main Window** (bottom left in the figure) can be used to view precise coordinates, DNA or amino acid sequence, BLAST results (shown in figure), and more. It is also where you will change start positions, add notes, or add/delete genes. The **Choose ORF Start Window** (bottom right in the figure) allows you to see all potential start codons for a given reading frame along with the associated Shine-Dalgarno scores, start positions, and resulting ORF lengths.

Using all of DNA Master's capabilities in concert facilitates fast and accurate genome annotation.

DRE Analysis for Labrace	uit from Field Library Patient	ne fasta					- X
				 3 			
+3							ШЦ
	and man in the						
+2		<u> </u>					
	H T T T					11 11 10	1 1 1 1 1
	╨╪╝╝╢╴╶╷╢╌╵╵┤╄	<u>─┴┠──║╶┠╷║──┴┴──┼╷┴┟╟╢╢╢╸╁╴┲╨╵╫╷┼└╶╵</u>				╨┵╖┧╴╅╫╖╥╶┠┙	4444
1 1 1 1 1 1 1 1 1	1 1 1 10 1 10 1		110	1.1.11	manna coma com	a contraction of the	1 11
<mark>┊</mark> ┟║ <u>╢╢╢╢</u> ╢╎╢╎╎╴╎╴╿╢╎╖	╶╓╓╸╜┟╴╎╜╙╴┫╓╖┼╵╜╶╢	┱┰┟╶╢┶╄╫╫┶╨╧╹╨┽═┺╧┶┉┉╄	ЦЦ				
	nin nin ni linno		LIT			11 11	1 11
┊┝╨╨╨╵╴╢╶╵┍╶╶┼┼┬╖╨╜	╨┉╓┼┼╓╷┧╴╓╶╶╻╓╷╷╓┉	╨┽┼╨╎╨╫╨╴┰╶╴╖╴╓╢╴┼╷┚╖╫╵╨╢		╨╨╨┰┶╓┿┥		<u> </u>	
				minter			1
10625 10937 1124	9 11561 11873			745 14057	14369 14681	14993 15305	15624
bp: 10629 H H H H D	► ► ORF 12413 • 13234	G+C 5 Window: 33 🚖 bp	. 1.5	10 11007	11001	1000	10024
List of features in documentation						ORFS DE	3 照 🥻 ?
DNA Extracted from Eacth Li	brary Dationnes fasta		DNR Chan	en ORE start			
Den in Fosturos D. Com			Starte : 11	OBE Start	12413 Edn 1 Edn 2 1	Cdn3 Length	
Uverview reduies Reference	s Sequence Documentation		Selected	1 ORF Stop	: 13234 5' End 58.3 75.0 :	33.3 36	oeument
Sort By Index -	Name Start Stop	Description Sequence Product Regions Blast Context		ORF Lengt	h:822 3'End 39.7 58.6 !	56.7 784	
Select Features Direct SQL	16 9031 9918	Score Target Description	ЦЦШ				12902
Type is All 💌	17 9978 10127	132 gp19 [Mycobacterium phage Barnyard] 119 gp18 [Mucobacterium phage Konstantine]	Shi	ne D algarno	Sequence of the Region	Start Start C	ORF
Name like	18 10138 10470	The grid (mycobaccelium phage Konstantine)	# SCO	re Space	Opstream of the Start	CTC 12412	sength
GenelD =	19 10491 11075		2 364	6	CCCCGGCAACGAAGGACATCCT	GTG 12449	786
Locus like	20 11079 11519		3 399	7	TTCTTTCCCGTGGAGCAAGCGA	ATG 12485	750
Start >	21 11512 12426		4 378	7	GTGGAGCAAGCGAATGGACCCG	ATG 12494 1	741
Length	22 12413 13234	BLAST Hit	5 345	9	GCGAATGGACCCGATGTTTTCT	GTG 12503 7	732
Begione N	23 13312 14010	GI 29566994 Export ALL	6 345	9	GGAAAAGCAAAAGCGCGCTGCG	ATG 12791 4	444
	24 14060 13262	Length 280 Delete	7 273	7	CGCTGCTCTGAAAAATCAAGCT	GTG 12878 S	357
%GC <	29 19371 16097	Max Score 132 Date 7/18/2011 Delete All	8 378	7	CGCCCCCGAGATTTTCGAAGAC	ATG 12977 2	258
CAI >	26 16060 16363	High-Scoring Pairs (HSP)	9 399	7	TTTCGAAGACATGGAAGTCGAT	TTG 12989 2	246
EC# like	27 16330 16761	HSP Data Alignment	10 273	8	CGGAACGGTTACTGTTTCGGGT	GTG 13019 2	216
Product like	29 17304 17705	Bit Score 55.5 Positives 91	11 294	7	GGCTCTCGCTGAGAAGTTCCCC	GTG 13208 2	27
Function like	30 17748 18134	Score 132 Identities 48					
FeatureID =	31 18140 18358	E-Value 2.5E-6 Similarity 50.6 %					11 B ?
Tag like	32 18355 19203	% Aligned 64.3 %	1 20				
T 10de Leased Factor	2	Query 90 - 267			10- 10-	F	
nide ignored reatures	Innert Doloto Bost Matidate	Target 93-272					
Select All Features			1.01-	9			
	1 - 50000 Position : 4200	5 🔽 Controls 🤉 Map 🔽 Map 🖓 Controls					
	2 2 242 1 24 3 35	海南 37 38 339 142 43 年11 年7月11日月1日 63 月11 117 1	-				
116 Features Live		70508 🚙 ?	0				

Creating a Spreadsheet of Current Gene Calls

- 1. Make sure your main window is open and selected. From the top menu, select **Genome** \rightarrow **Profile**.
- 2. Click **OK**, then save in the format you want.
- 3. Open the file you generated with Excel or a similar spreadsheet program.

Creating an ORF Map of Current Gene Calls

- 1. Make sure your main window is open and selected. From the top menu, select **DNA** \rightarrow **Export Map**.
- 2. Modify any options you'd like to, then click **Draw**.
- 3. A Windows Meta File (.wmf) image is generated, and can be opened with most image-viewing programs.

Appendix III

References for gene functions

Appendix III: References for gene functions

TM4 structural genes

TUBERCLE AND LUNG DISEASE : THE OFFICIAL JOURNAL OF THE INTERNATIONAL UNION AGAINST TUBERCULOSIS AND. Vol 79, Issue 2, Pages 63-73, 1998. Mycobacteriophage TM4: genome structure and gene expression. M E Ford, C Stenstrom, R W Hendrix, G F Hatfull PubMed ID: <u>10645443</u>

L5 integrase

JOURNAL OF BACTERIOLOGY. Vol 175, Issue 21, Pages 6836-41, 1993.

Mycobacteriophage L5 integrase-mediated site-specific integration in vitro.

M H Lee, G F Hatfull

PubMed ID: <u>8226625</u>

L5 Xis (gp 36)

MOLECULAR MICROBIOLOGY. Vol 35, Issue 2, Pages 350-60, 2000.

Identification and characterization of mycobacteriophage L5 excisionase.

J A Lewis, G F Hatfull

PubMed ID: <u>10652095</u>

Bxb1 serine integrase (gp 35) MOLECULAR MICROBIOLOGY. Vol 50, Issue 2, Pages 463-73, 2003. Mycobacteriophage Bxb1 integrates into the Mycobacterium smegmatis groEL1 gene. Amy I Kim, Pallavi Ghosh, Michelle A Aaron, Lori A Bibb, Shruti Jain, Graham F Hatfull PubMed ID: <u>14617171</u>

Bxb1 RDF (gp 46)PLOS BIOLOGY. Vol 4, Issue 6, Pages e186, 2006.Control of phage Bxb1 excision by a novel recombination directionality factor.

Pallavi Ghosh, Laura R Wasil, Graham F Hatfull

PubMed ID: <u>16719562</u>

D29 integrase

GENE. Vol 225, Issue 1-2, Pages 143-51, 1998.

Mycobacteriophage D29 integrase-mediated recombination: specificity of mycobacteriophage integration.

C E Peña, J Stoner, G F Hatfull

PubMed ID: <u>9931474</u>

Bethlehem DnaB intein

HE JOURNAL OF BIOLOGICAL CHEMISTRY. Vol 285, Issue 4, Pages 2515-26, 2010.

Splicing of the mycobacteriophage Bethlehem DnaB intein: identification of a new mechanistic class of inteins that contain an obligate block F nucleophile.

Kazuo Tori, Bareket Dassa, Margaret A Johnson, Maurice W Southworth, Lear E Brace, Yoshizumi Ishino, Shmuel Pietrokovski, Francine B Perler

PubMed ID: <u>19940146</u>

Che9c RecE and RecT (gp60 and 61)

MOLECULAR MICROBIOLOGY. Vol 67, Issue 5, Pages 1094-107, 2008.

Efficient point mutagenesis in mycobacteria using single-stranded DNA recombineering: characterization of antimycobacterial drug targets.

Julia C van Kessel, Graham F Hatfull

PubMed ID: <u>18221264</u>

L5 structural proteins

MOLECULAR MICROBIOLOGY. Vol 7, Issue 3, Pages 395-405, 1993.

DNA sequence, structure and gene expression of mycobacteriophage L5: a phage system for mycobacterial genetics.

G F Hatfull, G J Sarkis

PubMed ID: <u>8459766</u>

L5 Repressor (gp 71)

MOLECULAR MICROBIOLOGY. Vol 7, Issue 3, Pages 407-17, 1993.

Superinfection immunity of mycobacteriophage L5: applications for genetic transformation of mycobacteria.

M K Donnelly-Wu, W R Jacobs, G F Hatfull

PubMed ID: <u>8459767</u>

Bxb1 Repressor (gp 69)

MOLECULAR MICROBIOLOGY. Vol 38, Issue 5, Pages 971-85, 2000.

Transcriptional regulation and immunity in mycobacteriophage Bxb1.

S Jain, G F Hatfull

PubMed ID: <u>11123672</u>

L5 genes that are cytotoxic, (Set of genes that we know that the clones can not be transformed into a non-lysogen)

MICROBIOLOGY (READING, ENGLAND). Vol 154, Issue Pt 8, Pages 2304-14, 2008.

Identification of three cytotoxic early proteins of mycobacteriophage L5 leading to growth inhibition in Mycobacterium smegmatis.

Jan Rybniker, Georg Plum, Nirmal Robinson, Pamela L Small, Pia Hartmann

Cluster G repressors

MICROBIOLOGY (READING, ENGLAND). Vol 155, Issue Pt 9, Pages 2962-77, 2009.

Mycobacteriophages BPs, Angel and Halo: comparative genomics reveals a novel class of ultrasmall mobile genetic elements.

Timothy Sampson, Gregory W Broussard, Laura J Marinelli, Deborah Jacobs-Sera, Mondira Ray, Ching-Chung Ko, Daniel Russell, Roger W Hendrix, Graham F Hatfull

PubMed ID: <u>19556295</u>

Lysins A and B

MOLECULAR MICROBIOLOGY. Vol 73, Issue 3, Pages 367-81, 2009.

Mycobacteriophage Lysin B is a novel mycolylarabinogalactan esterase.

Kimberly Payne, Qingan Sun, James Sacchettini, Graham F Hatfull

PubMed ID: <u>19555454</u>

WhiB of tm4

MOLECULAR MICROBIOLOGY. Vol 77, Issue 3, Pages 642-57, 2010.

Insights into the function of the WhiB-like protein of mycobacteriophage TM4--a transcriptional inhibitor of WhiB2.

Jan Rybniker, Angela Nowag, Edeltraud van Gumpel, Nicole Nissen, Nirmal Robinson, Georg Plum, Pia Hartmann

PubMed ID: 20545868

Non-essential genes

PLOS ONE. Vol 3, Issue 12, Pages e3957, 2008.

BRED: a simple and powerful tool for constructing mutant and recombinant bacteriophage genomes.

Marinelli LJ, Piuri M, Swigonová Z, Balachandran A, Oldfield LM, van Kessel JC, Hatfull GF

PubMed ID: 19088849

Tapemeasure of TM4, defect in infection of stationary phage cells

MOLECULAR MICROBIOLOGY. Vol 62, Issue 6, Pages 1569-85, 2006.

A peptidoglycan hydrolase motif within the mycobacteriophage TM4 tape measure protein promotes efficient infection of stationary phase cells.

Mariana Piuri, Graham F Hatfull

Phage I3 (cluster C) promoters and some structural genes

GENE. Vol 143, Issue 1, Pages 95-100, 1994.

Structural proteins of mycobacteriophage I3: cloning, expression and sequence analysis of a gene encoding a 70-kDa structural protein.

G R Ramesh, K P Gopinathan

PubMed ID: <u>8200544</u>

INDIAN JOURNAL OF BIOCHEMISTRY & BIOPHYSICS. Vol 33, Issue 1, Pages 83, 1996.

Cloning and characterization of mycobacteriophage I3 promoters.

G R Ramesh, K P Gopinathan

PubMed ID: <u>8744840</u>

TM4 non-essential genes, D29 non-essential genes

PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA. Vol 94, Issue 20, Pages 10961-6, 1997.

Conditionally replicating mycobacteriophages: a system for transposon delivery to Mycobacterium tuberculosis.

S Bardarov, J Kriakov, C Carriere, S Yu, C Vaamonde, R A McAdam, B R Bloom, G F Hatfull, W R Jacobs

PubMed ID: <u>9380742</u>

D29 and L5 non-essential genes

JOURNAL OF MOLECULAR BIOLOGY. Vol 279, Issue 1, Pages 143-64, 1998.

Genome structure of mycobacteriophage D29: implications for phage evolution.

M E Ford, G J Sarkis, A E Belanger, R W Hendrix, G F Hatfull

PubMed ID: <u>9636706</u>

L5 Promoters

MOLECULAR MICROBIOLOGY. Vol 17, Issue 6, Pages 1045-56, 1995.

Transcriptional regulation of repressor synthesis in mycobacteriophage L5.

C E Nesbit, M E Levin, M K Donnelly-Wu, G F Hatfull

PubMed ID: <u>8594325</u>

Unpublished:

Bps gene 22 :

Mutations in this gene have expanded host range

Rosebush genes 32 and 42

Mutations in these genes have expanded host range

Cluster N integrases and repressors, Brujita Integrase and repressor

Cluster J capsids:

Identified through N-terminal sequencing

Giles non-essential genes identified through BRED

Appendix IV

Genome Maps:

L5

Cluster B map

Cluster C mal

Cluster D map

Cluster E map

Cluster F map

Cluster G map

Cluster I map

Cluster J map

Cluster K map

Cluster L map

Cluster O map

Giles map

Wildcat map














































Tail assembly































Appendix V: Etude Annotation

First, BLAST Etude against phagesdb.org.

Go to the phagesdb.org BLAST page. Paste in the Etude sequence or browse to the FASTA file on your computer. Turn off the low complexity filter. Press BLAST.



When Etude is BLASTed against phagesdb.org, it appears that there is similarity to the Cluster L1 phages, and to the Cluster A3 phages.

```
Query= etude
(14,998 letters)
```

		Color Ke	y for Align n e	nt Scores	
<40		40-50	50-80	80-200	>=2
2	ĸ	5K	7K	10K	12K
		·	· ·	·	
		·	: .		:
				:	
					· :=
			-		
					:
					:
					:

Distribution of 392 Blast Hits on the Query Sequence

Score E (bits) Value

Sequences producing significant alignments:

UPIE Complete Sequence, 73784 bp including 10 bp 3' overhang (TC	1.314e	+04 0.0
LeBron	1.178e	+04 0.0
JoeDirt Final Sequence, 74914 bp including 10 bp 3' overhang (TC	1.169e	+04 0.0
Microwolf Final Sequence, 50864 bp including 10 bp 3' overhang,	4022	0.0
Vix Complete Sequence, 50963 bp including 10 bp 3' overhang (CGG	3998	0.0
JHC117 Final Sequence, 50877 bp including 10 bp 3' overhang, Clu	3998	0.0
Bxz2	3998	0.0
Faith1 Complete Sequence, 75960 bp including 10 bp 3' overhang (1388	0.0
Rockstar Complete Sequence, 47780 bp including 10 bp 3' overhang	a bp including 10 bp 3' overhang (TC 1.314e+04 0.0 a bp including 10 bp 3' overhang (TC 1.178e+04 0.0 b bp including 10 bp 3' overhang (TC 1.169e+04 0.0 bp including 10 bp 3' overhang, 4022 0.0 bp including 10 bp 3' overhang, Clu 3998 0.0 bp including 10 bp 3' overhang, Clu 3998 0.0 160 bp including 10 bp 3' overhang (1388 0.0 160 bp including 10 bp 3' overhang (1388 0.0 17780 bp including 10 bp 3' overhang 232 2e-59 212 2e-53 204 4e-51 13bp, including 10 bp 3' overhang, Clu 137 8e-31	
Peaches	212	2e-53
Eagle	204	4e-51
LHTSCC Complete Sequence (51813bp, including 10bp 3' overhang: C	196	1e-48
George Final Sequence, 51578 bp including 10 bp 3' overhang, Clu	137	8e-31

Now we pull up Etude in phamerator, next to its closest matches. I will use LeBron and Bxz3 for now, because these two phages have annotations in GenBank already—which means that when I look for individual genes using BLAST on the NCBI website, I should see these genes, and they will be genes that have already been curated and well-examined by the annotators. I will also check Upie, JoeDirt, and Microwolf's draft annotations in phamerator.

Etude



Notice how the purple between the two genomes indicates that the nucleotide sequence similarity is very high between Etude and LeBron for the first seven genes in both genomes. I will start with calling these seven genes.

Next, I open DNA master, and load the Etude sequence from its fasta file.

->	File ->O	pen ->	FastA	Multip	ole Sec	juence	File

🞇 FastA Sequences from etude.fasta	
Sequences Pairwise Alignment Multiple Alignment	
Idx Description Length	Insert Sequence
1 etude 14998	Delete Sequence
	Update Sequence
	Verify updates
	Save FastA File
	Export Summary
Sequence Complement Hide Annotation	
AGCGACACTTCTCTCTGGAAATTCAGGCAAGAACATGAGGGGGGGT TAGCGCCCCTAAAACCCCTGGTAGGAGGGCGAAATCGTGGGTAGAGGA CGTGGTAAGGACCCGTCAGGCCTGGGGGGGGGG	
1 Sequences 14,998 bp	GC 🖹 Export 🤶

I then click "export" in the lower right corner, and "Create sequence from this entry only" from the menu that appears.

Extracted from Fas	tA Library etude.fasta		
Overview Features Refe	erences Sequence Documental	tion	
Genome		Taxonomy	Notes
Organism :		Domain :	Notes on the Genome
GenomelD : 0	PrototypeID : 0	Division :	Notes and I axonomy are only available
Length:0	Date Created : 10/4/2011	Family :	database
Replicons : 1	Genome Date : 10/4/2011		
Features : 0	GC Content : 0.00		
CALLADIE : Disestems DMT and 40000	E00001745		
Directory: DMT emp 40826	.086331745		
Replicon			
Replicon	_		
NCBI Date : 10/4/2011	NCBI Genome :		
Topology : Linear	Accession :		~
Length : 14998	Version :		Notes on the Replicon
Features : 95	<u>GI :</u>		<u>^</u>
ORFs: 94	Status : Current		
RNAs: 1	Deletion : Vulnerable		
GC Content : 0.00	Origin : 1		
ORF GC : 59.8 ± 6.8	Terminus : 29836		
GC3: 77.8 ± 10.3			
CAL: 0.5978 ± 0.1095			
Translation Table : Unspeci	fied: default to standard code	<u> </u>	<u> </u>
$\mathbf{H} \bullet \mathbf{H} \bullet $	N 1 - 14998 Posit	tion : 3881 🖉 Controls >> N	1ap 🔽 Map >> Controls
0 Features Live			14998 🛃 🤮

A DNA Master sequence file will be created:

This file is empty other than the imported sequence (viewable if you click the "Sequence" tab above.

Now I auto-annotate this file to generate and import the information from Glimmer, GeneMark, and Aragorn into the file.

Click Genome->Annotation-> Auto-Annotate

Auto-Annotate					
 Gene Calling Document tRNAs found with Aragorn Document ORFs found with Glimmer 3.02 analysis GeneMark HMM analysis Both analyses, combining them as follows: Favor Glimmer calls Favor GeneMark calls Exclude genes called by only one method Exclude genes called only by second method Maximum wait time 1 minute 	BLAST Searches Perform BLAST search on nr database Save hits with E-values less than 10E- Image: Save hits with E-values less than 10E-				
Examine and learn gene features	No BLAST scheduled				
✓ Autosave DNAM5 file as Extracted from FastA Libra Export alternative start codons Export Export a summary of BLAST hits Export Export a GC-Content map Export Export Frames map Export Height 200 ♀ Width 599 ♀ Sca Size Export Directory Z:\Mycophages\	ry etude.fasta_Annotated.dnam5 ort a Profile of features ort a list of Gray Holes and gene overlaps ort a Six-Frame map ort an ORF map le 1/2'' / kb v to 25 v kb per tier Browse				
Analysis Current Replicon All Analyses Annot All Replicons No Analyses	ate Export log file				
Status : Idle	2				

Uncheck all the analyses buttons and then click "annotate". As Etude is a relatively short piece of DNA, I will check the "BLAST" box at the upper right.

Once my genome is annotated and BLASTed, I will save it as Etude_annotated.dnam5.

Now I will generate the data from the programs outside of DNA master that I will need to review the auto-annotation.

GeneMark TB:

GeneMark is located at http://exon.gatech.edu/genemark/genemark_prok_gms_plus.cgi

There is also a link from the phagesdb website. Upload the etude.fasta file and use the Mycobacterium tuberculosis coding model (either strain is fine).

Sequence File upload: /Users/welkin/Documents/Mycophages/SEA phages/etude.fasta 	
Running Options	
Species: • Mycobacterium_tuberculosis_H37Rv +	Window size: 9 96 C bp
RBS model: O E.coli	Step size: • 12 : bp
Use alternate genetic code: Eukaryote (e.g. Yeast, ATG = only start) Mycoplasma (TGA = Tryptophan)	Threshold:
Output Options	
Graphical output options	Text output options
Generate PDF graphics (screen)	List open reading frames (ORFs) predicted as coding sequences
Generate PostScript graphics (email)	(CDSs)
Mark orfs on graph	List regions of interest
🗹 Mark regions on graph	List putative eukaryotic splice sites
Mark stop codons on graph	Write protein translations of ORFs
Mark start codons on graph	Write nucleotide transcripts of ORFs
Mark frameshifts on graph	Write protein translations of regions
Mark putative exon splice sites	Write nucleotide transcripts of regions
Print graph in landscape format	Write protein translations of putative exons
	Write nucleotide transcripts of putative exons
Email address (required for PostScript email output)	
Run	
(Start GeneMark) (Default)	

Opening the .pdf of the GeneMark output should show you a 5 page doument that begins like this:



Here are the first four and start of the fifth genes in Etude as called by GeneMark. Each tier represents a different reading frame, with upticks from the center horizontal in each

tier representing start codons in that frame and downticks representing stop codons in that frame. ORFs of significant length are shown as horizontal lines in each tier. Coding potential is shown by the wiggly trace lines. Areas that GeneMark has designated "regions of interest" are shown with gray bars through the coding potential. Sometimes these regions are actually genes, sometimes not. I find it easier to just ignore the gray bars completely.

Aragorn:

Aragorn is found at

http://130.235.46.10/ARAGORN/

or linked to from phagesdb.org.

Upload your FastA file, and change the default to "tRNA and tmRNA".

Björn Canbäck Bioi	nformatics ARAGORN ARWEN BRUCE optalign RAN
Home Projects Publi	cations Online services
Online serv	ices area and a constant and a constant and a constant and a constant and a constant and a constant and a const
ARAGORN	ARAGORN, tRNA (and tmRNA) detection in nucleotide sequences
ARWEN	Dean Lastatt an Australian specialist in stable RNAs, is the developer of ARAGORN. There is a version for download which requires
BRUCE	compilation with a C-compiler (for example with gcc -03 -ffast-math -finline-functions -o aragorn aragorn1.2.33.c; ignore any warnings about
optalign	trigraphs).
RAMI	Input sequence (both strands will be searched, max. 15 MB)
External links tRNAscan-SE Sprinzl compilation	Upload a fasta file with one or several sequences: /Users/welkin/Documents/Mycophages/SEA phages/SEA2010/etude.fasta Browse
	or select a genome from the list. Methanococcus jannaschii (NC_000909)
	Select options (see here for all options in the standalone version)
	Search for (default tRNA):
	Search allowing introns, 0-3000 bases (default no): no ;
	Sequence topology (default linear):
	Strand(s) (default both):
	Output format (default standard):
	Submit (Reset)

Then click "Submit"



The web-based Aragorn output shows a single tRNA that has a correctly-trimmed 3' end. I will come back to this when I get to the tRNA in my draft annotation.

tRNA-Scan SE:

tRNA-Scan SE is available at:

http://lowelab.ucsc.edu/tRNAscan-SE/ or linked to from phagesdb.org

I leave the default settings alone, and browse to my file:

Search Mode: Default	Source: Mixed (general tRNA model)
Format:	
Raw Sequence	
Sequence name (optional):	(no spaces)
Other (FASTA, GenBank, EMBL, GCG, IG)	
Paste your query sequence(s) here:	Run tRNAscan-SE
Queries are limited to a total of less than 5 million	nucleotides at any one time)
Queries are minied to a total of less than 5 minior	Clear Sequence
or submit a file: /Users/welkin/Documents/Mycophage	s. Browse
• Show results in this browser.	
Receive results by e-mail instead:	
Run tRNAscan-SE Clear Form	

Now I click "Run tRNAscan-SE.

The results are similar to Aragorn, however tRNAscan SE has called the tRNA with one extra base:

Results

Sequence Name 	tRNA #	tRNA Begin 	Bounds End 	tRNA Type 	Anti Codon	Intron Begin	Bounds End 	Cove Score
Your-seq	1	6242	6315	Leu	CAA	0	0	62.03
View tRNA								

I will evaluate these when I reach the tRNA in the genome sequence.

At this point, I might also make a genome map, however, this genome is so small I can visualize the entire thing at once in DNA Master, so I am going to skip the map.

Whole Genome overview:

DNA Etude	_annot	ated												
Overview	Feature	s Refere	ences	Sequence	Docum	entation								
Sort By	Index	•	•	Name	Start	Stop	^	Descrip	tion S	equence	Product R	egions Blast	Context	
Select Fe	atures	 Direct SQL	.) 🕨	· 1	83	466		Score	• T	arget Des	cription			
Tuno	- in 🛛		하나	2	475	849		>	578 g	p1 [Mycol	bacterium pha	ge LeBron]		
туре			ĽĽ	3	889	1053			560 g	p1 [Mycol	bacterium pha	ge JoeDirt]		=
Name	Іке		_ -	4	1059	2636			167 g	p1 [Mycol	bacterium pha	ge Faith1]		
GenelD	=		_ -	5	2626	4296								
Locus	like			6	4318	5373								
Start				7	5427	6038								<u>~</u>
Lenath				8	6136	6249		BLAST	T Hit —	0000574	~~			Export
Begions				0	6239	6317		Access GL	ion YP. אחר	_0038571 1360950	32			Export All
e oo			- H	10	6344	6507		Length	127	7				Delete
շնե			_H	11	6902	7260		MaxSc	ore 578	3	Date 10	/10/2011		Delete All
CAI	>		_ h	12	8315	9187		· High-S	coring	Pairs (HSF	9			
EC#	like			13	9411	10655		HSP	Data 4	Alignment	ĺ			
Product	like		- -	14	11265	11486		1	MCDC	PGRXXX	××××××××××××××××××××××××××××××××××××××	V VYTDSAMRAR	VAARDEMA	
Function	like		- -	15	11588	12100		1	1111					
FeaturelD			- [16	12299	12637		1	MGRG	RGKDPS	SPGGRSRDS	R PGTRSAWEAK	VAARPENA	QE YAVQ 🗧
Taa	iko [-[17	13031	13438		51	WEVE	KPNVWT	NOGMHAAGI	E TLIMRKGDAY	VYATFTWP	NG RIRT
Tay	iike j		-6				~	51	1111	111111	1111111		11111111	11 1111
🔲 Hide I	Ignored F	eatures				>		51	WEVE	KPNVWT	NQGMHAAGI	E TLTMRKGDAY	VYATFTWP	NG RIRT
Sele	ort All Fea	atures		Insert D	elete Po	ost Valida	ate	<u> </u>						
H H I	€		N 1-	14998		Position : 4	755		Control	s>> Map	🔽 Map >> (Controls		
	4	K	6	5	6		สก	11)	(12	(13	15	16) 17)	18) 19 20
21 Feature	s I	Live				2 141					`		1	4998 🔒 ?

My auto-annotation has 21 called genes and 1 tRNA. If I click on the BLAST tab for gene 1, I can see the scores of the alignments from GenBank and the actual alignment.

I can tell from the interactive map at the bottom of my Etude sequence file that there are some large gaps in my genome between genes 11 and 12 and 13 and 14. I will take a closer look at these areas when I reach them in my annotation. There is also a gene overlap with gene 8 and the tRNA. This will also need to be resolved.

Refining my annotation

I now open the Frames window:

Click \rightarrow DNA \rightarrow Frames

 \rightarrow Toggle on the features listed in my features table by clicking the "ORFs" button in the lower right-hand corner.



As you mouse over the Frames window, the lower left box will display in real-time the base pair coordinate of your cursor.

My screen isn't quite big enough to comfortably display both the frame window and my sequence window, so I will be flipping back and forth between the windows.

In the Frames window, I see my forwards transcribed ORFs in green, the reverse in red, and my tRNA in blue. The horizontal tiers represent the six-translational frames, and the full vertical lines within the tiers are stop codons while the half-vertical lines are start codons.

Before I start going through the genes one by one, I am going to validate the gene calls and renumber the genes. The auto-annotation function of DNA Master does not assign gene numbers to tRNAs, however, we do count them in our genbank files. To keep everything consistent, it is easiest to renumber genes initially and then again at the end of the gene identification process. As you grow more confident in your annotation ability, you may also want to renumber periodically as you add and delete genes from the annotation.

To renumber (and assign locus tags):

Click \rightarrow Validate (at the bottom of the central column)

(Etude	_ann	otat	ed															×
	Overview	Feat	ures	Reference	es	Sequence	Docume	ntation											
1	Sort By	Index	-	[I	ſ	Name	Start	Stop	^	Desc	ription	Sequence	Product	Regions	Blast	Context	Validation		
ł	Select Fe	atures	Dire	ect SQL		1	83	466		All O	RFs app	bear valid							1
1	T		All		Þ	2	475	849											1
I	туре	IS 			L	3	889	1053											
I	Name	like			L	4	1059	2636											
I	GenelD	=			L	5	2626	4296											
	Locus	like			L	6	4318	5373											
	Start	>	í—		L	7	5427	6038											
	Longth		<u> </u>		L	8	6136	6249											
ł	Lengtri		<u> </u>		L		6239	6317											
I	Regions	>	<u> </u>		L	9	6344	6607										~	
l	% GC	<			L	10	6647	6796		<								>	
I	CAI	>			L	11	6902	7360		Con	trol Ni	umbering							
	EC#	like	í –		L	12	8315	9187			Assian t	Vames bu or	ter of appe	arance					
1	Product	انلام	<u> </u>		┡	13	9411	10655			Assign I	Producto by ore	action apple						
l		iike Pi	-		ŀ	14	11265	11486			Assign r	-TODUCIS Dy I		pearance					
1	Function	like	<u> </u>		⊢	15	11588	12100			V Uve	rwrite existing	g values						
I	FeaturelD) =			ŀ	15	12299	12637		_		Uverwrite de	fault Name	es & Produc	ots only				
I	Tag	like			H	17	13031	13438	~		Assign r	new Locus T	ags by ord	ler of appe	arance				
I	□ Hide	lanore	d Feal	ures				>	_		locus I	ag Prefix E	TUDE	Search N	ICBI da	tabase			
	,		_	1	ľ	Insert De	lete Po	st Valida	ate	Re	assign I	Gene Data	Locate	Gray Holes	> 30)0 🔹 bp)		
	Sele	ont All F	-eatur	es	- 1						_		_			_			_
			ર∣ ▶		1 -	14998	P	osition : 5	315		Con	trols >> Map	Map :	>> Controls					
1	$1 \rangle 2 \rangle 3$		4		5	5	6	7	€	11		(12	(13		$\langle \langle 1 \rangle \rangle$	5 16	17 18	19	20)
	21 Feature	s	Live	e													14998	3	?

Check the box marked Overwrite existing values

Write the phage's name in the Locus Tag Prefix field. GenBank uses locus tags to assign a unique id to every gene in the database. We prefer to create our GenBank submission files with locus tags comprised of the phage's name and gene number already assigned, to prevent GenBank from assigning every gene a random number.

Click→Reassign Gene Data

DNA Etude	_ann	otate	ed																	×
Overview	Featu	ues	Reference	es	Sequence	Docume	ntation													
Sort By	Index	-	•	ſ	Name	Start	Stop	^	Ī	Description	Sequence	Product	Regio	ons	Blast	c	ontext	Validation	1	
Select Fe	atures	Dire	ect SQL		1	83	466		Ï	All ORFs app	Dear valid									71
Turn	.	All		Þ	2	475	849													1
туре	IS	AII		L	3	889	1053													
Name	like			L	4	1059	2636													
GenelD	=			L	5	2626	4296													
Locus	like			L	6	4318	5373													
Start				L	7	5427	6038													
Length				L	8	6136	6249													
				L	9	6239	6317	_	ļ											
Regions				L	10	6344	6607												~	
% GC	<			L	11	6647	6796			<									>	
CAI	>			ŀ	12	6902	7360			Control N	umbering									
EC#	like			ŀ	13	8315	9187			🔽 Assian I	Names bu oro	ter of appe	Parance							
Product	اندم			ŀ	14	9411	10655			Assign I	Producto by ore	rder of an		, 						
Freedor	11.0			ŀ	15	11265	11486			Assign	moducis by t	nuer or ap	pearan	ce						
Function	іке			H	15	11088	12100				awrite existing	j values / http://www.								
FeatureID) =			ŀ	17	12233	12637				Uverwrite de	fault Name	es & Pro	oducts	only					
Tag	like			ŀ	10	13031	13438	v		Assign i	new Locus T	ags by ord	ler of ap	opeara	ance					
🗖 Hide I	anored	Feat	ures				>			Locus I	ag Prefix E	TUDE	Searc	sh NU	BI dat	tabas	<u>e</u>			
			1	ľ	Insert De	elete Por	et Valid	ate		Reassign	Gene Data	Locate (Gray Ho	oles >	30)0 🛓	🖨 bp			
Sele	ont All F	eatur	es	-					1								_			_1
	 € €	₹₽		1-	14998	P	osition : 7	248		Con	trols >> Map	Map:	>> Cont	trols						
1 2 3		4		ŝ	;)	6	7	ol		12	(13	14		(16	6	17	18 19	20>	21)
21 Feature:	s	Live																14998	8	?

The genes have been renumbered (notice the tRNA is now gene 9, whereas before it didn't have a number), and the locus tags have been adjusted.

I will now start with Gene 1.

Gene 1:

We need to decide: if this gene is a gene, if it is really gene 1, and where its start is. To do this, I will examine five pieces of data: coding potential in GeneMark TB, Glimmer/GeneMark calls, ribosome binding site (RBS) scores, gene gap/overlap with preceding gene, and BLAST alignment with previously annotated genes.

Coding Potential: From looking at the GeneMark TB output, it appears that the coding potential starts in this genome around 200 or so bp. There aren't any ORFs upstream of the called Gene 1 with coding potential, so I am confident that this gene is, in fact, Gene1.

Glimmer/GeneMark: From the Notes field in my auto-annotation, I can tell that GeneMark and Glimmer have called this gene; Glimmer starting at bp 83, and GeneMark starting at bp 215. This means that the GeneMark call does not encompass all of the coding potential as shown by the GeneMark TB output. However, both programs called the gene, and there is good coding potential in the GeneMark TB output. So I am confident that this ORF is a gene, and now just need to resolve what the start coordinate should be.

DNA Etude	_anno	tated	d																
Overview	Featur	res F	Reference	s	Sequence	Docum	entatio	on											
Sort By	Index	-	4		Name	Start	Sto	p	^	Descripti	on	Sequence	Prod	uct Regions	s Blast	Contex	: Valida	ion	
Select Fe	atures	Direc	t SQL	Þ	1	83	466	;		Name	1		_	GenelD					
Тире	is Z	411	-	Ц	2	475	849			Тире		-		<u></u>					
туре			<u> </u>	Ц	3	889	105	i3		iype		2	<u> </u>	<u>u</u>	CTUDE				
Name	іке				4	1059	263	6		Start			83	Locus Lag	ETUDE.	_1			
GenelD	=			Ц	5	2626	429	16		Stop			466	Regions					1
Locus	like			Ц	6	4318	537	3		Length	384			Tag					
Start				Ц	7	5427	603	18		Direction	Forv	ard			·				
Lenath	Ξİ.			H	8	6136	624	9		Translativ	on Ta	ble Und	efined						Ţ
Desires				H	9	6239	631	7		EC Numb	on re	Die Jona	onnoa						
Regions				H	10	6344	660	<u>//</u>											
%GC	<			H	11	6647	675	16		Product									
CAI	>			H	12	6902	736	0		an1									
EC#	like			H	13	8315	918	97 		gp.									
Product	like [H	14	3411	100	00		Eurotion									
Function				H	10	11260	114	00		Tunction									
Function	iike			H	10	1000	121	27											
FeatureID	=			H	10	12233	120	20		 Notes									
Tag	like			۳	10	13031	1.54	100	~	Original	Glimn	ner call @	bp83 h	as strength 7.1	02: Genel	Aark calls	start at 2	15	
🔲 Hide I	anored	Featur	res	<				>											
Solo				1	nsert De	lete P	ost	/alidati	e										\sim
			NN NI		4000		Deski		100		Combo	da x x kitaa		an XX Control					
	a 2				4336		Positio	n : 14:	520 1/11		Jonth	us >> Maj	J ♥ M	ap >> Control					
1 X 2 X 9	6	4	X	-5	K	6 >	7	> <mark>11</mark> (0	N.	12>		(13		14	((16		18 1	9 > 2	<u>.0> 21></u>
21 Features	3	Live															1499	98	a <mark>?</mark>

RBS scores: Click on the first highlighted green bar in the "Frames" window, and then click the "RBS" button at the lower right of the Frames window. A new window will pop-up.

	hoose OR	F start								-		×
Start Seler	s : 10 cted : 1	ORF Start : ORF Stop ORF Length	:83 :466 h:384	5' End 3' End	Cdn 66.7 38.4	1 Cdn2 66.7 70.6	Cdn 69.0 54.1	3 Le) 1 2	ength 26 56	Doci	ument]
							_			<u> </u>	275	_
	Shine D	algarno	Sequence	e of t	the	Region	. S	tart	Start	OR	F	\mathbf{a}
#	Score	Space	Upstream	n of t	the	Start	c	odon	Positio	n Lei	ngth	
1	525	8	ACCCCTG	GTAGGI	AGGC	TAAATC	G	TG	83	384	4	
2	345	9	GAAGAAC	GCGCA	GGAA	TACGCG	G	TG	209	25	з	
з	420	8	CGCGCAG	GAATA(CGCG	GTGCAG	A	TG	215	25	z	
4	441	7	GAACGTT	TGGAC	CAAT	CAGGGG	A	TG	272	19	5	
-	441	0	CONCERC	- eree	.	CACACT	т	TC	205	17	1	⊻
										f	r e	?

The start at position 83, the one called by Glimmer, has a higher Shine-Dalgarno (RBS) score (525) than the GeneMark start at position 215 (420). The start at position 83 is yields the longest possible gene as well.

Gap/Overlap: Since it is gene 1, we can omit determining the gap or overlap with the upstream gene (as there isn't one!)

BLAST data: If I click on the BLAST tab (see below), I can see that the genes in GenBank that align well with my gene. Our top hit is (as expected from our phamerator view) to

LeBron gene 1. More importantly, when we look at the alignment, we see that "Query 1" aligns with "Sbjct 1". This means that we selected the same start codon that was chosen in the LeBron annotation. While it is not necessary to pick the same start codon as another closely related previously curated phage, it is necessary to examine all possible start codons, weigh the data, and make the best choice. Knowing that we have selected the same start codon as a similar phage makes it more likely that we have not accidentally missed a more appropriate start codon and that our reasoning is supported by previous examination of other annotators.

	Etude	_ann	otat	ed														_ 🗆 🗙
	Overview	Feat	ures	Reference	s	Sequence	Documer	ntation										
	Sort By	Index	-	•		Name	Start	Stop	^	D	escription	Sequence	Product	Reg	ions Blast	Conte	xt Valida	tion
	Select Fe	atures	Dire	ect SQL	Þ	1	83	466		Г	Score	Target Des	cription			1		<u> </u>
1	T		All			2	475	849		┢	578	ap1 (Mycol	pacterium	ohage	: LeBron]			1
I	туре	IS 	A	<u> </u>		3	889	1053		ŕ	560	ap1 (Mycol	oacterium	phage	JoeDirt]			
1	Name	like				4	1059	2636		F	167	ap1 (Mycol	oacterium	phage	Faith11			
L	GenelD	=				5	2626	4296		E		51-1-5						
	Locus	like				6	4318	5373		L								
	Start	>				7	5427	6038										
	Length					8	6136	6249		Г	BLAST Hit -						_	Europe I
ł	D ·					9	6239	6317	-	A	Accession YI	P_0038571	32				-	Export All
L	Regions	2				10	6344	6607			al 31 Jenoth 10)4360950)7					-	Delete
ł	% GC	<				11	6647	6796		L N	Aax Score 57	-7 78	Date	10/1	0/2011		-	Delete All
L	CAI	>				12	6902	7360		Ŀ	High-Scoring	Paire (HSP	on					
	EC#	like			_	13	8315	9187		h		Alianment	í					
1	Product	like			H	14	9411	10655		L,	пог рака	Alignment						(I
	Franking	10			H	15	11265	10100			1 MGR	GRGKXXX	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	XXX	XXTRSAWEAK	VAA	KPKNAQE	YAVQ: 🔨
ł	Function	іке			H	10	10000	1200			1 MGR	GRGKDPS	SPGGRSI	USR	PGTRSAWEAK	VAAI	KPKNAQE	YAVQ:
L	FeatureID	=			H	10	12233	12037										
	Tag	like			⊢	10	13031	13430	×		51 WEVI	EKPNVWT	NQGMHAA	GIE	TLTMRKGDAY	VYA:	FFTWPNG	RIRT
L	Hide I	anorea	d Feal	tures	<			>	_		51 WEVI	EKPNVWT	NOGMHAA	GIE	TLTMRKGDAY	VYA'	IIIIIII FFTWPNG	RIRT
				1	-	Insert De	lete Pos	t Valida	ite									►
-	Sele		eatur a La	1 29	_							1	E U					
		₩	₹₽		-	14998	P	osition : 88	390	_	_ I∕ Contro	ols >> Map	Map :	>> Co	ntrois			
ł	1 2 3		4		5	K	6 >	7 >₿₩	0 Ø	1	2>	(13	(14		((16	$ 17\rangle$	18	19 > 20> 21>
	21 Features	:	Live	Э													149	98 🛛 🛃 🤶

At this point, I am ready to make my decision. The BLAST data, RBS score, GeneMark TB coding potential, and Glimmer output all suggest that the best start for this gene is bp83.

Now we need to decide if this gene has a function. Click the "product" tab

DNA Etude	_ann	otat	ed																
Overview	Feat	ures	Referen	ices	Seque	nce	Docur	mentation											
Sort By	Index	-	[]	•	Name		Start	Stop	^	Ī	Description	Sequence	Product	Regions Bla	st	Conte	st V	/alidation	
Select Fe	atures	Dire	- ect SQL		1		83	466			128 Residues	MW = 14	.36 kd	Kyte Hydro =	-0.0	983			
T		All			2		475	849				pl = 8.47		OMH Hydro :	= -0.0	0246			
туре	IS 				3		889	1053			MGRGRGKDP	SSPGGRSR	DSRPGT	RSAWEAKVAAKI	KNA	QEYAU	7QMA	ESLGWE	VEK 📉
Name	like	<u> </u>		- -	4		1059	2636	_		PNVWTNQGM	HAAGIETL	TMRKGD.	AYVYATFTWPNC	RI	RTVDVI	RVNG	FHEDME	GSE
GenelD	=				5		2626	4296			RKDKRKAMK	RITRKACK	2						
Locus	like				6		4318	5373											
Start	>	<u> </u>		- -	7		5427	6038											
Length		<u> </u>		- -	8		6136	6249	-										
Destant		<u> </u>		- -	9		6239	6317	-	1									
Regions		<u> </u>		- -	10		6344	6607	-										
%GC	<				11		6647	6796	-										
CAI	>				12		6902	7360	-										
EC#	like	í –		- -	13		8315	9187	-										
Product	lika	i—		- -	14		9411	10655	-										
Emotion	10	<u> </u>		- -	15		11265	11486	-										
Function	іке	<u> </u>		- -	15		10088	12100	-										
FeaturelD	=			_ H	17		12233	12637	-11										
Tag	like			ŀ	10		13031	13430	~				_						
🔲 Hide I	Ignore	d Feal	tures		< []			>											
Sele	oct All F	Featur			Insert	Del	ete F	Post Valid	ate		LIVP	мсц	JGA	STDE	к	RН	0 1	I Q W	FΥ
सन्दि		al∍		1	- 14998			Position : 1	470	18	Contr	ols >> Map	🔽 Мар	>> Controls					
1 2 3 21 Feature:	s	4 Live	K		5	X	6	7	0(12	(13	14		16	17>	1	8 19 14998	20 21

Copy this sequence and paste into NCBI's BLASTP page.

http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blast p&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome

5	BLAST	Basic Local Alignment Search Tool
د `	Home	Recent Results Saved Strategies Help
► N	CBI/ BLAST/	blasto suite
bla	sta blasta	harden state
Dia	<u>sui</u> Diastp	
	Enter Qu	ery Sequence BLASTP programs search protein databases using a protein query. more
	Enter acces	sion number(s), gi(s), or FASTA sequence(s) 🕢 <u>Clear</u> Query subrange 😡
	>gp:83-466 MGRGRGKDPSSI	amino acid sequence): 127 residues GGRSRDSRPGTRSAWEAKVAAKPKNAQEYAVOMAESLG From
	WEVEKPNVWTN GFHEDMFGSERI	CHRAAGIETLITHRKGDAYYYATFTWENGRIRTYDYNN IDRRAMKELLEXYGK
		То
	Or, upload f	Browse)
	Job Title	
		Enter a descriptive title for your BLAST search 😡
	🗆 Align two	or more sequences 😣
	Choose	Coarch Sat
	Detabase	
	Omenion	Non-redundant protein sequences (nr)
	Organism Optional	Enter organism name or idcompletions will be suggested
		Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. 🥹
	Exclude	Models (XM/XP) Uncultured/environmental sample sequences
	Optional Entroz Ouor	
	Optional	y
		Enter an Entrez query to limit search 🐨
	Program	Selection
	Algorithm	blastp (protein-protein BLAST)
		O PSI-BLAST (Position-Specific Iterated BLAST)
		O PHI-BLAST (Pattern Hit Initiated BLAST)
		Choose a BLAST algorithm 🥹
	RIACT	Search database Non-redundant protein sequences (nr) using Blasto (protein-protein BLAST)
	DLAST	

By pressing the "BLAST" button at the bottom of the page, we get the following result:

✓ <u>¥alue</u> Lin 7e-69 1.4 C
✓ <u>value</u> Lin 7e-69 1.4 C
✓ <u>value</u> Lin 7e-69 C 1.4 C
→ <u>E</u> <u>value</u> Lin 7e-69 C 1.4 C
E Lin 7e-69 C 1.4 C
∠ <u>E</u> Lin 7e-69 C 1.4 C
E valueLin7e-69C1.4C
7e-69 C 1.4 C
1.4 G
7.5
9.9 <mark>G</mark>
7.5 9.9

Our top hit is (as expected from our Phamerator view) to LeBron gene 1. LeBron is the only hit listed above that has an acceptable E value, and therefore we discount the other hits and their possible functional assignments.

HHPred.

http://toolkit.tuebingen.mpg.de/hhpred

Copy the amino acid sequence into HHPred's sequence field. Make sure to remove the "Z" at the end (DNA Master represents stop codons with a Z in the product field).

	HOME			Le	ogin PDBalert	Personal Databases	Contact Imprint	Disclaimer Help
			Bioinformatics Tool Max-Planck Institute for Devel	kit opmental Biology				Quickfinder _
MAX-P	LANCK-GESELL	SCHAFT	Search Alignment Sequence Ar	alysis 2ary Structure	3ary Structure	Classification Utils		
Show	w results o	f						
job:			CS-BLAST HHblits HHpred HHsen	ser HMMER3 PSI-BLAST	PatternSearch Pr	otBLAST SimShiftDB		
	Show results	5	HHpred - Homology deter NEW: Official	ction & structure p server results for CASP	rediction by H structure prediction	MM-HMM compa	arison (Help)	
			Input					
Rece	ent jobs:							
	8237908	HHPR	Paste protein sequence or multiple	SLGWEVEKPNVWTNQGMHAAGIE	ILTMRKGDAYVYATFTWPNG	RIRTVDVRVNGFHEDMFGSERKI	KRKAMKEILEKYGK	
	3443134	HHPR	alignment					
	2933052	HHPR						
	8633021	HHPR		6				
	8646597	HHPR		(L.				
	1309395	HHPR	or upload a local file			(E	Browse	
	5901234	HHPR	Select input format	FASTA V				
	9593792	ннрк	Select input format					
	6939899	HHPR				Deach form	Cubmitiah	
	1341118	HHPR				Reset form	Submit job	
	5777435	HHPR	Search Options					
	1616175	HHPR						
	1208530	HHPR	Select HMM databases	Standard	G	ienomes		
	6741199	HHPR	(hold Ctr to select several)	pdb70_80ct11	Arabidopsis_thalla	na 🧴		
	8088012	HHPR		pdb_on_hold_6Oct11	Caenorhabditis_el	egans 🔍		
	9578553	HHPR		scop70_1.75	Drosophila_melan	ogaster		
	3212046	HHPR		pfamA v25.0	Mus musculus	÷		
	8324397	HHPR			Thes_mascalas	1		
	8188366	HHPR	MSA Generation Method	🙂 HHblits 🕖 Psiblast				
	5150140	HHPR	Max HHblits iterations	3 -				
	9166533	HHPR	Plaza Hindrids Relations					
	8706971	HHPR	Score secondary structure	🖲 yes 🔘 no 🔘 predict	ed vs predicted only			
	7668194	HHPR	Alignment and	🙆 local 🔘 clobal				
	9704715	HHPR	Alignment mode	Uluar U yivbar				

Click "Submit job" (at the far right, just above the beige bar labeled "Search Options")



Query Tue_Oct_11_17:53:21_+0200_2011 (seq=MGRGRGKDPS...MKEILEKYGK Len=127 Neff=1.0 Nseqs=1) Parameters score SS:yes search-local realign with MAP:no

	No	Hit	Prob	E-value	P-value	Score	SS	Cols	Query HMM	Templat	e HMM
	1	<pre>3adg_A F21M12.9 protein; HYL1,</pre>	84.0	3.4	0.00013	24.7	5.9	48	80-127	23-73	(73)
	2	1r91_A Glycine betaine-binding	79.9	0.69	2.7E-05	31.9	1.8	44	40-89	25-68	(309)
	3	2cpn_A TAR RNA-binding protein	72.0	4.2	0.00016	25.4	3.8	37	88-124	45-82	(89)
	4	luhz_A Staufen (RNA binding pr	66.8	7.3	0.00028	24.4	4.0	40	88-127	35-75	(89)
	5	31q0_A Proastacin; metallopept	65.8	2.7	0.00011	30.7	2.1	60	58-124	11-74	(235)
	6	2dix_A Interferon-inducible do	65.1	7.6	0.00029	23.8	3.8	37	88-124	36-73	(84)
	7	2h0e_A Transthyretin-like prot	59.9	5.5	0.00021	27.1	2.7	33	60-92	18-50	(121)
\square	8	1n91 & DHOT_LOV1 nutative blu	59 2	6.2	0 00024	22.2	2 5	33	77-113	3_35	11091

Only one of these alignments has a Probability score of above 80, and it from a small portion of our query to a protein in *Arabidopsis thaliana*. We will consider this not a match. Finally, the Hatfull-labeled maps also suggest that there is no known function for gene 1 in LeBron.

Gene 1 has no known function (NKF).

Our last task is to add our annotation rationale to the Notes field for this gene.

In the Click on the Description tab in the right-hand section of the Features tab.

In the Notes field, add your notes.

Things to include: The gene coordinates for your gene call. Is this the longest possible gene for this gene call? Is this the Glimmer/GeneMark call? What is the gap or overlap between this gene's start and the previous gene's stop? Does this start have the best RBS score? Does this gene match anything in GenBank when you BLAST it? If so, what? What is the alignment between the start that you chose and the closest GenBank match? Is there a known function?

Etude	_ann	otate	e d							×
Overview	Featu	ures	Reference	es	Sequence	Docume	ntation			
Sort By	Index	-	4		Name	Start	Stop	^	Description Sequence Product Regions Blast Context Validation	
Select Fe	atures	Dire	ect SQL	I	1	83	466		Name 1 GenelD	
Tune		All			2	475	849			
туре	15	A			3	889	1053			_
Name	like			Ц	4	1059	2636		Start 83 Locus Tag ETUDE_1	_
GenelD	=			Ц	5	2626	4296		Stop 466 Regions	1
Locus	like			Ц	6	4318	5373		Length 384 📰 Tag	-1
Start	>			H	7	5427	6038		Direction Forward	
Lenath	5			H	8	6136	6249		Translation Table Undefined	٦I
Begione				H	9	6239	6317		EC Number	-1
negions evice				H	10	6344	6607			5
% նԼ	<			Н	12	6047 6002	7260		Product	2
CAI	>			Н	12	8302 8315	9197		gp1	71
EC#	like			H	14	9411	10655			ā.
Product	like			H	15	11265	11486		Function	2
Function	like			H	16	11588	12100		1	5
FesturelD	_			F	17	12299	12637		N	
TEditalElD				F	18	13031	13438		Notes	-
Tag	іке			Г				¥	🖞 83-466. Glimmer call. Longest possible ORF. Includes all coding potential. Best 🛛 🖉	•
🔲 Hide I	gnored	l Feat	ures	<			>		RBS score (525). Aligns 1:1 with LeBron gene 1. NKF.	
Sele	rt ∆ll F	eatur	P\$		Insert De	lete Po:	st Valida	te		
RAA	@ [6	₹ Þ		1 - 1	14998	P	osition : 88	890	0 🔽 Controls >> Map 🔽 Map >> Controls	
	1	4	K	5	X	6	7 84	1		21)
21 Features	:	Live	;			- 71	2.005	1, 1	14998 🛃	?

It is important that you physically type in the gene coordinates into your comments, as in some cases I have received files in which people believe they have changed their start coordinates and were not actually able to. In case there are any discrepancies between the gene coordinates that you think you are choosing and the gene coordinates that are actually saved into the file, it is important that you write what your gene coordinate choices are into the notes here. It is also important that you report the BLAST alignments here, for two reasons: It is possible to generate spreadsheets of the gene data fields, including the notes, that can be very useful for genome checking. These spreadsheets will not include data from the BLAST tab. And if you ever accidentally lose your BLAST data (say, from parsing your documentation, or from corrupting your file) you'll have a record of what the alignment was without having to BLAST your entire genome again.

Post your changes to the notes field, either by clicking "Post", or by moving to gene 2 by clicking on the corresponding row in the central column.

Gene 2:

On the feature tab, click "2" in the central column.

In the Notes for gene 2, we can once again see that Glimmer and GeneMark have disagreed on the start for the gene (475 for Glimmer and 514 for GeneMark). However, as with gene 1, we can see that both programs have called the gene, and that there is good coding potential for the gene in the GeneMark TB output. So we will agree that this is a gene, and now just need to resolve its start.

Now we check our three criteria for start selection: coding potential, gene gap/overlap and RBS scores.

Coding Potential: the trace for the GeneMark TB coding potential doesn't start to rise until about bp 600 or so, so both the Glimmer and GeneMark start codons encompass all the coding potential.



Look at the Frames window for gene gap/overlap:

We can see here that the called start (the Glimmer start) represents the longest possible start for this gene—any extension would run into the upstream stop codon. There is no gene overlap, and there is a 9bp gap.

The RBS scores: Click in the box with the second green highlighted bar, and then click the RBS button on the lower right side of the frames window.

	A C	hoo	ose OR	F start														×
S	tart	s:9		ORF Sta	rt : 41	75			Cdn	1 Cdn	2 C	dn3	i Le	ngth		Joeu	ment	1
S	elec	cted	:1	ORF Sto	р:8	49	5	5' Enc	38.5	69.2	- 4	6.2	3	9		2000	ment	
				ORF Ler	igth :	375	3	l' Enc	41.1	68.5	6	2.2	3	34				
												Π					795	
		Sh:	ine D	algarn	o S	equ	lence	of	the	Regio	n	St	art	Start		ORF		^
#		Sc	ore	Space	U	pst	ream	of	the	Start	;	Co	don	Positi	on	Len	gth	
1		37	8	8	A	GTA	logge.	AAGI	AGGO	GGATI	Т	AT	G	475		375		
2		15	0	9	A	AAC	GCTG	ттсо	ACCI	GGCC1	Т	AT	G	514		336		
3		21	0	7	Т	GAC	GCTC	CGTO	GAGO	GTCCI	Т	ΤT	G	586		264		
4		50	4	7	С	GCT	CCGT	CGAC	GGTC	сттт	G	ΤT	G	589		261		_
5		25	2	7	C	CCT	TCCC	C T A T		CACAC	۰ ۲	тт	۰ ۳	e1 e		204		×
1																	2	
																- fi	8 🗎	?

Again, the Glimmer call at 475 has a higher score than the GeneMark call at 514.

Etude	_ann	otate	d												
Overview	Feat	ures	Reference	es	Sequence	Docume	entation								
Sort By	Index	-	4	ſ	Name	Start	Stop	^	Ī	Description	Sequence	Product Region	is Blast	Conte	xt Validation
Select Fe	atures	Dire	et SQL]		1	83	466		Г	Score	Target Desc	ription			<u> </u>
-		A 11		Ð	2	475	849		h	572	an2 (Mucoh)	acterium phage Jo	eDirt1		1
Туре	IS	AII		E	3	889	1053		ľ	560	an2 (Mycob)	acterium phage i e	Bronl		
Name	like			L	4	1059	2636		lŀ	437	ap2 (Mycob)	acterium phage Ea	aith11		
GenelD	=			L	5	2626	4296		lŀ	193	hypothetical	protein SVEN 39	85 (Streptor	mvces ·	
Locus	like			L	6	4318	5373		lŀ	191	ap10 (Mycot	bacterium phage () Jmegal	.,	
Start	5	, 		L	7	5427	6038		ľ		36	, ,			
Lawath				L	8	6136	6249			BLAST Hit					
Length	2			L	9	6239	6317	_		Accession A	EK07049				Export
Regions	>			L	10	6344	6607			GI 3 Lavetta 1	39781215				Export All
% GC	<			L	11	6647	6796			Length I May Score 5	24 72	Date 10/10/2	011		Delete
CAI	>			L	12	6902	7360		L				.011		Delete All
FC#	like	, 		Ŀ	13	8315	9187	-11		High-Scorin	g Pairs (HSP)	1			
Deadarat	11			Ŀ	14	9411	10655	-11		HSP Data	Alignment				
Product	шке			Ŀ	15	11265	11486	-11		1 MGD	TVKNAVP P	GLMAAGKEL WE	SVASERE	L DAPS	SRVLLLN ACRI. 🔼 📗
Function	like			Ŀ	16	11588	12100	- 11			11111111	IIIIIIIIIIIIIIIIII	 29173 9 8 9 9	 1 DADS	TULLN ACDT
FeatureID) =			Ŀ	17	12299	12637	- 11		1 100	I VILLANDE I	GEHAAGKEE WE	SUNDERE.	L DAP	SKULLEN ACKI.
Tag	like			ŀ	18	13031	13438	-		51 LDQ	EIDGRLL S	YNQRGDEVI NF	LISEHRQ	Q YTTI	LANILGK MGLG
- 		d En elu		l.	e [m]		3		1	51	IIIIII I	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII		 	
	ignored	Ireau	ues I	P		uu la				31 100	EIDGEDE 5	SINGKGDEVI NF	-DISERKQ	Q 1111	ANTIOR HOLD
Sele	ont All P	eature			Insert D	elete Po	Valio	late	L						
N 4 4	0	\mathbb{R}		1.	14998	F	Position :	1378	0	Cont	rols >> Map	Map >> Control	ols		
1 2 3		4	X	ł	5	6	7)	606		12)	(13	14	((16	17	18) 19) 20) 21)
21 Features	s	Live					2 1.0	5 11		/					14998 🛃 ?

Examine the BLAST tab:

Once again, the best match aligns 1:1 with JoeDirt and LeBron.

So we will pick the Glimmer call at 475 as our gene start, and enter the appropriate description into the notes. Now that we have an upstream gene, we will also write the gap/overlap in bp of this gene with the previous one.

Functional assignment: If we BLASTP this gene outside of DNA Master on the NCBI website, or examine Phamerator, or the Hatfull-approved genome maps with functions, we will see this gene is the small subunit of the terminase.

👺 Etude_annotated 📃 🔲 🔀																			
Overview	Features Reference			es	Sequence	Documentation													
Sort By	Index	-	4	ſ	Name	Start	Stop	^	Īſ	Description	Sequence	Product	Regions	Blast	Cor	ntext	Valida	ation	
Select Fe	eatures	Dire	ect SQL		1	83	466			125 Residues	MW = 13	.76 kd	Kyte Hy	ydro = -0	.0587				1
Ture		All		Ľ	2	475	849				pl = 7.75		OMH H	lydro = ·	0.0230				
туре	15			L	3	889	1053			MGD TVKNAVI	PELMAAG	KELWESV	ASERELI	APSRV	LLLNA	CRIA	DRLD	QLDQ:	BI 🔼
Name	like	<u> </u>		L	4	1059	2636			DGRLLSYNQ	RGDEVINP	LISEHRQ	QYTTLAN	IILCKM	GLGEI	PKAR	QENS	RWD E	LA
GenelD	=			L	5	2626	4296			KKRABRAAK	LAUASZ								
Locus	like			L	6	4318	5373												
Start	>	í –		L	7	5427	6038												
Length		<u> </u>		L	8	6136	6249												
Design		<u> </u>		⊢	9	6239	6317	_											
Regions	2	<u> </u>		⊢	10	6344	6607												
% GC	<			⊢	11	6647	6796												
CAI	>			⊢	12	6902	7360												
EC#	like			⊢	13	8315	3187												
Product	like	í—		H	14	3411 11005	10600												
Eurotion	lika	<u> </u>		ŀ	10	11203	12100												
		<u> </u>		F	17	12299	12637												
FeatureIL		<u> </u>		F	18	13031	13438												~
Tag	like			F	10	10001	10400	¥	ľ										
T Hide Ignored Features																			
Sele	Select All Features Insert De					lete Po:	st Valida	ate	Ľ	LIVP	мсц	JGA	STD	ЕК	Rŀ	1 0	NG	W	FΥ
M M P P P P P P P P P P P P P P P P P P																			
Zirreature	\$	LIVE	-														143	50	18

Click on the product tab in DNA Master (or on the gene in Phamerator

As in gene 1, copy and paste the amino acid sequence into the NCBI BLASTP page. The BLAST result show more hits this time, not only LeBron, but multiple other phages. This gene is the small subunit of the terminase, and so we must add the function into our annotation. The LeBron alignment is still "query 1 to sbjct 1", indicating that gene 2 of LeBron and our gene 2 of Etude use the same start codon.



Descriptions

Legend for links to other resources: 🛄 UniGene \Xi GEO 🖸 Gene 🧕 Structure 🛄 Map Viewer 👫 PubChem BioAssay

Sbjct 121 AQAS 124

P 003857133.1 C P 818311.1 C P 06825994.1 F P 002882537.1 F P 002781224.1 F P 003162104.1 F P 0036210.1 F P 00501003.1 F	pp2 [Mycobacterium phage LeBron] >gb]ADL70969.1] gp2 [Mycobacterium phage LeBroi gp10 [Mycobacterium phage Omega] >gb]AAN12654.1] gp10 [Mycobacterium phage Om ypothetical protein SSBG_02573 [Streptomyces sp. SPB74] >gb]EDY44611.1] hypothetical protein ROP_0320 [Rhodcoccus opacus B4] >db][BAYE475.1] ypothetical protein ROP_0320 [Rhodcoccus opacus B4] >db][BAYE375.1] ypothetical protein ROP_0320 [Rhodcoccus opacus B4] >db][BAYE375.1] ypothetical protein ROP_0320 [Rhodcoccus opacus B4] >db][BAYE375.1] hypothetic ypothetical protein Iden_2164 [Jonesia denitrificans DSM 20603] >gb]ACV09801.1] hyp	246 87.8 72.4 57.0 56.2	246 87.8 72.4 57.0	100% 89% 59%	5e-64	G
818311.1 g 06825994.1 F 002882537.1 F 002781224.1 F 003162104.1 F 706493.1 F 06501003.1 F	p10 [Mycobacterium phage Omega] >gb AAN12654.1] gp10 [Mycobacterium phage Om ypothetical protein SS8C_02573 [Streptomyces sp. SP874] >gb EDV44611.1 hypothet ypothetical protein Bcav_2527 [Beutenbergia cavernae DSM 12333] >gb ACQ80775.1 ypothetical protein ROP_40320 [Rhodococcus opacus B4] >db][BAH52279.1] hypothetical protein Long_12164 [Donesia denitrificans DSM 20603] >gb ACV098011.1 hyp	87.8 72.4 57.0 56.2	87.8 72.4 57.0	89%		
06825994.1 h 002882537.1 h 002781224.1 h 003162104.1 h 706493.1 h 06501003.1 p	vpothetical protein SSBG_02573 [Streptomyces sp. SPB74] >gb[EDV44611.1 hypothet vpothetical protein Bcav_2527 [Beutenbergia cavernae DSM 12333] >gb[ACQ80775.1] vpothetical protein ROP_0320 (Rhodococcus opacus B4) >db][BAH52279.1] hypothetic vpothetical protein Jden_2164 [Jonesia denitrificans DSM 20603] >gb[ACV09801.1] hyp	72.4 57.0 56.2	72.4 57.0	59%	4e-16	G
002882537.1 h 002781224.1 h 003162104.1 h 206493.1 h 06501003.1 p	hypothetical protein Bcav_2527 [Beutenbergia cavernae DSM 12333] >gb[ACQ80775.1] hypothetical protein ROP_40320 [Rhodococcus opacus B4] >dbj[BAH52279.1] hypothetic hypothetical protein Jden_2164 [Jonesia denitrificans DSM 20603] >gb[ACV09801.1] hyp	<u>57.0</u> <u>56.2</u>	57.0	0010	2e-11	
002781224.1 h 003162104.1 h 06493.1 h 06501003.1 p	hypothetical protein ROP_40320 [Rhodococcus opacus B4] >dbj BAH52279.1 hypothetic hypothetical protein Iden_2164 [Jonesia denitrificans D5M 20603] >gb ACV09801.1 hyp	56.2		55%	7e-07	G
03162104.1 h 06493.1 h 6501003.1 p	hypothetical protein Jden_2164 [Jonesia denitrificans DSM 20603] >gb ACV09801.1 hyp		56.2	75%	1e-06	G
06493.1 H 6501003.1 p		53.5	53.5	69%	8e-06	G
6501003.1 p	hypothetical protein RHA1_roub562 [Rhodococcus jostii RHA1] >gb[ABG98335.1] hypoth	50.1	50.1	56%	9e-05	G
	phage terminase, small subunit, P27 family [Micrococcus luteus SK58] >gb EFD51948.1	48.5	48.5	66%	2e-04	
7714862.1	conserved hypothetical protein [Corynebacterium pseudogenitalium ATCC 33035] >gb EF	41.6	41.6	83%	0.029	
55890.1 g	p25 [Mycobacterium phage Wildcat] >gb ABE67630.1 gp25 [Mycobacterium phage Wil	40.4	40.4	82%	0.074	G
6832385.1 C	conserved hypothetical protein [Rhodococcus equi ATCC 33707] >gb[EFG59276.1] conse	37.7	37.7	65%	0.43	
3646336.1	hypothetical protein BbifN4_04215 [Bifidobacterium bifidum NCIMB 41171] >ref YP_003	37.7	37.7	73%	0.43	
85333.1 s	serine 3-dehydrogenase [Mycobacterium smegmatis str. MC2 155] >gb ABK70390.1 ser	37.7	37.7	55%	0.44	G
03886949.1 s	odium/hydrogen exchanger [Cyanothece sp. PCC 7822] >gb[ADN13674.1] sodium/hydr	34.7	34.7	37%	4.4	G
4402134.1	MSHA biogenesis protein MshM [Vibrio cholerae TMA 21] >gb EE015293.1 MSHA biogen	34.3	34.3	56%	5.1	
64435.1 P	PREDICTED: similar to sterile alpha motif domain containing 4 isoform 4 [Canis familiaris	33.9	33.9	61%	5.9	UG
02894263.1 t	ransketolase [Tolumonas auensis DSM 9187] >gb ACQ94677.1 transketolase [Tolumona	33.5	33.5	71%	9.0	G
06399327.1 a	acvi-CoA dehvdrogenase domain protein [Micromonospora sp. L5] >reflYP_003833223.1	33.5	33.5	41%	9.7	
Select All	Get selected sequences Distance tree of resu	<u>its Multi</u>	pie alignme	ent		
> <u>r</u> e	ef YP_003857133.1 G gp2 [Mycobacterium]	phage Le	Bron]			
gb i Lengt GENI	ADL70969.1 9p2 [Mycobacterium phage L th=124 E ID: 9711608 2 gp2 [Mycobacterium phage	e LeBron	1]			

 Query
 1
 MGDTVKNAVPPGLMAAGKELWESVASERELDAPSRVLLLNACRIADRLDQLDQEIDGRLL MGD VKN VPPGLMAAGKELWESVASERELDAPSRVLLLNACRIADRLDQLDQEIDGRLL
 60

 Sbjct
 1
 MGDGVKNTVPPGLMAAGKELWESVASERELDAPSRVLLLNACRIADRLDQLDQEIDGRLL
 60

 Query
 61
 SYNQRGDEVINPLISEHRQQYTTLANILGKMGLGELPKAKQENSRWDELAKKRAERAAKA SYNQRGDEVINPLISEHRQQYTTLANILGKMGLGELPKAKQENSRWDELAKKRAERAAKA
 120

 Query
 121
 AQAS
 124

We are confident enough with the BLASTP and map assignments that it is not necessary to run HHPred.

Add detailed annotation notes as in gene 1. Make sure that you include the gene gap/overlap, the functional assignment, and the source of your functional assignment.
DNA Etude	_anno	otate	d															<
Overview	Featu	ires	Reference	es	Sequence	Documer	ntation											
Sort By	Index	-	4	ſ	Name	Start	Stop	^	Descriptio	on Seq	uence	Prode	uct Regions	Blast	Context	Valida	tion	
Select Fe	atures	Dire	ct SQL		1	83	466		Name	2		_	GenelD					
Tunn	. [All		<u>1</u>	2	475	849		Tune	-			CI					
туре	15	AI	<u> </u>	L	3	862	1053		Type	lus			<u>ui</u>					-
Name	like			L	4	1059	2636		Start			475	Locus Tag	ETUDE	_2			
GenelD	=			L	5	2626	4296		Stop			849	Regions				1	
Locus	like			L	6	4318	5373		Length	375			Tag					1
Start	S			L	7	5427	6038		Direction	Forward				,				I
Longth				L	8	6136	6249		Translatia		Undefi	ned						1
Lengur				L	9	6239	6317		EC Numb	in Lable	Jonden	neu					•	1
Regions				L	10	6344	6607		EUNUMD	er								1
% GC	<			L	11	6647	6796		Deadurat									
CAI	>			ŀ	12	6902	7360		Product									1
EC#	like			ŀ	13	8315	9187		gpz									
Product	اللاحم			ŀ	14	9411	10655		Function								<u>×</u>	
F				┝	15	11265	11486		Terminas	e emall	subunit							1
Function	like			┝	15	11588	12100		r cirilinda	ic, smail	sabariik							
FeatureID	=			ŀ	17	12299	12637		Matas								<u></u>	
Tag	like					13031	13438	~	475-849.	Gene g RBS sc	ap 9bp. I ore 375	Glimm Ímucł	er start. Long higher than i	est possib GeneMar	ole ORF, a k start). Bl	ll coding AST alio	potential A	
, nuer	gnoreu	reau	ares (ľ					1:1 with a	JoeDirt.	[erminas	se, sm	all subunit(BL	AST, GFF	H Map)			
Sele	ot All F	eature	×	_	Insert De	ete Pos		te										
H H I	(€			1 -	14998	P	osition : 99	992		ontrols >	> Map	M	ap >> Control	s				
1 2 3 21 Features	5	4 Live	X	ę	5	6	7) 🕅	o]≬	12		13		14	((16	17	18) 149	19 20 20 2 38 🛃	1) ?

Since this is a Hatfull Map approved function, we will add it to the Function field as well:

Gene 3: Look at the coding potential trace in the GeneMark TB output. The coding potential after Gene 2 shows a smaller peak in the top tier following the gene 2 peak.



Glimmer has decided to call this ORF a gene, GeneMark has decided to omit it.

DNA Etude	_anno	otate	d												•	(
Overview	Featu	ires	Reference	s	Sequence	Docume	ntation											
Sort By	Index	-	4		Name	Start	Stop	^	Descripti	on	Sequence	Prod	uct Regions	Blast	Context	Valida	ion	
Select Fe	atures	Dire	ect SQL		1	83	466		Name	3		_	GenelD					
Turne	i	All		L	2	475	849		Turne		c		CI.					
туре	n l	~"	<u> </u>	Ŀ	3	889	1053		Type		3	<u> </u>	<u>u</u>	C T UD C	_			
Name	іке			L	4	1059	2636		Start			889	Locus Lag	ETUDE	_3			_
GenelD				L	5	2626	4296		Stop			1053	Regions					1
Locus	like			L	6	4318	5373		Length	165			Tag					
Start	>			L	/	5427	6038		Direction	Forv	vard	=						
Lenath	S Í			┝	8	6136	6249		Translati	on T.	able Undel	ined						T
Begions				┝	10	6239	6317		EC Num	her								
e cc				⊢	11	6344 CC/7	6607											
շնե				⊢	12	6902	7260		Product									
CAI	>			⊢	13	8315	9187		gp3									
EC#	like			H	14	9411	10655											
Product	like [F	15	11265	11486		Function									
Function	like			F	16	11588	12100											
FeatureID				F	17	12299	12637											
T	- I			F	18	13031	13438		Notes									
Tag	іке ј						_	~	Original	Glimr	mer call @bj	o 889 I	has strength 6	.25 ** not	called by	GeneMa	rk	
🔲 Hide I	gnored	l Feat	ures	4			>											
Sele	et All F	eatur	P4		Insert De	elete Po:	st Valida	ate										
M M M	(⊕ [⊝	₹Þ		1 -	14998	P	osition : 1-	4708		Contr	ols >> Map	💌 M	ap >> Control	s 🔤				
		4	X	5	;)	6	7 84	00	12		(13		14	1 (16	17	18) 1	9 20) (21)
21 Features	3	Live	;			- 71	2 100									1499	38 🕴	a <u>?</u>

Since we see some coding potential in this frame and not in any others, and this ORF nicely fills a gap in the genome between gene 2 and gene 4, we are going to call this gene.



Now we need to pick a start codon. Examine the frames window:

There are at least four possible starts for gene 3 that will not overlap with gene 2 (we can't overlap gene 2 at all in this case, as genes 2 and 3 are in the same frame. Gene 2's stop codon would prevent translation of gene 3 from any earlier start.)

Now we use our five pieces of data to determine which start of the four possible starts we like the best.

Coding Potential: the earliest blip in the GeneMark TB coding potential trace is about 900 bp, so all four starts encompass all the coding potential.

Gene gap/overlap: The best start here is the longest start, which leaves no gap between genes. However, the tandem starts (start 2 and start 3) leave a fairly small gap as well, either 10 or 13 bp.

RBS scores:

B	NA C	hoose OR	F start									X
\$	Start	s:5	ORF Start :	850		Cdn 1	Cdn2	Cdn3	Lei	ngth	Document	1
\$	Selec	cted : 1	ORF Stop	: 1053	5' End	100.0	100.0	66.7	9		Docament	-
			ORF Length	n : 204	3' End	49.2	57.8	62.5	- 19	93		
E											898	
		Shine D	algarno	Sequen	ce of t	the R	egion	Sta	rt	Start	ORF	Γ
ŧ	ŧ	Score	Space	Upstre	am of t	the S	tart	Cod	on	Position	Length	
	L	247	6	TAAAGC	CGCTCA	GCGT	CCTAG	GTG		850	204]
2	2	247	6	TCAGGC	GTCCTA(GGTGC	ceece	ATG		859	195	
3	3	312	6	GGCGTC	CTAGGT	GCCGG	GGATG	ATG		862	192	
4	ł	252	7	GTCCAT	TCGTTC	GGGCC	GGGGT	GTG		889	165	
5	5	399	7	CTACTG	GTATAG(GCGCG	CTCAT	ATG		1006	48	
												?

According to DNA master, starts 1 and 2 have exactly the same score, which is only very slightly lower that start 4 (the Glimmer start). Start 3 has the best score of 312.

At this point, we need to weigh whether we minimize the gap (start 1 at bp 850) or pick the best SD score (start 3 at bp 862). An interesting biological discussion!

I will check the BLAST results, too:

DNA Etude	_annota	ted												
Overview	Features	Refe	erences	Sequence	Docume	ntation								
Sort By	Index 💽	•	•	Name	Start	Stop	^	Description	Sequence	Product	Regions	Blast	Conte	xt Validation
Select Fea	atures D	irect SC	3L) [1	83	466		Score	Target Des	cription				
Tupe	is All		TH.	2	475	849		▶ 28	gp3 [Mycot	acterium p	hage UPIE	[]		
Neere			<u>_</u> <u>-</u> _	3	889	1053		27	5 gp3 (Mycob	acterium p	hage Joe[)irt]		
Name	пке		— ŀ	4	1059	2636		27	4 gp3 (Mycob	acterium p	hage LeBr	on]		
GenelD	<u> </u>			5	2626	4296		17	5 gp3 (Mycob	oacterium p	hage Faith	1]		
Locus	like			6	4318	5373								
Start	>			/	5427	6038		1						
Length			— ŀ	8	6136	6249		- BLAST Hi	AEK07547					Export
Begions	T in		— ŀ	10	6233	6317		GI	AEKU7547 339781718					Export All
e cc			— ŀ	11	6644	6796		Length	63					Delete
76 GL			— ŀ	12	6902	7360		Max Score	282	Date	10/10/201	1		Delete All
CAI			I	13	8315	9187		High-Scor	ng Pairs (HSF	ŋ				
EC#	like			14	9411	10655		HSP Data	Alignment					
Product	like			15	11265	11486		1 MVC	PCCDENT G	TRAKERR	OM BBRA	GARLAF	YDYM	YERAHM BOYLD
Function	like		— ľ	16	11588	12100		1 +11	1111111	111111				
FeatureID	- 1		— ľ	17	12299	12637		10 VVC	PCCDEWT G	TRAKEKR	QW RREA	GAELAE	YDYW	YRRAHM RDYLD
Tag	like		-[18	13031	13438		51 NGT	D					
ray	iike j		— I.				~	51	I					
🔲 Hide I	gnored Fe	atures				2		60 NGD	D					
Sele	ot All Feat	ures		Insert De	elete Po:	st Valida	ite	<u> </u>						
RAA			1	- 14998	P	osition : 44	445	Cor	itrols >> Map	🔽 Map >	> Controls			
	4	L, N	· · · ·	5	6	7 84	อไส	12	(13)	(14		6 (16	17	18) 19 20 21)
21 Features	s Li	ve				2.003		-7						14998 🛃 🤶

This indicates that we are 9 amino acids short of the called start in the best BLAST hit, that of UPIE. 9 amino acids is 18 bp, or equivalent to the start at 862.

Without wet lab evidence, we can't really say. So at this point, I will choose the best SD score at (bp 862) and the BLAST alignment. I could just as easily choose to minimize the gap, as the SD score is not too terrible for start 1 and we frequently see phage genes that start and stop practically on top of each other in this manner. Both are valid choices.

Change the gene start in the description tab:

 \rightarrow Write the new coordinate in the start field, and then click the calculator icon to change the gene length and post the change to the database.

Etude	_annot	tate d												_	
Overview	Feature	es Refere	ences	Sequence	Docume	ntation									
Sort By	Index	- i	┛	Name	Start	Stop	^	Description	Sequence	Prod	luct Regions	Blast	Context	Validation	
Select Fe	atures	 Direct SQL	пĒ	1	83	466		Name 🔤		_	GenelD				· 1
Tune	- in 🗖		÷1	2	475	849		Turne I	20.0						
туре			<u>⊔</u> ⊔	3	862	1053		Type It	.05	<u> </u>	<u>u</u>				
Name	like		_ -	4	1059	2636		Start		862	Locus Lag	ETUDE_	3		
GenelD	=			5	2626	4296		Stop		1053	Regions				1
Locus	like			6	4318	5373		Length 1	92		Tag				
Start			- -	7	5427	6038		Direction F	orward	2					
Lenath	БÌГ		- -	8	6136	6249		Translation	Table Unde	fined					
Begione			- -	9	6239	6317		EC Number	, abio Terres						
e co			— H	10	6344	6607									
% GC			_ H	12	6902	7260		Product							
CAI	>		_ h	12	8315	9187		gp3							~
EC#	like			14	9411	10655									~
Product	like 🗌		- -	15	11265	11486		Function							
Function	like 🗍		- -	16	11588	12100									~
FeaturelD			- h	17	12299	12637									\sim
T			- F	18	13031	13438		Notes							-
Tag	іке						~	Original Gli	mmer call @b	p 889	has strength 6	6.25 ** not	called by I	GeneMark	~
🔲 Hide I	Ignored F	Features				>									
Sele	ort All Fe	atures		Insert De	elete (Po:	st) Valida	ite								\sim
H 4 4			X 1-	14998	P	osition : 44	145	🔽 Co	ntrols >> Map	💌 M	lap >> Control	s 🔤			
	4	<u>k</u>		5	6	7 84	o]≬]	12	(13		14	1 (16	17>	18 19	20)21)
21 Feature:	s	Live				2 141 5		~	,					14998	3 ?

Now reBLAST the gene: click the BLAST tab.

Etud	e_ann	otate	ed			Ū				
Overview	Feat	ures	Reference	s I !	Sequence	Docume	entation			
Sort By	Index	-	•		Name	Start	Stop	^	Ī	Description Sequence Product Regions Blast Context Validation
Select F	, eatures	Dire	ect SQL		1	83	466			There are no BLAST results for this feature
Turn					2	475	849			Blast this gene Blast ALL Genes Clear All
туре	IS 19		<u> </u>		3	862	1053			
Name	like			Ŀ	4	1059	2636			
GenelD	=			Ľ	5	2626	4296			
Locus	like				6	4318	5373			
Start	>				7	5427	6038			
Length	5	i –			8	6136	6249			
Bagiana		<u> </u>		H	9	6239	6317		ļ	
negions	-	<u> </u>		H.	10	6344	6507			
% GC	<			H.	11	6647	7260			
CAI	>			H.	12	8315	9197			
EC#	like			H	14	9411	10655			
Product	like			H.	15	11265	11486			
Function	like	i –		H	16	11588	12100			
Featural		<u> </u>		H.	17	12299	12637			
T caluier	<u> </u>	<u> </u>		H.	18	13031	13438			
lag	like			Г				¥		
🔲 Hide	Ignore	d Feat	tures	<			>			
Sel.	ect All I	Featur	ee	lı	nsert De	elete Po	ost Valida	ate		
		al⊾		1 - 1	4998	F	Position : 1	492	0	Controls >> Map 🔽 Map >> Controls
1/2/8		4	N	5	N	6	7 \ BM	สก	11	
21 Feature	es 🛛	Live	•	-	- 4	- 7		-11	1	14998 🔮 ?

Click "Delete all". Click yes in the box that pops up that asks you if you really want to do this. Then click "BLAST this gene"

A new window will appear, showing your BLAST request status. When it finishes, it will load your data and look like this:

BLAST	search for 862	- 1053 (3)		
Retrieve	XML Results Text F	Results Save to Database		
Score	Description		^	BLAST Hit
330	gp3 [Mycobacterium	n phage UPIE]		Accession AEK07547
324	gp3 [Mycobacterium	n phage JoeDirt]		GI 339781718
323	gp3 [Mycobacterium	n phage LeBron]		Length 63
223	gp3 [Mycobacterium	n phage Faith1]		Max Score 330
74	hypothetical protein	[Plasmodium vivax Sal-1]		
73	hypothetical protein	SCHCODRAFT_237055 [Schizophyllum commune h		
72	hypothetical protein	Mnod_2475 [Methylobacterium nodulans ORS 2060		HSP Coverage Map
70	lyase family protein [Achromobacter xylosoxidans A8]		
70	hypothetical protein	Osl_19228 [Oryza sativa Indica Group]	~	
20		DACCELL 2020 M ID A CHURCH HILL THE DOMA		
E-Value	Starts	Bit Score 131.721 Querv 1.63		Positives 63
5.2E-40	1	Score 330 Target 1-63		Identity 63
		E-Value 5.2E-40 Length 63		Similarity 100.0
		% Aligned 100.0		Gaps 0
		1 MQSIRSGRGV VCPCCDEWTG TRAKEKRQ	WR I	REAGAELAEY DYWYRRAHMR 📶
		1	 ז סזו	
		i ingerkokov vereopievio inadanką	wr. 1	ABAGABBABT PTWTKRAIIIR
		51 DYLDSLEIGN GDD		
		51 THE STREET STREET		
		<u> </u>		~
				?

Click the tab that says "Save to Database"

BLAST search for 862 - 1053	3 (3)
Retrieve XML Results Text Results	Save to Database
Maximum E-Value of HSPs to save	Ignore Definitions including the following terms
1.2E-23 💌	
Save 4 Values	
	?

Only the top four hits have reasonable E values, so we will only save those four to our genome database. I will click "save 4 values". Then I close the window.

DNA Etude	_annotat	ed													
Overview	Features	Reference	s Sequence	Docume	ntation										
Sort By	Index 👻] •	Name	Start	Stop	^	D	escription	Sequence	Product	Regions	Blast	Contex	t Validation	
Select Fe	atures Dir	ect SQL	1	83	466		Г	Score	Target Des	cription					~
Тире	is All	-	2	475	849		Þ	330	gp3 [Mycob	acterium p	ihage UPII	E]			
Nere			<u>▶</u> 3	862	1053			324	gp3 (Mycob	acterium p	hage Joel	Dirt]			=
Name	пке		4	1059	2636			323	gp3 (Mycob	acterium p	hage LeB	ron]			
GenelD	<u> </u>		5	2626	4296			223	gp3 (Mycob	acterium p	hage Faith	h1]			
Locus	like		6	4318	5373		Г								
Start	>		7	5427	6038		L								~
Length			8	6136	6249		Γ.	BLAST Hit						Ev	roort
Decience			9	6239	6317			Accession A	EKU7547					Exp	ort All
negions			10	6344	6507		Ľ	an an an an an an an an an an an an an a	3					De	elete
% GC	<		10	6647	5795		N	Aax Score 3	30	Date	10/11/20	11		Dele	ste All
CAI	>		12	6902	7360		F	High-Scorin	α Pairs (HSE	ท					
EC#	like		13	0411	3187		h	HSP Data	Alianment	í					
Product	like		14	11005	11400		h								
Eunction	like		10	11588	12100			1 MQS1	RSGRGV V	CPCCDEW	TG TRAK	EKRQWF	R REAGA	LELAEY DYW:	YR 🔷 📗
Function			17	12299	12637			1 MQSI	RSGRGV V	CPCCDEW	TG TRAK	EKRQWE	REAGA	ELAEY DYW	YR
FeatureID	'=		18	13031	13438										=
Tag	like		- 10	10001	10400	~		51 DYLD	SLEIGN G						
🔲 Hide I	Ignored Fea	itures	<		>			51 DYLD	SLEIGN G	DD					
Sele	- ct All Featu	res	Insert De	elete Po	st Valida	ate									
n ad a			1 - 14998		Position : 4	813	-	Cont	nls >> Man	Map >	> Controls				
						ສຸດ			40	1 100		6 Leo		40 40	20 20
1 / 2 / 3/	4		>)	• >		<u>o</u> lá I	1	<u>-</u> /	13	14		6 46	112	18/19	20/21/
21 Features	s Liv	e												14998	- I 🗟 <u>/</u>

Back in my main genome window, the BLAST data has been altered for gene 3 (I had to click gene 2 and then back to gene 3 for it to load into the window):

Now when we BLAST the amino acid sequence of gene 3, it matches the UPIE annotation gene 3, with the "query 1" aligning with the "sbjct 1". This makes me feel better about the start that I chose, for even though both starts were good choices, it is nice to be consistent with a similar genome. That way, in the future, any wet bench data that we get about this gene from one phage will be easily applied to similar genomes.

Check for functions via BLAST, Phamerator, HHPred, and the Hatfull Maps. None of these return a known or likely function (the best HHPred match is to a zinc-finger protein in Homo sapiens).

Add detailed annotation info to the notes on the Description tab:

DNA Etude	annota	ted												
Overview	Features	References	Sequence	Docume	ntation									
Sort By	Index 💌		Name	Start	Stop	^	Description	Sequenc	e Prod	uct Regions	Blast	Context	Validatio	n
Select Fe	atures Di	rect SQL	1	83	466		Name 3			GenelD				· ·
Turce			2	475	849		Tupo	DC		CI				
туре	18 OII		<u>(</u> 3	862	1053		iype ju	.05		<u>u</u> 	ETUDE			
Name	пке		4	1059	2636		Start		862	Locus Lag	ETODE_	_3		
GenelD	-		5	2626	4296		Stop		1053	Regions				1
Locus	like		6	4318	5373		Length 19	92		Tag				
Start	>		7	5427	6038		Direction Fr	orward	- 🗧					
Lenath		ŀ	8	6136	6249		Translation	Table Un	defined					_
Begions		ŀ	3	6239	6317		EC Number	1 4010 1						
w cc		ŀ	11	6544	6607									
շնե		ŀ	12	6902	7360		Product							
CAI		ŀ	13	8315	9187		gp3							~
EC#	like		14	9411	10655									~
Product	like	ŀ	15	11265	11486		Function							
Function	like		16	11588	12100									~
FeatureID	- i	ľ	17	12299	12637									\sim
T		—— ľ	18	13031	13438		Notes							-
lag	іке	-				~	862-1053.	previous ge	ene gap 1	3bp. Gene ca	alled by Gli	immer not	GeneMark	
🔲 Hide I	gnored Fea	atures			>		neither pro	gram chose Thas best B	e start. Inc IBS score	cludes all codi • Aligns 1·1 w	ng potenti vith LIPIF c	al. Notion tene 3 Ni	igest possi (F	ble
Sele	ct ∆ll Feati	ITES .	Insert De	elete Po:	st Valida	te	J					,		<u></u>
RAA	[⊕ [⊜] i		- 14998	P	osition : 48	13	Cor	ntrols >> Ma	ap 🔽 M	ap >> Control	s en e			
1 2 8	4	1	5	6	7	16	12	(13	17	14	6 (16	17	18 19	20 21
21 Features	s Liv	/e					7						14998	3?

Gene 4:

Back to the coding potential trace in GeneMark TB. Gene four is in frame three. It looks like the coding potential starts around bp ~980 or 990. According to the frames window, there are two possible starts for this gene:



1053 and 1059 (the start called by both Glimmer and GeneMark).

We already know that both starts encompass all the coding potential, and that both have minimal gaps. The final piece of data is the RBS score:

DNA	Choose OF	RF start									×
Sta	rts : 43	ORF Start :	1059		Cdn 1	Cdn2	Cd	In3 Le	ngth	Document	1
Sel	ected : 1	ORF Stop	: 2636	5' End	50.0	100.0	50	.0 6	_		-
		ORF Lengt	n: 1578	3' End	46.6	66.3	58	.9 1	576		
	Shine D	algarno	Sequer	nce of t	the R	egion		Start	Start	ORF	^
#	Score	Space	Upstre	am of t	the S	tart		Codon	Position	h Length	
1	357	8	AGATTO	GTAATG	GTGAT	GATTA		GTG	1053	1584	
2	483	8	GTAATG	GTGATG	ATTAG	TGGCA		ATG	1059	1578	
з	315	8	TACATO	CCTAGAI	AGACG	ACGGG		ATG	1230	1407	
4	294	7	TCTGGG	CCGCAC	TAAAG	AGGGT		TTG	1317	1320	-
5	420	7	CCCTTT	CTCCTC		CCCAC		CTC	1005	1202	
										ATG 🖻	
										TAĂ	🥻

Here we come to one of those gray areas in annotation. The RBS score of the first start is lower, but not much lower. The gap between genes is smaller with the first start, but not much smaller. And both algorithms selected the second start.

BLAST: The data from the BLAST tab indicates that the algorithms have selected the same start as the genes already in GenBank.

So I am going to pick the Glimmer/GeneMark start.

Functional assignment:

Copy and paste the amino acid sequence into a BLASTP pane at NCBI.



to for links to other re	ources: 🖬 Unigene 🖬 GEU 🖬 Gene 📷 Structure 🖬 Map Vièwer 🛂 PubChem BioAssay					
Sequences producir	g significant alignments:					
Accession	Description	Max score	Total score	Query coverage	🛆 <u>E value</u>	Links
YP_003857135.1	gp4 [Mycobacterium phage LeBron] >gb ADL70971.1 gp4 [Mycobacterium phage LeBro	1083	1083	100%	0.0	G
YP_655891.1	gp26 [Mycobacterium phage Wildcat] >gb ABE67631.1 gp26 [Mycobacterium phage Wil	242	242	85%	8e-62	G
(P_002781225.1	hypothetical protein ROP_40330 [Rhodococcus opacus B4] >dbj BAH52280.1 hypothetic	221	221	82%	2e-55	G
ZP_03927248.1	phage Terminase [Actinomyces urogenitalis DSM 15434] >gb EEH65874.1 phage Termin	216	216	91%	8e-54	_
<u>/P_001800806.1</u>	hypothetical protein cur_1412 [Corynebacterium urealyticum DSM 7109] >emb[CAQ053]	211	211	88%	2e-52	G
<u>2P_06185208.1</u>	putative phage terminase, large subunit [Mobiluncus mulieris 28-1] >gb EEZ90351.1 pu	197	197	85%	3e-48	
<u>(P_07452355.1</u>	possible phage-related terminase [Mobiluncus mulieris ATCC 35239] >gb[EFM46107.1] p	195	195	84%	1e-47	
P_003490422.1	putative phage terminase [Streptomyces scable] 87.22] >emb[CBG/18/9.1] putative pha	194	194	88%	20-47	G
Alignment Select All	S Get selected sequences Distance tree of results Multiple alignme	<u>nt</u>				
Select All	S Get selected sequences Distance tree of results Multiple alignme	<u>nt</u>				
Select All	<u>S</u> <u>Get selected sequences</u> <u>Distance tree of results</u> <u>Multiple alignme</u> <u>YP_003857135.1</u> G gp4 [Mycobacterium phage LeBron]	<u>nt</u>				
Select All	<u>S</u> <u>Get selected sequences</u> <u>Distance tree of results</u> <u>Multiple alignme</u> <u>vp_003857135.1</u> G gp4 [Mycobacterium phage LeBron] <u>70971.1</u> G gp4 [Mycobacterium phage LeBron]	<u>nt</u>				
Select All	S Get selected sequences Distance tree of results Multiple alignme YP_003857135.1] G gp4 [Mycobacterium phage LeBron] 70971.1] G gp4 [Mycobacterium phage LeBron] 525	nt				
Select All Select All	<u>S</u> <u>Get selected sequences</u> <u>Distance tree of results</u> <u>Multiple alignme</u> <u>TP_003857135.1</u> G gp4 [Mycobacterium phage LeBron] <u>70971.1</u> G gp4 [Mycobacterium phage LeBron] <u>525</u> D: 9711610 4 gp4 [Mycobacterium phage LeBron]	<u>nt</u>				
Select All > <u>ref</u> <u>gb ADI</u> <u>GENE 1</u>	<u>S</u> <u>Get selected sequences</u> <u>Distance tree of results</u> <u>Multiple alignme</u> <u>YP_003857135.1</u> G gp4 [Mycobacterium phage LeBron] <u>70971.1</u> G gp4 [Mycobacterium phage LeBron] <u>525</u> <u>D: 9711610 4</u> gp4 [Mycobacterium phage LeBron]	<u>nt</u>				
Select All > <u>ref</u> <u>gb</u> ADI Length- <u>GENE 1</u> Score Identif	S Get selected sequences Distance tree of results Multiple alignme WP_003857135.1 G gp4 [Mycobacterium phage LeBron] 70971.1 gp4 [Mycobacterium phage LeBron] 525 D: 9711610 4 gp4 [Mycobacterium phage LeBron] = 1083 bits (2801), Expect = 0.0, Method: Composition. ties = 525/525 (100%), Positives = 525/525 (100%), Gape	nt al matrix a = 0/525	adjust. (0%)			
Select All > ref gb ADT Length= GENE J Score Identia Query	S Get selected sequences Distance tree of results Multiple alignme WP_003857135.1 G gp4 [Mycobacterium phage LeBron] 70971.1 g gp4 [Mycobacterium phage LeBron] 525 D: 9711610 4 gp4 [Mycobacterium phage LeBron] = 1083 bits (2801), Expect = 0.0, Method: Composition. ties = 525/525 (100%), Positives = 525/525 (100%), Gap 1 MTVIPSIPTDRTVESESDLWTPIDEKAREWSDKGLIGAOKPRLSNYPTF	nt al matrix a = 0/525 FTSLEDDGM	adjust. (0%) DF 60			
Select All Select All Select All Construction Construction Score Identia Query Sbjct	S Get selected sequences Distance tree of results Multiple alignme Multiple alignme (Multiple alignme) (Multiple alignme) (M	nt al matrix a = 0/525 TSLEDDGM TSLEDDGM	adjust. (0%) DF 60 DF 60 DF 60			
Select All > refi gh ADI Length- GENE 1 Score Identi Query Sbjct	S Get selected sequences Distance tree of results Multiple alignme TP 003857135.1] G gp4 [Mycobacterium phage LeBron] 70971.1] G gp4 [Mycobacterium phage LeBron] 525 D: 9711610 4 gp4 [Mycobacterium phage LeBron] = 1083 bits (2801), Expect = 0.0, Method: Composition. ties = 525/525 (100%), Positives = 525/525 (100%), Gap 1 MTVIPSIPTORTVESESDLWTPIDEKAREWSDKGLIGAQKPRLSNYPTFI MTVIPSIPTORTVESESDLWTPIDEKAREWSDKGLIGAQKPRLSNYPTFI 1 MTVIPSIPTORTVESESDLWTPIDEKAREWSDKGLIGAQKPRLSNYPTFI 1	nt al matrix s = 0/525 TTSLEDDGM TTSLEDDGM	adjust. (0%) DF 60 DF 60 DF 60			
Select All > ref[gb AD] Length= GENE 1 Score Identi Query Sbjct Query	S Get selected sequences Distance tree of results Multiple alignme [YP_003857135.1] G gp4 [Mycobacterium phage LeBron] 70971.1] gp4 [Mycobacterium phage LeBron] 525 D: 9711610 4 gp4 [Mycobacterium phage LeBron] = 1083 bits (2801), Expect = 0.0, Method: Compositionaties = 525/525 (100%), Fositives = 525/525 (100%), Gapi 1 MTVIPSIPTDRTVESESDLWTPIDEKAREWSDKGLIGAQKPRLSNYPTF1 MTVIPSIPTDRTVESESDLWTPIDEKAREWSDKGLIGAQKPRLSNYPTF1 1 MTVIPSIPTDRTVESESDLWTPIDEKAREWSDKGLIGAQKPRLSNYPTF1 1 MTVIPSIPTDRTVESESDLWTPIDEKAREWSDKGLIGAQKPRLSNYPTF1 1 IEAAGYNLLPWORALFRASLGRTKEGLWSARQVCLVPPQQCKTELLEAA 1 TEAAGYNLLPWORALFRASLGRTKEGLWSARQVCLVPPQQCKTELLEAA 1 TEAAGYNLLPWORALFRASLGRTKEGLWSARQVCLVPPQQCKTELLEAA	nt 11 matrix 3 = 0/525 TSLEDDGM TSLEDDGM TSLEDDGM TSLEDDGM	adjust. (0%) DF 60 DF 60 NE 120 NE 120			

Gene 4 is the large subunit of the terminase, which is part of the DNA packaging machinery (helps to stuff the DNA into the new phage head). We got a conserved domain hit and numerous phage hits. Once again, our best match is to LeBron, and we once again align perfectly, with the Query 1 matching the Sbjct 1. The assignment is supported by the Hatfull Maps, and running HHPred is not necessary.

I will make the appropriate notes in the Notes field and Function field.

Gene 5:

Gene 5 is the easiest gene by far that we have looked at. Both Glimmer and GeneMark call this gene, the GeneMark TB coding potential starts around ~2650, and there is only one start codon (at 2626) that neither overlaps gene 4 too much and encompasses all the coding potential. In fact, there is only one start codon in the correct frame for this gene. Notice there is a small 10bp overlap between genes now. This is OK, overlaps need to be much larger before we discount them.



BLAST tab results indicate that this gene is a perfect match 1:1 with LeBron gene 4.

Done!

Paste the amino acid sequence into a BLAST p pane at NCBI.

This protein is a phage portal protein and forms a dodecameric ring at the vertex of the capsid that the DNA is threaded through and that the tail then joins to. We once again match the LeBron gene call perfectly. The Hatfull Maps support this assignment.

I will write the appropriate notes in the Notes field and Function field.

Gene 6:

Again, both the Glimmer and GeneMark calls agree on a single start that does not have any close starts near it in the same frame and is the longest possible start for this gene.



The BLAST tab data shows a 1:1 alignment with LeBron's gene 6. The Hatfull Map shows that this gene is the capsid maturation protease (frequently found after the portal gene in phage genomes). This protease cleaves the scaffolding protein in the immature

capsid (also called the procapsid) and allows the phage capsid to expand to its mature size during assembly and DNA packaging.

I will write the appropriate Notes in the Notes field, and Function field.

BLASTing the sequence at NCBI shows that we once again match LeBron gene 6's start exactly.

Gene 7:

Coding potential: The coding potential for this gene begins around 5450.



Gene 7 has two possible start choices:



The start that Glimmer and GeneMark have selected at 5427 and the TTG start at 5400. Both starts include all the coding potential. GeneMark never calls TTG starts and Glimmer undercalls them, so we must take that into consideration when deciding which start to pick (not a good time to say, "well, all things being equal, we will take the algorithms' call," because TTG starts are **NOT** equal from the point of view of the programs.)

When we look at the two starts in context of gap closing and SD scores,

DH	e C	hoose ORI	F start											×
S S	tart ele:	s : 7 cted : 1	ORF Start : ORF Stop ORF Length	5352 : 6038 n : 687	5' End 3' End	Cdn 1 22.2 39.7	Cdn2 33.3 63.1	Cdn 44.4 66.5	3 L0 - 2 i E	ength 27 310	<u> </u>	<u>D</u> ocu	ment	
		Shine D	algarno	Sequenc	e of t	the P	legion	ເ ຮ	tart	Sta	art	ORF		^
#		Score	Space	Upstream	m of t	the S	tart	С	odon	Pos	ition	Len	gth	
1		378	7	CGGGACA	GTCAC	GCGTT	TTCAA	. Т	TG	540)0	639		
2		462	7	TGGCTTA	AGTGA.	AGGAA	AATTA	. A	TG	542	27	612		
3		294	7	GAAGCTG	TCTAA	GGCTG	AGCGC	A	TG	574	12	297		
4		435	9	GTCTAAG	GCTGA	GCGCA	TGGAG	: A	TG	574	18	291		
E.		420	٥	COTTOCC	CACCT	ACCTO	ACCAC		TC	677	17	267		\mathbf{r}
1	1												>	
												- fi	1 🗎	?

the TTG start at 5400 has a score of 378 and the ATG start has a SD score of 462. However, the size of the gap left between the stop codon of gene 6 is either 24bp or 54bp.

BLAST data: The BLAST tab shows that the longer gene start has been called for the genes already in GenBank (our alignment has a mismatch of 1:10).

Given that 24bp is already a sizable gap for a phage genome and the other GenBank phages ue the longer start, we will pick the TTG start at 5400.

Change the gene start on the Description tab, and click the calculator button to recalculate gene length and to post the change to the database.

ReBLAST the gene through DNA Master to make sure that you see the correct alignment.

Functional assignment: the NCBI results indicated that this is a likely scaffolding protein in another phage. While this is not labeled as a scaffolding protein on the Hatfull Maps, synteny certainly supports this assignment. We will assign this gene the function of "Scaffolding", but won't enter it into the function field.

Gene 8:

Gene 8 is a very small gene (38 res) that exactly matches the beginning of LeBron gene 8 in the BLAST tab alignments. Normally, I would include this gene in an annotation, even though it is very small because it shows that a gene has been truncated and therefore is a good example of genome mosaicism and recombination. Unfortunately, gene 8 also overlaps a tRNA. Generally we do not see CDS and tRNA overlaps, except for possibly a few bases at the 3' end of both of them. The positioning of the tRNA, and the truncation of gene 8 suggests that the tRNA interrupted the original gene 8. I will therefore delete the called gene 8 from the auto-annotation. I am not going to renumber the genes again until I am done with the annotation.

Now we know from our phamerator alignment that Etude no longer has a high degree of similarity to LeBron after gene 8. But there are some lines in the phamerator map indicating that gene 9 of Etude is similar to something farther to the right in the LeBron genome. Slide the Etude map to the right in relation to LeBron and you will see:



LeBron is still the bottom genome, but you can't see the title any more because it is all the way over at the left end of the genome. From the map above, it looks like the next 8kbp are very similar to LeBron, with the final two kbp having no similarity.

Gene 9 the tRNA:

Click on the product tab:

DNA Etude	_annot	iate	d						
Overview	Feature	es I	Reference	s	Sequence	Docume	ntation		
Sort By	Index	-	4		Name	Start	Stop	^	Description Sequence Product Regions Blast Context Validation
Select Fe	atures	Direc	et SQL	L	1	83	466		Anticodon : Complement of TTG
T	. A				2	475	849		Show Aragom v1.1 Structure Aragom Web Site
туре	is pa		<u> </u>	L	3	862	1053		aa 🗸
Name	like				4	1059	2636		c i
GenelD	=				5	2626	4296		a 71+
Locus	like				6	4318	5373		g-c =
Start					7	5427	6038		t-a
				Þ	9	6240	6314		c-g
Length				L	10	6344	6607		c-g t+g
Regions	>			L	11	6647	6796		g-c
%GC	<			L	12	6902	7360		tc tg
CAI				L	13	8315	9187		a accc a
EC#	lika [L	14	9411	10655		t cgg tggg c
				L	15	11265	11486		g !!! g tt
Product	like			L	16	11588	12100		g gcc t
Function	like			L	17	12299	12637		caaa y c c-att
FeatureID	= [L	18	13031	13438		d-c d
Tag	like [L	19	13469	14065		t-a
		_		L	20	14179	14577		c-g a-t
🔲 Hide I	gnored F	Featu	ires	L	21	14628	14975	~	c a
Sele	ct All Fe	ature	is	<			>		ta
					Insert De	lete) Po:	t Valida	ite	caa 💌
RAA	(@ (Q			1 -	14998	P	osition : 10	0975	i 🔽 Controls >> Map 🔽 Map >> Controls
1 2 3	4	1	X	ł	5)	6	7) 14	00	
21 Features	3	Live					2 10		14998 🛃

If you check the box at the top marked "Show Aragorn v1.1 Structure" the folded view of the tRNA will appear. Examine the top stem loop and 3' end of the tRNA: does the stem loop have seven base pairs? No, it has eight. Is the 3' end sequence after the stemloop NCCA? No, It is ngct. This means that we will need to look at the tRNA outputs from the other programs and trim the tRNA appropriately.

Web-Based Aragorn:



the output for Web-based Aragorn is much better: a seven base pair top stem loop, and on the 3' end there is a discriminator base (the A), followed by a single C from the CCA. This follows our tRNA rule: the trimming the CCA part of the sequence once it deviates from CCA.

The tRNAscan SE output is:

Results

Sequence Name 	tRNA #	tRNA Begin 	Bounds End 	tRNA Type	Anti Codon	Intron Begin	Bounds End 	Cove Score
Your-seq	1	6242	6315	Leu	CAA	0	0	62.03
View tRNA								

Here the tRNA sequence begins two bases after the web-based Aragorn output, and ends one base after it does, generating a slightly different secondary structure.



In this case, I prefer the web-based Aragorn, that has the beginning of the CCA sequence that is found at the end of canonical tRNAs. So I will adjust the tRNA coordinates to those and click the calculator button to recalculate the gene length. I will also add the tRNA amino acid and anti-codon to the function and notes fields:

Etude	_anno	otate	ed												[
Overview	Featu	ıres	Refere	ences	Sequence	Docume	entation										
Sort By	Index	-		┛	Name	Start	Stop	^	Descriptio	Sequence	Prod	uct Regions	Blast	Context	Validatio	n	
Select Fe	atures	Dire	ect SQL	n [1	83	466		Name [9	_	GenelD					1
Turne	. [All		-1	2	475	849		Turne	-	_	CL					
туре	si 	AII		Ľ	3	862	1053		Type	thna	_	<u>u</u>					
Name	like				4	1059	2636		Start	6	5240	Locus Tag	ETUDE	_9			
GenelD	=				5	2626	4296		Stop	6	5314	Regions					1
Locus	like				6	4318	5373		Length	'5		Tag					
Start	ЪÍ			-1	7	5427	6038		Direction F	orward		-					
Length	H			-1	▶ <u>9</u>	6240	6314		Translation		ined						
Lengin	H			-1	10	6344	6607		FC Number		ineu						
Regions				_	11	6647	6796			۶ ۲							
% GC	<				12	6902	7360		Destat								
CAI	>				13	8315	9187		Product								
EC#	like			-	14	9411	10655		l gha								2
Broduct	like [-	15	11265	11486										
Floquet					16	11588	12100		Function	lass							
Function	like			_	17	12299	12637		UTINA LEC	i (caa)							2
FeatureID	=				18	13031	13438										\leq
Tag	like				19	13469	14065		Notes INDIA Lor	(app) web bo	and Ar	agorn call					
E Hidel	-	East			20	14179	14577		INA Let	i (caa)web-ba:	sed An	agoin call.					
1 miller	gnoreu	ii ca	uies		21	14628	14975	~									
Sele	et All F	eatur	es	- I			>										
					Insert (D	elete) Po	st Valida	ate _									\mathbf{N}
N	\	21		1	- 14998	F	Position : 1	4270		ontrols >> Map	<u>м</u>	ap >> Controls					
1 2 3		4	, Y		5	6	7) k	60	12	13		14	(5) (16	6 17	18 1	9) 2	0) (21)
21 Feature:	s	Live	•			/	7 10	5-19	/						1499	18	a ?

Gene 10: It seems pretty clear from GeneMark TB coding potential, from our phamerator alignment, and that the next gene is a reverse gene, rather than a forward gene like all the previous ones.



There are actually three separate ORFs in the fourth frame above, the first two corresponding to two different auto-annotated genes, with the final ORF overlapping

with the forward ORF in the second frame. Both algorithms decided to call the forward ORF as a gene over the third reverse ORF (you can't pick both; they overlap almost completely). The coding potential of the forwards ORF is much better than the third reverse ORF—as shown by the higher, more extensive peaks, so we will also pick that gene when we get to gene 12.

Back to gene 10: when a forward gene and reverse gene meet stop codon to stop codon (or end of tRNA to stop codon as genes 9 and 10 do), it is not necessary to leave much space between the two ends of the genes. Several bases is sufficient. However in the opposite case, when the forward and reverse genes meet start codon to start codon, it is necessary to leave at least 50-60 bases between the two starts. This is because there will be a promoter for the RNA polymerase preceding the start codon of each gene. Since gene 10's stop codon has plenty of room after the tRNA transcript ends, we don't need to worry about this here (See below). We will need to worry about it when we select the starts for both gene 10 and gene 11.



in the frames window, the deleted gene is still highlighted. Click the green "refresh" arrow at the lower right side (next to the ORFs button) to update the frames window.



Gene 9 only has one real choice for a start codon, and it was selected by both the GeneMark and Glimmer aligorthms.

BLAST check:

Distribution of 6 Blast Hils on the Query Sequence @ Under over to see the define, did to show alignments Color key for alignment scores 1 Oper Key for alignment scores 1 Color key for alignment scores 1 Oper Key for alignment scores 1 Oper Key for alignment scores 1 Oper Key for alignment scores Color Key for alignment scores Color Key for alignment scores Oper Key for alignment scores Color Key for alignment scores Mark score for Score		No putative conserved domains have be	en detected				
Mouse over to see the define, dick to show alignments Color key for alignment scores 0 0 0 0 0000 >>>2000 1 0 20 30 40 50 60 70 80 2 0 1 0 20 30 40 50 60 70 80 Structure in the define, dick to show alignments Color key for alignment scores Color key for alignments Structure in the define i		Distribution of 6 Blast Hits on the Query So	equence 😡				
Color key for alignment scores Superior		Mouse over to see the define, click to show alignments					
Add-do 30-80 80-200 >>2000 1 10 20 30 40 50 60 70 80 construction 1 10 20 30 40 50 60 70 80 construction 1 10 20 30 40 50 60 70 80 construction 1 10 20 30 40 50 60 70 80 construction 1 10 20 30 40 50 60 70 80 construction 1 100 100 100 100 100 100 44 44 10 100 100 44 44 10 100 100 100 44 44 10 100 100 44 44 10 100		Color key for alignmen	tscores				
Criptions Y Observe General Studue Map Vewer M PubChem BioAssay Sequences producing significant alignments: UniGene G GEO G Gene Studue Map Vewer M PubChem BioAssay Y Observe Concernage M (Wycobacterium phage LeBron) - gb(ADL71077.11 gp124 [Mycobacterium phage L 180 100% 44-44 C Y Observe Concernage M (Wycobacterium phage LeBron) - gb(ADL71077.11 gp124 [Mycobacterium phage L 180 100% 44-44 C Y Observe Concernage M (Wycobacterium phage LeBron) - gb(ADL71077.11 gp124 [Mycobacterium phage L 180 100% 44-44 C Y Observe Concernage M (Wycobacterium phage LeBron) - gb(ADL71077.11 gp124 [Mycobacterium phage L 180 100% 44-44 C Y Observe Concernage M (Wycobacterium phage LeBron) - gb(ADL71077.11 gp124 [Mycobacterium phage L 180 100% 44-44 C Y Observe Concernage M (Wycobacterium phage LeBron) - gb(ADL71077.11 gp124 [Mycobacterium phage L 180 100% 44-44 C Y Observe Concernage M (Wycobacterium phage LeBron) - gb(ADL71077.11 gp124 [Mycobacterium phage L 180 100% 44-44 C Y Observe Concernage M (Wycobacterium phage LeBron) - gb(ADL71077.11 gp124 [Mycobacterium phage L 127.4 </th <th></th> <th><40 40-50 50-30</th> <th>80-200</th> <th>>=200</th> <th></th> <th></th> <th></th>		<40 40-50 50-30	80-200	>=200			
Image: Structure in the second sequences in the		Query					
and of links to other resources: UnGene © GEO © Gene © Structure © Map Viewer © PubChem BioAssay Sequences producing significant alignment: <u>Accession in the protein in 100011001288, R2601_22966 (Pelagibaca bermudensis HTCC2601</u> <u>27.0359224.1</u> hypothetical protein in 100011001288, R2601_22966 (Pelagibaca bermudensis HTCC2601 <u>27.0359224.1</u> hypothetical protein in 1000180, R2601_22966 (Pelagibaca bermudensis HTCC2601 <u>27.0359224.1</u> hypothetical protein in 1000180, R2601_22966 (Pelagibaca bermudensis HTCC2601 <u>27.0359224.1</u> hypothetical protein in 1000180, R2601_22966 (Pelagibaca bermudensis HTCC2601 <u>27.0359224.1</u> hypothetical protein in 1000180, R2601_22966 (Pelagibaca bermudensis HTCC2601 <u>27.0359224.1</u> hypothetical protein thurbro08, 22840 (Bacillus thuringiensis servour berliner ATCC 1079 <u>14.6</u> <u>69%</u> <u>0.029</u> <u>0.27</u> <u>17.035921.11</u> hypothetical protein Birdinger, Bacillus create NRRL 2338) > emb[CAM05291.1] pyl <u>33.5</u> <u>33.5</u> <u>18.6</u> <u>33.5</u> <u>18.6</u> <u>33.5</u> <u>18.6</u> <u>53.5</u> <u>18.6</u> <u>53.5</u> <u>19.6</u> <u>19.6</u> <u>19.6</u> <u>19.6</u>		1 10 20 30 40 50	60 70	80			
Accession 200382241 1000500000000000000000000000000000000							
Implicities Introduction							
Accession Geo Geone Structure Map Vewer PubChem BioAssay Accession Acc							
d for links to other resources: U UnGene E GEO C Gene Structure W Map Viewer & PubChem BioAssay iequences producing significant alignments: TY 0.0382246.1 pp124 (Mycobacterium phage LeBron) > gb)(ADL71077.1 gp124 (Mycobacterium phage L 22 0.044272.1 hypothetical protein 1100011001288, R2601_22966 (Pelagibaca bermudensis HTCC2601 C226 77.8 91% 44-13 c27.055993.1 conserved hypothetical protein R0580199 . pp01149.21(1) + pyothetical protein bthur0008, 22840 (Bacillus thuringiensis servour berliner ATCC 1079 41.6 499% 0.029 (Pr 002271161 hypothetical protein bthur0008, 22840 (Bacillus thuringiensis servour berliner ATCC 1079 41.6 899% 0.029 (Pr 002271611 hypothetical protein bthur0008, 22840 (Bacillus thuringiensis servour berliner ATCC 1079 41.6 538.5 59% 0.27 (C 97.001108215.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb)(CAM05291.1 pyruvate carboxylase [Saccharopolyspora er							
d for links to other resources: U UnGene C Gene Structure W Map Viewer S PubChem BioAssay kequences producing significant alignments: Accession 122.01447/27.1 123.01 123							
d for links to other resources: U UniGene C GEO C Gene Structure Map Viewer M PubChem BioAssay ded of links to other resources: U UniGene C GEO C Gene S Structure Map Viewer M PubChem BioAssay resources producing significant alignments: Accession D Description P2_003857245.1 pp124 (Mycobacterium phage LeBron) >gb)ADL71077.1 gp124 (Mycobacterium phage L 180 180 100% 4e-44 C 22_0144717.1 hypothetical protein 1100011001288, R2601_22966 (Pelagibace bermudensis HTCC2601 27.8 77.9 91% 4e-13 27.0 91% 4e-14 C 27.0 91% 4e-14 C 27.0 91% 4e-13 27.0 91% 4e-13 27.0 91% 4e-13 27.0 91% 4e-14 C 27.0 91% 4e-14 C 27.0 91% 4e-13 27.0 91% 4e-13 27.0 91% 4e-14 C 27.0 91% 4e-14 C 27.0 91% 4e-13 27.0 91% 4e-14 C 27.0 91% 4e-14 C 27.0 91% 4e-13 27.0 91% 4e-13 27.0 91% 4e-14 C 27.0 91% 4e-13 27.0 91% 4e-14 C 27.0 91% 4e-13 27.0 91% 4e-13 27.0 91% 4e-13 27.0 91% 4e-13 27.0 91% 4e-14 C 27.0 91% 4e-1	rintione						
d for links to other resources: U UniGene G GEO G Gene S Structure M Map Viewer S PubChem BioAssay iequences producing significant alignments: Accession Description Max score Total score Query coverage E value Link P2 00385245.1 p124 (Mycobacterium phage LeBron) >pb/ADL71077.1 (g124 (Mycobacterium phage L 180 180 100% 44-44 C P2 00385245.1 p124 (Mycobacterium phage LeBron) >pb/ADL71077.1 (g124 (Mycobacterium phage L 180 180 100% 44-44 C P2 00385245.1 opstate protein 110010288, R2640 (22056) (Pelagibaca bermudensis HTCC2601 27.8 91% 4e-13 6e-10 P2 0045427.11 hypothetical protein Roseibium sp. TrichSKD4) >pb/EF031303.11 conservery to 62.4 67.4 91% 6e-10 P2 0025218.1 hypothetical protein Bthur008_22840 (Bacillus thuringiensis servar berliner ATCC 1079 38.5 38.5 89% 0.27 C P2 00130215.1 pyruvate carboxylase (Saccharopolyspora erythraea NRRL 2338) >emb/CAM05291.1 (pr) 33.5 57% 9.1 C	riptions						
equences producing significant alignments: Accession Description Hax score Total score Query coverage 2 Evalue Uit Accession 0 Description Hax score 180 100% 4e-44 C P_01444717_1 gp124 [Mycobacterium phage LeBron] >gb]ADL71077.1] gp124 [Mycobacterium phage L 180 180 100% 4e-44 C P_0144717_1 thypothetical protein 1100011001288, R2601_22966 [Pelagibaca bermudensis HTCC2601 27.8 77.8 91% 4e-13 C P_0152211.1 hypothetical protein Rosebilum sp. TrichSK04 pb]FF031030.1] conserved for 46.4 67.4 <	d for linke to other rea	auroes 🕖 UniGona 🧧 GEO 🖸 Gona 🗧 Structura 🕅 Man Viewar 🦉 PubCham BioAceau					
Bequences producing significant alignments: Accession Description Hax score Total score Query coverage 2 E value Lin 97_003857245.1 gp124 (Mycobacterium phage LeBron) >gb1ADL7107.1 gp124 [Mycobacterium phage L 180 180 100% 4e-44 C 20.04547.12.1 hypothetical protein 1100011001288. R2601_22966 [Pelagibace bermudensis HTCC2601 72.8 77.8 91% 4e-13 27_003857245.1 posterical protein 11000011001288. R2601_22966 [Pelagibace bermudensis HTCC2601 72.8 77.4 91% 6e-10 27_005993.1 conserved hypothetical protein [Roseiblum sp. TrichSKD4] >gb1E7031303.11 [conserved r 67.4 67.4 91% 6e-10 28_00122211.1 hypothetical protein BCH308197_B0021 [Bacillus turninglemise servora heriner ATCC 1079 91.6 41.6 69% 0.029 VP_00102215.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb[CAM05291.1] py 33.5 33.5 57% 9.1 C	d for links to other rea						
Description Max score Total score Query coverage E value Lin Accession 124 (Mycobacterium phage LeBron) >gb1/ADL71077.1 (gp124 (Mycobacterium phage L 180 180 100% 4e-44 C VP_00357246.1 hypothetical protein 1100101288, R2601.22966 (Pelagibace bermudensis HTCC2601 27.8 77.8 91% 4e-13 CP_005993.1 conserved hypothetical protein 10001288, R26240 (Bacillus thuringiensis servorar berliner ATCC 1079 41.6 69% 0.029 PU0250158.1 hypothetical protein 1000197_80.021 (Bacillus churingiensis servorar berliner ATCC 1079 38.5 89% 0.27 C P00250158.1 prozent buthor008 (Saccharopolyspora erythraea NRRL 2338) >emb[CAM05291.1] py 33.5 57% 9.1 C							
19.00387245.1 p1242 (Mycobacterium phage LeBron) > gb)ADL71077.1 (p124 (Mycobacterium phage L 180 180 180 100% 44-44 C 29.0144717.1 hypothetical protein (10001288,R2601_22966 (Pelagibaca bermudensis HTCC2601_27.8 77.8 91% 4e-13 C 29.0144717.1 hypothetical protein (Roselibum sp. TrichSKO4) > gb)EF03103.1 (conserved h 67.4 67.4 91% 6e-10 27.01102211.1 hypothetical protein bthur0008, 22840 (Bacillus thuringiensis servors befiner ATCC 1079 41.6 89% 0.29 19.00226128.1 hypothetical protein bthur0008, 22840 (Bacillus thuringiensis servors befiner ATCC 1079 41.6 89% 0.27 C 19.00236128.1 pyruvate carboxylase (Saccharopolyspora erythraea NRRL 2338) > emb CAM05291.1 pyr 33.5 57% 9.1 C	Accession	Description	Max score	Total score	Query coverage	🛆 E value	Links
22.0144717.1 hypothetical protein 1100011001288_R2601_22966 [Pelagibaca bermudensis HTCC2601 27.8 77.8 91% 4e-13 22.0144717.1 conserved hypothetical protein [Roseibium sp. TrichSKD4] >gb[EF031303.1] conserved h 67.4 91% 6e-10 22.014271.1 hypothetical protein [Roseibium sp. TrichSKD4] >gb[EF031303.1] conserved h 67.4 91% 6e-10 27.0102211.1 hypothetical protein BKH0080_2240 [Bacillus turingiensis servars berliner ATCC 1079 41.6 41.6 89% 0.29 VP_000198216.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb[CAM05291.1] py 33.5 33.5 57% 9.1 C	YP_003857246.1	gp124 [Mycobacterium phage LeBron] >gb[ADL71077.1] gp124 [Mycobacterium phage L	180	180	100%	4e-44	G
<u>CP2_07599311</u> conserved hypothetical protein [Roselbium sp. TrichSKD4] > objEF021303.1] conserved h <u>67.4</u> <u>67.4</u> <u>91%</u> <u>66-10</u> <u>CP2_07599311</u> hypothetical protein [Roselbium sp. TrichSKD4] > objEF021303.1] conserved h <u>41.6</u> <u>41.6</u> <u>89%</u> <u>0.029</u> <u>CP_010221111</u> hypothetical protein BCH308197_B0021 [Bacillus cereus H3081.97] > gb]ACI30565.1] co <u>38.5</u> <u>89%</u> <u>0.27</u> <u>C</u> <u>CP_00108216.1</u> pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] > emb[CAM05291.1] pyl <u>33.5</u> <u>57%</u> <u>9.1</u> <u>C</u>		hypothetical protein 1100011001288_R2601_22966 [Pelagibaca bermudensis HTCC2601	77.8	77.8	91%	4e-13	
2P. 041/02211.11 hypothetical protein bth/ur0008, 22840 (Bacillus thuringiensis servors berliner ATCC 1079 41.6 B9% 0.029 Hypothetical protein bth/ur008, 22840 (Bacillus thuringiensis servors berliner ATCC 1079 41.6 B9% 0.27 C Hypothetical protein bth/ur008, 22840 (Bacillus thuringiensis servors berliner ATCC 1079 38.5 38.5 B9% 0.27 C Hypothetical protein bth/ur008, 22840 (Bacillus thuringiensis servors berliner ATCC 1079 33.5 38.5 B9% 0.27 C Hypothetical protein bth/ur008, 22840 (Bacillus thuringiensis servors berliner ATCC 1079 33.5 38.5 B9% 0.27 C Hypothetical protein bth/ur008, 22840 (Bacillus thuringiensis servors berliner ATCC 1079 33.5 38.5 B9% 0.27 C	ZP_01444717.1				010/	6e-10	
mpozzeszuszi hypothetical protein BCH308197_B0021 [Bacillus cereus H3081.97] >gb]ACI30565.1] co 38.5 89% 0.27 C vP_001108216.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb]CAM05291.1] pyr 33.5 33.5 57% 9.1 C Imments ct All Get selected sequences Distance tree of results Multiple alignment	ZP_01444717.1 ZP_07659993.1	conserved hypothetical protein [Roseibium sp. TrichSKD4] >gb EFO31303.1 conserved h	67.4	67.4	9170	00-10	
VP_001108215.1 pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb[CAM05291.1] pyi 33.5 33.5 57% 9.1 [C Imments ct All Get selected sequences Distance tree of results Multiple alignment	ZP_01444717.1 ZP_07659993.1 ZP_04102211.1	conserved hypothetical protein [Roseibium sp. TrichSKD4] >gb[EF031303.1] conserved f hypothetical protein bthur0008_22840 [Bacillus thuringiensis serovar berliner ATCC 1079	<u>67.4</u> <u>41.6</u>	67.4 41.6	89%	0.029	_
Iments at All Get selected sequences Distance tree of results Multiple alignment	<u>ZP_01444717.1</u> <u>ZP_07659993.1</u> <u>ZP_04102211.1</u> <u>YP_002267188.1</u>	conserved hypothetical protein (Roseibium sp. TrichSKO4) > gb EF031303.1] conserved h hypothetical protein bthur0008_22840 (Bacillus thuringiensis serovar berliner ATCC 1079 hypothetical protein BCH308197_B0021 (Bacillus cereus H3081.97) >gb ACI30565.1] co	67.4 41.6 38.5	67.4 41.6 38.5	91% 89% 89%	0.029 0.27	G
Imments ct All Get selected sequences Distance tree of results Multiple alignment	<u>ZP_01444717.1</u> <u>ZP_07659993.1</u> <u>ZP_04102211.1</u> <u>YP_002267188.1</u>	conserved hypothetical protein (Roseibium sp. TrichSKO4) > gb EF031303.1] conserved h hypothetical protein bthur0008_22840 (Bacillus thuringiensis serovar berliner ATCC 1079 hypothetical protein BCH308197_B0021 (Bacillus cereus H3081.97) >gb ACI30565.1] co	67.4 41.6 38.5	67.4 41.6 38.5	91% 89% 89%	0.029 0.27	G
ct All Get selected sequences Distance tree of results Multiple alignment	<u>ZP_01444717.1</u> <u>ZP_07659993.1</u> <u>ZP_04102211.1</u> <u>YP_002267188.1</u> <u>YP_001108216.1</u>	conserved hypothetical protein [Roseibium sp. TrichSKO4] >glp[EF031303.1] conserved i hypothetical protein btur0080.22840 [Bacillus thuringlensis servora herliner ATCC 1079] hypothetical protein BCH306197_B0021 [Bacillus cereus H3061.97] >glp[ACI30565.1] co _pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb[CAM05291.1] pyi	67.4 41.6 38.5 33.5	67.4 41.6 38.5 33.5	91% 89% 89% 57%	0.029 0.27 9.1	G
	<u>2P.014447771</u> <u>2P.07659993.1</u> <u>2P.04102211.1</u> <u>YP.002267188.1</u> <u>YP.001108216.1</u>	conserved hypothetical protein [Roselibium sp. TrichSKO4] >gb[EF031303.1] conserved h hypothetical protein btur/0008_22840 [Bacillus thuringiensis serovar berliner ATCC 10/9 hypothetical protein BCH308197_B0021 [Bacillus cereus H3081.97] >gb]ACI30565.1[o pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb[CAM05291.1] pyl	67.4 41.6 38.5 33.5	67.4 41.6 38.5 33.5	91% 89% 57%	0.029 0.27 9.1	G
	2P 07659993.1 2P 07659993.1 2P 04102211.1 YP 002267188.1 YP 001108216.1 mments cct All <u>Get selectr</u>	conserved hypothetical protein [Roselbium sp. TrichSK04] >gb[F6031303.1] conserved h hypothetical protein bthur0008_22840 [Bacillus thuringlensis serovar berliner ATCC 1079 hypothetical protein BCH305197_B0021 [Bacillus cereus H3081.97] >gb]ACI30555.1] co pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb[CAM05291.1] pyl ed sequences Distance tree of results Multiple alignment	67.4 41.6 38.5 33.5	67.4 41.6 38.5 33.5	9179 89% 89% 57%	0.029 0.27 9.1	G
	2P 0765993.1 2P 0765993.1 2P 04102211.1 YP 002267188.1 YP 001108216.1 nments ct All <u>Get select</u>	conserved hypothetical protein [Roselbium sp. TrichSK04] >gb[EF031303.1] conserved h hypothetical protein bthur0008_22840 [Bacillus thuringiensis serovar berliner ATCC 1079 hypothetical protein SCH306197_B0021 [Bacillus cereus H3081.97] >gb]ACI30565.1 [o pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb[CAM05291.1] pyl ed sequences Distance tree of results Multiple alignment	67.4 41.6 38.5 33.5	67.4 41.6 38.5 33.5	91.79 89% 89% 57%	0.029 0.27 9.1	G
	2P 0765993.1 2P 0765993.1 2P 04102211.1 YP 002267188.1 YP 001108216.1	conserved hypothetical protein [Roselbium sp. TrichSKO4] splp[F031303.1] conserved i hypothetical protein btur0080.22840 [Bacillus thuringlenis servora berliner ATCC 1079 hypothetical protein BCH306197_B0021 [Bacillus cereus H3061.97] >gb]ACI30565.1] co _pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb]CAM05291.1[pyi ed sequences Distance tree of results Multiple alignment	67.4 41.6 38.5 33.5	67.4 41.6 38.5 33.5	9178 89% 89% 57%	0.029 0.27 9.1	G
>□ref YP 003857246.11 G gp124 [Mycobacterium phage LeBron]	2P 0765993.1 2P 0765993.1 2P 04102211.1 YP 002267188.1 YP 001108216.1 mments cct All Get selects	conserved hypothetical protein [Roselbium sp. TrichSKO4] -gb[EF031303.1] conserved i hypothetical protein btur0008.22840 [Bacillus thurnignenis servora berliner ATCC 1079 hypothetical protein BCH308197_B0021 [Bacillus cereus H3081.97] >gb]ACI30565.1] co pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb[CAM05291.1] pyl ed sequences Distance tree of results Multiple alignment	67.4 41.6 38.5 33.5	67.4 41.6 38.5 33.5	9176 89% 89% 57%	0.029 0.27 9.1	G
* <u>Cref YP 003857246.1</u> G gpl24 (Nycobacterium phage LeBron) gb/AD/21077.11 G gpl24 (Nycobacterium phage LeBron)	ZP 014497171 ZP 076599931 ZP 076599931 ZP 076599931 ZP 04109211.1 YP 002267188.1 YP 002106216.1 YP 00106216.1 Contents	conserved hypothetical protein [Roseibium sp. TrichSKO4] splp[F031303.1] conserved i hypothetical protein btur0080.22840 [Bacillus thuringlenis servora berliner ATCC 1079] hypothetical protein BCH306197_B0021 [Bacillus cereus H3061.97] >gh]ACI30565.1] co pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb]CAM05291.1] pyi ed sequences Distance free of results Multiple alignment	67.4 41.6 38.5 33.5	67.4 41.6 38.5 33.5	91 % 89% 89% 57%	0.029 0.27 9.1	6
Cref[TP_003857246.1] G gp124 [Mycobacterium phage LeBron] gb]ADL71077.1] G gp124 [Mycobacterium phage LeBron] ength=97	2P 01447/171 2P 07659993.1 2P 07659993.1 2P 00120211.1 YP 001108216.1 YP 001108216.1 YP 001108216.1 YP 001108216.1 Cect All Get selecth	conserved hypothetical protein [Roselbium sp. TrichSKO4] splp[FO31303.1] conserved f hypothetical protein bturNo08.22840 [Bacillus thurnignens servour berliner ATCC 1079 hypothetical protein BCH308197_B0021 [Bacillus cereus H3081.97] >gb ACI30555.1] co pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb CAM05291.1] pyl ed sequences Distance tree of results Multiple alignment 46.1] C gp124 [Nycobacterium phage LeBron] G gp124 [Nycobacterium phage LeBron]	67.4 41.6 38.5 33.5	67.4 41.6 38.5 33.5	92.56 89% 89% 57%	0.029 0.27 9.1	6
<pre>pref [YP 003857246.1] gp124 [Mycobacterium phage LeBron] gb]ADL71077.1] gp124 [Mycobacterium phage LeBron] dergth=07 GRME ID: 9711726 124 gp124 [Mycobacterium phage LeBron]</pre>	2P 01447/171 2P 07659953.1 2P 07659953.1 2P 0022618.1 YP 0022618.1 YP 00108216.1 YP 00108216.1 Get soloch Cref (YP 0038572 gb (ADL71077.1) Get Soloch Gent Die 9711726	conserved hypothetical protein [Roselbium sp. TrichSKO4] =gb[EF031303.1] conserved f hypothetical protein btur0008_22840 [Bacillus thurnigness servora Feliner ATCC 1079 hypothetical protein BCH308197_B0021 [Bacillus cereus H3081.97] >gb]ACI30565.1] co pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb[CAM05291.1] py] ed sequences Distance tree of results Multiple alignment 44.1] G gp124 [Mycobacterium phage LeBron] gp124 [Mycobacterium phage LeBron] 124 [gp124 [Mycobacterium phage LeBron]	67.4 41.6 38.5 33.5	67.4 41.6 38.5 33.5	92.99% 89% 89% 57%	0.029 0.27 9.1	G
<pre>[YP_003857246.1] G gpl24 [Mycobacterium phage LeBron] agb[ADL71077.1] G gpl24 [Mycobacterium phage LeBron] agrgUn=87 GENE ID: 9711726 124 gpl24 [Mycobacterium phage LeBron] Score = 180 bits (457), Expect = 4e-44, Methodi compositional matrix adjust. Identities = 8787 (1008), Positives = 8787 (1008), Gaps = 0/87 (08)</pre>	22-01449/171 22-07659993.1 22-07659993.1 22-06102211.1 YP_002267188.1 YP_002267188.1 YP_0022618.1 YP_002618.1 YP_00	<pre>conserved hypothetical protein [Roselbium sp. TrichSKO4] gpl;EF031303.11 conserved f hypothetical protein btur0080.22840 [Bacillus thurnignens servour berliner ATCC 1079 hypothetical protein BCH308197_B0021 [Bacillus cereus H3081.97] >gb ACI30565.1 co pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb CAM05291.1 pyl ed sequences Distance tree of results Multiple alignment 46.1] gpl24 [Mycobacterium phage LeBron] gpl24 [Mycobacterium phage LeBron] 124 gpl24 [Mycobacterium phage LeBron] 125 (1005), Foxpert = 4 e-44, Wethod: Compositional matrix adjust. 87 (1005), Foxpert = 4 e-47 (1005), Cage = 0/67 (05)</pre>	52.4 41.6 38.5 33.5	67.4 41.6 38.5 33.5	9176 89% 89% 57%	0.029 0.27 9.1	G G
<pre>>Cref YP 003857246.1 G gpl24 (Mycobacterium phage LeBron) gb/ADL712077.1 G gpl24 (Mycobacterium phage LeBron) dength=B0 comparison of the state</pre>	22 014447171 22 0155993.1 22 0155993.1 22 01502211.1 27 00225188.1 YP 00225188.1 YP 002108216.1 YP 001108216.1 YP 001108216.1 YP 00108216.1 Sector 1 (1) (1) (1) (1) (1) (1) (1) (1) (1) (<pre>conserved hypothetical protein [Roselbium sp. TrichSkO4] splp[FO31303.1] conserved f hypothetical protein bthr0080 .22840 [Bacillus thuringlenes Servora Perimer ATCC 1079] hypothetical protein BcH3008197_B0021 [Bacillus cereus H3081.97] >gh]ACl30565.1] co pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb]CAM05291.1] pyi ed sequences Distance tree of results Multiple alignment def.il G gp124 [Hycobacterium phage LeBron] gp124 [Mycobacterium phage LeBron] igp124 [Mycobacterium phage LeBron] is (457), Expect = 4e-44, Methodi Compositional matrix adjust. 57 (1005), Fositives = 87/87 (1008), Cape = 0/87 (08) UPTFUEREQUENTFIFTERGOMENTIFLEVVOLUMOVARMEDIZINET 60</pre>	52.4 41.6 28.5 33.5	67.4 41.6 38.5 33.5	91% 89% 89% 57%	0.029 0.27 9.1	G
<pre>></pre>	2P_01447/11 2P_07659993.1 2P_07659993.1 2P_0762993.1 P_00226188.1 YP_00226189.1 YP_00226189.1 YP_00226189.1 YP_00226189.1 YP_00226189.1 YP_00226189.1 YP_00226189.1 YP_00226189.1 YP_00226189.1 YP_00226189.1 YP_0026189.1 YP_0	<pre>conserved hypothetical protein [Roselbium sp. TrichSkO4] splp[FO31303.1] conserved f hypothetical protein bthr0080 .22840 [Bacillus thuringlenes Servora Perimer ATCC 1079 hypothetical protein BcH3008197_B0021 [Bacillus cereus H3081.97] >gh]ACl30565.1] co pyruvate carboxylase [Saccharopolyspora erythraea NRRL 2338] >emb]CAM05291.1] pyi ed sequences Distance free of results Multiple alignment def.il [] gp124 [Hycobacterium phage LeBron] gp124 [Mycobacterium phage LeBron] igp124 [Mycobacterium phage LeBron] igp124 [Mycobacterium phage LeBron] sc (457), Expect = 4e-44, Methodi Compositional matrix adjust. 57 (1005), Fositives = 87/87 (1008), Cape = 0/87 (08) WiDPTIEERCLOPENTTIFIERGEMENTIFILEVVOLIAKEVARMED(TINFT 60 WiDPTIEERCLOPENTTIFIERGEMENTIFILEVVOLIAKEVARMED(TINFT 60)</pre>	52.4 41.6 28.5 33.5	67.4 41.6 38.5 33.5	91% 89% 89% 57%	0.029 0.27 9.1	G

We still are similar to LeBron, only now we are similar to gene 124. We still align perfectly with the same start codon as selected in the LeBron annotation.

In the annotation notes, when you are calculating the "gap/overlap" number, you should still look at the start codon of gene 10, only now you should compare to the stop codon of gene 11, because we are going in the reverse direction. The reason why we include these gap or overlap numbers is to see how well the start we chose fills out the genome from this gene to the neighboring one. Since we can't change the stop codon, the only way to fill the genome is by changing the start codon. This number provides an extra reference as to how closely the genes are called in your annotation. LeBron gene 124 has no known function.

Genes 11 and 12: As mentioned above, gene 11 is a reverse gene while gene 12 is a forward gene. As they will be "head to head" (so to speak), we need to take care in choosing the starts for each of them that we leave at least 50-60 bases between the two genes.



Currently, if we accept the two GeneMark calls for the genes, we will just barely squeak by with our minimum of 60 bases (6841 to 6902). Notice, however, in the above figure, that it is not possible to extend gene 12 to start any closer to gene 11 (there aren't any more start codons in the same frame any closer to gene 11 than the one already called). On the other hand, Gene 11 has four possible start codons, including the one used in the Glimmer call, which is way back at position 6796. From looking at the GeneMark TB coding potential trace, it is pretty clear that the Glimmer start does not encompass all the coding potential, so we will eliminate this choice as a possible start. We do still have three more starts to check: the one selected by GeneMark and the two immediately after it. If we check the RBS scores for all three starts; we find that the GeneMark start also has the highest RBS score. So we will accept the GeneMark call for gene 11. Gene 12 really only has one possiblility for a start, and it is the one called by both programs above.

BLAST check: gene 11 aligns with LeBron gene 125 query 1 to sbjct 1. Gene 12 aligns with LeBron gene 126 query 22 to sbjct 22 (which still counts as picking the same start, just the beginnings of the genes are not as similar as the later portions.)

Gene 13: While there are some blips in the GeneMark TB coding potential, none of them align well within an open reading frame (if the peaks did fit better into an ORF I would be likely to include them as genes). So we will leave a fairly large gap between gene 12 and gene 13. This is also what the algorithms suggest.

Both the Glimmer and GeneMark calls suggest that gene 13 begins at postion 9187, but this start leaves a huge gap between gene 14 and 13 (gene 13 is reverse, so we compare its start to the upstream gene). The BLAST data suggests that this gene is much shorter than the other similar entries in GenBank (UPIE 1:59).



There are five more possible starts for gene 13 between 9300 and 9400.

DNA C	hoose OR	F start			
Start: Selec	s : 33 cted : 1	ORF Start : ORF Stop ORF Length	9400 Cdn 1 Cdn2 Cd : 8315 5' End 50.0 75.0 50. 1: 1086 3' End 66.7 80.0 10	n3 Length .0 12 0.0 16	Document
Ш					9358
	Shine D	algarno	Sequence of the Region	Start Start	ORF
#	Score	Space	Upstream of the Start	Codon Position	Length 📃
1	357	8	GCTAGGACCCCGTACCGAAGCG	GTG 9373	1059
2	525	8	TACCGAAGCGGTGCGGGGGTCCT	TTG 9361	1047
з	420	8	GGTGCGGGGTCCTTTGCTTTGC	GTG 9352	1038
4	143	6	GCGGGGTCCTTTGCTTTGCGTG	GTG 9349	1035
5	210	7	AGGCCCCGCTAACCGCGTCGGT	TTG 9253	939
6	483	7	ACACCACACCAGGAGGAACACC	ATG 9187	873
7	315	8	AACCGACACTGATATTCAGTAC	GTG 9148	834
8	357	8	GTTCGAAAGCTTCGCGGAATTC	GTG 9124	810
9	273	7	GGCCAGCGTCAAGGCTAAGGGC	ATG 8938	624
10	399	7	CAGCGTCAAGGCTAAGGGCATG	GTG 8935	621 💌
					11日 🕄 🕄

The best RBS score goes to the TTG at 9361. We will pick this start.

BLAST check: while we now match UPIE, we are substantially longer than LeBron,(Query 59 aligns with Sbjct 1)—which is interesting. The two phages seem very similar according to phamerator (all that purple between the genomes), so why wouldn't the LeBron annotators have chosen the longer start that we did? While it is not necessary to match the starts for all genes 1 to 1 with a similar phage in genbank, I was still puzzled. To solve this, I actually loaded the LeBron sequence into the web-based GeneMark TB coding potential site, and examined the two outputs side by side. It turns out that LeBron has a point mutation in this area which causes an extra stop codon in this frame, and the start that we chose for Etude is not a possibility in LeBron. The point mutation is not a large enough sequence difference to show up as another color in phamerator. I am happy to proceed with our start selection.

Gene 14: Is another reverse gene, this time in the fifth tier in the GeneMark TB coding potential view. We will ignore the peak in the forward third tier because we can't call

both of them and the reverse one is larger and lovelier. There is only one possible start codon for gene 14.

BLAST Check: we match LeBron gene 128, with the start codons aligning perfectly.

Gene 15: while there is a teensy little peak in the forward direction in the second tier GeneMark TB coding potential trace nicely centered in an ORF, I am more inclined to omit it for simply being too small. It is also important to get a feel for you phage genome—are all the genes very tightly packed? Do multiple genes have very small blips of coding potential? You need to think about these things when you are making your decision. There is also a trend in most phage genomes for clusters of genes to be transcribed in the same direction, and Etude is no exception. And since we are going from a reverse gene 14 to another reverse gene --supported by the algorithms calls-- we will skip the teeny blip in the second tier.

However, the best way to really be sure would be to add this gene into your annotation, and BLAST the protein sequence to see if there are any similar genes in GenBank.

Gene 16 is the larger peak in the fourth tier after the smaller peak in the fifth tier, then gene 17 is forward again in the second tier(the peak in the top tier does not align well with an ORF).

Back to Gene 15:

There are really only two choices for a start for gene 15; either the GeneMark call or the Glimmer call. The GeneMark call has a higher RBS score so we will pick the GeneMark call—both for SD score and for being a longer gene.

When we do the BLAST check, we match LeBron gene 130 perfectly at the start.

At this point, it is worth revisiting whether or not we want to include the little forward ORF with the blip of coding potential to see if it aligns with LeBron gene 129. Our gene 15 matches LeBron 128, and our gene 16 matches LeBron 130. Our phamerator alignment indicates that there is still the highest level of sequence similarity between these genomes in this region, but again, a point mutation resulting in the loss of a start or stop codon would not be enough to change the nucleotide identity color in phamerator. And again, it is not necessary to make the genome annotations match, but it is worth looking at the data. When in doubt, BLAST the potential gene to see if there are any matches in GenBank. I will leave it as a exercise to see if LeBron gene 129 exists intact in Etude, and if so, should it be included in the annotation.

Gene 16:

Gene 15 has a stop codon almost immediately to the right the start codon called by Glimmer and Genemark (remember, we are in the reverse direction, so the gene is transcribed right to left). This means that the start codon selected by the two algorithms is already the longest possible start that can be selected for this gene.

The SD score is nice and high, and the start encompasses all the GeneMark TB coding potential. I am happy to accept the call as is.

BLAST check: we match gene 131 from LeBron with a perfect start codon alignment.

Gene 17: Gene 17 is the peak shown in the fourth tier of the GeneMark TB coding potential. Since we are once again switching from reverse to forward, we need to make sure that we leave at least 60 bases between the two gene calls.

There are three possible starts for gene 17 before a stop codon appears in the frame, including the Glimmer start:

All three start encompass all the coding potential shown in the GeneMark TB trace; however, the SD score is best for the second of the three starts; plus this gene call fills our genome gap a bit better without getting in the way of promoter sites. We will pick the second start, at 12242.

BLAST check: this gene aligns with UPIE, but not with LeBron. Even though phamerator shows that this sequence is still similar to LeBron, and the LeBron annotation does not call this gene, we are going to rely on our own data and Glimmer calls. It is possible that LeBron has another point mutation in this region, eliminating this ORF. We can check on the LeBron genome's coding potential in the GeneMark program again, if we want to.

Genes 18-21 The remainder of the genes are no longer similar to LeBron, but instead appear to match the A3 cluster phages.



Gene 18: There is a fairly big (400bp) gap between the end of gene 17 and the beginning of the Glimmer/GeneMark calls for gene 18. While unusual in phage genomes, this is OK. There is no coding potential blips anywhere in the GeneMark TB coding potential traces between the two genes, and while there is an earlier alternate start codon at 12827, its RBS score is low, while the RBS score for the Glimmer/GeneMark call is very high! So we will pick the Glimmer/GeneMark call.

BLAST check: when we BLAST the amino acid sequence using BLASTp, we find that we align perfectly with Bxz2 gene 22, with the Query 1 aligning with Sbjct 1.

Gene 19: We know from the GeneMark TB coding trace that gene 18 and gene 19 are in the same frame, so the stop codon of 18 precludes any start codon for gene 19 earlier in the genome. The start codon selected by Glimmer and GeneMark is already the longest possible gene call for the gene, encompasses all the coding potential, and has a RBS good score. We will accept the algorithms' calls.

BLAST check:

We match the Bxz2 gene 23 perfectly, with a Query 1 aligning to Sbjct 1.

Functional assignment: This is the major tail subunit of the phage.

Gene 20: There are two possible starts to Gene 20, the one called by Glimmer and GeneMark, and the one upstream at 14116 (another TTG). If you check the coding potential GeneMark TB output, you will see that the Glimmer and GeneMark calls do

not encompass all the coding potential. When we check the SD scores, we get a higher score for the TTG start than the Glimmer/GeneMark start. We will pick the 14116 start.

BLAST check: This gene matches Microwolf gene 25 perfectly (aligns Query 1 to Sbjct 1).

Functional assignment: this is the first of the two tail assembly chaperones (the equivalent of G in phage lambda).

Gene 21: The GeneMark call for this gene overlaps with the end of gene 20, while the Glimmer call leaves a large gap. The GeneMark call encompasses all the coding potential while the Glimmer call does not. This is quite a large overlap.

BLAST check: This gene matches Bxz2 gene 24 perfectly, but begins in the middle of the equivalent gene for Peaches and Eagle. This is because of its function: this is the second of the two tail assembly chaperones, and actually begins at 14116 and then frameshifts into the remainder of this gene, creating the G-T fusion.

Note about this function:

Bxz2 gene 23 is a tail assembly chaperone, (gene "G" in phage lambda), and with its partner gene 24 (gene "T" in lambda) makes a fusion protein that helps assembly the tail of the phage correctly. Both the first protein, the "G" like protein, and the fusion "G-T" like protein are produced, however, the second gene product, the T-like protein is not made on its own but only as part of the fusion. In the flexible tailed phages, the tail assembly chaperones frequently precede the tapemeasure gene (generally the longest gene in the genome), and are characterized by a "slippery sequence" that allows the ribosome to shift translational frame during protein synthesis. The ribosome will "slip" back a base, causing a -1 frameshift. Another way to think about it is that the ribosome reads the same base twice. The slippery sequence is generally rich in As but can begin with Gs as well. There are numerous examples of the G-T frameshift in phamerator (look for any phage genome that has two genes that begin at the same start codon, with one of them being longer than the other and followed by the longest gene in the genome). Notice I put it in the Etude annotation in phamerator already. In the phages that we have studied, the G-T slippery sequence almost always occurrs at the C-terminal end of the G protein (in our case Etude gp 20).

To find the coordinate, the easiest way is to find a closely related phage that already has the frameshift correctly annotated. If we look in our BLAST hits, you will see that phage Peaches is similar to Etude, and has the frameshift already correctly annotated in phamerator. While Peaches and Etude do not have a ton of nucleotide sequence similarity between them, notice the tail assembly proteins are in the same pham.



In the picture above, I have clicked on the first, shorter of the two purple Peaches genes (the G equivalent), and then highlighted the C-terminal portion of the protein sequence where frameshifting is likely to occurr.

In this next picture, I have clicked on the longer, correctly annotated frameshifted G-T fusion protein, and highlighted the same amino acids as in the shorter previous gene. We can see that the the sequence matches perfectly until the glycine following SPG.





Finally, I look for the equiavlent area in Etude:

Again, above I have clicked on the first of two tail assembly genes, and then looked at the C-terminal end where frameshifting is likely to occurr. Now, we go back into our six-frame translation of the sequence.

~	1,		-	-	~	~	-	~	~		~		• •	<u> </u>	-		1	14	-	~	~	~	11		~	10	~	14	•	1	-	-		~	~
- (2	Р	н	D	F	F	K D) (2 1	v	v	v	s	E	Р	A	P	Q	D	- I	: F	r I	5 I	1 1	ν	7 E	: F	Q	G	Ε	L	D	Е	L	1
	A	S	2	3	R	v	Q	R	L	С	R	R	E	R	2	5 (3	Т	Ρ	R	Н	v	G	R	R	G	R	Т	ន	G	*]	Ρ	RI	RI	P
Ş	2	P	G	E	A	E	F	۱ s	3 (P	A	*	L	М	Ν	Т	A	Т	A	. *	2	5 1	5	r s	5 (5 F	R R	т	À	. W	I	s	G	Т	\$
	s	R	(3	K	Р	s	A	R	Ρ	Ρ	D	*	*	3	[]]	R	R	L	Р	s	R	*	Р	L	G	D	v	R	R (G :	S	Q (G 1	H
Ρ	A		G	G	s	R	A	L	A	R	L	I	: 1		E	Y	G	D	С	L	v	A	D	L	ឃ	Ε	Т	Y	G	V	D	L	R	D	Ι
СС	CAG	СС	GGG	GGG	AAG	cce	GAGC	GC	ГСG	ссс	GCC	TGA	TT	GAT	GAJ	TA(CGG	CGA	СТG	ССЛ	AGI	rcgo	CTGI	ACCI	гсто	GGGI	IGAC	GTA	CGG	CGT	GGA	тст	CAG	GGA	CAY
. :	:		۱.,		:	•••		.:				:			:		•••		. : .		1	:		.		:			. : .			:		.	
GC	ЭТC	GG	ссс	ccc	TTC	GGC	TCG	CG1	AGC	GGG	CGG	ACT	AA(CTA	СТІ	TAT	GCC	GCT	GAC	GGA	TCI	GCC	GACT	rggj	AGAC	ссст	CTG	CAT	GCC	GCA	сст.	AGA	GTC	ссто	GT.
	W	G	I	Р.	s	A	s	R	Ε	G	A	Q	Ν	I	F	7	V	A '	V	A	*	D	s	v	Е	Р	L	R	v	A :	H	I	E I	Ρī	σ
G	L		R	Ρ	F	G	L	A	R	G	G	្ទ	; (2	Н	I	R	R	s	G	L	R	Q	G	R	Ρ	s	Т	R	R	Р	D	*	Р	С
0	3	A	Ρ	F	L	F	A S	. :	ŏ.	A	R	R	I	S	S	Y	Ρ	S	Q	9 F	2	Г I	1 3	5 I	R Q	2 2	; v	Y	P	Т	S	R	L	S	1
Р	A		s	Р	Р	s	S	Р	V	s	S	5	5 I	2	s	Y	R	s	A	R	G	s	т	Q	R	ន	A	A	A	s	s	s	G	D	G

The dark blue bar above above is from gene 20, the first tail assembly protein; while the purple start codon is from the gene 21 call. You can also see here that gene 20 is in the first translational frame while gene 21 is in the third translational frame. (a -1 shift). Now we look for our Peaches-like sequence towards the C-terminal end of gene 20—and it is seven bp from the right-hand end of the picutre. The nucleotide sequence for those amino acids is GGGGGAA—a slippery sequence if ever there was one. Now to go from the top frame, gene 20; into the bottom frame to gene 21; we pick a nucleotide that will be included twice: once as the third base in a triplet codon for gene 19; and once as the first base in a triplet codon for gene 20. If we use the Peaches annotation as a guide, the amino acid sequence in the fusion protein should go from P-G-E-A to P-G-G-S. This means that the nucleotide that gets counted twice is the third G in the glycine codon of PGEA (and the first G of the second glycine codon in PGGS). This is nucleotide number 14553.

) verview	Features	Reference	ces	Sequence	Docume	ntation		
Sort By	Index 🔹	i I	٩Ľ	Name	Start	Stop	~	Description Sequence Product Regions Blast Context Validation
Select Fe	atures n			1	83	466	-	Name 21 GanalD
-			īΓ	2	475	849		
Туре	IS All	<u> </u>	![3	862	1053		Type CDS 🔄 🛅
Name	like			4	1059	2636		Start 14628 Locus Tag ETUDE_21
GenelD	=			5	2626	4296		Stop 14975 Regions 2
Locus	like		1	6	4318	5373		Length 348 🔳 Tag
Start				7	5427	6038		Direction Forward
Land				9	6240	6314		
Length				10	6344	6607		
Regions	>			11	6647	6796		EC Number
% GC	<			12	6902	7360		
CAI			Ĺ	13	8315	9361		Product
EC#	like		L	14	9411	10655		gp21
				15	11265	11486		
Product	like			16	11588	12100		Function
Function	like			17	12299	12637		
FeatureID	=			18	13031	13438		
Tag	like			19	13469	14065		Notes
				20	14179	14577		Original Glimmer call @bp 14628 has strength 2.90; GeneMark calls start at 14529
Hide I	gnored Fe	atures	Ĩ	21	14628	14975	~	
Sele	ct All Feat	ures	<			>		

On the description tab for gene 21, change the number in the regions field from 1 to 2:

Now click the Regions tab:

Enter in the coordinates for the upstream regions, followed by any number you like in the lengths field. Click "Insert" at the bottom of the regions tab, and then enter the

coordinates of the second region (it will first appear on top, but will later be correctly reordered):

DNA Etude	_anno	otate	ed																-		×
Overview	Featu	ues	References	s !	Sequence	Docume	ntation														
Sort By	Index	-			Name	Start	Stop	^	D	escription	Se	quence	Pro	duct	Regions	Blast	Context	Vali	idation		
Select Fe	atures	Dire	ect SQL	ŀ	1	83	466			Start	1	Stop	Γ	Lenath	1		 				-
Turne	. [All			2	475	849		*	1455	53	. 149	975	1							
туре	IS	AII	<u> </u>		3	862	1053			1417	79	145	553		1						
Name	like				4	1059	2636		F		-										
GenelD	= [5	2626	4296														
Locus	like				6	4318	5373														
Chart					7	5427	6038														
Start					9	6240	6314														
Length	>			Ī	10	6344	6607														
Designed			ŀ			00.17	0700														

click "assign lengths"

	ıde_a	nnota	ted														
Overv	iew F	eatures	References	Sequence	Docume	ntation											
Sort E	By Ind	ex 🔹		Name	Start	Stop	^	D	escription	Se	equence	Product	Region	s Blast	Contex	t Validation	1
Selec	t Featu	es D	irect SQL	1	83	466		Г	Start		Stop	Leng	th				~
Туре	is	All	-	2	475	849	-		141	79	145	53	375				
Name	e lik	e 🗀		3	1050	1053	-	Ŀ	145	53	149	75	423				
Cana				4	1059	2030	-	L									
	- U		ŀ	6	4318	5373	-	L									
Locu	s lik			7	5427	6038		L									
Start	>			9	6240	6314		L									_
Leng	th 🔉			10	6344	6607		L									
Regio	ons 🔉			11	6647	6796											
% GC	<			12	6902	7360											
CAL		i –	[13	8315	9361											

The correct numbers will be calculated.

Now return to the description tab, and adjust the start coordinate accordingly.

1000	1.1.18.28	1.189	24	1.	24	C. Cattan	02	11.18	1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1	12-124	S. G. Carladian	189	16.5 B. C. 15	1685-189	11-124	10.00
🔀 Etude	_annot	tated														X
Overview	Feature	es Ref	erences	Sequence	Docume	ntation										
Sort By	Index	•	•	Name	Start	Stop	^	Descriptio	n Sequence	Produ	uct Regions	Blast	Context	Validatio	on]	
Select Fea	atures	Direct S	QL)	1	83	466		Name	21	_	GenelD					
Tune	is A	.11	-il-	2	475	849		Tune	rns.	Ţ	61					
Name	like [3	862	1053		Chart	14179	<u> </u>	Loous Tog	ETUDE	21			—
CanalD			—ŀ	4	1059	2636		Start	14173	075	EUCUS Tay	ETODE_	_21			_
GenelD			ŀ	5	2626	4296		Stop	14	975	Regions	<u> </u>				2
Locus	like		h	7	4310 5427	6038		Length	798		Tag					
Start	>		ŀ	9	6240	6314		Direction	Forward	=						
Length	>			10	6344	6607		Translatio	n Table Undefi	ned						-
Regions	>		— ŀ	11	6647	6796		EC Numb	er							
%GC	< T		— I	12	6902	7360										
CAL	s i		-[13	8315	9361		Product								
EC#	like [-[14	9411	10655		gp21								<u>^</u>
			—L	15	11265	11486										\sim
Product	іке			16	11588	12100		Function								
Function	like		H	17	12299	12637										2
FeatureID	=			18	13031	13438		Matas								
Tag	like		_ -	19	13469	14060		Original (Glimmer call @hr	1462	8 has strength	2.90: Ge	eneMark c	alls start a	14529	
🔲 Hide I	anored F	Features	. h	20	14175	14975		- inginiari			e nae eneriga	2.00, 00				
Cala			1 6	2 (m)	14110	14010										
Seler	Ct All Fe	atures	」 ┞		I. n.											
				Insert De	ele Po		ne j	1		_		_				
K K K	& (2)) 🕅 1-	14998	F	osition : 2	241		iontrols >> Map	Ma Ma	ap >> Controls					
1 2 3	4			5	6	7 >	0	12	(13	К	14	(5 (16	17	18 1	9 20	\rangle 21 \rangle
20 Features	3	Live												1499	38	3 ?

Enter your gene Notes.

Finally, validate and renumber all your genes.

DNA Etude	_annota	ted						
Overview	Features	Refere	ences	Sequence	Docume	ntation		
Sort By	Index 🔻	·		Name	Start	Stop	^	A Description Sequence Product Regions Blast Context Validation
Select Fe	atures Di	- rect SQL	лE	1	83	466		All ORFs appear valid
÷			ΞĹ	2	475	849		
Туре	IS All		⊒∟	3	862	1053		
Name	like		_ [4	1059	2636		
GenelD	=			5	2626	4296		
Locus	like			6	4318	5373		
Start			- [_	7	5427	6038		
Length			-1-	8	6240	6314		
Length			-1-	9	6344	6607		
Regions			_ -	10	6647	6796		
% GC	<			11	6902	7360		
CAI	>		- -	12	8315	9361		
EC#	like		- -	13	9411	10655		Control Numbering
Deadurat	Bira [14	11265	11486		
Floquet				15	11588	12100		Assign Names by order of appearance
Function	like		_ -	16	12299	12637		Assign Products by order of appearance
FeatureID) =			17	13031	13438		Overwrite existing values
Tag	like			18	13469	14065		🔲 Overwrite default Names & Products only
🖂 Hide I	r Ianored Fe	atures	Ŀ	19	14179	14075		Assign new Locus Tags by order of appearance
j maci	ignorea i ei	itares (H	20	14179	14975	~	Locus Tag Prefix ETUDE Search NCBI database
Sele	ect All Featu	ires				2		Beassign Gene Data Locate Gray Holes > 300 A
				Insert De	elete Po:	st Valida	ite	
N M M	(x 1-	14998	P	osition : 95	539	i39 🔽 Controls >> Map 🔽 Map >> Controls
1 2 3	4			5	6 >	7 >	9 (

If you have not been reBLASTing all your genes, now is a good time to delete all the BLAST hits from your file and do a new complete genome BLAST. Then start your QC.

-Review your notes (correct format? All the information?)

-check those gene gaps and overlaps one last time. Did you miss any?

Finally save your final file (yourphagename_final.dnam5 is a good name), and send it off to Pitt for review.