Table of Contents

1 Introduction to DNA Master 7 1.1 DNA Master overview. 7 1.2 Installation 7 1.3 Quick Start Guide 7 1.4 DNA Master program structure 7 1.4 DNA Master program structure 7 Analysis programs running within DNA Master 8 1.4.1 Glimmer 8 1.4.2 GeneMark 9 1.5 Setting Preferences 10 1.5 Set befault Translation Table & Template Insertion (Local Settings Tab) 11 1.5.2 Set totar codon choices 12 1.5.4 Set start codon choices 13 1.5.5 Choose a default location for saving files 13 1.5.6 Finishing up your Preference settings 14 1.6 Getting help 14 1.7 Checking for updates 15 1.8 Event Manager 17 2.1 Overview 17 2.1 Overview 20 3.1 Overview 21 2.2 BLASTing your sequence into D	H	ow t	to use this guide	5
1.1 DNA Master overview 7 1.2 Installation 7 1.3 Quick Start Guide 7 1.4 DNA Master program structure 7 Analysis programs running within DNA Master 8 1.4.1 Glimmer 8 1.4.2 GeneMark 9 1.4.3 Aragorn 10 1.5.5 Setting Preferences 10 1.5.1 Set tofault Translation Table & Template Insertion (Local Settings Tab) 11 1.5.2 Set default values for BLAST searches 13 1.5.5 Choose a default location for saving files 13 1.5.6 Finishing up your Preference settings 14 1.6 Getting help 14 1.7 Checking for updates 15 1.8 Event Manager 15 1.8 Event Manager 17 2.1 Overview 23 2.3 Cluster assignment of your phage database 17 2.4 BLASTing your sequence against the mycobacteriophage database 17 2.5 Inporting your DNA sequence into DNA Master	1	Int	troduction to DNA Master	7
1.2 Installation 7 1.3 Quick Start Guide 7 1.4 DNA Master program structure 7 Analysis programs running within DNA Master 8 1.4.1 Glimmer 8 1.4.2 GeneMark 9 1.4.3 Aragorn 10 1.5 Setting Preferences 10 1.5.1 Set tool preferences 10 1.5.2 Set color preferences 11 1.5.3 Set start codon choices 12 1.5.4 Set default values for BLAST searches 13 1.5.5 Choose a default location for saving files 13 1.5.6 Finishing up your Preference settings 14 1.6 Getting help 14 1.7 Checking for updates 15 1.8 Event Manager 15 1.8 Event Manager 17 2.1 Overview 17 2.2 BLASTing your sequence against the mycobacteriophage database 17 2.3 Ilupter assignment 20 3.1 Overview 2				
1.3 Quick Start Guide 7 1.4 DNA Master program structure 7 Analysis programs running within DNA Master 8 1.4.1 Glimmer 8 1.4.2 GeneMark 9 1.4.3 Aragorn 10 1.5 Setting Preferences 10 1.5.1 Set Default Translation Table & Template Insertion (Local Settings Tab) 11 1.5.2 Set color preferences 12 1.5.4 Set default values for BLAST searches 13 1.5.5 Choose a default location for saving files 13 1.5.6 Finishing up your Preference settings 14 1.6 Getting for updates 15 1.8 Event Manager 15 2 Provisional Cluster assignment of your phage 17 2.1 Overview 17 2.3 Cluster assignment 20 3 Inporting your sequence against the mycobacteriophage database 17 2.3 Cluster assignment 20 3 Inporting your DNA sequence into DNA Master 23 3.1 Overview				
1.4 DNA Master program structure 7 Analysis programs running within DNA Master 8 1.4.1 Glimmer 8 1.4.2 GeneMark 9 1.4.3 Aragorn 10 1.5 Setting Preferences. 10 1.5.2 Set color preferences. 11 1.5.3 Set start codon choices 12 1.5.4 Set default values for BLAST searches. 13 1.5.5 Choose a default location for saving files 13 1.5.6 Finishing up your Preference settings. 14 1.7 Checking for updates 15 1.8 Event Manager. 15 1.8 Event Manager. 17 2.1 Overview 17 2.2 BLAST ing your sequence against the mycobacteriophage database. 17 2.1 Overview 23 3.1 Overview 23 3.1 Overview 23 3.1 Overview 23 3.1 Overview 23 3.2 Where do 1 get my phage genome sequence from?		1.3		
Analysis programs running within DNA Master 8 14.1 Glimmer 8 14.1 GeneMark 9 14.3 Aragorn 10 15.5 Setting Preferences 10 15.1 Set color preferences 11 15.2 Set color preferences 11 15.3 Set start codon choices 12 15.4 Set default values for BLAST searches 13 15.5 Choose a default location for saving files 13 15.6 Finishing up your Preference settings 14 16 Getting help 14 17 Checking for updates 15 18 Event Manager 15 18 Event Manager 17 2.1 Overview 17 2.2 BLASTing your sequence against the mycobacteriophage database 17 2.3 Cluster assignment 20 3 Importing your phage genome sequence into DNA Master 23 3.1 Overview 23 3.2 Where do 1 get my phage genome sequence from? 23 3.3 <td></td> <td>1.4</td> <td>•</td> <td></td>		1.4	•	
1.4.1 Glimmer 8 1.4.2 GeneMark 9 1.4.3 Aragorn 10 1.5 Setting Preferences 10 1.5.1 Set color preferences 11 1.5.2 Set otor preferences 12 1.5.4 Set default values for BLAST searches 13 1.5.5 Choose a default location for saving files 13 1.5.6 Finishing up your Preference settings 14 1.6 Getting help 14 1.6 Getting for updates 15 1.8 Event Manager 15 1.8 Event Manager 17 2.1 Overview 17 2.1 Overview 17 2.2 BLASTing your sequence against the mycobacteriophage database 17 2.3 Cluster assignment 20 3 Importing your phage genome sequence into DNA Master 23 3.1 Overview 23 3.2 Where do l get my phage genome sequence from? 23 3.3 Importing your DNA sequence into DNA Master 24				
1.4.3 Aragorn 10 1.5 Setting Preferences 10 1.5.1 Set Default Translation Table & Template Insertion (Local Settings Tab) 11 1.5.2 Set color preferences 11 1.5.3 Set start codon choices 12 1.5.4 Set default values for BLAST searches 13 1.5.5 Choose a default location for saving files 13 1.5.6 Finishing up your Preference settings 14 1.6 Getting for updates 15 1.8 Event Manager 15 2 Provisional Cluster assignment of your phage 17 2.1 Overview 17 2.2 BLASTing your sequence against the mycobacteriophage database 17 2.3 Cluster assignment 20 3 Importing your phage genome sequence into DNA Master 23 3.1 Overview 23 3.2 Where do I get my phage genome sequence from? 23 3.3 Importing your DNA sequence into DNA Master 24 3.4 Reverse-complementing your sequence 26 4 Performing and vi		•		
1.5 Setting Preferences 10 1.5.1 Set Default Translation Table & Template Insertion (Local Settings Tab) 11 1.5.2 Set color preferences 11 1.5.3 Set start codon choices 12 1.5.4 Set default values for BLAST searches 13 1.5.5 Choose a default location for saving files 13 1.5.6 Finishing up your Preference settings 14 1.6 Getting help 14 1.7 Checking for updates 15 1.8 Event Manager 15 1.8 Event Manager 17 2.1 Overview 17 2.1 Overview 17 2.1 Overview 17 2.2 BLASTing your sequence against the mycobacteriophage database 17 2.3 Cluster assignment 20 3 Importing your phage genome sequence into DNA Master 23 3.1 Overview 23 3.2 Where do I get my phage genome sequence from? 23 3.3 Importing your DNA sequence into DNA Master 24 3.4 <td></td> <td>1.4</td> <td>ł.2 GeneMark</td> <td>9</td>		1.4	ł.2 GeneMark	9
1.5.1 Set Default Translation Table & Template Insertion (Local Settings Tab) 11 1.5.2 Set color preferences 11 1.5.3 Set start codon choices 12 1.5.4 Set default values for BLAST searches 13 1.5.5 Choose a default location for saving files 13 1.5.6 Finishing up your Preference settings 14 1.6 Getting help 14 1.7 Checking for updates 15 1.8 Event Manager 15 2 Provisional Cluster assignment of your phage 17 2.1 Overview 17 2.2 BLASTing your sequence against the mycobacteriophage database 17 2.1 Overview 20 3 Importing your phage genome sequence into DNA Master 23 3.1 Overview 23 3.2 Where do I get my phage genome sequence from? 23 3.3 Importing your DNA sequence into DNA Master 24 3.4 Reverse-complementing your sequence 26 4 Performing and viewing a rapid automated annotation of your genome. 27 4.1		1.4	1.3 Aragorn	10
1.5.2 Set color preferences. 11 1.5.3 Set start codon choices 12 1.5.4 Set default values for BLAST searches. 13 1.5.5 Choose a default location for saving files 13 1.5.6 Finishing up your Preference settings. 14 1.6 Getting help 14 1.7 Checking for updates 15 1.8 Event Manager 17 2.1 Overview 17 2.2 BLASTing your sequence against the mycobacteriophage database 17 2.3 Cluster assignment 20 3 Importing your phage genome sequence into DNA Master 23 3.1 Overview 23 3.2 Where do I get my phage genome sequence from? 23 3.3 Importing your DNA sequence into DNA Master 24 3.4 Reverse-complementing your sequence 26 4 Performing and viewing a rapid automated annotation of your genome. 27 4.1 Overview 27 4.2 Running Auto-Annotate 27 4.3 Saving your file 29 <td></td> <td>1.5</td> <td>8</td> <td></td>		1.5	8	
1.5.3 Set start codon choices 12 1.5.4 Set default values for BLAST searches 13 1.5.5 Choose a default location for saving files 13 1.5.6 Finishing up your Preference settings 14 1.6 Getting help 14 1.7 Checking for updates 15 1.8 Event Manager 15 2 Provisional Cluster assignment of your phage 17 2.1 Overview 17 2.2 BLASTing your sequence against the mycobacteriophage database 17 2.3 Cluster assignment 20 3 Importing your phage genome sequence into DNA Master 20 3 Importing your DNA sequence into DNA Master 23 3.1 Overview 23 3.2 Where do I get my phage genome sequence from? 23 3.3 Importing your DNA sequence into DNA Master 24 3.4 Reverse-complementing your sequence 26 4 Performing and viewing a rapid automated annotation of your genome. 27 4.1 4.1 Overview 27 4.2 Running				
1.5.4 Set default values for BLAST searches				
1.5.5 Choose a default location for saving files 13 1.5.6 Finishing up your Preference settings 14 1.6 Getting help 14 1.7 Checking for updates 15 1.8 Event Manager 15 2 Provisional Cluster assignment of your phage 17 2.1 Overview 17 2.2 BLASTing your sequence against the mycobacteriophage database 17 2.3 Cluster assignment 20 3 Importing your phage genome sequence into DNA Master 23 3.1 Overview 23 3.2 Where do I get my phage genome sequence from? 23 3.3 Importing your DNA sequence into DNA Master 24 3.4 Reverse-complementing your sequence 26 4 Performing and viewing a rapid automated annotation of your genome. 27 4.1 4.1 Overview 27 4.2 Running Auto-Annotate 27 4.3 Saving your file 29 4.4 Looking at the output of your automated annotation 30 4.4.1 Viewing features in the				
1.5.6 Finishing up your Preference settings 14 1.6 Getting help 14 1.7 Checking for updates 15 1.8 Event Manager 15 2 Provisional Cluster assignment of your phage 17 2.1 Overview 17 2.2 BLASTing your sequence against the mycobacteriophage database 17 2.3 Cluster assignment 20 3 Importing your phage genome sequence into DNA Master 23 3.1 Overview 23 3.2 Where do I get my phage genome sequence from? 23 3.3 Importing your DNA sequence into DNA Master 24 3.4 Reverse-complementing your sequence 26 4 Performing and viewing a rapid automated annotation of your genome. 27 27 4.1 Overview 27 27 4.2 Running Auto-Annotate 27 4.3 Saving your file 29 4.4 Looking at the output of your automated annotation 30 4.4.1 Viewing features in the Feature Table 31 4.4.2 Viewing feat				
1.6 Getting help 14 1.7 Checking for updates 15 1.8 Event Manager 15 1.8 Event Manager 17 2 Provisional Cluster assignment of your phage 17 2.1 Overview 17 2.2 BLASTing your sequence against the mycobacteriophage database 17 2.3 Cluster assignment 20 3 Importing your phage genome sequence into DNA Master 23 3.1 Overview 23 3.2 Where do I get my phage genome sequence from? 23 3.3 Importing your DNA sequence into DNA Master 24 3.4 Reverse-complementing your sequence 26 4 Performing and viewing a rapid automated annotation of your genome. 27 27 4.1 Overview 27 4.2 Running Auto-Annotate 27 4.3 Saving your file 29 4.4 Looking at the output of your automated annotation 30 4.4.1 Viewing features in the Feature Table 31 4.4.2 Viewing features in the Feature Table 33<				
1.7 Checking for updates 15 1.8 Event Manager 15 1.8 Event Manager 15 2 Provisional Cluster assignment of your phage 17 2.1 Overview 17 2.2 BLASTing your sequence against the mycobacteriophage database 17 2.3 Cluster assignment 20 3 Importing your phage genome sequence into DNA Master 23 3.1 Overview 23 3.2 Where do I get my phage genome sequence from? 23 3.3 Importing your DNA sequence into DNA Master. 24 3.4 Reverse-complementing your sequence 26 4 Performing and viewing a rapid automated annotation of your genome. 27 4.1 Overview 27 4.2 Running Auto-Annotate 27 4.3 Saving your file 29 4.4 Looking at the output of your automated annotation 30 4.4.1 Viewing features in the Feature Table 31 4.4.3 Viewing ORFs in the Frames window 34 4.5 Running the BLAST function				
1.8 Event Manager 15 2 Provisional Cluster assignment of your phage 17 2.1 Overview 17 2.2 BLASTing your sequence against the mycobacteriophage database 17 2.3 Cluster assignment 20 3 Importing your phage genome sequence into DNA Master 23 3.1 Overview 23 3.2 Where do I get my phage genome sequence from? 23 3.3 Importing your DNA sequence into DNA Master 24 3.4 Reverse-complementing your sequence 26 4 Performing and viewing a rapid automated annotation of your genome. 27 27 4.1 Overview 27 4.2 Running Auto-Annotate 27 4.3 Saving your file 29 4.4 Viewing features in the Feature Table 30 4.4.1 Viewing ORFs in the Feature Table 31 4.4 Viewing ORFs in the Frames window 34 4.5 Running the BLAST function 37			0 1	
2 Provisional Cluster assignment of your phage 17 2.1 Overview 17 2.2 BLASTing your sequence against the mycobacteriophage database 17 2.3 Cluster assignment 20 3 Importing your phage genome sequence into DNA Master 23 3.1 Overview 23 3.2 Where do I get my phage genome sequence from? 23 3.3 Importing your DNA sequence into DNA Master 24 3.4 Reverse-complementing your sequence 26 4 Performing and viewing a rapid automated annotation of your genome. 27 4.1 4.1 Overview 27 4.2 Running Auto-Annotate 27 4.3 Saving your file 29 4.4 Looking at the output of your automated annotation 30 4.4.1 Viewing the sequence in the Sequence tab 33 4.4.3 Viewing ORFs in the Frames window 34 4.5 Running the BLAST function 37				
2.1Overview172.2BLASTing your sequence against the mycobacteriophage database.172.3Cluster assignment203Importing your phage genome sequence into DNA Master233.1Overview233.2Where do I get my phage genome sequence from?233.3Importing your DNA sequence into DNA Master243.4Reverse-complementing your sequence264Performing and viewing a rapid automated annotation of your genome.274.1Overview274.2Running Auto-Annotate274.3Saving your file294.4Looking at the output of your automated annotation304.4.1Viewing the documentation304.4.2Viewing features in the Feature Table314.4Viewing ORFs in the Frames window344.5Running the BLAST function37		1.0	LVent Manager	
2.1Overview172.2BLASTing your sequence against the mycobacteriophage database172.3Cluster assignment203Importing your phage genome sequence into DNA Master233.1Overview233.2Where do I get my phage genome sequence from?233.3Importing your DNA sequence into DNA Master243.4Reverse-complementing your sequence264Performing and viewing a rapid automated annotation of your genome. 274.1Overview274.2Running Auto-Annotate274.3Saving your file294.4Looking at the output of your automated annotation304.1.1Viewing the documentation304.2Viewing features in the Feature Table314.3Viewing the sequence in the Sequence tab334.4Viewing ORFs in the Frames window344.5Running the BLAST function37	2	Pr	ovisional Cluster assignment of your phage	
2.2BLASTing your sequence against the mycobacteriophage database172.3Cluster assignment203Importing your phage genome sequence into DNA Master233.1Overview233.2Where do I get my phage genome sequence from?233.3Importing your DNA sequence into DNA Master243.4Reverse-complementing your sequence264Performing and viewing a rapid automated annotation of your genome. 27274.1Overview274.2Running Auto-Annotate274.3Saving your file294.4Looking at the output of your automated annotation304.4.1Viewing features in the Feature Table314.4.3Viewing the sequence in the Sequence tab334.4Viewing ORFs in the Frames window344.5Running the BLAST function37				
2.3 Cluster assignment 20 3 Importing your phage genome sequence into DNA Master 23 3.1 Overview 23 3.2 Where do I get my phage genome sequence from? 23 3.3 Importing your DNA sequence into DNA Master 24 3.4 Reverse-complementing your sequence 26 4 Performing and viewing a rapid automated annotation of your genome. 27 27 4.1 Overview 27 4.2 Running Auto-Annotate 27 4.3 Saving your file 29 4.4 Looking at the output of your automated annotation 30 4.1 Viewing the documentation 30 4.2 Viewing features in the Feature Table 31 4.3 Viewing ORFs in the Frames window 34 4.5 Running the BLAST function 37				
3 Importing your phage genome sequence into DNA Master 23 3.1 Overview 23 3.2 Where do I get my phage genome sequence from? 23 3.3 Importing your DNA sequence into DNA Master 24 3.4 Reverse-complementing your sequence 26 4 Performing and viewing a rapid automated annotation of your genome. 27 4.1 Overview 27 4.2 Running Auto-Annotate 27 4.3 Saving your file 29 4.4 Looking at the output of your automated annotation 30 4.1 Viewing the documentation 30 4.2 Viewing features in the Feature Table 31 4.3 Viewing the sequence in the Sequence tab 33 4.4 Viewing ORFs in the Frames window 34 4.5 Running the BLAST function 37				
3.1Overview233.2Where do I get my phage genome sequence from?233.3Importing your DNA sequence into DNA Master243.4Reverse-complementing your sequence264Performing and viewing a rapid automated annotation of your genome. 274.1Overview274.2Running Auto-Annotate274.3Saving your file294.4Looking at the output of your automated annotation304.4.1Viewing the documentation304.4.2Viewing features in the Feature Table314.4.3Viewing the sequence in the Sequence tab334.4.4Viewing ORFs in the Frames window344.5Running the BLAST function37				
3.2Where do I get my phage genome sequence from?233.3Importing your DNA sequence into DNA Master243.4Reverse-complementing your sequence264Performing and viewing a rapid automated annotation of your genome. 274.1Overview274.2Running Auto-Annotate274.3Saving your file294.4Looking at the output of your automated annotation304.4.1Viewing the documentation304.4.2Viewing features in the Feature Table314.4.3Viewing the sequence in the Sequence tab334.4.4Viewing ORFs in the Frames window344.5Running the BLAST function37	3	Im	porting your phage genome sequence into DNA Master	
3.3 Importing your DNA sequence into DNA Master.243.4 Reverse-complementing your sequence264 Performing and viewing a rapid automated annotation of your genome. 274.1 Overview274.2 Running Auto-Annotate274.3 Saving your file294.4 Looking at the output of your automated annotation304.4.1 Viewing the documentation304.4.2 Viewing features in the Feature Table314.4.3 Viewing the sequence in the Sequence tab334.4.4 Viewing ORFs in the Frames window344.5 Running the BLAST function37		3.1		
3.4 Reverse-complementing your sequence 26 4 Performing and viewing a rapid automated annotation of your genome. 27 27 4.1 Overview 27 4.2 Running Auto-Annotate 27 4.3 Saving your file 29 4.4 Looking at the output of your automated annotation 30 4.4.1 Viewing the documentation 30 4.4.2 Viewing features in the Feature Table 31 4.4.3 Viewing the sequence in the Sequence tab 33 4.4.4 Viewing ORFs in the Frames window 34 4.5 Running the BLAST function 37		3.2	Where do I get my phage genome sequence from?	23
4 Performing and viewing a rapid automated annotation of your genome. 27 4.1 Overview 4.2 Running Auto-Annotate 4.3 Saving your file 4.4 Looking at the output of your automated annotation 30 4.4.1 4.4.2 Viewing the documentation 31 4.4.3 4.4.3 Viewing the sequence in the Sequence tab. 33 34.4.4 4.5 Running the BLAST function		3.3	Importing your DNA sequence into DNA Master	24
4.1 Overview274.2 Running Auto-Annotate274.3 Saving your file274.3 Saving your file294.4 Looking at the output of your automated annotation304.4.1 Viewing the documentation304.4.2 Viewing features in the Feature Table314.4.3 Viewing the sequence in the Sequence tab334.4.4 Viewing ORFs in the Frames window344.5 Running the BLAST function37		3.4	Reverse-complementing your sequence	26
4.1 Overview274.2 Running Auto-Annotate274.3 Saving your file274.3 Saving your file294.4 Looking at the output of your automated annotation304.4.1 Viewing the documentation304.4.2 Viewing features in the Feature Table314.4.3 Viewing the sequence in the Sequence tab334.4.4 Viewing ORFs in the Frames window344.5 Running the BLAST function37	1	Do	rforming and viewing a rapid automated apportation of your ge	nome 27
4.2Running Auto-Annotate274.3Saving your file294.4Looking at the output of your automated annotation304.4.1Viewing the documentation304.4.2Viewing features in the Feature Table314.4.3Viewing the sequence in the Sequence tab334.4.4Viewing ORFs in the Frames window344.5Running the BLAST function37	т			
4.3 Saving your file294.4 Looking at the output of your automated annotation304.4.1 Viewing the documentation304.4.2 Viewing features in the Feature Table314.4.3 Viewing the sequence in the Sequence tab334.4.4 Viewing ORFs in the Frames window344.5 Running the BLAST function37				
4.4 Looking at the output of your automated annotation304.4.1 Viewing the documentation304.4.2 Viewing features in the Feature Table314.4.3 Viewing the sequence in the Sequence tab334.4.4 Viewing ORFs in the Frames window344.5 Running the BLAST function37				
4.4.1Viewing the documentation304.4.2Viewing features in the Feature Table314.4.3Viewing the sequence in the Sequence tab334.4.4Viewing ORFs in the Frames window344.5Running the BLAST function37				
 4.4.2 Viewing features in the Feature Table				
 4.4.3 Viewing the sequence in the Sequence tab				
4.4.4Viewing ORFs in the Frames window344.5Running the BLAST function37			0	
4.5 Running the BLAST function				
			-	

5	Gathering additional information for refining your annotation	
	5.1 Generating a six-frame translation	41
	5.2 Generating a provisional genome map in DNA Master	
	5.3 Generating a graph of coding potential using GeneMark	
6	Phamerator & other Tools to assist with annotation	
	6.1 Phamerator	49
	6.1.1 Overview	
	6.1.2 Why Phamerator is useful to you at this stage of your annotation	
	6.1.3 How did my genome get into Phamerator already?	51
	6.1.4 Making Phamerator maps	
	6.1.5 Understanding and using the genome maps made by Phamerator	
	6.1.6 Viewing nucleotide sequence similarities in Phamerator	
	6.1.7 Other Phamerator features6.1.8 Saving Phamerator maps	
	6.2 Starterator	
	6.2.1 Overview	
	6.2.2 Interpreting Starterator results:	
7	8 1 1 8	
	7.1 Overview	
	7.2 The Guiding Principles	63
8	Gene by gene: evaluating and improving your draft annotation	
-	8.1 Overview	
	8.2 Button-pushing mechanics reserved for Section 9	
	8.3 Decision Tree for evaluating the draft annotation	
	8.4 Evaluating protein-coding gene calls	
	8.4.1 Is the designation of this ORF as a gene well-supported?	
	8.4.2 Is the called start site for this gene the best possible choice?	
	8.5 Checking gaps in the draft annotation for uncalled genes	
	8.6 Finding and refining tRNA and tmRNA genes	
	8.7 Completing your annotation refinement	
9	The mechanics of making changes to your annotation	
	9.1 Overview	
	9.2 Making common changes to your annotation	
	9.2.1 Deleting a gene	
	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	83
	9.3.1 Posting changes	
	9.3.2 Validating your annotation	
	9.3.3 Renumbering & formatting annotated features	
	9.3.4 Re-BLASTing a gene	
	9.4 Making less common changes to your annotation	
	9.4.1 Annotating programmed translational frameshifts	
	9.4.2 Annotating introns	

9.4.3 Annotating wrap-around genes	
9.5 Predicting tRNA and tmRNA genes	
9.5.1 Running web-based Aragorn (version 1.2.36)	
9.5.2 Running tRNAscan-SE (version 1.23)	
9.5.3 tRNA secondary structure and end determination	
9.5.4 Entering a tRNA in DNA Master	
9.5.5 Identifying and annotating tmRNA genes	
9.6 Documenting your gene calls	
10 Assigning gene functions	105
10.1 Overview	
10.2 Using bioinformatic tools to assign gene function	
10.2.1 BLASTP	
10.2.2 Conserved Domain Database (CDD)	
10.2.3 HHpred	
10.3 Other ways to assign gene function	114
10.3.1 Synteny	
10.3.2 Prior functional assignments	115
10.3.3 Phamerator	115
11 Merging and checking annotations	117
11.1 Merging overview	
11.2 Merging multiple annotations into a single file	
11.3 Checking an annotation	
12 Submitting final files for review and GenBank submission	
12.1 Details of your final DNA Master (.dnam5) file	
12.1 Details of your mar DNA Master (.unani5) me	
12.3 Details of your author list	
12.4 Details of your cover sheet	128

How to use this guide

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

There are several different ways you can use this guide.

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

A NOTE ON CLASSROOM PRAGMATICS

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

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

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

AN IMPORTANT NOTE ABOUT THIS GUIDE'S SYNTAX

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

File → Open → Archived DNA Master file

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

e Manager Tools '	Window H	Help	
New	Ctrl+N	rit Frittingt	A Mighter
Open		Archived DNA Master file	Ctrl+Alt+O
Import Merge Close	Ctrl+W	GCG-Formatted File GenBank-Formatted File	Ctrl+O
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		Phylip Multiple Sequence File	
Preferences	Ctrl+P	Sequence from Accession Number	

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

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

New Ctrl+N Directories GCG Export New Features Translation Shine Dalgamo Colors Validation Bevert Import Click to specify Map colors Click to specify Six Frame colors Click to specify Six Frame colors Enter the specify Six Frame colors En	File Edit Genome DN	IA Tools	Window Help	Internet Local Settings Automation Phylogeny Timed Events Miscellaneous	04
Import Reverse Forward CDS Frame 1 Apply Close Ctrl+W Close All CDS Frame 2 Apply Save as DNAM5 File Ctrl+S ImRNA CDS Frame 3 Apply Export GCG file Ctrl+E ImRNA CDS Frame 4 Export Wattoparse Ctrl+P ORF Export Apollo XML File Export Apollo XML File Export Apollo XML File Forward RNA Quit Ctrl+Q Sine Dalgamo Frame 1 Errore 1 Frame 1 Frame 1 Office and file File File Frame 3 Click to specify Frame Plot colors Frame 3 Frame 3 Frame 3 Offing Site Frame 3 Frame 3 Frame 3 Frame 3 Frame 3		Ctrl+N		Directories GCG Export New Features Translation Shine Dalgamo Colors Validation	n Bevert
Close Cliffwire Cliffwire Apply Close All IfRNA ImRNA CDS Frame 3 Export Save as DNAMS File Ctrl+E CDS Frame 4 CDS Frame 4 Export Export GGG file Ctrl+E DRF CDS Frame 6 Export CDS Frame 6 Export Preferences Ctrl+P ORF CDS Frame 6 Export Apolo XML File Export Apolo XML File CDS Frame 6 Export Apolo XML File CDS Frame 6 Export Apolo XML File Export Apolo XML File<		,	素	Reverse Forward CDS Frame 1	s <u>C</u> ancel
Save as DNAMS File Ctrl+S Export GCG file Ctrl+E Export Apollo XML File Export Split Files Quit Ctrl+Q Quit Ctrl+Q		Ctrl+W			Apply
Export Apollo XML, File DRF CDS Frame 6 Export Apollo XML, File PseudoDRF CDS Frame 6 Preferences Ctrl+P ORF Reverse RNA Quit Ctrl+Q Stite Click to specify Frame Plot colors Frame 1 Frame 1 Frame 1 Repeat Binding Site Frame 3			·		Export
Autoparse Preferences Ctrl+P Dulitices Ctrl+Q Quit Ctrl+Q Shine Dalgarno Frame 1 Terminator Frame 2 Binding Site Difformed 3 Difformed and site Difformed 3	Export Apollo XML File	Ctrl+E	14		
Preferences Ctrl+P Quit Ctrl+Q PrevidoRNA 10 Site -35 Site	Export Split Files		-	PseudoORF Forward RNA	
Quit Ctrl+Q Shine Dalgarno Click to specify Frame Plot colors Terminator Terminator Frame 1 Repeat Frame 3 Binding Site Offenentiation	Preferences			-10 Site	
Terminator Frame 2 Repeat Frame 3 Binding Site Site	Quit	Ctrl+Q	19. Ť	Click to specify Frame Plot colo	rs
Binding Site	And Street	1		Terminator Frame 2	
		-		Frame 3	
	- 14 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1	111	1997	Oligonucleotide	

1 Introduction to DNA Master

1.1 DNA Master overview

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

1.2 Installation

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

1.3 Quick Start Guide

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

1.4 DNA Master program structure

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

The Feature Table

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

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

The Documentation

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

Interaction between the two

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

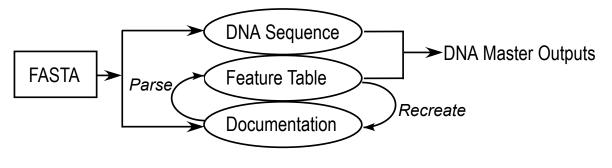


Figure 1.1

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

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

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

IMPORTANT TO REMEMBER:

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

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

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

Analysis programs running within DNA Master

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

1.4.1 Glimmer

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

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

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

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

http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer 3.cgi

1.4.2 GeneMark

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

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

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

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

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

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

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

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

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

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

A graph of the heuristic model can also be generated for comparisons. Use this version of the website <u>http://opal.biology.gatech.edu/GeneMark/heuristic_hmm2.cgi</u>

Note:

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

1.4.3 Aragorn

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

http://130.235.46.10/ARAGORN/

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

1.5 Setting Preferences

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

To get to the Preferences Window, select:

File → Preferences

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

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

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

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

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



Figure 1.2

1.5.2 Set color preferences

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

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

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

Then set the values as shown below.

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

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

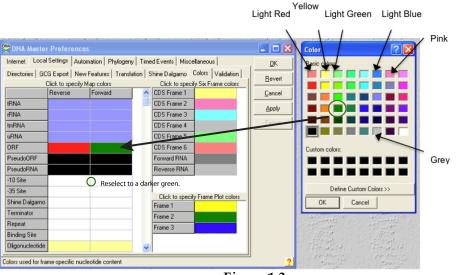


Figure 1.3

1.5.3 Set start codon choices

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

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

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

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

Figure 1.4

1.5.4 Set default values for BLAST searches

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

File → Preferences [Internet] [[Blast]]

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

🞇 DNA Master Preferences		
Internet Local Settings Automation F	Phylogeny Timed Events Miscellaneous	<u>0</u> K
Email REBase NCBI Gene Predict	ion Blast	Revert
Communicating with Remote BLAST Ser	ver	
QBlast Server	http://blast.ncbi.nlm.nih.gov/Blast.cgi	<u>C</u> ancel
☑ Include "ignore self" terms when initia	ting BLAST form	Apply
Number of hits to request from server	100 🜲	Export
Number of hits expected	10 🚖	
 Saving BLAST results to Local Database 		
Save hits with E-Values smaller than 10E-	3 🔹	
🔽 Save a minimum number of	1 thits, regardless of E-Value	
Save a maximum number of	100 🚖 hits, regardless of E-Value	
Query Timing when performing Genome	BLAST	
Delay between multiple BLAST queries	4 🚖 seconds	
Send requests before retrieving for	40 🚖 seconds	
Limit queries to a maximum number of	20 🔶 pending queries	
Limit wait time for BLASt results to	10 minutes	

Figure 1.5

1.5.5 Choose a default location for saving files

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

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

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

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

Figure 1.6

1.5.6 Finishing up your Preference settings

Once you have finished setting your DNA Master preferences:

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

The Preferences Window will close.

1.6 Getting help

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

👺 Extracted from FastA Library Timshel.fasta									
Overview Features References Sequence Documentation									
Sort By	Index	-	•		Name	Start	Stop		Description Sequence Product Regions Blast Context
Select Fe	atures	Dire	ct SQL	▶	1	408	704		Name 1 GenelD
-	· .	All			2	743	1177		
Туре		All	<u> </u>		3	1254	1577		Type CDS 💽 🖬
Name	like			L	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	ъÍ			L	7	4377	5294		Direction Forward
				L		5374	5449		
Length				L	8	5547	6899		Translation Table Undefined
Regions	>			L	9	6896	7309		EC Number
% GC	<				10	7306	8283		2 1 1
CAI	5			L	11	8479	10005		Product
EC#	like [12	10002	11459		gp1
					13	11456	12469		_
Product	like			L	14	12520	13020		Function
Function	like				15	13055	14014		
FeatureID	- [16	14080	14268		
Tag	like			L	17	14276	14650		Notes
-		- ·		I.				∠	Original Glimmer call @bp 408 has strength 2.35; GeneMark All calls start at 297
Hide I	gnored	Featu	ures	Ľ		- 1	· ·	L	
Sele	ot All Fr	eature	~	_	Insert De	lete Pos	t Valida	ate	
🔣 📢 🗶 🕀 🕨 🕨 🕨 1 - 50000 Position : 48037 🔽 Controls >> Map 🔽 Map >> Controls									
4 5 6	7 8	Bho l	11 12 1	N.	shilled b	25	26 27	N	1321 34 MT&R 44 MAY 49 605 611666 66 61 6 6 1 6 1 6 1 6 1 6 1 6 1
89 Features	<u>, , , , , , , , , , , , , , , , , , , </u>	Live	1.2	ZU		<u> </u>	<u>/</u>		53278 (🕃 ?
				_			_		

Figure 1.7

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

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

1.7 Checking for updates

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

Help → Update DNA Master

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

1.8 Event Manager

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

2 Provisional Cluster assignment of your phage

2.1 Overview

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

http://phagesdb.org/clusters/

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

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

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

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

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

2.2 BLASTing your sequence against the mycobacteriophage database

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

- Go to 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 Nucleotide BLAST

Go to Protein BLAST

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

Choose program to use and database to search

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

Advanced Options

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

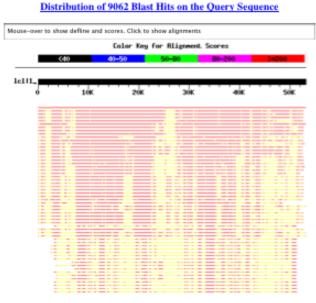
Filter 📋 Low complexity 📄 Mask for lookup table only	
Expect 10 🗘 Matrix BLOSUM62 🛟 🗆 Perform ungapped	alignment
Query Genetic Codes (blastx only) Bacterial (11)	•
Database Genetic Codes (tblast[nx] only) Bacterial (11)	\$
Frame shift penalty for blastx No OOF	

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.





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

	Score	12	
Sequences producing significant alignments:	(bits) \	Jalue	
Echild complete sequence, 53159 bp including 10bp overhang (CGGT	1.01	54e+05	0.0
Turbido Complete Sequence, 53169 bp including 10 bp (CGGTCGGTTA).		34a+04	0.0
Jern complete sequence, 53163 bp including 10 bp 3' overhang (CG.		01e+04	0.0
Whabigail7 complete sequence, 53167 bp including 10 bp 3' overha.		19#+04	0.0
Bugsy complete sequence, 49937 bp including 10 bp 3' overhang (C.		51e+04	0.0
Heffalump complete sequence, 53085 bp including 10 bp 3' overhan.			
ChipMunk complete sequence, 53932 bp including 10 bp 3' overhang.			
EvilGenius complete sequence, 53935 bp including 10 bp 3' overha.	7452		
SemperFi complete sequence, 53235 bp including 10 bp 3' overhang.	4244	0.0	
Pukovnik	2661	0.0	
Trixie Complete Sequence, 53526 bp including 10 bp 3' overhang (
D29	260		
Loser complete sequence, 53486 bp including 10 bp 3' overhang (C.			
AnnaL29 complete sequence, 53253 bp including 10 bp 3' overhang .			
	2460		
RedRock			
Jaquared complete sequence, 52967 bp including 10 bp 3' overhang.			
Serenity complete sequence, 52000 bp including 10 bp 3' overhang.			
15	156-		
Odin complete sequence, 52807 bp including 10 bp 3' overhang (CG.			
ADZIY complete sequence, 52519 bp including 10 bp 3' overhang (C.	1111	0.0	
Che12	98	7 0.0	
Jeffabunny Complete Sequence, 48963 bp including 10bp 3' overhan.	913	2 0.0	
Harmer	913	2 0.0	
CloudWang3 complete sequence, 52873 bp including 10 bp 3' overha		2 0.0	
Artemis2UCIA complete sequence, 52344 bp including 10 bp 3' over		0.0	
Blue7 Complete Sequence, 52288 bp including 10 bp 3' overhang (C.			
Zaka complete sequence, 52122 bp including 10 bp 3' overhang (CG.			
Gladiator Final Sequence, 52213 bp including 10 bp 3' overhang (
		0.0	
Fibonacci complete sequence, 52462 bp including 10 bp 3' overhan.		0.0	
EagleEye complete sequence, 52974 bp including 10 bp 3' overhang.			
PackMan Complete Sequence, 51339 bp including 10 bp 3' overhang .			
Hyxus Complete Sequence, 53425 bp including 10 bp 3' overhang (C.			
Kazan complete sequence, 52160 bp including 10 bp 3' overhang (C.			
JewelBug complete sequence, 50341 including 10 bp 3' overhang (C.	661	0.0	
Syxis complete sequence, 51250 bp including 10 bp 3' overhang (C.		0.0	
HcFly complete sequence, 52502 bp including 10 bp 3' overhang (C.		0.0	
DaVinci Final Sequence, 51547 bp, 10 bp 3' Overhang (CGGTCGGTTA).		0.0	
EricB Complete Sequence, 51702 bp including 10 bp 3' overhang (C.	653	0.0	
Catalina complete sequence, 53411bp including 10bp overhang (CGG.			
Alma Complete Sequence, 53177 bp including 10 bp 3' overhang (CG.			
Horpher26 complete sequence, 51294 bp including 10 bp 3' overhan.			
LittleGuy complete sequence, 51178 bp, including 10 bp 3' overha.			
Wile Complete Sequence, 51308 bp including 10 bp 3' overhang (CG.			
arra combrara Sadnanca, prove ob rucronrud to ob 2. overneud (co-	39	e-10	-

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.

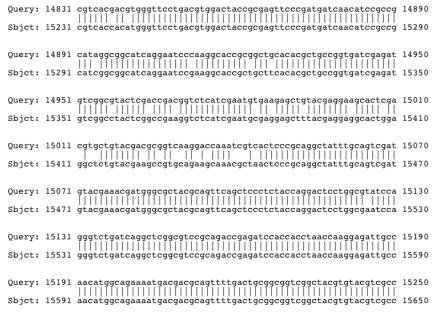


Figure 2.4

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	show defline and	scores. Click to	show alignment	s			
			for Alignmen				
	<4 0	40-50	50-80	80-200	>=200		
1.114						_	
1c111_	10K	20K	30K	40K	50K		
						K1	
						 K2, K3, K4	, K5
				1			
		=	-				
Sequences pr	oducing sign	ificant ali	ignments:		Sco (bi	ore E ts) Value	
BEEST Comple JAWS Complet BarrelRoll C Angelica Fin	nal Sequence, te Sequence, omplete Sequ al Sequence te Sequence,	59906 bp i 59749 bp ir ence, 59672	including 1 cluding 11 bp includ	1 bp 3' over bp 3' over ing 11 bp 3	chang (C chang (CT 3' overha	$\frac{4.766e+04}{3.606e+04}$ $\frac{3.010e+04}{3.008e+04}$ $\frac{2.906e+04}{2.906e+04}$	0.0 0.0 0.0 0.0 0.0
CrimD Final	Sequence, 59 Sequence, 61	798 bp, 11	bp 3' over	hang (CTCG1	AGGCAT),	$\frac{1.826e+04}{1193}$ 0.0	0.0
			Figure ?	5			

Figure 2.5

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

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

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

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

3 Importing your phage genome sequence into DNA Master

3.1 Overview

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

3.2 Where do I get my phage genome sequence from?

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

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

A NOTE ON FILE TYPES

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

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

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

> Giles Complete Genome Sequence, 53746 bp GGCAGACTTTTTTTGCGCGGCGGCGCGCGCGCGCGCGCCGCCCCC GCCGGGTCGGAGGCGGCCGAATGACGCCACCTCGGGCCGCGGTGGCCGAC ACGCCGGATACGCCCGCAGAGGGCAAATCAGGGGCCAAAACGCGGGCCAA

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 Windo	w <u>H</u> elp		
New	Ctrl+N	Hit Pristant	Frighter 1
Open	Þ	Archived DNA Master file	Ctrl+Alt+O
Import Close Close All	► Ctrl+W	GCG-Formatted File GenBank-Formatted File	Ctrl+0
Save as DNAMS I Export GCG file Export Apollo XM Export Split Files	Ctrl+E	Previously downloaded NCBI file Entrez ASN.1 Sequence File Entrez XML Sequence File Apollo XML File	ctrl+Alt+A
Autooperso	<	FastA Multiple Sequence File	Ctrl+Alt+F
 Autoparse Preferences 	Ctrl+P	Phylip Multiple Sequence File	
Utilities	•	Sequence from Accession Numb	er
Quit	Ctrl+Q	Here a	1 Hand

File \rightarrow Open \rightarrow FastA Multiple Sequence File

Figure 3.1

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

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

👺 FastA Sequence	s from Echild.f	asta			1. (A.	11-2-19
Sequences Pairwise	Alignment Multiple	Alignment Site Analysis			Shere at	- Sure
Idx Description			Length	Insert Sequence	2 the	-det.
 Echild complete s 	equence, 53159 bp	including 10bp overhang (CGG	TC(53159	Delete Sequence	the second second	*
				Update Sequence	at a	ille a
				Verify updates	and a	There
				Save FastA File	and the	reither
				Export Summary	anti-	
Sequence		Complement Hide	Annotation		the state	in the second
AGCCGGCGCAGGAGCI CAAGATCCCCGCTCAI	CCTCACCCAGCCT CCCGGGTAGTTGTA	AAGTCGATCGGAGTCCCG	2. Select	<u>^</u>		
		CCTAGGTATGGCTACCCA CCCGTCGTCGGTGTTGCC		Create Seq	uence from this en	try only
		ACCGCAGGGAGGACCGCC			arate sequences fo	
CGCCGCCCGAGGCGG AGGACGAATACCTCC	TGGGATGCGGAAA ATAGTGATCTACG	GCTCGCACTCATTCGCAT TCTACGAGCCGTGGGATG CCACTTGCTCGGTGGGTG	1. Click	Catenate a	ate sequences for Il entries into a sin(filtering for valid Ol	gle sequence
		AGGGCCTAAAGGCCCGAA CGCCCACTAGGGCGCTCG		Create Seq	uence with entries	as replicons
AAGTAATACCGGCCT:	GAGGGCCGGTTAT	CTGACCCGGCAACCGCCG 🔽			100	The Fra
1 Sequences 53	159 bp Create	es a sequence form		GC 🖹 Export ?	and a second	Fritting
and a	and a	and a	Cart 11	Charles -	and a	"Allerer

Figure 3.2

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

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

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

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

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

Overview	Features Refere	ences S	Sequence	Docu	mentation	1									
Feature		1	• F	> 19	Position	: 866 - 1	234 (369	bases)	Raw		BLASTN	BLASTF	Add F	eature	1 62.
1	TGCGGCCGCCCC	ATCCT	TACGGO	TTTCC	AGTCGA	TCGGAG	TCCCGA	GCCGGCGC	AGGAGC	GC C	TCACCCA	GCCTCTG	TGCGCC	CCCCA	GGA
93	CGCAAGATCCCC	GCTCA	CGCGGGT	AGTTGT	ATGGGC	TAATCO	GCAAAC	GGCCTCTG	AGGCCG	CGA	GACCAAT	GTCACAC	CAGGTO	GTGG	LTG
185	TTATTGACGCAC	GCGTC	GTTAAG	AGGACA	TGGCCT	AGGTAT	GGCTAC	CCAAACTT	AGATTC.	AAA	ACCAGTO	CCCTGGC	CCCCGT	CGTC	GGT
277	GTTGCCGCTCCT	rggcggi	CGCGGGG	CCAGGI	CCACCA	CCGCAG	GGAGGA	CCGCCATG	AAGATC.	ATC	CGCTCGC	TEGEEGG	GGCGCT	CGCA	стс
369	ATTCGCATCGCC	GCCCG	AGGCGGC	TGGGAT	GCGGAA	ATCTAC	GAGCCG	TGGGATGA	GGACGA	ATA	CCTCCT#	TAGTGAT	CTACGO	CACTI	FGC
461	TCGGTGGGTGTC	AAGTG	ATACTCA	TGTATO	TAGTTA	TTGAGO	GCCTAA	AGGCCCGA	ATAAGA	GCC	GCACAGO	CGGCTCT	CTAAGA	AGCGCO	CCA
553	CTAGGGCGCTCG	GAAGTA	ATACCGO	CCTTG	GGGCCG	GTTATO	TGACCC	GGCAACCG	CCGGGT	стт	CTGCCGC	GCCCAGT	GCCCCC	GCTCA	ATA
645	GAAGGGGTGAGG	CAACCI	TGTACO	GCACTO	GCTCGA	GTGCCI	ACTGGG	CCTCGCAG	CCGGGG.	AAG	TTCGACO	TTCTGAA	CCTGCG	GATGA	ACG
737	TTCCCGAGCACG	TCGCCI	GCACGAC	ATCCCC	GATCTT	CGGCCC	TCTGAC	TTCGTTCC	GGCCAA	сст	CGCAGCO	TGGAACA	TGCCAC	GCCAT	rcg
829	CGAATACGCCGC	CATTT	CGGGCGG	CGCACT	CCACTT	CITTOCI	TGACGA	TTACCOTT	TCGAGA	CCG	TCTGGTO	GTCCCCC	GAGCGT	ICTTCI	FCC
921	CCCGCGTGCAAG	GCGGTTI	GGGGCCC	CTCTGA	CCCCCG	ATTTCA	GCGTGT	GGCGAGAA	ATGCCG	CGG	GTAGCGO	AGGTCTG	GAATAC	CTATO	CGC
1013	AGCCGCTGGTGC	GGCGCI	TATTGO	CAATCI	CAGGGG	ATCGAG	GTAATT	CCCACGGT	CGGTTG	GGG	TEGACEC	GACACAT	ACGAAT	TCTG	TT
1105	TGACGGCCTGCC														
1197	AAGAACTCGTCT	IGGAGA	ACGCAGO	CGAAGA	CCCTTT	TEGTCT	ACGGCC	GGCTCAGA	TTCTGC	GAC	GAAATAG	ACCTTCC	CGAGGT	rccgg	GAG
1289	TACCCGACCTTC														
1381	AGCCGCGCCCCGA	AAGGG	GGTGGC	GGTGCO	ACAACT	CCCGGA	GGCGGT	GGATCTGG	AGGCGG	TGG.	AGGCGGG	GGAAAGG	GCAAAG	GCTC	CGC
1473	AAGCGCCTCAAC	CGGTG	GCGTAAA	GAACCO	CGGAGG	CGGCGC	ATCCGG	CTCCGGTG	GTGGCG	GTA	GCCAGAT	GCCGAGC	AAGCCC	CAAACO	CGT
1565	GGCCGGATAACC	TACTO	AGAAGT	CEGACA	AGCGCC	TEGACA	CAGACO	AGCGTCGA	GCCGAC	AAC	ATCGCCC	COGTCAA	CCCGCG	GTGGA	AC
1657	GAGCGCACGCCC														
1749	CGACGTAACTGC	GCTGCI	CAAGEE	CGAGAA	CCTCAA	TGACA	CTACAC	CCACATGT	CGTCCG	CGA	AGTEGAT	CGATGAG	CACGGO	CAACCO	CGG
1841	CTCAGTGGGACT	PACCTCI	ACGGCT	CCCGCC	ACAGAA	TCATGO	ACGAGA	TCAACGCC	CGAGCT	GCG	TTGTGGG	COGAGGO	AGCCCG	AGGG	тс
1933	ATCCGCACGACC														
2025	GTGGGACGCCAG														
2117	ACCTGTCGAAGT														
2209	AAGCTCGGTGAG														
2301	CCCTGAGTTCAA														
2393	CGACAACCGAGC														
2485	GTCCCTACGCGA														
2577	ACAGGEGAGTAC														
2669	ACCATGAGCTGG														
2009	GATETGEGAGET														
5701	ACCOUNTER	LOCAUT	POTALCO	000000				LCCCCCCCC					CALCO	COCCO	100
a lat a		N 1-5	0000		Position	- 1		Controls >>	Man 🔽	Man	>> Contri	als			
					1 0010011	11	100								
1 Features	Live												53	159	2

Figure 3.4

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

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

3.4 Reverse-complementing your sequence

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

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

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

4 Performing and viewing a rapid automated annotation of your genome

4.1 Overview

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

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

4.2 Running Auto-Annotate

• As shown in **Figure 4.1**, go to:

Genome \rightarrow Annotation \rightarrow Auto-Annotate

DNA DNA	Master						
File Ec	lit Genome DNA Tools Wi	ndow Help					
1			from FastA Li	hrary Echil	d fasta		
28312	Analyze all gone starts						
	Auto-Annotate		ures Reference	s Sequence	Documenta	ition	
1	Blas Table BLAST All Genes		– –	Name	5'End L	ength	Description
100	Coding Capacity		Direct SQL	Þ			Name
1362	Codon Spacing		AI V				Туре
	Features	•	<u> </u>				
		Ctrl+Alt+G					5'End
628 3	Karlin's Dinucleotides						3'End
	Learn Features	Ctrl+Alt+L					Length
1	Mutational Bias						Direction
		Ctrl+Alt+H	I				Translation Ta
	Origin prediction		L				EC Number
	Predict stop						EUNUMBER
		Ctrl+Alt+P					Product
1.5	Rearrange						Product
	RNASeq Run Length	•	i1				
		Ctrl+Alt+T	I				Function
12.2		Ctrl+Alt+K	L				Function
1.13	tRNA Profile	COTT PICT R					
	and an internet		4				Notes -
1 De	the fit of the	Tag like					Notes
1	anistican's paris	Hide Ignore	dEasturas				
28312	and a	i inde ignore	d i edities				
	the site of the	Select All	Features				
1	and the second						
11	and the stand						
1824	Here Here						
TE	and the Press			<		>	
7	pristing to paris			Insert D	elete Post	Validate	
828 J	the state			1 - 50000		ition : 1	Contr
	11 the 1	(Pos	ition : 1	UV Contri
- 13	in the second	_					
-	0	Features	Live Whole	e genome anal	ytical tools		

Figure 4.1

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

	Annotate					
Current	Replicon	🔽 Export	log file	Annotate	1	
C All Rep Gene Call		& Comments	BLAST	Searches Auto	mated Analyses	:]
🔽 Docur	nent tRNAs	found with /	Aragom			
Docur	nent ORFs	found with			1	
	mmer 3.02 a					•
	neMark HM					
	th analyses Favor Glim	, combining	them as to	swollc		
		iner cans ieMark calls				
		enes called b				
				cond method		
		e 4 minutes		•		
		o comments		gene		
✔ Examin	ne and learr	n gene featu	res			
Status : Idl			No BLAST	ſ		?
🚆 Auto-	Annotate					
		Export	log file	Annotate	1	
🔿 All Rep				Annotate		
Gene Cali	ng Tags I	& Comments	BLAST	Searches Autor	nated Analyses	
Locus T	ags					
Prefix	DNAM	_			•	
Start	1	1			۷.	
) ·	-				
Spacing	1	•		×		
SSC: CP:	SD: SCS:	Gap: Blast	: LO: ST:	F: FS:		~
🔛 Auto-	Annotate	,				
Curren			log file			
		A Export		Annotate		
C All Rep			RIAST	Annotate		1
C All Rep Gene Cal	dicons ling Tags	& Comments		Searches Autor	nated Analyses	
C All Rep Gene Cal	vicons ling Tags m BLAST s	& Comments earch on NC	BI nr data	Searches Autor	nated Analyses	
C All Rep Gene Cal Perfor Save	licons ling Tags m BLAST s hits with E-	& Comments earch on NC values less t	:Bl nr data han 10E	Searches Autor base	nated Analyses	
C All Rep Gene Cal Perfor Save	licons ling Tags m BLAST s hits with E-	& Comments earch on NC	:Bl nr data han 10E	Searches Autor base 3 🔹	nated Analyses	
C All Rep Gene Cal Perfor Save V R	vicons ling Tags m BLAST s hits with E- egardless o	& Comments earch on NC values less t	:BInrdata han 10E es, save	Searches Autor bjee 3 🔹 1 🔹 hit(s)	nated Analyses	
C All Rep Gene Cal Perfor Save V R	viicons Iing Tags m BLAST s hits with E- egardless o mit number	& Comments earch on NC values less t f their E-valu	181 nr data han 10E- es, save saved to	Searches Autor bjee 3 🔹 1 🔹 hit(s)		
C All Rep Gene Cal Perfor Save V R V L Delay bet	viicons Img Tags Im BLAST s hits with E- egardless o mit number ween multip	& Comments earch on NC values less th f their E-valu of hits to be	BInrdata han 10E es, save saved to ueries	Searches Autor bare 3 \$ 1 \$ hil(s) 100 \$ hits 4 \$ seco		
C All Rep Gene Cal Perfor Save V R V Li Delay bet	Vicons Ing Tags m BLAST s hits with E- egardless o mit number ween multip wests before	& Comments earch on NC values less th f their E-valu of hits to be ble BLAST qu e retrieving fo	BInrdata han 10E es, save saved to ueries	Searches Autor byte 3 \$ 1 \$ 100 \$ hits 4 \$ seco	ds	
C AI Rep Gene Cal Perfor Save V R V Li Delay bet Send requ	Vicons Ing Tags m BLAST s hits with E- egardless o mit number ween multip uests before Annotate	& Comments earch on NC values less th f their E-valu of hits to be ble BLAST qu e retrieving fo	IBI nr data han 10E es, save saved lo ueries r	Searches Autor bare 3 \$ 1 \$ hil(s) 100 \$ hits 4 \$ seco	ds	
C AI Rep Gene Cal Perfor Save I♥ R I♥ Li Delay bet Send req OHP Auto C Curren	Vicons Ing Tags m BLAST s hits with E- egardless o mit number ween multip uests before Annotate t Replicon	& Comments earch on NC values less th f their E-valu of hits to be ble BLAST qu e retrieving fo	IBI nr data han 10E es, save saved lo ueries r	Searches Autor bare 3 \$ 1 \$ hil(s) 100 \$ hits 4 \$ seco	ds	
C All Rep Gene Cal Perfor Save V R V Li Delay bet Send requ C All Rep	Vicons Ing Tags Im BLAST s hits with E- egardless o mit number ween multip uests before Annotate t Replicon Vicons	& Comments earch on NC values less th f their E-valu of hits to be ble BLAST qu e retrieving for e E Export	iBInrdata han 10Er es, save saved o veries r t log file	Searches Autor	ds nds	
C AI Rep Gene Cal Perfor Save V R V Li Delay bet Send req C Auto C AI Rep Gene Cal	Vicons Ing Tags m BLAST s hits with E- egardless o mit number ween multip ween multip ween multip constants Annotate t Replicon Sicons ling Tags	& Comments earch on NC values less th f their E-valu of hits to be ole BLAST qu a retrieving for e IV Export & Comments	IBI m data han 10E es, save saved lo ueries r t log file t log file	Searches Autor tape 3 ÷ 1 ÷ hit(s) 100 ÷ hits 4 ÷ seco 100 ÷ seco Annotale Searches Autor	ds Inds	
C AI Res Gene Cal Perfor Save V R V Li Delay bet Send requ C AI Res Gene Cal V Auto	Vicons Ing Tags m BLAST s hits with E- egardless o mit number ween multip uests before Annotats t Replicon vicons t Replicon vicons save DNAM	& Comments earch on NC values less th f their E-valu of hits to be ole BLAST qu a retrieving for e IV Export & Comments	BI nr data han 10E es, save saved lo ueries r log file BLAST neen_Blas	Searches Autor befet 3 ÷ 1 ÷ hil(s) 100 ÷ hils 4 ÷ jacco 100 ÷ sept Annetale Searches Auto	ds Inds Inated Analyses am5	
C AI Rep Gene Cal Perfor Save V R V C Cal Send requ C Curren C AI Rep Gene Cal V Auto C Expo	Vicons Ing Tags m BLAST s hits with E- egardless o mit number ween multip uests before Annotate t Replicon officins tags save DNAM rt alternative	& Comments search on NC values less th f their E-valu of hits to be ale BLAST qu le retrieving for e e for Export & Comments 15 file as SF	IBI nr data han 10E/ es, save saved to veries rr t log file t BLAST neen_Blas	Searches Autor bere 3 1 1 thils 100 thils 4 1 seco 100 thils 4 thils Searches Autor Searches Autor ted Annotated dr Export a Profile	ds Inds Inded Analyses am5 of features	1
C AI Rep Gene Cal C Perfor Save V R V R V C Delay bet Send req C Curren C AI Rep Gene Cal V Auto C AI Rep Gene Cal V Auto C AI Rep Gene Cal	Vicions Ing Tags m BLAST s hits with E- egardless o mit number ween multip uests before chanotate t Replicon vicions ling Tags save DNAM rt a summar rt a GC-Cor	& Comments earch on NC values less th f their E-valu of hits to be ele BLAST qu is retrieving for e training for Export & Comments to file as SF e start codor y of BLAST itent map	IBI nr data han 10E es, save saved o veries rr log file : BLAST neen_Blas ns F hits F	Searches Autor by 5 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	ds nds mated Analyses am5 of features àray Holes and me map	1
C AI Rep Gene Cal C Perfor Save V R V R V C Delay bet Send req C Curren C AI Rep Gene Cal V Auto C AI Rep Gene Cal V Auto C AI Rep Gene Cal	Vicons Ing Tags m BLAST s hits with E- egardless o mit number ween multip uests before Annotato t Repticon Vicons ling Tags save DNAM rt atternativ rt a GC-Cor rt Frames m	& Comments earch on NC values less til f their E-valu of hits to be ble BLAST qu is retrieving for e is retrieving for the ST qu & Comments to file as is is e start codor y of BLAST itent map rep	IBI nr data han 10E es, save saved o veries rr t log file t BLAST neen_Blas ns F F F	Searches Autor april 3 ÷ 1 ÷ hit(s) 100 ÷ hits 4 ÷ seco 100 ÷ seco Annotate Searches Autor Export a set of Export a	ds mated Analyses am5 of features array Holes and me map map	1
C AI Rep Gene Cal C Perfor Save V R V R V C Delay bet Send req C Curren C AI Rep Gene Cal V Auto C AI Rep Gene Cal V Auto C AI Rep Gene Cal	Vicons Ing Tags m BLAST s hits with E- egardless o mit number ween multip uests before Annotato t Repticon Vicons ling Tags save DNAM rt atternativ rt a GC-Cor rt Frames m	& Comments earch on NC values less th f their E-valu of hits to be ele BLAST qu is retrieving for e training for Export & Comments to file as SF e start codor y of BLAST itent map	IBI nr data han 10E es, save saved o veries rr t log file t BLAST neen_Blas ns F F F	Searches) Autor april 3 ÷ 1 ÷ hit(s) 100 ÷ hits 4 ÷ second 100 ÷ second Annotate Searches Autor Export a Sire Fir Export a Sire Fir Scale 1/2" /	ds mated Analyses am5 of features siray Holes and me map map kb	1
C All Repairs Gene Cal Save V R V R V R V C V Delay bet Send req C Auto C C All Rep Gene Cal Expo C Expo Expo Expo	inicons ing Tags m BLAST s hits with E- egardless o and the segardless o equivalence of the segardless equivalence of the segardless equivalence of the segardless equivalence of the segardless (200 €)	& Comments earch on NC values less til f their E-valu of hits to be ble BLAST qu is retrieving for e is retrieving for e is comments to file as [SF e start codor y of BLAST itent map rep	IBI nr data han 10E es, save saved o veries rr t log file t BLAST neen_Blas ns F F F	Searches) Autor april 3 ÷ 1 ÷ hit(s) 100 ÷ hits 4 ÷ second 100 ÷ second Annotate Searches Autor Export a Sire Fir Export a Sire Fir Scale 1/2" /	ds mated Analyses am5 of features array Holes and me map map	1
C Al Report Gene Cal Gene Cal Save V R V R V C C Cal Send req C Altro C Carene C Altro C Carene C Altro C Carene C Altro C Carene C Altro C Expo C Carene C Altro C Carene C C	inion ing Tags m BLAST s egardess o egardess o mit number egardess o m	& Comments earch on NC values less ti f their E-valu of hits to be lee BLAST qu e retrieving for e tertieving for c E Comments t5 file as SF e start codor y of BLAST tert map width 21	BBIn ridate han 10E es, save saved o ueries r log file log file blaST r r r r r r r r r r r r r r r r r r r	Searches) Autor april 3 ÷ 1 ÷ hit(s) 100 ÷ hits 4 ÷ second 100 ÷ second Annotate Searches Autor Export a Sire Fir Export a Sire Fir Scale 1/2" /	ds nated Analyses am5 of features äray Holes and map map tab tab tab tab tab tab tab tab tab tab	1
C AI Rep Gene Cal Gene Cal V Reformed V Reformed Save V Reformed Save V Reformed Save V Cal Feed Cal Feed Cal Feed Cal Feed Cal Feed Cal Feed V Reformed Cal Feed V Reformed Save V Reformed Save Save V Reformed Save V Reformed Save V Reformed Save V Reformed Save Save V Reformed Save Save Save Save Save Save Save Save	inion ing Tags m BLAST s egardess o egardess o mit number egardess o m	& Comments exerch on NC Values less II if their E-valued of hits to be else BLAST q estettieving for to file as [Sif] & Comments & C	BB m data han 10E es, save saved o veries r log file save Save r r log file save Save r r r r r r r r r r r r r r r r r r r	Searches Autor april 3 ÷ 1 ÷ hit(s) 100 ÷ hits 4 ÷ seco 100 ÷ seo Annotate Searches Autor Export a Sire Fire Export a Sire Fire Scale 1/2" / Sire to Z5	ds nated Analyses am5 of features äray Holes and map map tab tab tab tab tab tab tab tab tab tab	gene overlaps

Figure 4.2

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

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

SOME NOTES ON AUTO-ANNOTATE OPTIONS

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

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

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

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

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

4.3 Saving your file

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

File → Save as DNAM5 file

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

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

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

4.4 Looking at the output of your automated annotation

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

Extracted from Fas	tA Library Echild.fasta			×
Overview Features Refe	erences Sequence Documenta	tion		
Genome Organism : GenomeID : 0 Length : 0 Repticons : 1 Features : 100 CAT Table : <u>Directory : DMTemp 41564</u> Repticon	PrototypeID : 0 Date Created : 10/17/2013 Genome Date : 10/17/2013 GC Content : 0.00	Taxonomy Domain : Division : Family :	Notes Notes on the Genome Notes and Taxonomy are only available when sequences are loaded from the database	
Replicon NCBI Date : 10/17/2013 Topology : Linear	NCBI Genome : Accession :			~
Length : 53159 Features : 100 ORFs : 100 RNAs : 0 GC Content : 0.00 ORF GC : 63.4 ± 2.8 GC3 : 83.6 ± 6.4 CA: 1 : 0.000 ± 0.0000 Translation Table : Unspeci	Version : GL: Status : Current Deletion : Vulnerable Origin : 0 Terminus : 0 fied: default to standard code		Notes on the Replicon	
Image: Section 1 Image: Section 1	1 - 50000 Posi 13 1 1 26 26	ion : 1 Controls >> 1 29 21 31 1 1 43	fap IZ ⁻ Map >> Controls IA M <so im="" imsin,="" imsin,<br="" msin,="">IS3159 5</so>	

Figure 4.3

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

This map is dynamic and can be manipulated as follows:

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

4.4.1 Viewing the documentation

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

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

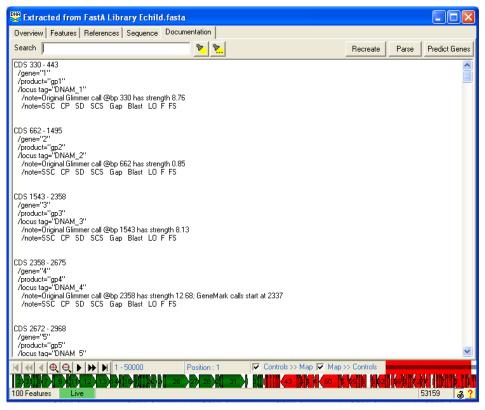


Figure 4.4

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

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

4.4.2 Viewing features in the Feature Table

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

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

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

Extracted from FastA Li	brary Echild. fasta	👺 Extracted from FastA Library Echild.fasta
Overview Features Reference	Sequence Documentation 1. Right click	Overview Features References Sequence Documentation
Sort By Index 💌 🤇	Name 5'End Length A Description Sequence Product Regions Blast Context	Sort By Index 👻 📕 Tag Name 5'End 3'End Length 🔺 Description Sequence Product Re 🚺
Select Features Direct SQL	Name 1 GenelD	Select Features Direct Set. DNAM_1 1 330 443 114 Name 1 GenelD
Type is All 💌	2 662 834 Two CDS G	Type is DNAM_2 2 662 1495 834 Type CDS S G
Name like	ORF Analysis 330 Locus Tag. DNAM 1	Name like DNAM 4 4 2358 2675 318 5 End 330 Locus Tag
GenelD =	Align 5 BLAST (local) 443 Regions 1	GenelD = DNAM_5 5 2672 2968 297 3'End 443 Regions
Locus like	6 BLAST (nr) 🕅 Tag	Locus like DNAM_6 6 3007 3477 471 Length 114 📰 Tag
Start	Choose Start	Start DNAM_7 7 3559 4524 966 Direction Forward
Length		DNAM_8 8 4639 4797 159
Regions >	3 Conv Condinate List to Cinhoard	
% GC <	10 Copy Condinate List to File	Regions > DNAM_10 10 6083 6544 462 EL Number
	12 Widen Feature List 2. Click	DNAM 12 12 7541 0220 1700 Broduct
	13 9995 1464 901	CAI 2 DNAM 13 13 9325 10788 1464 gp1
EC# like	14 10785 885	EC# like DNAM_14 14 10785 11669 885
Product like	15 11720 513 Function	Product like DNAM_15 15 11720 12232 513 Function
Function like	16 12262 957	Function like DNAM_16 16 12262 13218 957
FeatureID =	17 13288 183	FeatureID = DNAM_17 17 13288 13470 183
Tag like	18 13473 378 Notes 19 139,47 195 Original Glimmer call @bp 330 has strength 8.76	Tag like DNAM_18 18 13473 13850 378 Notes DNAM_19 19 13947 14041 195 Driginal Gimmer call @bp 330 has
Hide Ignored Features	19 13847 195 Original Glimmer call @bp 330 has strength 8.76 20 14038 369 SSC CP SD SCS Gap Blast L0 F FS	strength 8.76
	21 1406 336	SSC UP SD SUS Gap Blast LU F
Select All Features	22 14751 426	Select All Features DNAM_21 21 14405 14/41 335 FS
	23 15194 594	DNAM_23 23 15194 15787 594
	24 15899 408 🗸	DNAM_24 24 15899 16306 408
		DNAM_25 25 16420 16731 312 🗸
	Insert Delete Post Validate	Inset Delete Post Validate
	- 12500 Position : 1 🔽 Controls >> Map 🔽 Map >> Controls	H H H Controls >> Map I Map >> Controls
2 3 4		
00 Features Live	53159 🚜 ?	100 Features Live 53159 🔮

Figure 4.5

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

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

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

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

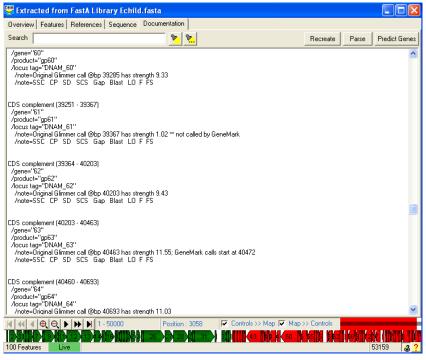


Figure 4.6

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

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

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

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

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

4.4.3 Viewing the sequence in the Sequence tab

Click on the [Sequence] tab.

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

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

Overviev	w Features Referenc	es Sequence Documentation		
Feature		💌 F 🏷 🌃 Position : 1	Raw 🕨 BLAST	N BLASTP Add Feature 100.0 3 00.0
Move	F ORF 330 (1)	est Match at 3'End Ove	rall OverallMatch	
	F ORF 662 (2)	orwards		Oligo
4 1	F ORF 1543 (3) F ORF 2358 (4)	atchstring		
	F ORF 2672 (5)	ackwards		
	F ORF 3007 (6)		FCCCGAGCCGGCGCAGGAGCGCCTCACC	
	F ORF 3559 (7)		GCAAACGGCCTCTGAGGCCGCGAGACCA	
	F ORF 4639 (8)		GCTACCCAAACTTAGATTCAAAACCAG	
	F ORF 4797 (9)		JGAGGACCGCCATGAAGATCATCCGCTC	
	F ORF 6083 (10)		JAGCCGTGGGATGAGGACGAATACCTCC	
71	F ORF 6541 (11) F ORF 7541 (12)	ACTCATGTATCTAGTTATTGAGG	GCCTAAAGGCCCGAATAAGAGCCGCACA	GGCGGCTCTCTAAGAGCGCCCA
	F ORF 9325 (13)	ACCGGCCTTGAGGGCCGGTTATC	FGACCCGGCAACCGCCGGGTCTTCTGCC	GCGCCCAGTGGCGCGGCTCATA
	F ORF 10785 (14)	GTACGGCACTCGCTCGAGTGCCT.	ACTGGGCCTCGCAGCCGGGGAAGTTCGA	CGTTCTGAACCTGCGGATGACG
53	F ORF 11720 (15)	ACGACATCCCCGATCTTCGGCCG	ICTGACTTCGTTCCGGCCAACCTCGCAG	CGTGGAACATGCCACGCCATCG
47	F ORF 12262 (16)	GGCGGCGCACTCCACTTCTTCCT	FGACGATTACCGTTTCGAGACCGTCTGG	rcgtcccccgagcgtcttctcc
41	F ORF 13288 (17)	GGCCGCTCTGACGCCCGATTTCA	GCGTGTGGCGAGAAATGCCGCGGGTAGC	GCAGGTCTGGAATACCTATCGC
	F ORF 13473 (18)	ATTGGCAATCTCAGGGGATCGAG	TAATTCCCACGGTCGGTTGGGGTCGAC	CGGACACATACGAATTCTGCTT
	F ORF 13847 (19)	GGAAATGTCGCCATTTCCTGCCT	AACCCTCCGATCGAAGCAAGAGGATCGC	GAGCTGTTCACCCGAGGCGTGC
	F ORF 14038 (20) F ORF 14406 (21)	BCAGCCGAAGACCCTTTTGGTCT	ACGGCCGGCTCAGATTCTGCGACGAAAT	AGACCTTCCCGAGGTCCGGGAG
	F ORF 14406 [21]		GATGGGCAAGCGCGGAGGGAGCGGAGG	
	F ORF 15194 (23)	TEGREGETERCARAACTCCCGGA	GCGGTGGATCTGGAGGCGGTGGAGGCG	SCGGAAAGGGCAAAGGCTCCGC
	F ORF 15899 (24)		ATCCGGCTCCGGTGGTGGCGGTAGCCAG	
	F ORF 16420 (25)	24	CAGACCAGCGTCGAGCCGACAACATCGC	
693 I	CARCECACECCCC	CGATCCGAACACCGATCGTCAGTGGAAC		
787		TGCCCAAGCCCGAGAACCTCAATGACAA		
881		CTCGACGGCTCCCGCGACAGAATCATGG		
001 975		GGAGGGCGGCGGCGGTCACATCTTCAAC		
975 069		CGTCGTGGACCAACCAGGCGTCGGAGTG		
163				
		GTACGCAACCGCACGCCAGAGGAGATCA		
257		GATCCGCCAGGTGTTCCTCGAGGGCTGG		
351		CGATCTACGAAGCGGCAGTGAAATGGGC.		
445		ACCTGTACCCGGCCGAGGCCGACTTCCA		
11		GTCTACCCGGCACGCAATGCCGATGACC		
		1 - 50000 Position : 3058	Controls >> Map 🔽 Map >> Cor	trois
2.316	7 9 111 12 13	4116 22 26 27 28 2 31		66211 66611V2VV 111V1116 11

Figure 4.7

4.4.4 Viewing ORFs in the Frames window

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

DNA → Frames

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

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

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

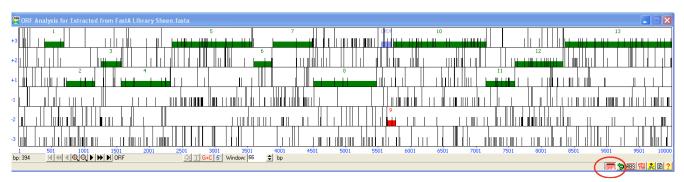


Figure 4.8

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

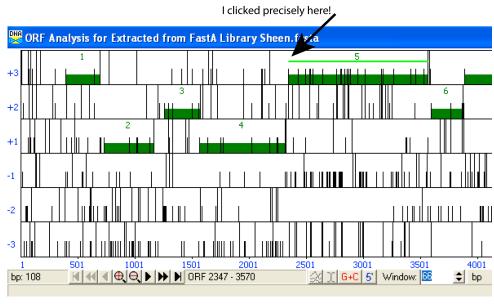


Figure 4.9

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

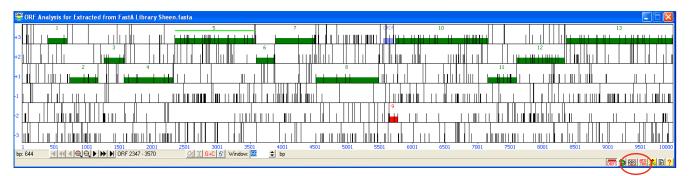


Figure 4.10

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

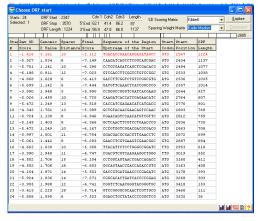


Figure 4.11

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

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

ORF Analysis for Extract	ed from FastA I	Library Sh	een.fasta																	
				5		_	7				4	10	10) 					13	
						6			11 11	ЦЩ	╧╇╝╧┿╋					12				
												և հայ հ				11		hellburk.		
2								ЦШШ	ШЦ		-11₽4	шци , ш		╷╫╫╫╷╷	Ψ.			ЩИЦИ		╨┶┶┸┷┿┹
		, , , , , , , , , , , , , , , , , , ,	du i								h			հւհերհա						
					II		<u>.</u>													╷╨─┼─┴┬┴─╵
	. III - 1		- mahar	I	ա տես	l l l i	dalah dalah	dan Li				l n de rect			սևսեե	i i II i	Ihlii			հու հահոտ՝
╵┞╶╷╝╌╝┟╶╽╢╴┧╓	╷╷╷║└──┬╴┥		_			цпт				ш	┯┯╋┯┯╋					└╥╆╌╌┾╀╌╀	╷┼╨╹╹╢╹			
2 1 1 1 1 1 1 1	llein hid				1111							Li II			шпр	r II II an dalah k	հեսեւ		Ш	
	╵╵┛╵║╴╶╿╶		1 11 11	пп							┶┼┼╫╨	┍╺╹╸╴┤┤					┉╻┅╷			┶┰╦┶╦┢╌┦╝┯┙
3 m 1 m 1 m m m		տոհենտ						Lп		, uu	الم العام ا									
1 501 1001	1501	2001	2501			501	4001	4501		DNA (Choose OR	lF start								
	H ORF 234				5 Window		bp	-1001			rts : 24	ORF Start			1 Cdn2 Cd	SD	Scoring Ma	trix Kib	ler6	
Position of internal starts										Sele	ected : 1			5'End 62.1	41.4 86	.2 87	-	nt Matrix Kar		<u>D</u> ocument
Contraction Contraction Contraction								and the second second	28172635			ORF Leng	th:1224	3'End 55.9	47.8 86	.5 1137 ^{Sp}	acing weigi	ic middline jusces	in the second second	2805
🞇 Extracted from FastA Lil										St o	a Raw SD	Genomic	Spacer	Final	Semienc	e of the Regio	n Start	Start	ORF	2005
Overview Features Reference	s Sequence Do	cumentation								#	Score	Z Value	Distanc			a of the Start	_	Position		
Sort By Index -	Tag	Name	5' End	3 End	Length 🔥	Descript	on Sequen	ce Prodi	uct Regio	1	-1.418	3.161	10	-2.112		GAACAGGAGATAGC		2347	1224	
Select Features Direct SQL	DNAM_1	1	397	693	297	Name	5	<u> </u>	GenelD	2	-5.927	1.034	8	-7.149	CAAGATC.	AGCCTCGGCAGCGA	C GTG	2434	1137	
Type is All 💌	DNAM_2	2	732	1166	435	Туре	CDS	-	GI	3	-5.701	1.141	10	-6.396		GATCAGCTCGACAC		2494	1077	
Name like	DNAM_3	3	1259	1576	318	5'End	1000	2347	Locus Ta	4	-6.188	0.911	12	-7.023		TTCGGTCTGTCCGG		2533	1038	
GenelD =	DNAM_4 DNAM_5	4	1566 2347	2318 3570	753	3'End	<u> </u>	3570	Regions	8 6	-4.668	1.628	6	-6.413		GGTCTGTCCGGCGT AAACTCACCGGCCC		2536 2557	1035	
	DNAM 6	6	3599	3877	279		1004		-	7	-3.990	1.948	5	-5,990		GGGTCGATCACGAA		2644	927	
	DNAM_7	7	3889	4512	624	Length			Tag	8	-5.026	1.459	10	-5.720		ACCATCGAGAACAT		2764	807	
Start >	DNAM_8	8	4509	5477	969	Direction		2		9	-5.472	1.249	13	-6.518	CACCATC	GAGAACATCATGAC	G ATG	2770	801	
Length >	DNAM_86		5551	5628	78		on Table Un	idefined		10	-4.343	1.781	13	-5.389		AACGAACAGTCCAA		2803	768	
Regions >	DNAM_87		5629	5704	76	EC Num	ber			11	-5.724	1.130	8	-6.946		TCCAACATGTCGTT		2812	759	
% GC <	DNAM_9 DNAM 10	9	5635 5731	5772 7155	138	Product				12 13	-5.145	1.403	8 10	-6.366		TTCGTCCTGAACCC TCGGACGACGCGAC		2836	735	
CAI	DNAM_10 DNAM 11	11	7152	7595	444	gp5				14	-5.997	1.249	11	-6.754		GCGACGTTGAACCT		2872	699	
EC# like	DNAM 12	12	7592	8332	741					15	-5.061	1.442	8	-6.283		GTGAAGTCGAACGG		2887	684	
Product like	DNAM_13	13	8359	10059	1701	Function				16	-4.663	1.630	10	-5.358	TTACATT	CTCCTGGGCCGGAT	C TTG	2953	618	
Function like	DNAM_14	14	10056	11552	1497					17	-3.990	1.948	11	-4.747	CGACGTC	GTCAAGAAGGCTGG	G TTG	3019	552	
FeatureID =	DNAM_15	15	11549	12562	1014					18	-4.502	1.706	15	-6.104		GTAACCGACCAGAC		3160	411	
Tag like	DNAM_16	16	12621	13130	510	Notes	Glimmer call (abo 2347	has streng	19	-4.502	1.706	18	-6.803		ACCGACCAGACCGT		3163	408	
Hide Ignored Features	DNAM_17	17	13160	14149	990	SSC: C	P: SD: SCS:	Gap: Bl	ast: LO: F:	20 21	-4.154	1.036	14	-5.501		ATGAACCCCGAGAT CTGATCACCCCGGA		3178	393	
Colori All Conturno	Insert Delete	Post Vali	date							22	-3.905	1.988	12	-4.741		AACGGGTACGGGTG		3418	153	
	- 50000	Position :	1	Contro	ls⇒> Map 🔽	Map XX Dr	elate			23	-3.413	2.220	18	-5.714		GCGCAACTCGTTGC		3460	111	
						Map // 44		. K.L.	AND IN SU	24	-5.808	1.090	6	-7.553	GGAGCTG	CTATACCCCGGCTC	G ATG	3535	36	
87 Features		FF761 2	1 28	12/1		TUPIL 44	PIR RIS	- RR	PIR VR											a nia niemi rodi
	4		-	22.418	Contraction (184)	Server 1		(Sec. 4)		_										

Figure 4.12

4.5 Running the BLAST function

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

Genome → BLAST All Genes

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

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

Figure 4.13

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

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

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

Saving Results Locally	
Save hits with E-Values smaller than 10E- 3	
I I	
✓ Limit number of hits saved to 100 ★ Hits	
Execute BLAST Query	
C BLAST all protein-coding genes 🔽 Skip genes already analyzed	
C BLAST all RNA-encoding genes Clear previous BLAST results	
BLAST pre-selected set of genes BLAST all genes	
Blast All Cancel	~
Sent 30 requests (90 total) to the QBlast server	~
Retrieving completed results	
Continuing to send Sent 10 requests (100 total) to the QBIast server	
All requests have now been sent to the QBlast server	_
Retrieving completed results	
Waiting for 0 results	
Genome BLAST has been completed	
	~
	?

Figure 4.14

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

Dverview	Featu	ures Refe	rences	Sequence Docu	mentation				
Sort By	ndex	-	┛	Tag	Name	5'End	3'End	Length 🔼	Description Sequence Product Regions Blast Context
Select Fea	atures	Direct SG	1L)	DNAM_1	1	330	443	114	Score Target Description
Туре	is	All	Ī	DNAM_2	2	662	1495	834	808 hypothetical protein First_005 [Mycobacterium phage First]
			<u> –</u>	DNAM_3	3	1543	2358	816	799 gp6 [Mycobacterium phage Turbido]
Name	like		_	DNAM_4	4	2358	2675	318	675 hypothetical protein [Mycobacterium hassiacum]
GenelD	=		_	DNAM_5	5	2672	2968	297	666 gp5 [Mycobacterium phage PackMan]
Locus	like			DNAM_6	6	3007	3477	471	663 gp5 [Mycobacterium phage D29]
Start	>			DNAM_7	7	3559	4524	966	
Length	>		-1	DNAM_8	8	4639	4797	159	BLAST Hit Export
-			-1	DNAM_9	9	4797	6086	1290	Accession TF_007677414
Regions	>		_	DNAM_10	10	6083	6544	462	GI 472437721 Export All Length 156 Delete
% GC	<			DNAM_11	11	6541	7518	978	Max Score 808 Date 10/18/2013 Delete All
CAI	>			DNAM_12	12	7541	9328	1788	High-Scoring Pairs (HSP)
EC#	like			DNAM_13	13	9325	10788	1464	HSP Data Alignment
Product	like		-1	DNAM_14	14	10785	11669	885	
			-1	DNAM_15	15	11720	12232	513	Bit Score 315.8 Identities 154 Score 808 %Identity 98.72
Function			_	DNAM_16	16	12262	13218	957	E-Value 0.0E0 Positives 156
FeatureID	=			DNAM_17	17	13288	13470	183	Length 156 %Similarity 100.00
Tag	like			DNAM_18	18	13473	13850	378	% Aligned 100.0 % Gaps 0
Hide Id	-	Eastures		DNAM_19	19	13847	14041	195	Query 1 - 156 Target 1 - 156
	-			DNAM_20	20	14038	14406	369	Target 1 - 156
Selec	et All F	eatures			- 1	1		×	
-				Insert Delete	Post Valid	ate			J
	€.		H 1	- 50000	Position : 2	7801 J	Control	s >> Map 🔽	Map >> Controls

Figure 4.15

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

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

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

Figure 4.16

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

4.6 Re-opening an archived (saved) file

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

File → Open → Archived DNA Master file

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

5 Gathering additional information for refining your annotation

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

5.1 Generating a six-frame translation

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

Genome → Six-frame translation

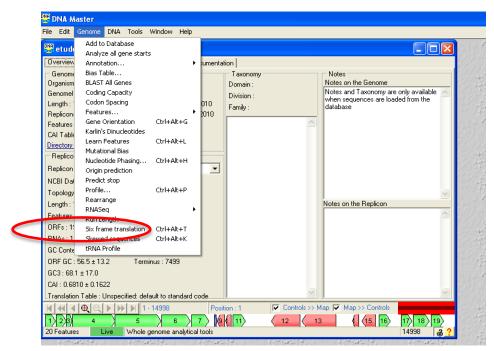


Figure 5.1

The six-frame translation window will open.

👺 Six Frame Translation of Etude_annotated	
View Map Export Map	
	PVKP PSS RQA CCCGTCAAGC : GGGCAGTTCG GTL GDL VR*A
 W W A V S 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 P G G R S R 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 D L V G G L G T V V P A R A R P G R R L P P N R R T R R N T R C R W P 116 CCTGGTGGGCGGTCCGGCAGGCGCTGGGAGGGGAGGGGA	K P R E S L G K A S GAAAGCCTCG CTTTCGGAGC R F G R S L R F A E
L G G * E Å E R L D Q S G D Å R R W Y R D F D D Å Q G R C V R V C D V W E V E K P N V W T N Q G M H Å Å G I E T L T M R K G D Å Y V Y Å T F V C D V D C D T D C C T D V W C D V T D C N D D C V D C V D	HLA TWP ORFS
	$\frac{1}{1}$

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	Frame Translation of Etude_annotated
View M	ap Export Map
1	S D T S L S G N S G K N M R G V S A P K T P G R R L N R G * R T W * G P V K P W W A V S A T L L S L E I Q A R T * G G L A P L K P L V G G * I V G R G R G K D P S S P G G R S R R H F S L W K F R Q E H E G G * R P * N P W * E A K S W V E D V V R T R Q A L V G G L G ACCGACACTTCTCTGTGGAAGTCAGGGGGGGTAGGGGGGGTGGGGGGGG
	SRCKERQFNLCSCSPP*RG*FGQYSALDHTSSTTLVR*ARTPPR GQSSRHALGLGGEGCRQTEERAGIRGADGRKPRLGG*EAERLDQ DSRPGTRSAWEAKVAAKPKNAQEYAVQHAESLGWEVEKPNVWTN TVVPARARPGRRLPPNRRTRRNTRCRWPKASVGRLRSRTFGPI
133	GGACAGTCGTCCCGGCACGCCTCGGCCTGGGAGGCGAAGGTTGCCGCCAAACCCGAAGAACGCGCAGGAATACGCGGTGCAGATGGCCCGAAAGCCTCGGTTGGGAGGTTGAGAAGCCGAACGTTTGGACCAA
14866	CCTGTCAGCAGCGCGGGCCGGACCCTCCGCTTCCAACGGCGGTTGGGCGTCTGCGGCGCCACGCTTCGGGCGCCAGCCGCCAACCCTCGGACCCTCCGAACGGCGGCGGCGGCGGCGGCGCAGCCTCCGGACCCTCCGGACCCTCCGGACCCTCCGGACCCTCCGGACCCTCCGAACGGCGGTTGGGAACCTGGTT P C D D R C A S P R P P S P Q R W V S S R A P I R P A S P R F G R N P P Q S A S R K S W S L R G P V R E A Q S A F T A A L G F F A C S Y A T C I A S L R P Q S T S F G F T Q V L P V T T G A R A R G P L R L N G G F R L V R L F V R H L H G F A E T P L N L L R V N P G
	S G D A R R W Y R D F D D A Q G R C V R V C D V H L A * W P H S Y C * R A G * R L S R G Q <mark>Q C H 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</mark> 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

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 • Landscape Mode Export RTF File
•	▶ Position : 13 Font 🖲 💌 륝 ?

Figure 5.4

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

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



Figure 5.5

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

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

5.2 Generating a provisional genome map in DNA Master

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

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

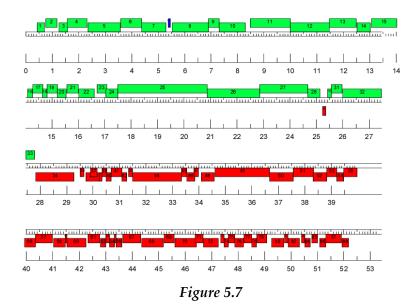
Export ORF Map		
Map Format	ORF Labels	Export Options
Scale 1/2" / kb 💌	Center Label Name : repA 🗨	 Entire Sequence
Tiers 3 🚖	Size Adjust 🔘 None	C Region Selected in Sequence Form
AutoSize 25 🜩 kb per tier	 Trim to fit Shrink to fit 	Specified Region :
		From 1 🖨 🔳
Estimated		To 53278 🔿 🗐
Dimensions	Size Adjust C None C Trim to fit	
🔽 Stagger ORFs	 Shrink to fit 	Length 53278
Divide By Direction	Text Size Default 👻	🗹 Draw ORFs 🛛 🗖 Draw Promoter
Central Marker 100 bp Ruler 💌	,	🔽 Draw RNAs 📄 Draw Oligo
Bottom Ruler 500 bp Ruler 💌		🔽 Draw in Color
		Show image when complete
Label Bottom Ruler		🕥 Draw Map
		?

DNA → Export Map

Figure 5.6

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

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



5.3 Generating a graph of coding potential using GeneMark

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

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

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

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

- <u>http://exon.gatech.edu/GeneMark/genemarks.cgi</u> . Also found on the Links page of <u>http://phagesdb.org</u> as GeneMark (version 2.5).
- Once on the site, Select '**Browse**', then find and select your sequence file. This is the same YourPhage.fasta file that you imported into DNA Master.
- From the 'Select Species' dropdown box, select 'Mycobacterium_smegmatis mc¹⁵⁵' (assuming you are annotating a mycobacteriophage genome).
- In the **'Output Format'** section, check **'LST**' under 'Output format for gene prediction and 'PDF' under output options. You can ignore the optional email option (See **Figure 5.8**).
- Click on the 'Start GeneMark' button at the middle left (See Figure 5.8).

Ger	neMark.hmm proka	aryotic						
Prokaryotic GeneMark.hmm version 1 Alexander Lukashin and Mark Borod GeneMark.hmm: new solutions for gene Nucleic Acids Research (1998) 26, pp 11	finding.							
Prokaryotic GeneMark.hmm version 2 John Besemer, Alexandre Lomsadze a GeneMarkS: a self-training method for r sequence molfs in regulatory regions. Nucleic Acids Research (2001) 29, pp 20	prediction of gene starts in micro	robial genomes. Implicatio	ns for finding					
This webpage provides access to gene p sets of pre-computed species specific alg species is available <u>here</u> . These paramet unsupervised training on each genome. <u>Browse GeneMark.hmm prokarvoti</u>	orithm parameters (model para ter sets were derived by applica	meters). The list of curren	tly supported					
Input sequence and Select species								
	Enter sequence (FASTA or multi FASTA format) or, upload file: Select species Mecobacterium_smegmatis_MC2_155							
Action Start GeneMark.hmm Reset								
Options								
Output format for gene prediction	Output options	Optional: results by E-mail						
GFF GFF GFF GFF GFF GFF Coding potential graph (not for multi FASTA) GFF GFF								
Advanced options	· · · ·							
 Switch off gene start related motif(s) 	1							
	Contact Us Home							

Figure 5.8

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

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



Figure 5.9

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

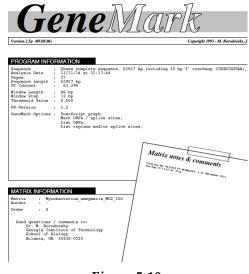


Figure 5.10

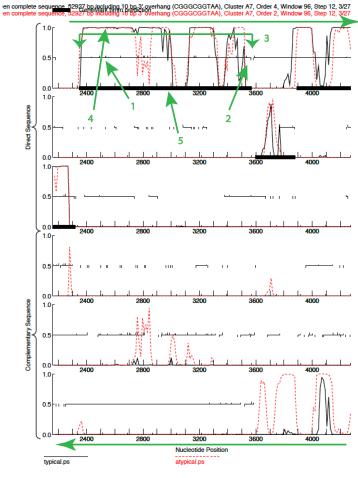


Figure 5.11

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

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

6 Phamerator & other Tools to assist with annotation

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

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

6.1 Phamerator

6.1.1 Overview

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

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

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

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

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

Our pipeline works as follows (See Figure 6.1):

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

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

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

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

See the Phamerator Help Menu for the User Manual.

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

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

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

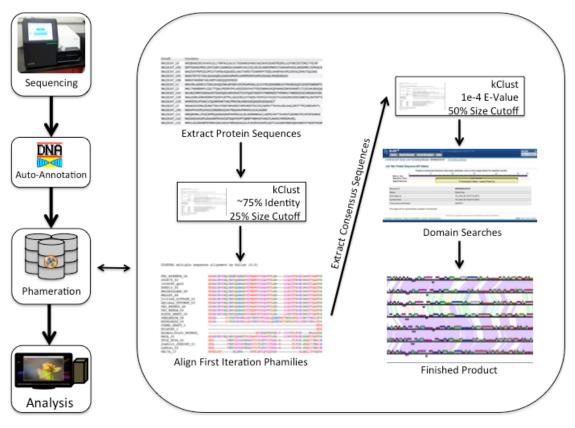


Figure 6.1

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

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

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

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

6.1.3 How did my genome get into Phamerator already?

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

6.1.4 Making Phamerator maps

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

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

Figure 6.2

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

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

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

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

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

🖴 🔺		ត				
		The second secon				
ources	Map Pham Circle Goog	gie maps				
Phages	Name	Length (bp)	GC %	Number of Genes	Cluster	
Phams	THIDBUIL-DRAFT	100327	00.793590		J	
Genes	Optimus	109270	60.790702	230	J	
Domains	Anaya	60835	66.394345	98	К1	
	CrimD	59798	66.876819	95	К1	
	JAWS	59749	66.617015	94	К1	
	BEEST_draft	59906	66.555938	91	К1	
	BarrelRoll-DRAFT	59672	66.615833	95	К1	
	Angelica	59598	66.388134	94	К1	
	Adephagia	59646	66.606311	94	K1	
	TM4		68.113719		К2	
	Pixie	61147	67.303384	100	К3	
	Fionnbharth-DRAFT	58076	68.007439	93	К4	
	Larva-DRAFT	62991	65.295042	95	К5	
	Ender-DRAFT	74731	58.796216	111	L1	
	JoeDirt	74914	58.780735	126	L1	
	LIDIE	73704	FO 703334	4.9.9		

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



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



Figure 6.4

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

6.1.5 Understanding and using the genome maps made by Phamerator

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

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

e ⊡ Genome Map File View Color	
© Q Q	
<u> </u>	
The second states in the secon	
	- 17
n)F)))
click on a gene	click on a gene

Figure 6.5

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

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

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

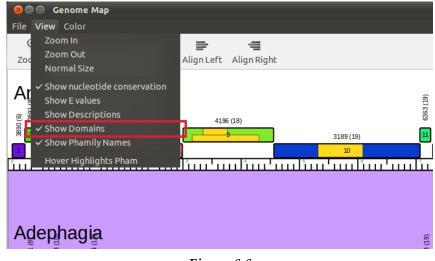


Figure 6.6

Lots of information is displayed on Phamerator maps.

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

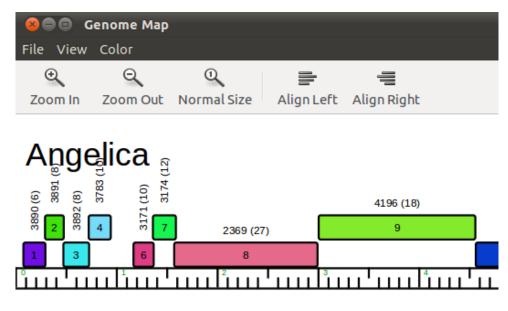
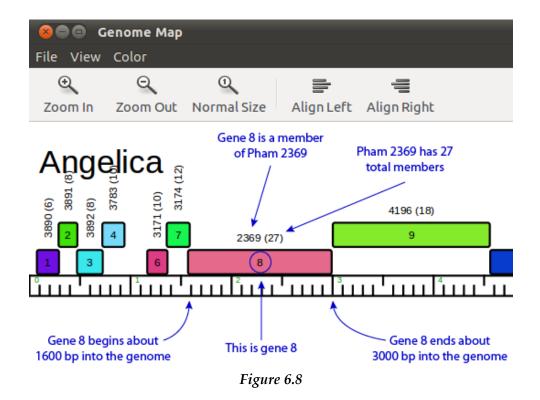


Figure 6.7

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

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

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



6.1.6 Viewing nucleotide sequence similarities in Phamerator

A NOTE ON TWO DIFFERENT TYPES OF SIMILARITY

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

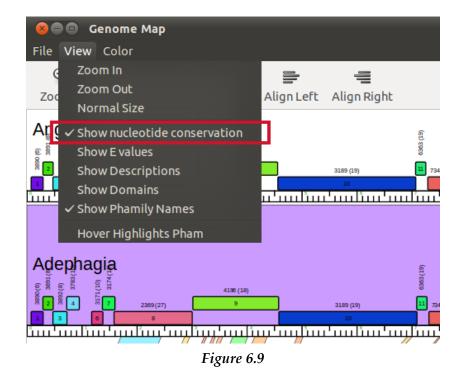
Protein similarity is a comparison of the **amino acid sequence** of two **proteins**. It is often determined by BLASTP or ClustalW. On Phamerator maps, protein similarity is shown by similarly colored gene boxes.

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

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

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

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



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

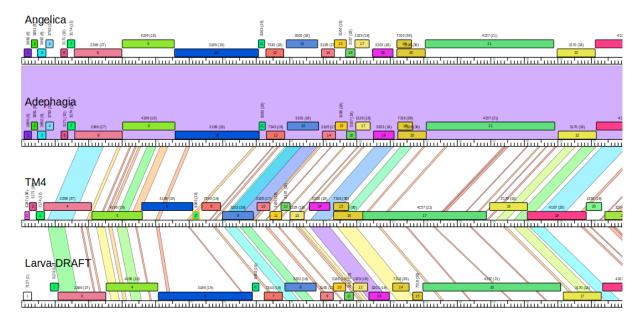


Figure 6.10

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

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

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

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

6.1.7 Other Phamerator features

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

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

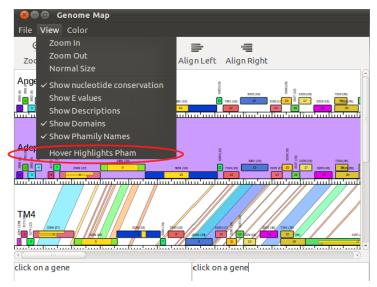


Figure 6.11

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

6.1.8 Saving Phamerator maps

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

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

6.2 Starterator

6.2.1 Overview

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

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

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

6.2.2 Interpreting Starterator results:

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

l î î	N 91 9	^o	N 9	.*
Track 1	Track 1	Track 1	Track 1	Track 1
Track 2	Track 2	Track 2	Track 2	Tyack 2
Track 3	Track 3	Track 3	Track 3	Jack 3
Track 4	Track #	Track 4	Track 4	Track 4
Track 5	Track 5	Track 5	Track 5	Track 5
Track 6	Track 6	Track 6	Track 6	Track 6
Track 7	Track	Tr <mark>a</mark> ck 7	Track 7	lýack 7
Track 8	Track B	Track 8	Track 8	Track 8
Track 9	Tragk 9	Track 9	Track 9	gack 9
Track 10	Tragk 10	Track 10	Track 10	Track 10
Track 11	Track 11	Track 11	Track 11	Cack 11
Track 12	Track 12	Track 12	Track 12	Track 12
Track 13	Track 13	Track 13	Track 13	Track 13
Track 14	Track 14	Track 14	Track 14	P Track 14
Track 15	Track 15	Track 15	Track 15	Track 15
Track 16	Track 16	Track 16	Track 16	Track 16
Track 17	Track 17	Track 17	Track 17	Track 17
Track 18	Tragk 18	Track 18	Track 18	lyack 18
Track 19	Track 19	Track 19	Track 19	lýack 19
Track 20	Track 20	Track 20	$\frac{\text{Track 20}}{12}$	Jack 20

Figure 6.12

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

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

Pham 1523 Report

 Track 1 : Ardmore_5, Taj_5, Tweety_gp5, Shauna1_5, Mutaforma13_5, Wee_gp5, SG4 5

Track 2 : Florinda 5

Track 3 : Ruby_Draft_4, MisterCuddles_Draft_4, Girr_Draft_4

- Track 4 : Brocalys Draft 6, Saal 5
- Track 5 : Cabrinians_Draft_5
 Track 6 : Spartacus_5, Hades_Draft_5
- Track 7 : Che8_5
 Track 8 : SuperGrey_Draft_5, Bipolar_Draft_5, Batiatus_Draft_5, Ovechkin_Draft_5
- Track 9 : GUmbie_5, Llij_5, Mantra_Draft_5, PMC_5, Dante_Draft_5
- Track 10 : ShiLan_5
- Track 11 : Dorothy_5, Inventum_Draft_5, Daenerys_5, Pacc40_5
 Track 12 : Bubbles123_Draft_5
- Track 13 : Llama 5
- Track 14 : DotProduct_5
 Track 15 : Empress_Draft_5
- Track 16 : SiSi 5
- Track 17 : OlympiaSaint Draft 6
- Track 18 : Hamulus 5
- Track 19 : Fruitloop_5
 Track 20 : Ibhubesi_5

Figure 6.13

Sisi gene 5 is Track 16 of the graph.

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

Suggested Starts:

SiSi_5, (6, 4435) Gene Information: Gene: SiSi_5 Start: 4435, Stop: 5028 Candidate Starts for SiSi_5: [(1, 4292), (2, 4313), (4, 4379), (6, 4436), (7, 4448), (9, 4490), (10, 4583), (11, 4739), (12, 4796), (14, 4880)]

Figure 6.14

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

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

7 Guiding Principles of Bacteriophage Genome Annotation

7.1 Overview

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

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

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

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

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

7.2 The Guiding Principles

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

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

GUIDING PRINCIPLES OF BACTERIOPHAGE GENOME ANNOTATION

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

CONTINUED...

GUIDING PRINCIPLES

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

8 Gene by gene: evaluating and improving your draft annotation

8.1 Overview

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

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

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

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

8.2 Button-pushing mechanics reserved for Section 9

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

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

8.3 Decision Tree for evaluating the draft annotation

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

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

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

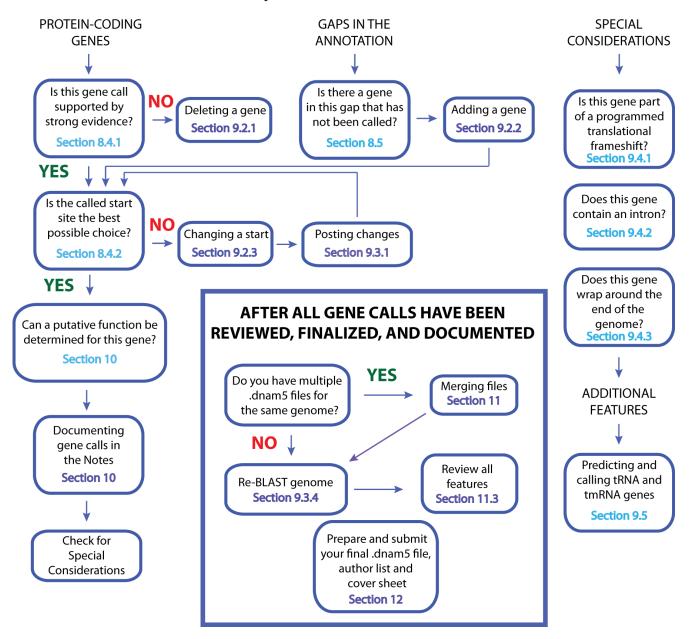


Figure 8.1

8.4 Evaluating protein-coding gene calls

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

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

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

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

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

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

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

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

🔛 Sheen_	Blas	ted								
Overview	Featu	res Reference	es	Sequence Doc	umentation					
Sort By In	ndex	- 4		Tag	Name	5' End	3' End	Length	^	Description Sequence Product Regions Blast Context
Select Fea	tures	Direct SQL		SHEEN_1	1	397	693	297	_	Name 5 GenelD
		All 🔻	L	SHEEN_2	2	732	1166	435		
	· .	AII	L	SHEEN_3	3	1259	1576	318		
Name	like		Ł	SHEEN_4	4	1566	2318	753		5'End 2347 Locus Tag SHEEN_5
GenelD	=	$\langle \rangle$	I	SHEEN_5	5	2347	3570	1224		3'End 3570 Regions 1
Locus	like [SHEEN_6	6	3599	3877	279		Length 1224
Start	, [L	SHEEN_7	7	3889	4512	624		Direction Forward
			L	SHEEN_8	8	4509	5477	969		
-	>		L	SHEEN_9	9	5551	5628	78		Translation Table Undefined
Regions	>			SHEEN_10	10	5629	5704	76		EC Number
% GC	<		L	SHEEN_11	11	5635	5772	138		
CAL	> [L	SHEEN_12	12	5731	7155	1425		Product
	like		L	SHEEN_13	13	7152	7595	444		gp5
				SHEEN_14	14	7592	8332	741		
Product	like			SHEEN_15	15	8359	10059	1701		Function
Function	like [SHEEN_16	16	10056	11552	1497		<u>^</u>
FeatureID	= 1			SHEEN_17	17	11549	12562	1014		
	like			SHEEN_18	18	12621	13130	510		Notes
-				SHEEN_19	19	13160	14149	990	6	Original Glimmer call @bp 2347 has strength 10.85
Hide Ignored Features										
Color	3 A II C			Insert Delete	Post Valid	ate				M
H H I										Map >> Controls
86 Features	[7]s)	12)114 15 Live	X	16)17)119)22	22)8]	29 30	31 3	34	36	\$```` \$````\$``````````````````````````

Figure 8.2

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

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

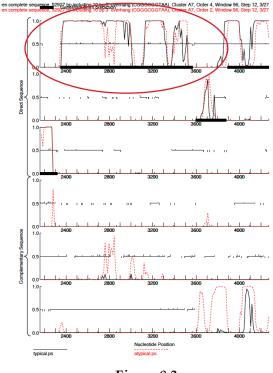


Figure 8.3

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

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

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

iort By	ndex	•	4	Tag	Name	5'End	3'End	Length	^	Description Sequence Product Begions Blast Context
elect Fea	atures	Direct SOL		SHEEN_1	1	397	693	297		Score Target Description
		Price out	- E	SHEEN_2	2	732	1166	435	0	1089 putative structural protein [Mycobacterium phage HINdeR]
,,,-		All 🔽	1	SHEEN_3	3	1259	1576	318		1066 gp5 [Mycobacterium phage Timshel]
ame	like			SHEEN_4	4	1566	2318	753	Ξ	571 gp27 [Mycobacterium phage Hinstel]
enelD	-		10	SHEEN_5	5	2347	3570	1224		568 gp27 [Mycobacterium phage Toto]
ocus	lke		- 0	SHEEN_6	6	3599	3877	279		568 gp27 [Mycobacterium phage Bask21]
tart	>		- []	SHEEN_7	7	3889	4512	624		
enoth	÷		- 🗆	SHEEN_8	8	4509	5477	969		BLAST Hit
	_		- [_	SHEEN,- High-Sc				1		Accession YP_008051857
egions	>			SHEEN, HSP Da	ata Alignm	ent		-		GI 508179911 Export A
GC	<			SHEEN, Bit Score		Identities	s 203	38		Length 406 Delete Max Score 1089 Date 11/11/2014 Delete A
AI	>			SHEEN, Score	1089	%Identity		425	\mathbf{N}	111 L 0 1 (100)
C#	lke		- _	SHEEN, E-Value		Positives		144	. 1	High-Secting Pairs (HSP) 3.
			- -	SHEEN Length	400	%Similari Gaps	ity 68.59 3	'41		High Data Alignment
	lke		- -	SHEER, O	3 - 401		3	701	/	3 LSLGTKLPRL IQWGDDKVIR KISLGSDVIW TAADPDLAQG LPDQD
unction	like			SHEEN Target	4 - 401	2.		197	1	1 ++ + + + + + ++ ++ + + 4 IILGDKPIRR ISWCSOB-IK KVSLGNSLLW AAEDPDMSOA NKTAI
eatureID	-			SHEEN_17	11	11345	16306	10TA	1	TIBODAPIAR 1000008-IN NUBBROBBN AREPIDINOR MAINI
ag	lke		- -	SHEEN_18	18	12621	13130	510	1	53 TLESTFGLSG VMFKLTGPVG KYEAAFGSTA GRALALMONF RAGSI
		d Features	Ŀ	SHEEN_19	19	13160	14149	990		51 + + ++ + +
HIDE IS	inoted	i reatures		SHEEN_20	20	14229	14390	162		53 APRIQERIPS VSERISSEAS RIARAFOILE SEPERISSER COST
Selec	t All F	eatures		SHEEN_21	21	14394	14768	375		103 GQAALQQIAA GNIEYDATID TYVSGYPNGD KITIENINTM TSGLY
			ŀ	SHEEN_22	22	14765	14920	156	Λ.	101 + + ++ + + + +++
				SHEEN_23	23	14917	15300	384	- \	103 SQAVWEQIDK GTLHLDDTVN QWVPGLPSGN LIKIPHLLQM TSGLY
				SHEEN_24	24	15303	15647	345		153 MSFMLMFVLN PMTPWSDDAT LNLVKSNGVY FQPGKGYKYA NSNY
				SHEEN_25	25	15660	16109	450		
				SHEEN_26	26	16124	16708	585		153 TSFMLNFULN PQTAWSAANT LALAKANPLA HPPGTKYAYT MSQYI
			-	SHEEN_27	27	16821	17186	366	~	203 LETVICETIS DILINDVVKK AGLMNTTWPT AGGUERGGIP TPAAN
				Insert Delete	Post Valio	late				
				50000	Position :	1 0700	T. Cantual	a sa Man	-	Map >> Controls
44 4	a (<u> </u>	J.	50000						Map >> Controls

Figure 8.4

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

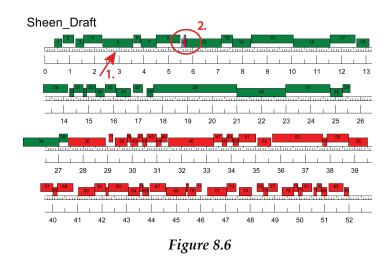
- 1. The first match to Sheen's 5th gene under the parameters of BLASTp in DNA Master is to a putative structural gene in mycobacteriophage HINdeR.
- 2. Evaluate the E-Value. The score is 0.0E0, which is not exactly identical but close. (The E-value on NCBI's BLASTp (default parameters) is 1e-122). Why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

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

verview	Featu	iles	References	Sequence D	ocumentation				
ort By	Index	-	┛	Tag	Name	5'End	3' End	Length 📥	Description Sequence Product Regions Blast Context
elect Fe	atures	Dire	ect SQL	SHEEN_1	1	397	693	297	Name 5 GenelD
уре	is [All		SHEEN_2	2	732	1166	435	
		AII		SHEEN_3	3	1259	1576	318	
lame	like			SHEEN_4	4	1566	2318	753	5'End 2347 Locus Tag SHEEN_5
ienelD	=			SHEEN_5	5	2347	3570	1224	3) nd 3570 Regions
ocus	like			SHEEN_6	6	3599	3877	279	Length 1224 📓 Tag
tart	ЪÍ		— L	SHEEN_7	1	3889	4512	624	Direction Forward
			L	SHEEN_8	8	4509	5477	969	
ength	>		L	SHEEN_9	9	5551	5628	78	Translation Table Undefined
legions	>			SHEEN_10	10	5629	5704	76	EC Number
GC	<			SHEEN_11	11	5635	5772	138	
AI			L	SHEEN_12	12	5731	7155	1425	Product
C#	like		L	SHEEN_13	13	7152	7595	444	gp5
			L	SHEEN_14	14	7592	8332	741	
roduct	like			SHEEN_15	15	8359	10059	1701	Function
unction	like			SHEEN_16	16	10056	11552	1497	
eatureID				SHEEN_17	17	11549	12562	1014	
aq	like		—- C	SHEEN_18	18	12621	13130	510	Notes
ay	ince			SHEEN_19	19	13160	14149	990	Original Glimmer call @bp 2347 has strength 10.85
Hide I	gnored	Feat	ures					~	SSC: CP: SD: SCS: Gap: Blast: LO: F: FS:
Cala			~ _	Insert Delete	e Post Valid	ate			
	⊕ ∈		H 1 -	50000	Position : 4	7418	Control	k SS Man 🔽	Map >> Controls

Figure 8.5

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



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

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

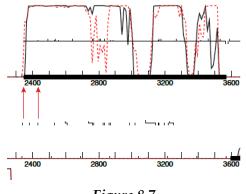
GUIDANCE: Most gene calls will pass this stage. Exceptions are genes that are called by only one program, have little or no coding potential, have very weak or no BLAST matches, are too short, and/or substantially overlap other genes.

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

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

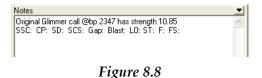
This can be a tricky, but the simplest way to answer is to address the following questions.

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





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



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

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

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

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

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



Figure 8.9

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

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

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

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

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

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

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

😤 DNA Haster	
File Tools Window Help	
🖞 ORF Analysis for Sheen_Blasted	
	9 10 12 15
	14 14 14 14 14 14 14 14 14 14 14 14 14 1
2 4 8	
	▆▝▓▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖▖
	է վերահանակեր անձառանուլի է վերկվերության առաջորդանությո
	_{┶┍┍╅╋} ╫╫╫┺╧╧╧╼╋╧╋╢╫╢╫╧╧┲╬╦╤┿┇┍┿┿┿╖╧╫╢┷╢╢╋╺╢╢╸╛╢╝╧╶╢╝╝
	😤 Choose ORF start
sha lan dalamahan a a manah lammil lala 🖉 👘 👘 👘 👘 👘 👘	Stats: 24 DRF Stat : 2347 Cdn 1 Cdn 2 Cdn 3 Lengt SD Scoring Matrix Kibler6 Selected: 10 DRF Step : 3570 5 End 0.0 33.3 100.0 5
1 417 633 1249 1665 2061 2497 2913 3329 3745 4161 4577 4993 54	Selected:10 0RF Step::3570 5 End 0.0 33.3 100.0 9 000 0RF Length:1224 3 End 57.3 49.0 87.0 789 Spacing Weight Matrix Kadin Medium
tp: 765 H 44 4 € € ▶ ▶ ▶ 0 0FF 2347 - 3570	2396
Position of internal stats	Star Raw 5D Genomic Spacer Final Sequence of the Region Start Start ORF
Dverview *688/65 References Sequence Documentation	# Score Z Value Distance Score Upstream of the Start Codon Position Length
SortBy Index - 1 Tag Name S'End 3'End Length A Description Sequence Product Regions Blast Context	1 -1.410 3.161 10 -2.112 TGACACCGAACADGAGATAGCC GTG 2347 1224
Select Features Direct SOL SHEEN_1 1 397 693 297 None 8 GenetD	2 -5.927 1.024 8 -7.149 CARGATCAGCTCGGCAGCGAC GTG 2434 1137 2 -5.701 1.141 10 -6.296 CCTGCGAGATCAGCTCGACAGG ATG 2494 1077
Type is AI V SHEEN_2 2 732 1166 435 SHEEN 3 3 1259 1576 318 Type CDS V GI	4 -6.199 0.911 12 -7.023 GTCSACCTTCSGTCTGTCCGGC GTG 2533 1039
Name ike SHEEN 4 4 1566 2319 753 5'End 4509 Locus Tag SHEEN_8	5 -4.669 1.629 6 -6.412 GACCTTCGGTCTGTCCGGCGTG ATG 2536 1035
GenelD - SHEEN.5 5 2347 3570 1224 3/End 5477 Regions 1	6 -5, 699 1.142 6 -7.444 0ATOTTCAAACTCACCOGCCC0 0TG 2557 1014
Locus Bee SHEEN_6 6 3599 3677 279 Length 569	7 -3.990 1.948 5 -5.990 CC000CC00TC0ATCAC0AA0 ATG 2644 927
Start SHEEN_7 7 3889 4512 624 Distribut Forward	8 -5.026 1.459 10 -5.720 CAAGATCACCATCGAGAACATC ATG 2764 807
▶ SHEEN_8 8 4509 5477 959	9 -5.472 1.249 13 -6.518 CACCATCGAGAACATCATGACG ATG 2770 801
Length 2 SHEEN_3 9 9551 9528 78 Translation Lable Orderhold ** Regional SHEEN_10 10 9529 9704 76 EC Number	10 -4.343 1.781 13 -5.389 OCTOTACAACGAACAGTCCAAC ATG 2803 768 11 -5.724 1.130 8 -6.946 CGAACAGTCCAACATGTCGTTC ATG 2812 789
2.GC C SHEEN_11 11 5635 5772 138	12 -5.145 1.403 8 -6.366 CCTCAACATGCCTCTGAACCCC ATG 2836 735
CAU SHEEN_12 12 5731 7155 1425 Product	13 -5.472 1.249 10 -6.167 CCCDTOTCGGACGACGACG TTG 2863 708
	14 -5.997 1.001 11 -6.754 GGACGACGCGACGTTGAACCTC GTG 2872 699
ECH Bee SHEEN 14 14 7592 8332 741	15 -5.061 1.442 8 -6.283 GAACCTCGTGAAGTCGAACGCC GTG 2887 684
Product likeSHEEN_15 15 8359 10059 1701Function	16 -4.663 1.630 10 -5.358 TTACATTCTCCTGGGCCGGATC TTG 2953 618
Function Ree SHEEN_16 16 10056 11552 1497	17 -3.990 1.948 11 -4.747 CGACGTCGTCAAGAAGGCTGGG TTG 3019 552
SHEEN 18 19 12221 12120 510 Notes	18 -4.502 1.706 15 -6.104 CCC000CAGTAACCGACCAGACC GTG 3160 411 CHOME SJ 2 VHUM 19 -4.502 1.706 18 -6.003 GDCAGTAACCGACCAGACCGTG ATG 3163 400
Tag like SHEEN 19 19 19100 14149 990 Dising Gimmer cal (Rop 4809 has strendth 5.52	20 -4.154 1.071 14 -5.501 GACCOTGATGAACCCCCGAGATC ATG 3178 393
F Hide Ignored Features SSC: DP: SD: SCS: Gap: Bladt: LD: F. FS.	21 -5.924 1.036 14 -7.271 CGGCACACTGATCACCCCGGAG ATG 3268 203 Genomic SD scores
e daas a Contras Inset Delete Post Valdate	22 -3.905 1.908 12 -4.741 COUTCAACOUTACOUTCO GTO 3418 153
	23 -2.412 2.220 10 -5.714 GTCCGGGGGGGAACTCGTTGCG ATG 2460 111
	24 -5.808 1.090 6 -7.552 GGASCTGCTATACCCCSGCTCG ATG 2525 26 💜 🚩
66 Features Live Stefane rap of starts and stops 527 2 2	🕺 🖬 🖬 🖼 🔛 🕐

Figure 8.10

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

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

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

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

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

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

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

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

Figure 8.11

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

Notes

Original Glimmer call @bp 2347 has strength 10.85 SSC: 2347 - 3570 CP:only start to capture all coding potential SD: best scores SCS: Agrees with all Gap: a gap of 30 Blast: First match is to HINdeR and Timshell with 3:4 alignment L0:YES! F: structural protein FS: phagesDB w/ 68%pos. and 50% ID to HINdeR but still requires further investigation

Figure 8.12

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

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

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

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

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

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

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

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

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

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

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

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

8.5 Checking gaps in the draft annotation for uncalled genes

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

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

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

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

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

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

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

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

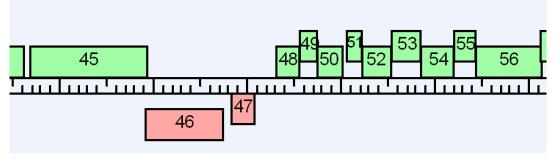


Figure 8.13

8.6 Finding and refining tRNA and tmRNA genes

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

8.7 Completing your annotation refinement

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

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

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

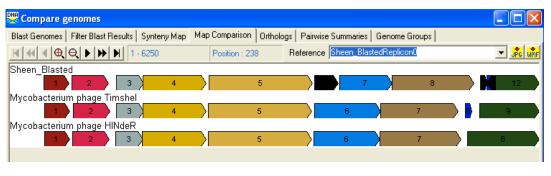


Figure 8.14

Several important steps remain.

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

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

9 The mechanics of making changes to your annotation

9.1 Overview

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

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

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

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

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

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

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

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

• Predicting tRNA and tmRNA genes

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

• Documenting your gene calls

9.2 Making common changes to your annotation

9.2.1 Deleting a gene

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

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

9.2.2 Adding a gene

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

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

Figure 9.1

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

Echild_	10.	18.13											
) verview	Featu	ires Reference	ces	Sequence Docu	imentation								
Sort By In	ndex	•	┛□	Tag	Name	5' End	3'End	Length	^	Descriptio	on Sequence Pr	duct Regions	Blast Context
Select Feat	tures	Direct SQL		DNAM_1	1	330	443	114		Name	2	GenelD	
. .	. r	All	Ĩ	DIIAM_2	2	662	1495	834			-	1	
Туре і	is	All 💌		DNAM_3	3	1543	2358	816		Туре	CDS 💽	<u>61</u>	
Name I	like			DNAM_4	4	2358	2675	318		5' End	663	2 Locus Tag	DNAM_2
GenelD	= [DNAM_5	5	2672	2968	297		3' End	149	Regions	
- 	ika İ			DNAM 6	6	3007	3477	471		مالا من ما	004 E	al -	



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

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

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

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

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

- Saving your file while still in Insert Mode.
- ² 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).

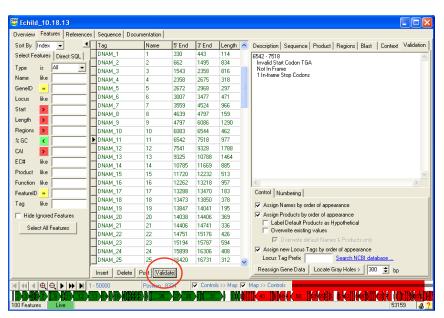


Figure 9.3

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".

Echild	_10.1 Featu			es Ì	Seguence		entation]		
Sort By	Index	•		-	Name	5'End	Length	^	Description Sequence Product Regions Blast Context Validation
Select Fe	atures	Dire	ect SQL		1	330	114	-	All ORFs appear valid
		· · ·		Г	2	662	834		
Туре		All	-	Г	3	1543	816		
Name	like			Г	4	2358	318		
GenelD	= [Г	5	2672	297		
Locus	like			Г	6	3007	471		
Start				E	7	3559	966		
				L	8	4639	159		
Length	>			L	9	4797	1290		
Regions	>				10	6083	462		
% GC	<			Þ	11	6541	978		
CAI	5			L	12	7541	1788		
EC#	like	_		L	13	9325	1464		
				L	14	10785	885		
Product	like			L	15	11720	513		×
Function	like			L	16	12262	957		<u><</u>
FeatureID	=			L	17	13288	183		Control Numbering
Tag	like	-		L	18	13473	378		Assign Names by order of appearance
-		-		L	19	13847	195		
Hide I	gnored	Feat	ures	L	20	14038	369		Assign Products by order of appearance Label Default Products as Hypothetical
Sele	ct All F	eatur	es	L	21	14406	336		Caper Derauk Producks as hypothetical Overwrite existing values
		_		L	22	14751	426		Overwrite default Names & Products only
				L	23	15194	594		Assign new Locus Tags by order of appearance
				L	24	15899	408	~	Assign new Locus Tags by order or appearance Locus Tag Prefix Search NCBI database
				<			>		
					Insert D	elete Po	st [Valid	ate	Reassign Gene Data Locate Gray Holes > 300 🖕 bp
• • •	€ €		H H	1 -	50000	F	Position : 1	_	Controls >> Map 🔽 Map >> Controls
2) 3) 00 Feature	ХġХ	Live	12 13	ł	Dellik)	2 26	27 28	3	31) MINIM 43 MAR 4 50 16 K5116 BK21 K611 K214 111116 11 53159 ⊜

Figure 9.4

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

9.3.3 Renumbering & formatting annotated features

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

To renumber your features:

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

rerview ort By	ndex	-	Reference	-	Sequence Doo Tag	Name	5'End	3' End	Length	~	Description	Sequence	Product	Begions	Blast	Context	Validation
elect Fea		_	st SQL		ECHILD 1	1	330	443	114	-	All ORFs app		Troduct	riegione	Digot	1 contone	
DISCUTO		_	a sur l		ECHILD 2	2	662	1495	834		All Unit's app	ear valiu					
/pe	is	All	-		ECHILD 3	3	1543	2358	816								
ame	like			H	ECHILD 4	4	2358	2675	318								
enelD	=			H	ECHILD 5	5	2672	2968	297								
cus	like			H	ECHILD_6	6	3007	3477	471								
art .				П	ECHILD_7	7	3559	4524	966								
	>			П	ECHILD_8	8	4639	4797	159								
ngth	>			H	ECHILD_9	9	4797	6086	1290								
gions	>				ECHILD_10	10	6083	6544	462								
GC	<			F	ECHILD_11	11	6541	7518	978								
	5				ECHILD_12	12	7541	9328	1788								
	_			П	ECHILD_13	13	9325	10788	1464								
#	like			П	ECHILD_14	14	10785	11669	885								
oduct	like				ECHILD_15	15	11720	12232	513								
nction	like			П	ECHILD_16	16	12262	13218	957		<						>
atureID	- 1				ECHILD_17	17	13288	13470	183		Control Nu	umberina					
3	like			П	ECHILD_18	18	13473	13850	378			- 1					
-					ECHILD_19	19	13847	14041	195		Assign N						
Hide Ig	gnored	d Featu	ires		ECHILD_20	20	14038	14406	369		Assign F						
Selec	t All F	eature	s		ECHILD_21	21	14406	14741	336			el Default Pro write existino		ypothetica			
- 0101					ECHILD_22	22	14751	15176	426		\sim						
					ECHILD_23	23	15194	15787	594			Overwrite del					
					ECHILD_24	24	15899	16306	408		Assign n						
					ECHILD_25	25	16420	16731	312	×	Locus T	ag Prefix	CHILD	Sarch N	CBI data	abase	
					nsert Delete	Post Valid	ate			_	Reassign (Gene Data	D cate G	ray Holes :	> 300	🗦 🕏 bp	
					ISON D'OICIC		ate					_					

Figure 9.5

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

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

9.3.4 Re-BLASTing a gene

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

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

Verview Features Reference Sort By Index -	Name	Start	Stop	~	Description Sequence Product Regions Blast Contes
	1 IName	117	254	-	a contract of the second
Select Features Direct SQL		251	481		Score Target Description
Type is All 🔻	3	478	645		1447 putative terminase gp4 [Mycobacterium phage T
Name like	4	645	914	Ξ	1405 gp8 [Mycobacterium phage JAWS]
GenelD =	- 4	911	1273		1402 gp8 [Mycobacterium phage CrimD]
	6	1273	1273		1396 gp7 [Mycobacterium phage Pixie]
Locus like	7	1985	2224		1395 gp8 [Mycobacterium phage Adephagia]
Start >	8	2221	2493		J
Length	▶ 9	2465	3910		BLAST Hit Accession NP 569740 Export
Regions >	10	3922	5508		GI 18496890 Export A
	11	5489	8248		Length 474
	12	8245	8442		Max Score 1447 Date 10/5/2011 Delete A
CAI >	13	8534	9106		High-Scoring Pairs (HSP)
EC# like	14	9165	10094		HSP Data Alignment
Product like	15	10234	10608		
Function like	16	10611	10967		1 MGLGFDRWQD DLGKLICAKR PDGLYAADMF GMS
	17	10954	11214		36 MGLGFDLUQD DLGKLICAKR DDGLYAADMF AMS
FeatureID =	18	11211	11636		
Tag like		- Prine P	11000	~	51 LCVAAVDRPL TVIWTAHRTR TAAETFKSMQ GF3 51 + +
Hide Ignored Features	<		>		86 LCINTPNT TVIWTAHRTR TAAETFRSMQ GL3
Select &II Features	Insert	Delete Po	st Valida	te	
	1 - 50000	F	Position : 1		Controls >> Map V Map >> Controls

Figure 9.6

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

Fionn	bhart	h_c	hecked_	W	P					
Overview	Featu	ues	Reference	es	Sequence	Documer	ntation			
Sort By	Index	-	•		Name	Start	Stop	^	Description Sequence Product Regions Blast	Context
Select Fe	atures	Dire	ect SQL		1	117	254		There are no BLAST results for this feature	
T	is	All		L	2	251	481		Blast this gene Blast ALL Genes Cle	ar All
Туре		AII		L	3	478	645			<u></u>
Name	like			L	4	645	914			
GenelD	-			L	5	911	1273			
Locus	like			L	6	1273	1959			
Start				L	7	1985	2224			
				L	8	2221	2493			
Length	>			Þ	9	2465	3910			
Regions	>			L	10	3922	5508			
% GC	<			L	11	5489	8248			
CAI				L	12	8245	8442			
EC#	like			L	13	8534	9106			
				L	14	9165	10094			
Product	like				15	10234	10608			
Function	like				16	10611	10967			
FeatureID					17	10954	11214			
Tag	like			L	18	11211	11636			
-				L				~		
Hide	Ignored	Feat	ures	Ľ			>			
Sele	oct All F	eatur	es		Insert De	elete Pos	t Valida	ite		
H 4 4	 €, €	ע ∢	H H	1.	50000	P	osition : 3	7354	4 🔽 Controls >> Map 🔽 Map >> Controls	
6 9	10)	11)	h /n 4)]]]h /n	X	22	23 24 2	26 28	29	8 31 8 / 136 11444 454 1 11 15 61 6 64 6 7 /	74
95 Feature	s	Live						-71	580	76 🛃 ?

Figure 9.7

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

BLAST	search for 246	5 - 39	10 (9)				
Retrieve [XML Results Text	Results	Save to Datab	ase			
Score	Description			<u>^</u>	BLAST Hit -		
1567	putative terminase g	jp4 (My	cobacterium phaj	ge TM4]		Accession N	IP_569740
1474	gp8 [Mycobacteriun	n phage	JAWS]			GI 1	8496890
1471	gp8 [Mycobacteriun	n phage	e CrimD]			Length 4	74
1466	gp8 [Mycobacteriun	n phage	e Adephagia]			Max Score 1	567
1463	gp6 (Mycobacteriun	n phage	e Larva]				
1456	gp7 [Mycobacteriun	n phage	e Pixie]				
1449	gp8 [Mycobacteriun	n phage	e Angelica]			HSP Co	verage Map
1442	gp9 [Mycobacteriun	n phage	e Anaya]			1101 00	rolago hiap
1151	terminase [Segnilipa	irus rotu	indus DSM 4498	5]			
1040	0.001			-			
E-Value 0.0E0	Starts 1	Bit So Score E-Val	9 1567	P) Query Target Length % Aligned	1 - 453 1 - 446 453 94.1	Positive Identity Similarity Gaps	340
		1	MSNSTATLTD	VARHVVAPTG	IVSTGFSAV	R ATCRHMGLGF	DRWQDDLGKI 🔨
		1	++ ++ MNHSTATLSE	+ VARHVIAPQG	+ + IVSTAWPSV	 R ATCGAMGLGF	DLWQDDLGKI
		51	ICAKRPDGLY	AADMFGMSIP	RQTGKTYLL	G AIVFALCVAA	VDRPLTVIW]
		51 51				+ +	+
		51	ICARADDGLY	AADMFAMSIP	RUIGKTYLL	G ALVFALCIKT	PN11V101
		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		
Retrieve XML Results Text Results	Save to Database	
	Ignore Definitions including the following terms	2
		~
		?

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 10³) that also gives you a useful number of matches (at least 10 or so). If you only have E-values higher than 10³, 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).

-	Features	References	Sequence				
Sort By In		╵╶┙	Name	Start	Stop	<u>^</u>	Description Sequence Product Regions Blast Context
Select Feat	tures Dir	ect SQL	1	117	254		Score Target Description
Туре і	is All		2	251	481		1567 putative terminase gp4 [Mycobacterium phage T
	like	F	3	478	645		1474 gp8 [Mycobacterium phage JAWS]
		ŀ	4	645	914	_	1471 gp8 [Mycobacterium phage CrimD]
	-	ŀ	5	911	1273		1466 gp8 [Mycobacterium phage Adephagia]
Locus	like		6 7	1273 1985	1959 2224		1463 gp6 [Mycobacterium phage Larva]
Start	>	-	8	2221	2224		1
Length	>	k	8 • 9	2465	2493 3910		BLAST Hit Export
Regions	, –	—_ Ľ	10	3922	5508		Accession NP_569740 Export All Export All
		ŀ	11	5522	8248		Length 474 Delete
%GC	<	ŀ	12	9465 8245	8442		Max Score 1567 Date 10/8/2011 Delete All
CAI	>		13	8534	9106		High-Scoring Pairs (HSP)
EC#	like 🛛	F	14	9165	10094		HSP Data Alignment
Product I	like 🗍	—- F	15	10234	10608		
Function	like		16	10611	10967		1 MSNSTATLTD VARHVVAPTG IVSTGFSAVR AT(^ 1 ++ ++ + + +
		ŀ	17	10954	11214		1 MNHSTATLSE VARHVIAPQG IVSTAWPSVR ATC
	-	ŀ	18	11211	11636		
Tag I	like					¥	51 ICAKRPDGLY AADMFGMSIP RQTGKTYLLG AIX 51
☐ Hide Ig	nored Fea	tures 【			>		51 ICAKRDDGLY AADMFAMSIP RQTGKTYLLG ALL
Select	t ∆ll Featu		Insert De	elete Po:	st Valid	ate	J
a fact a fe			50000	Р	osition : 2	3111	Controls >> Map 🔽 Map >> Controls

Figure 9.10

9.4 Making less common changes to your annotation

9.4.1 Annotating programmed translational frameshifts

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

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

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

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

Determine the precise location of the shift

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

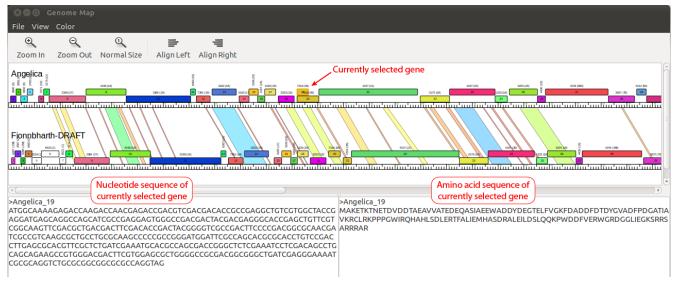


Figure 9.11

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

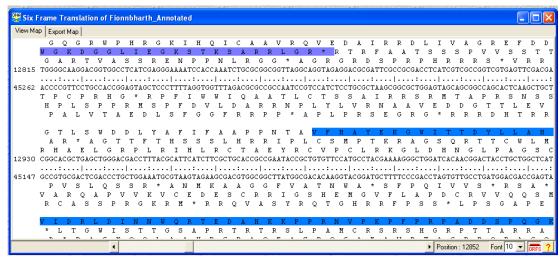


Figure 9.12

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

👺 Six Frame Translation of Fionnbharth_Annotated
View Map Export Map
G Q G R W P H R G K I H Q I C A A V R Q V E D A I R R D L I V
<mark>W G K D G G L I E G K S T K S A R R L G R *</mark> R T R F A A T S S G A R T V A S S R E N P P N L R G G * A G R G R D S P R P H R
12815 TGGGGCAAGGACGGTGGCCTCATCGA <u>GGGAAAA</u> TCCACCAAATCTGCGCGGCGGTTAGGCAGGTAGAGGACGCGATTCGCCGCGCGCCTCATCG
452.62 ACCCCGTTCCTGCCACCGGAGTAGCTCCCTTTTAGGTGGTTTAGACGCGCCGCCAATCCGTCCATCTCCTGCGCTAAGCGGCGCTGGAGTAGC
TPCPRHG*RPFIWWIQÀATLCTSSÀIRRSRM HPLSPPRMSPFDVLDARRNPLYLVRNÀAVED
PALVTAEDLSFGGFRRPP*APLPRSEGRG*R
GTLSWDDLYAFIFAAPPNTA <mark>VFHAYEKGWIT</mark>
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 R H A E L G R P L R I H L R C T A E Y R C V P C L R K G L D H
12930 CGGCACGCTGAGCTGGGACGACCTTTACGCATTCATCTTCGCTGCACCGCCGAATACCGCTGTGTTCCATGCCTACGAAAAGGGCTGGATCAC
45147 GCCGTGCGACCCGACCCTGCTGGAAATGCGTAAGTAGAAGCGACGTGGCGGCTTATGGCGACACAAGGTACGGATGCTTTTCCCGACCTAGTG
PVSLQSSR*ANMKAAGGFVATNWA*SFPQIV
V Å R Q Å P V V K V C E D E S C R R I G S H E M G V F L Å P D R C Å S S P R G K R M * R R Q V Å S Y R Q T G H R R F P S S *

Figure 9.13

Annotate the frameshift in DNA Master

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

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

Figure 9.14

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

Extrac	cted fr	om Fas	t <mark>a Li</mark> b	rary Fionn	bharth.	fasta							
Overview	Feature	es Refe	erences	Sequence	Documer	ntation							
Sort By	Index	•	•	Name	Start	Stop	^	Description Se	equence Pr	oduct Regi	ons Blas	t Cor	ntext
Select Fe	atures	Direct SI	oli [18	11211	11636		Start	Stop	Length			
Turn				19	11717	12328		* 12431	12844		1		
Туре		.0	-1	20	12431	12880				L	.i		
Name	like			21	12991	13311							
GenelD	=			22	13314	17489							
Locus	like		—L	23	17508	18623							
Start			—L	24	18623	20410	-						
			—L	25	20411	20929							
Length				26	20926	22023							
Regions	>			27	22035	22280							
% GC	<			28	22273	24582							
CAI	ЪÍ		-[29	24586	25590							
EC#			-0	30	25612	25989							
	like		_[31	26020	27630							
Product	like			32	27630	28421							
Function	like			33	28459	28857							
FeatureID			- 6	34	28854	29189							
Tag	like		-[35	29186	29446							
							~						
🔲 Hide I	gnored F	Features				>	J.	I	1	1			
Sele	ot All Fe	atures		Insert De	elete Pos	t Valid	ate	Insert	Delete	Assign	Lengths		
H + 4	 € Q	b b	N 93	76 - 15625	P	osition : 1	5098	Controls	s>> Map 🔽	Map >> Con	rols		
14	\rightarrow	15	16	17 18		9		20 21		2	2		
95 Feature:	s	Live	- 4	a							58	3076	a ?

Figure 9.15

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

🞇 Extrac	cted f	rom	FastA L	ibı	ary Fionn	bharth.1	fasta							
Overview	Featu	ues	Reference	es	Sequence	Documer	ntation							
Sort By	Index	-	[- E	4	Name	Start	Stop	^	Description 9	Sequence F	roduct	Region	s Blast	Context
Select Fe	atures	Dire	ect SQL		18	11211	11636		Start	Stop	Leng	th [~
Turn	is	All			19	11717	12328		1243		_	414		
Туре		AII	•	L	20	12431	12880		1284	4 1331	1	468		
Name	like				21	12991	13311		- ····					
GenelD	=				22	13314	17489					1		
Locus	like			L	23	17508	18623							
Start				L	24	18623	20410				~			
				L	25	20411	20929					Iculat		
Length	>			L	26	20926	22023					ength	IS	
Regions	>			L	27	22035	22280							
% GC	<			L	28	22273	24582							
CAI	>			L	29	24586	25590							
EC#	like			L	30	25612	25989							=
				L	31	26020	27630							
Product	like				32	27630	28421							
Function	like				33	28459	28857							
FeatureID) =			L	34	28854	29189							
Tag	like			L	35	29186	29446							
-				I.				×						
🔲 Hide I	Ignored	l Feat	tures	4			>		1	1				
Sele	nt All F	eatur	es		Insert De	lete Pos	t Valid	ate	Insert	Delete		Assign Le	engths	
444	⊕ [€	€ ו	H	93	76 - 15625	P	osition : 1	3717	Contro	ols >> Map 🔽	Map >	> Contro	ols 💼	
14	\rightarrow	-	15 16	Ŋ	17 18		9		20 21	X		22		
95 Feature:	s	Live	e										580	76 🛃 ?

Figure 9.16

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

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

Figure 9.17

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

Fionnbharth_checked_	WP											
Overview Features Reference	s Sequence	Docume	ntation									
Sort By Index 💌 🔳	Name	Start	Stop 🔥	1	Description	Sequence	Product	Regions	Blas	t Context		
Select Features Direct SQL	5	911	1273	1	Start	Stop	Lengt	h I				~
	6	1273	1959		124			414				
	7	1985	2224	н	128	44 1331	1	468				
Name like	8	2221	2493									
GenelD =	9	2465	3910									
Locus like	10	3922	5508									
Start >	11	5489	8248									
Length	12	8245	8442									
	13	8534	9106									
Regions >	14	9165	10094	н								
%GC <	15	10234	10608									
CAI >	16	10611	10967	н								
EC# like	17	10954	11214	н								
Product like	18	11211	11636									=
	19	11717	12328	н								
Function like	20	12431	12880	н								
FeatureID =	▶ <u>21</u>	12431	13311									
Tag like	22	13314	17489									
Hide Ignored Features	23 24	17490 18623	18623 20410									
	24	20411	20410									~
Select All Features	2	20411	20323									
	<		>									
	Insert D	elete Po	st Validate		Insert	Delet	. /	Assign Ler	ngths			
	1 - 50000	F	osition : 2502	28	Con	rols >> Map 🖪	Map>	> Controls				
		-	23 2	_	2 26 28		-			45 46 4 1 5 5 565 1	5 64) 59 71 A	74
95 Features Live	211.11 A 768	1 22	<u>A^23</u> 2	Ξ.	1/20/11 ZC	123/01 ST	P9PRI	°/11180	(7 %	**/******************	P/IP*/IUP?/U*/// 580	
55 reatures Live											000	176 🛛 😸 🙎

Figure 9.18

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

9.4.2 Annotating introns

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

9.4.3 Annotating wrap-around genes

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

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

9.5 Predicting tRNA and tmRNA genes

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

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

9.5.1 Running web-based Aragorn (version 1.2.36)

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

ARAGORN, tRNA (and tmRNA) detection

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

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



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

Type: Both (tRNA & tmRNA)

Allow introns: no

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

Strands: both

Output format: standard

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

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

Figure 9.20

The principles underlying Aragorn are described in:

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

9.5.2 Running tRNAscan-SE (version 1.23)

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

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

Improve detection of prokaryotic tRNAs to "yes"

Analyze sequences using COVE only to "yes"

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

Disable pseudogene checking to "yes"

Strict or relaxed tRNAscan-SE mode to "relaxed"

Strict or relaxed EufindtRNA mode to "relaxed"

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

Save secondary structure results file to "yes"

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

Run Reset tRNAscan-SE 1.23 ? Detection of transfer RNA genes advanced options * Sequence File (EDIT) CLEAR) paste db upload (Browse...) No file selected. \$ select >Rey Final Sequence, 83724 bp, 11 bp 3' overhang (ACCCCATGCAA), Singleton, GPCL454, 19 Primers ATCGGGCCTTCTCTCCCGGCACTTTTGGGCCCGAGACCCTTCGATTTCAA ACTCCCTGGTCGAGACGAATGAGTGCTCGAAAACAGCTGGTGGCACCTCG GTTGAGGGGTTTGCTGTATGCAAAAAACCCGCCCCTCCCGATGCAGGAAG Search Mode options Improve detection of prokaryotic tRNAs (-P) ? Yes Select archeal-specific covariance model (-A) ? No 🛟 Analyze sequences using Cove only (-C) ? Yes 🛟 Show both primary and secondary structure components to covariance model bit score (-H) ? Yes Disable pseudogene checking (-D) ? Yes Special options Strict or relaxed tRNAscan mode (-t) ? Relaxed (R) EufindtRNA mode (-e) ? Relaxed (R) Specify Alternate Cutoffs / Data Files options Cove cutoff score for reporting tRNAs (-X) ? 0 Output options Save secondary structure results file (-f) ? Yes 🖨 References :Fichant, G.A. and Burks, C. (1991) Identifying potential tRNA genes in genomic DNA sequences, J. Mol. Biol., 220, 659-671. de C. D. and Durkin, D. (1004) DNA annual anglisis units angliser medale. Nucl. Anida Dan, 22, 2020 2000

Figure 9.21

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

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

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

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

Mobyle @Pasteur

tRNAscan Search or All [mo	Welcome	Forms	Data Bookmarks	Jobs	Tutorials						
Programs	Overview			x	trnascan - 10/16/13 16:34:04 🛛 🗙						
 sequence nucleic pattern 			L6:49:24 x	3253830	5022955						
• trnascan			upda	te get l	help back to form remove job						
Data Bookmarks [overvi Sequence : Rey.fasta Jobs [overvi	parameters	Sequence File (DNA Sequence)									
<pre> trnascan - 10/16/13 16:18:53 trnascan - 10/16/13 16:34:04 trnascan - 10/16/13 16:49:24 </pre>	>Rey F. ATCGGG TTTGCT ACTCCC GTTGAG AGCGGC CGAGAT CCCTTC GACCTG TCACGG	CCTTCTCTCT STGGTACCAT GGTCGAGACO SGGTTTGCTG TTCTGAGCCCC CGGTGATCTT SCCCCCTGGC SCAGCTGGTG TCTAGACGTA	CCGGCACTTTTGGCCGAG TGGGAGGGGGCTTGGGGG TGGGAGTGCTCCGAAAACA TATGCAAAAAACCGCCCCCCCCC TTTTGAGCGGCTACTGGC CATGATCACGCCCATGGC CCATGATCACGCCCATGGC TTCGACCACGCCCCAGG STCACCCGGCATGATGTCC	CCCTTCG TGGGATG CTGGTGG CCCGATG GTCGGCG GGATAGT ATGAGTC TCCGCCC	IGAACCAA ICACCTCG ICAAGAAC ICAAGCAG ICGAAGCC ICGCTGCC ICGCTTCG						
	Improve		f prokaryotic tRNAs (- using Cove only (-C) ?		Je						
			and secondary structu checking (-D) <u>?</u> True		onents to covariance model bit score (-H) ? True						
	Special Strict or		IAscan mode (-t) ? R	laved (F	n						
			e) ? Relaxed (R)	iuxeu (r	·/						
	Specify	Alternate Cu	toffs / Data Files options								
			r reporting tRNAs (-X)	-							

Figure 9.22

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

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

Mobyle @Pasteur

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

Figure 9.23

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

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

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

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

9.5.3 tRNA secondary structure and end determination

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

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

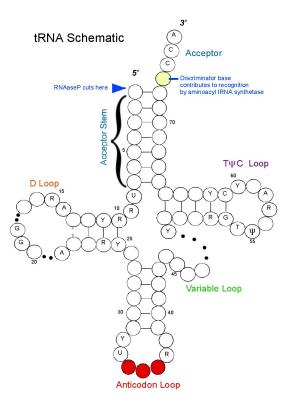


Figure 9.24

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

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

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

In Summary:

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

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

In our annotations, we will include:

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

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

9.5.4 Entering a tRNA in DNA Master

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

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

Description Sequence Product Regions Blast Context
Name 128 GenelD
Type tRNA 💌 🚹
5'End 61696 Locus Tag REY_128
3'End 61773 Regions 1
Length 78 🔳 Tag
Direction Forward 🚘
Translation Table Undefined
EC Number
Product
tRNA Thr
~
Function
~
Notes
tRNA-Thr (agt), COVE = 58.76, also called by Aragorn. Ends trimmed to August Dased Aragorn calls.

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.

	Feat	ures F	References		equence				
Sort By S	Start	-	_ <u> </u>	_	Name	5'End		Length	Description Sequence Product Regions Blast Context
Select Fea	itures	Direc	sQL	(C		55	444	390	Name 1 GenelD
Type	is	All			2	413	2089	1677	Type CDS - GI
Name	like		— <u> </u>	C		2106	3569	1464	5'End 55 Locus Tag DONOVAN 1
	_	<u> </u>			4	3560	4894	1335	
GenelD	=		ŀ		5	4896	5261	366	3'End 444 Regions
Locus	like		-	4	6	5299 6243	6243 6590	945 348	Length 390 📑 Tag
Start	>		ŀ		8	6594	6956	348	Direction Forward 🚘
Length	>	í —	— ŀ	ľ		6956	7306	353	Translation Table Undefined
Regions	>	<u> </u>	— H		10	7322	7645	324	EC Number
% GC	<	<u> </u>	—- ŀ		11	7645	7914	270	
		<u> </u>	— I		12	7914	8291	378	Product
CAI	>		ŀ	- 12	13	8288	8725	438	gp1
EC#	like		- 1		14	8819	9817	999	
Product	like				15	9933	10532	600	Function
Function	like	<u> </u>	ŀ	t	16	9933	10851	918	
FeatureID	-	<u> </u>	— ľ	t	17	10865	14632	3768	
	like	<u> </u>	— ľ	C	18	14629	16374	1746	Notes
ſag]		t	19	16374	18110	1737	Original Glimmer call @bp 55 has strength 12.44; GeneMark calls start
Hide Ig	nore	d Featu	res	C	20	18133	19026	894	151 SSC: 55-444 FWD, CP: all covered, SD: 483, highest, SCS: GL called,
Selec	t All F	eature:	s [21	19027	21288	2262	GM called shorter but I went with GL due to higher SD and BLAST
			- [22	21288	22325	1038	matches, Gap: 1st gene, Blast: Q1:S1 gp1 BigNuz, LO: longest, F: Terminase small, FS: Blast P on phagesdb.
					23	22340	22489	150	Commission and a rot progradu.
				C	24	22506	22868	363	~
				In	sert Del	ete Post	Validate	1	
[44] 4 [€l	∋I⊾I	DE NI	. 47	162	Po	sition : 1		▼ Map >> Controls

Figure 9.26

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

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

An example of good Notes:

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

10 Assigning gene functions

10.1 Overview

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

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

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

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

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

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

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

http://phagesdb.org/glossary/

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

10.2 Using bioinformatic tools to assign gene function

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

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

10.2.1 BLASTP

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

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

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

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

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

Some Pointers:

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

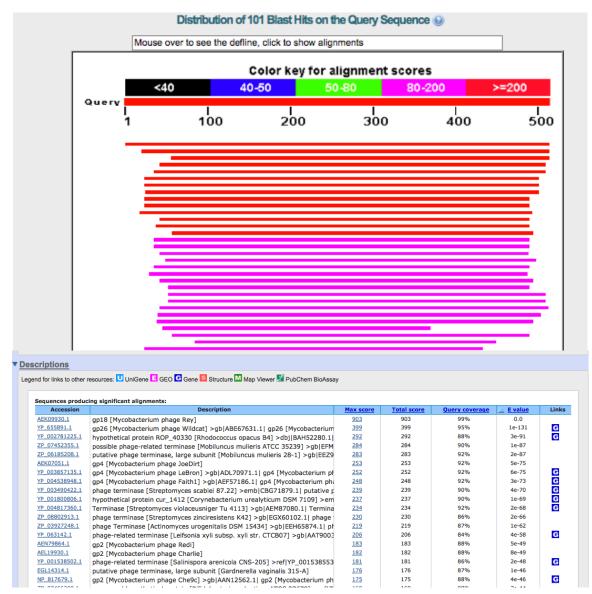
When assigning functions using BLASTP you should consider the following points.

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

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

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

Figure 10.1 is an example of a good BLASTP match, generated using NCBI's web-based BLASTP, where a putative function can be assigned. In the graphical portion of the results, there are many matches in red (the color for the highest match scores) that extend over the entire length of our query sequence. In the list of matches, we can see that all of the E values are well below 10⁴. Many of the hits have a Description that involves terminases. We can now say, with some confidence, that the protein we BLASTEd is a terminase.





BLASTing a protein on phagesdb.org:

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

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

Distribution of 107 Blast Hits on the Query Sequence

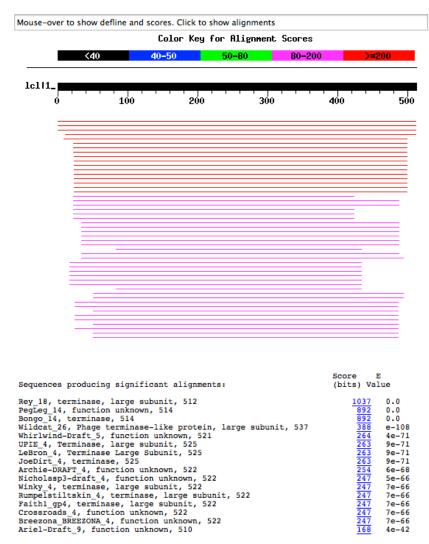


Figure 10.2

10.2.2 Conserved Domain Database (CDD)

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

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

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

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

An example of a reliable Conserved Domain hit reported by BLASTP on the NCBI server might look like: If you hover your cursor over these boxes with the mouse, a pop-up window will appear that tells you about the conserved domain. When you click on the pink Terminase_1 box in the provided example, you would detect an E value of 2e10⁴⁵.

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

Figure 10.3

The same gene in a Phamerator map might look like:

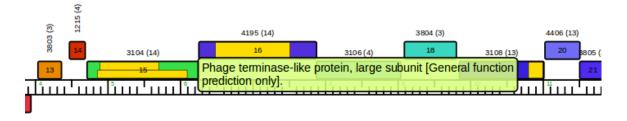


Figure 10.4

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

• <u>e</u>	Graphic Summary		
,	Show Conserved Domains		
		Putative conserved domains have been detected, click on the image below for detailed results.	
	Query seq.	1	
	Multi-domains	Phage_prot_Gp6	

Figure 10.5

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

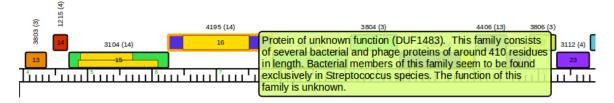


Figure 10.6

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

10.2.3 HHpred

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

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

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

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

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

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

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

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

- Copy your sequence and paste it in the appropriate window (A & B)
- Choose 3 databases available in the list (they can be simultaneously used): pdb70_[currentdate], PfamA_[currentdate],, and tgrfam_[currentdate], (C.) See <u>http://toolkit.tuebingen.mpg.de/hhpred/help_params#databases</u> for more information.
 - Pdb70 is the Protein Data Base that contains all publicly available 3D crystal structures of proteins.
 - PfamA is a database of curated alignments of genetically mobile domains found in signaling, extracellular and chromatin-associated proteins
 - TGRFAMs is a collection of curated protein families that provides a tool for identifying functionally related proteins based on sequence homology.
- Name the job. We suggest phage name_gene number OR phage name_STOP coordinate (D.)
- Click the Submit button (E.)
- Check the status of the job by its color. (F.)

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

Figure 10.7

The results can look like Figure 10.9. Note the following

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

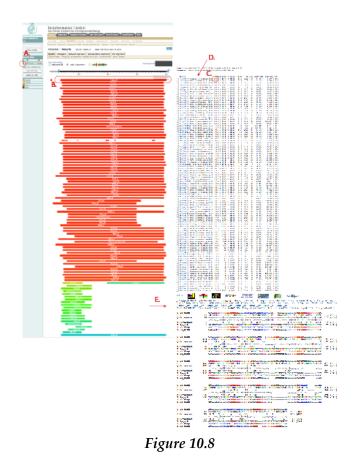


Figure 10.8: (This representation has been compressed.)

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

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

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

http://toolkit.tuebingen.mpg.de/hhpred/help_faq#correct%20match

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

Another example of an informative report from HHpred is below.



The top hit, to 2k5k_A, has a Probability score over 90, and it is an uncharacterized protein. The rest of the matches have low probabilities (80 or below), and high E values. So even though the other matches are to phosphatases, and one might be tempted to write "phosphatase", this would not be a supportable functional prediction for this protein.

10.3 Other ways to assign gene function

10.3.1 Synteny

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

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

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

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

10.3.2 Prior functional assignments

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

10.3.3 Phamerator

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

11 Merging and checking annotations

11.1 Merging overview

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

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

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

11.2 Merging multiple annotations into a single file

- Collect the files you'd like to merge into a single directory. Remember that these must all be from an identical DNA sequence (i.e., the same phage genome).
- You may want to include a newly made auto-annotated only genome. Go back to phagesDB.org to obtain the FASTA file and quickly auto-annotate. Use this as your reference file in the merged data. If the sequence has been corrupted in any of the student files, the merge will not work. This is an excellent quality control measure!
- Open DNA Master.
- Go to File → Process→Merge DNAM5 files

File	Edit Genome DNA	Tools Wir	idow Help	
	New	Ctrl+N	parating of	the star
	Open	•	1) A thirthe	1) 1 494
	Import	•	S. C. Fritzenster	S States
	Process	•	Merge DNAM5 Fil	es
	Close	Ctrl+W	Annotate Multiple	Files
	Close All		for the last	a frain
	Save as DNAM5 File	Ctrl+S	Pristant !!	the star
	Export Files	•	the state	1) 19
✓	Autoparse		10 million	
	Preferences	Ctrl+P	the start	1 Marine
	Utilities		Continue and	1 Section
	ounces	,	franke to.	A Strait
	Quit	Ctrl+Q	Pristing 1	Prest.

Figure 11.1

• A new window will open, as shown below.

Merge DNAM5 Files			
Select Files	Airmid_final.dnam5 Airmid_spotcheck_autoAnnol Arbiter_draft.dnam5 BarrelRoll_autoAnnotated.dn. BarrelRoll_final.dnam5 Benedict_spotcheck_autoAn Bongo_final.dnam5 Benedict_spotcheck_autoAn Bongo_final.dnam5 BPBiebs31_final.dnam5 Charlie_final2.dnam5 Cuco_autoAnnotated.dnam5 Cuco_autoAnnotated.dnam5 Cuco_final.dnam5 Drago_final.dnam5 Drago_final.dnam5 EIph10_final.dnam5 EIph10_final.dnam5 EIph10_final.dnam5 ERICBQueensborough_final. ERICBQueensborough_final. ERICBQueensborough_final. EUDan_final.dnam5 HelDan_final.dnam5 HelDan_final.dnam5	Add Selected Files Remove Selected Clear All Analyze Features Catenate Notes ERICBQueensborough_final.dnam5 EricB_NCCU_final.dnam5	
) 2 Files	List DNAM5 files only		?

Figure 11.2

- In the left column, browse to the directory on your computer that contains the DNA Master (.dnam5) files that you want to merge.
- In the center column, click on files that you want to add to your merged file.
- Click the 'Add Selected Files' button. The files will then appear in the empty white box on the right. You can browse to additional directories (if necessary) to add additional files.
- Once all of the files that you would like to merge added, they will be listed in the white, check the box marked "**Catenate Notes**".
- Click the 'Analyze Features' button.
- The window will open a new tab, [Merge Files].

Merge DNAM5	Files	
Select Files Merge	Files	
[Select Features] pr	esent in at least 1 🜩 files	Merge Files
Toggle Selected	Clear Selections	Export Summary
Feature	1 2	
488 - 1177		
1333 - 1926		
1090 - 1926		
1926 - 2147		
2483 - 2953		
3038 - 4000		
4000 - 4665		
4662 - 5360		
5403 - 5474		
5480 - 5552		
5595 - 5666		
5697 - 5855		
5855 - 7066		
7066 - 7500		
7524 - 9278		×
2 Files		2

Figure 11.3

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

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

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

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

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

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

Figure 11.4

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

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

CHANGE STREET		1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1	
Merged Sequence			
Overview Features Reference	s Sequence Documer	ntation	
Sort By Index 👻 🔳	Name Start	Stop 🔼	Description Sequence Product Regions Blast Context
Select Features Direct SQL	1 488	1177	Name 1 GenelD
Tune is All	2 1333	1926 🗧	
	3 1926	2147 🚽	Type CDS <u>GI</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 690 🔳 Tag
Start	7 4662	5360	Direction Forward
	8 5403	5474	
Length	9 5480	5552	Translation Table Undefined
Regions >	10 5595	5666	EC Number
%GC <	11 5697	5855	
CAI D	12 5855	7066	Product
	13 7066	7500	gp1
EC# like	14 7524	9278	
Product like	15 9275	10738	Function
Function like	16 10813	11619	
FeatureID =	17 11668	12198	×
Tag like	18 12231	13202	Notes 🗾
- ,		~	Feature identified in 2 of 2 files
Hide Ignored Features	<		Notes from EricB :
Select All Features	Insert Delete Pos	t Validate	
	1 - 50000 P	osition : 17878	3 🔽 Controls >> Map 🔽 Map >> Controls
121 N 5 8 7 1 2 1 1 4 1 2 1 1	180000000000000000000000000000000000000		MANY 43 TEENE 6K 51 TEE EE 167 ATE M62214 ATEMAN 6A AND A
112 Features Live	2716111667661 27 /291 23	2200-2200 PJ	INCLUS BARRIES IN CARDEN LINE OF DUBLIES OF
112 Features Live			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".

	Merge	d Se	auence											
Overview				s S	equence	Documer	ntation							
Sort By	Start	.	•	IN	Vame	Start	Stop	^	Descripti	on Seque	nce Prod	uct Regions	Blast	Con 4 +
Select Fe	atures	Dire	ect SQL	1		488	1177		Name	4	<u> </u>	GenelD		
		All		2	2	1090	1926			<u> </u>				
Туре		AII		2	2	1333	1926		Туре	CDS	_	<u>GI</u>		
Name	like			3	}	1926	2147		Start		2483	Locus Tag	ERICB_4	
GenelD	=			▶ 4	ļ	2483	2953		Stop		2953	Regions		1
Locus	like			5	i	3038	4000		Length	471		Tag	<u> </u>	
Start	>	, 		6		4000	4665		Direction	Forward		-	,	
				7		4662	5360				💻			
Length	>			8	}	5403	5474			on Table U	Inderined			_
Regions	>			9)	5480	5552		EC Numb	er				
% GC	<			1	0	5595	5666							
CAI	>			1	1	5697	5855		Product					
EC#	like	<u> </u>		_	2	5855	7066		gp4					\simeq
				_	3	7066	7500							\sim
Product	like			1	4	7524	9278		Function					
Function	like			1	5	9275	10738							<u>^</u>
FeaturelD) =			1	2	9296	10738							\sim
Tag	like			1	6	10813	11619		Notes					
-								~	Feature	identified in	2 of 2 files			<u>^</u>
Hide	Ignored	d Fea	tures	<			>	_	Notes fro	om EricB :				
Sele	oct All P	eatur	.ee	In	isert De	lete Pos	t Valida	ite	<u> </u>					<u> </u>
H H I			• • • • • • • •	- 62	250	Po	osition : 25	5590		Controls >> I	Map 🔽 M	ap >> Control	5	
		1	2		3	-	4		5	K	6	7	💦 📴 🚺	11) 12
112 Featur	es	Live	e		0 1								51702	2 🔒 ?

Figure 11.7

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

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

11.3 Checking an annotation

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

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

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

Go to: Genome → Profile

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

🔛 Genome Profile : Airmid	
Basics Codon Bias Dinucleotide Profile Merge Start and stop coordinates Start and stop codons and their flanking bases Start and stop codons and their flanking bases GC contents GC & AT skew Base counts Locus Tag, GeneID and GI tag Custom integer and string tags Protein Features (pl, Hydrophobicity, etc) Amino Acid Counts Window for Coiled-Coil Z8 aa AC Correlation Export Product Export Product Export Eunction	Genes All genes Selected genes All ORFs All RNAs Replicons Current Replicon All Replicons <u>DK</u> <u>Cancel</u> Load into Excel
	2

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 10²)? Do they match the Hatfull-approved data (where appropriate)? Is there a source listed for a function (HHpred, BLASTP, CDD, GFHmap, other)? If there is no known function, is "NKF" written?

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

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

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

12 Submitting final files for review and GenBank submission

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

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

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

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

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

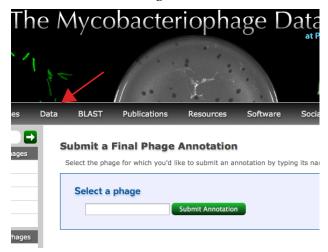


Figure 12.1

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

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

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

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

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

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

Sort By	Index	TT L	Ē	Tag	Name	5'End	3'End	Length	~	Descriptio	n Sec	uence Proc	luct Re 4
Select Fe		Direct SQL	F	SHEEN_1	1	397	693	297		Name	5		
			1E	SHEEN_2	2	732	1166	435			-		<u>GenelD</u>
Туре	is 4			SHEEN_3	3	1259	1576	318		Туре	CDS	•	<u>61</u>
Name	like			SHEEN_4	4	1566	2318	753		5' End		2347	Locus Tag
GenelD	= [Þ	SHEEN_5	5	2347	3570	1224		3' End		3570	Regions
.ocus	like			SHEEN_6	6	3599	3877	279		Length	1224		Tag
Start	ъŕ		Ľ	SHEEN_7	7	3889	4512	624		Direction			, og
				SHEEN_8	8	4509	5477	969					
.ength				SHEEN_9	9	5551	5628	78				Undefined	
Regions	>			SHEEN_10	10	5629	5704	76		EC Numb	er		
% GC	<			SHEEN_11	11	5635	5772	138					
CAL			L	SHEEN_12	12	5731	7155	1425		Product			
EC#	like [SHEEN_13	13	7152	7595	444		gp5			~
				SHEEN_14	14	7592	8332	741					~
Product	like			SHEEN_15	15	8359	10059	1701		Function			•
Function	like			SHEEN_16	16	10056	11552	1497					~
FeatureID	= [SHEEN_17	17	11549	12562	1014					~
Гад	like [SHEEN_18	18	12621	13130	510		Notes			•
-	1			SHEEN_19	19	13160	14149	990		Uriginal U strength		call @bp 2343	7 has 🛛 💆
Hide I	gnored	Features		SHEEN_20	20	14229	14390	162		SSC: 23	47 - 357	0 CP:only sta	
Sele	ct All Fe	eatures		SHEEN_21	21	14394	14768	375				g potential SE ses with all G	
				SHEEN_22	22	14765	14920	156				match is to HI	
				SHEEN_23	23	14917	15300	384	*			alignment LO	
				Insert Delete I	Post Valida	te				structural	protein	FS: phagesD	B W/

Figure 12.2

12.2 The second "minimalistic" Final .dnam5 file

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

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

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

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

Ň	Notes			
	D-ala-D-ala decarboxylase			
1	Figure 12.4			

- Recreate documentation (Section 1.4).
- Save file as "YourPhageName_Final.dnam5".

12.3 Details of your author list

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

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

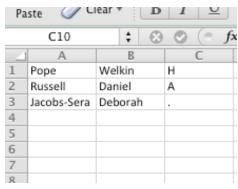


Figure 12.5

12.4 Details of your cover sheet

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

Acknowledgements

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

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

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

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

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

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

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