

What's New for School Year 2013-14 in Phage Genome Annotation

Created by DJS December, 2013.

The purpose of this document is to target the most prominent changes and/or updates to DNA Master Annotation Guide. Happy Annotating!

Change in final file Format (Section 12.1)

This year, the final file format requested has changed (again). A final .dnam5 file is one that has the following properties (Figure 1).

🔀 Echild_10.18.13						
Overview Features References	Sequence Docur	nentation				
Sort By Index 👻 🔳	Tag	Name	5' End	3'End	Length 🔼	Description Sequence Product Regions Blast Context Validation
Select Features Direct SQL	ECHILD_1	1	330	443	114	Name 9 GenelD
	ECHILD_2	2	662	1495	834	
Type is All 🔳	ECHILD_3	3	1543	2358	816	
Name like	ECHILD_4	4	2358	2675	318	5'End 4797 Locus Tag ECHILD_9
GenelD =	ECHILD_5	5	2672	2968	297	3'End 6086 Regions 1
Locus like	ECHILD_6	6	3007	3477	471	Length 1290 🔳 Tag
Start N	ECHILD_7	7	3559	4524	966	Direction Forward
	ECHILD_8	8	4639	4797	159	
Length	ECHILD_9	9	4797	6086	1290	Translation Table Bacteria and Plant Plastid Lode
Regions >	ECHILD_10	10	6083	6544	462	EC Number
% GC <	ECHILD_11	11	6541	7518	978	
	ECHILD_12	12	7541	9328	1788	Product
FC# Ba	ECHILD_13	13	9325	10788	1464	gp9
EC# IKe	ECHILD_14	14	10785	11669	885	×
Product like	ECHILD_15	15	11720	12232	513	Function
Function like	ECHILD_16	16	12262	13218	957	
FeatureID =	ECHILD_17	17	13288	13470	183	· · · · · · · · · · · · · · · · · · ·
Tag like	ECHILD_18	18	13473	13850	378	Notes
	ECHILD_19	19	13847	14041	195	Original Glimmer call @bp 4797 has strength 8.27
Hide Ignored Features	ECHILD_20	20	14038	14406	369	higher than most, SCS all agree, Gap One bp overlap (4797) Blast Best
Select All Features	ECHILD_21	21	14406	14741	336	hit is to LysA of Mycobacteriophage First, aligns 1:1. Aligns 1:1 with 10
	ECHILD_22	22	14751	15176	426	other mycobacteriophages. LU Yes. F Lys A F5 Blast and Phamerator data
	ECHILD_23	23	15194	15787	594	LysA
	ECHILD_24	24	15899	16306	408	
	ECHILD_25	25	16420	16731	312 🗸	
	Insert Delete F	Post Valida	ite			×
	- 50000	Position : 22	2810	Control	s>> Map 🔽	Map >> Controls
	AND	ADD as	Vary 28	N 60 21		III an III New Con III alkable bekatti kalkitti alkitti alkitti alkitti kalkitti
100 Features Live		7H 20	1-1 20	710 31		

Figure 1

- 1. It must be named "YourPhageName_Final.dnam5", which will help distinguish it from other versions you may have been working on.
- 2. It must contain one entry and set of notes per feature. That means that if you have merged multiple files, you need to have evaluated the data from each source, come to a decision, and deleted erroneous or repetitive versions of each feature. The notes for each feature should contain **everything** listed in **Section 9.6** about proper documentation of your gene calls. You may have to delete some notes, or even rewrite some notes from scratch to meet this criterion.

- 3. All features must be validated (Section 9.3.2).
- 4. All features must be re-numbered if necessary (Section 9.3.3).
- 5. All features must be re-BLASTed (Section 9.3.4).
- 6. Any functions are noted in the Notes fields, along with their source (**Section 9.3.3**). If it is determined that a particular gene has no function, include NKF (no known function) in the notes along with the sources for that determination.

Note: Our request for a different file format is because your final submission file is used to create a final version in Phamerator AND a GenBank file. These formats have differences. The file you submit is modified for both.

Local BLASTp

There are lots of avenues to explore to determine the functions of the genes (blastp at NCBI, Phamerator, publications, HHPred to name a few). Did you know that phagesDB has a BLASTp function? You can BLASTp any protein of your choice against a protein database derived from Phamerator. The protein data at PhagesDB contains the most curated Hatfull lab data. The newest entries tend to have the most complete data. Go to the Home Page of PhagesDB and click the BLAST tab in the top Banner and choose BLASTp (Figure 2). Happy Blasting!



Figure 2

Widen Feature Table

The default for the **Feature Table** includes **Name**, **5' End** and **Length**. **Right Click** on '**Name'** and choose "Widen Feature Table" (See Figure 3). In this view, the Feature table includes **Tag**, **Name**, **5'End**, **3'End**, and **Length**.



Choosing Start Data

The Choose start window is dramatically new and updated! The algorithms to score Shine Dalgarno sequences have been modified to include lots of choices. These choices are still under evaluation. In the meantime, you can find the "Old DNA Master" data in the SD Scoring Matrix and Spacing Weight Matrix in the upper right corner of the Choose ORF Start window. Look for more information coming soon!

DNA (hoose OR	F start				/				_	
Start Sele	s : 23 cted : 1	ORF Start : ORF Stop	6541 : 7518 5	Cdn 1 5' End 55.6	Cdn2 Cdn3 Length 38.9 83.3 54	SD Sco Spacing	ring Matr Weight	ix Kibler	6	•	Explore
_	1	Uhr Lenga	1:3/6 3	5 End 65.6	38.6 87.0 324		, nogra			<u> </u>	117317
Sta	Baw SD	Genomic	Spacer	Final	Sequence of the Rev	gion	Start	Start	OBF		
#	Score	Z Value	Distance	Score	Upstream of the St:	art	Codon	Position	Length		ī
1	-2.210	2.823	6	-3.608	CCTGGCGGAAACCAGAGG.	ATCG	ATG	6541	978		_
2	-2.323	2.769	14	-3.478	AGTCAACAGGTGGATCCG	GGTC	ATG	6595	924		
з	-5.224	1.383	9	-6.145	GTTCACAGTCCACGGCAC	CGGG	ATG	6784	735		
4	-4.041	1.948	16	-5.439	GGGTCCGGGTCTCCCCGC	CGAC	GTG	6820	699		
5	-4.291	1.829	5	-5.814	CCCCGCCGACGTGGCGCG	GGAC	GTG	6832	687		
6	-7.572	0.261	12	-8.531	CTACCCTGCCGCCGCCTT	CCCG	ATG	6889	630		
7	-3.225	2.338	7	-4.526	CGACAAGGCCATCGCGGA	GCTG	GTG	6925	594		
8	-4.463	1.746	10	-5.317	CGGTAACAACGACGAGTT	CTCG	ATG	6973	546		
	hoose OR	F start									
Star	s:23	ORF Start	6541	Cdn 1	Cdn2 Cdn3 Length	SD Sco	ring Mati	rix DBD	NA Master	-	Explore
Sele	cted:1	ORF Stop	: 7518 5	5' End 55.6	38.9 83.3 54			Kibler	5		Decument
		ORF Lengt	h:978 (3' End 65.6	38.6 87.0 924	Spacing	g weight	Matrix Kibler	6	-	Document
								Kibler Kibler	7		7354
Sta	Raw SD	Genomic	Spacer	Final	Sequence of the Re	gion	Start	Start Kible	104		^
#	Score	Z Value	Distance	Score	Upstream of the St	art	Codon	Posit Kibler	10B		
1	17	n/a	8	357	CCTGGCGGAAACCAGAGG.	ATCG	ATG	6541 Lusto	m NA Master	~	
2	28	n/a	7	588	AGTCAACAGGTGGATCCG	GGTC	ATG	6595	924		
3	13	n/a	8	273	GTTCACAGTCCACGGCAC	CGGG	ATG	6784	735		
4	15	n/a	8	315	GGGTCCGGGTCTCCCCGC	CGAC	GTG	6820	699		
5	22	n/a	9	330	CCCCGCCGACGTGGCGCG	GGAC	GTG	6832	687		
6	13	n/a	8	273	CTACCCTGCCGCCGCCTT	CCCG	ATG	6889	630		
7	25	n/a	9	375	CGACAAGGCCATCGCGGA	GCTG	GTG	6925	594		
8	27	n/a	6	351	CGGTAACAACGACGAGTT	CTCG	ATG	6973	546		

Figure 4

Event Manager

Though this isn't new or different, it is often overlooked. Go to Tools -> Event Manager. This window gives you more information when DNA Master is not working. Use the DNA Master Help Menu to evaluate the various events.

DNA Master Genome Manager and Compare Genomes functions

DNA Master allows you to load phages genomes into your own local database, and then perform several advanced bioinformatic analyses through the "compare genomes" function. The simplest one may be the 'Map comparison' of your phage genome with other genomes of your choice.

1. Load files

In order to compare files you will need to load files into Genome Manager. There are two kinds of files that you will want to load:

a. Current DNA Master files

To load a current file into your genome manager, with your .dnam5 file open, click "Genome \rightarrow Add to Database".

b. GenBank files.

To retrieve genomes already in GenBank, click "Tools→Genome Manager" (Figure 5). The genome manager has a number of tabs, the left-most one labeled "Browse". This tab allows you to view all the files in your local database. The fourth tab from the left is labeled "Retrieve". From this tab, you can search NCBI for GenBank files and import them directly into your local database or open them as .dnam5 files.

N 1	、	
👑 DNA Master	Genome Manager	
File Manager Tools Window Help	Browse Search Taxonomy Retrieve rRNA Genes F	avorites and Clipboard Tools
File Manager Tools Mitcideu. Lieb Genome Manager CrtH-6 Biss Marager CrtH-8H BLAST Library: Manager CrtH-8H-8 BLAST Library: Manager CrtH-8H-8 BLAST Library: Manager CrtH-8H-4H Genetic Code Manager Local Blast Tropector PPC Reader Protease Manager Query Manager Reference Tables Tehet Text Editor Tree Viewer Automated Analyses BLAST (local) Genome Comparison Reverse Translate Submit to GenBank	Browers Search Taxonomy Retrieve IRNA Genes F Organian Mycobacterium phage Che12 Mycobacterium phage Che8 Mycobacterium phage Che8 Mycobacterium phage Che9 Mycobacterium phage Che9 Mycobacterium phage Coeper Mycobacterium phage Coeper Mycobacterium phage Coeper Mycobacterium phage Couthouse Mycobacterium phage Couthouse Mycobacterium phage D05 Mycobacterium phage D19 Mycobacterium phage D05 Mycobacterium phage D19 Mycobacterium phage D18 Mycobacterium phage D19 Mycobacterium phage	avorites and Clipboard Tools Servorites and Clipboard Tools Servorites and Clipboard Tools Servorites and Clipboard Tools Data Directory <u>DwN00311</u> NCB Directory Genome Date 2/3/2012 Features Learned Date Created 11/22/2013 CAI Table Total Features 100 Genome Length 51547 Average %GC 61.55 GenomeD 311 Replicon 1 rRNA prototype 0 Replicon Description Replicon Breakpoints Replicon Notes Replicon Description Replicon Breakpoints Replicon Notes Replicon Complete genome RepliconID 0 Translation Table Bacteria and Plant Plastid Code ▼ Date 2/3/2012 Gi 33978272 Length 51547 Accession Version JF3970321
DOS Tools Web Tools Lawrence Lab	Complete genome Open Genome Delete Genome Delete Replicon	Features 100 Is circular DRFs 97 Is the primary replicon RNAs 3 Is a calenated replicon Is replaced by a more recent version Is protected and cannot be deleted
	255 Genomes 100 archaea Live	?

Figure 5

Using Figure 6 as a guide, follow this procedure to load Genome Manager with GenBank files.

- A. Go to Retrieve -> Fetch By accession
- B. Enter the name of the phage of interest. There are a few points about the Mycobacteriophages that you will want to keep in mind: PBI submits mycobacteriophage genomes to GenBank as Mycobacteriophage [DaVinci], when GenBank curates the file as a reference sequence, GenBank names the project Mycobacterium phage DaVinci. You made to look for both. In addition, there can be more than file for any given phage. There can be the submitted file and a the reference sequence file.. We recommend the submitted file. In addition, if the file was revised, there can be additional files.
- C. When you find the files of interest, click the "keep in Mind' button.
- D. That action places the Accession Number in the left box on the page.
- E. Change the Fetch by Accession action to Save in local database.
- F. Change windows to the Browse window.
- G. You are ready to compare your genomes of interest. You will start by placing those genomes on a Clipboard.

Α.	
Genome Manager	
Browse Search Taxonomy Retrieve NA Genes Fa	avorites and Clipboard Tools
Complete Genome BefSeg Entry Fetch By Accession P	Partial Genome
Complete denome interseq Entry Control NCDI (or	Musekashirekasa DeVinci P
Fetch by Accession Search NCB Tot	mycobacteriophage Davinci D. Snow 100
	Keep in Mind Open Found : 1
Auto-save riles Accession Length	Description
JF937092 51547	Mycobacterium phage DaVinci, complete genome
Separate Genomes	
JF937092 🔼 🔄	1.
	7
Genome Manager	
Browse Search Taxonomu Betrieve RNA Genes Fa	avorites and Clinboard Tools
Urganism 🔷	Genome Description Urganism Laxonomy Lags, Notes etc.
Mycobacterium phage Cherz	Organism Mycobacterium phage DaVinci
Mucobacterium phage Chego	Data Directory DWN00311\ NCBI Directory
Mucobacterium phage Che9d	Genome Date 2/3/2012 Eastures Learned
Mycobacterium phage Civit	Date Counted 11/02/2012 CALTable
Mycobacterium phage Colbert	Date created 11/22/2013 CALLADIE
Mycobacterium phage Cooper	Total Features 100 Genome Length 51547
Mycobacterium phage Corndog	Average %GC 61.55 GenomeID 311
Mycobacterium phage Courthouse	Replicons 1 rRNA prototype 0 🔊 🐉
Mycobacterium phage D29	rRNAs 0 User TaxonID
Mycobacterium phage DD5	
Mycobacterium phage DaVinci	Heplicon Description Heplicon Breakpoints Replicon Notes
Mycobacterium phage Doom	Replicon Complete genome
Mycobacterium phage Dreamboat	RepliconID 0 Translation Table Bacteria and Plant Plastid Code
Mycobacterium phage ET08	
	Figure 6

Note: you can add all and any file from NCBI into your Genome Manager. Once files are available you can organize them various ways. We will use the Clipboard for this task.

2. Collect your genomes for comparison

To collect your genomes for comparison, select Favorites and Clipboard from the Genome Manager menu. Once there Add the genomes of interest to the Clipboard (Figure 7). Then you may close that window.

🞇 Genome Manager			
Browse Search Taxonomy Retrieve rRNA Genes Fa	avorites and Clipboard	Tools	
All Genomes Selected		Favorite Genomes	
Mycobacterium phage Bx21 Mycobacterium phage Bx21 Mycobacterium phage Bx22 Mycobacterium phage Catera Mycobacterium phage Catera Mycobacterium phage Chel Mycobacterium phage Collett	Add All > Add All > Remove All Show GenomeID		
Mycobacterium phage Cooper Mycobacterium phage Courthouse Mycobacterium phage Courthouse Mycobacterium phage D23 Mycobacterium phage D25 Mycobacterium phage D4/inci Mycobacterium phage Deambaat Mycobacterium phage E108 Mycobacterium phage E108 Mycobacterium phage E1010 Mycobacterium phage E1010 Mycobacterium phage E10110 Mycobacterium phage Euphoria Mycobacterium phage Euphoria	Add All Add All Remove Remove All Select Unselect	Genome Clipboard : 4 Genomes E Echild, 10.18.13 Mycobacterium phage Pukovnik Mycobacterium phage Turbido Mycobacterium phage DaVinci	
Mycobacterium phage Fang		Recent Genomes	
Mycobacterium phage Fuilloop Mycobacterium phage Gumball Mycobacterium phage Gumball Mycobacterium phage Halo Mycobacterium phage Hope Mycobacterium phage Hope Mycobacterium phage JC27 Mycobacterium phage JC117 Mycobacterium phage J3per Mycobacterium phage J3per Mycobacterium phage J3per	<u>Clear</u> Reinitialize		
,	T: a series	. 7	
	Figure	27	

3. Compare Genomes

a. Preference Settings: To perform genome comparisons between genomes within your genome manager, you first must enter a release code into a field in the program preferences. Click "File→Preferences", and then click the far right tab labeled "Miscellaneous". At the bottom of this window, there is a field labeled "Release Code". Enter "Watson" into this field, and then click "Apply". Then click "OK", and close the window (Figure 8).

🞇 DNA Master Preferences	
Internet Local Settings Automation Phylogeny Timed Events Miscellaneous	<u>o</u> k
Display Show error messages when parsing	<u>R</u> evert
C Query to save CSV files upon loading into Excel	Cancel
▼ Display "Instant" oligos on Sequence Form between 9 🖕 and 61 🗼 bases	
Initial width of CDS maps on Sequence Forms 50000 bases	Export
Show CDS maps on other forms when sequence windows is less than 250000 bases	2.1
Codex Users Tables	
Percent of Codons to Use 0.750 Minimum length of genes 180	
Maximum Number of Codons 15000 Resample for correlations 1000	
Minimum Number of Codons 5000	
Edit Privileges	
C Lock Fields in Genome Manager	
Other Release Code Watson	
Show Welcome Screen On StartUp ✓ Query FastA Load of large files	
<u> </u>	?

Figure 8

- b. Go to Tools -> Genome Comparison-> Manual from the main menu.
- c. The window that appears should look like the one in Figure 9.
- d. Click Clipboard and the genomes that you selected to place on the clipboard will fill the top right field of this window.



- e. Then click Analyze. This process will take a few minutes (It is dependent on the number of genomes that you have selected. A % completion window appears at the bottom right of the window.
- f. Once complete, a new window replaces the last one with a menu as depicted in Figure 10. Choose Map comparison from this menu. Figure 10 contains 3 separate windows. The last one is a graphical gene comparison map, similar to the maps seen in Phamerator. Comparative analysis is based on W.G. Blast & Semiglobal Alignment.

🞇 Compare g	genomes										
Blast Genomes	Filter Blast Results	Synteny Map	Map Corr	pariso	n Orthologs	Pairwise	e Summaries	Genome Grou	ps		
Analysis W.G. BLAST & Semiglobal Alignment 🗾 🗖 Analyze replicons separately											
Output Calculate and filter results											
Organism		Replicons	Features	^		[Orga	nism		Sel. Rep.	Sel. Feat.	
Mycobacteriu	ım phage Elph10	1	14	5	Add 🕨	Echi	d_10.18.13		1	100	
Mycobacteriu	ım phage Rey	1	17	'5	Add All 🕨	Мусс	obacterium pł	nage Pukovnik	1	89	
Mycobacteriu	ım phage George	1	8	13		Мусс	obacterium pł	nage Turbido	1	95	_
Mycobacteriu	ım phage Solon	1	8	6	Clipboard H	 Myco 	obacterium pł	nage DaVinci	1	100	
Mycobacteriu	ım phage Tiger	1	8	6							
Mycobacteriu	ım phage LittleCherry	1	8	18							
Bruin		1	14	3							
Nala		1	14	9							
PhatBacter		1	15	i0							
Compare g	enomes										
Blast Genomes	Filter Blast Besults	Suntenu Man	Map Com	narisor	1 Orthologe	Painwise	Summaries	Genome Grou	ine l		
biasc denomes		уунсону мар	indp oom	panoo	· Oluloiogs		5 Juninaries		ihe l		0 0
<u> </u>		22333 - 11222	2333 Pos	tion : 1	N	eterence	pelect a ge	nome			JPG WITE
						*	Moh Pukov	8. I 3Hepliconu nikComplete de	nome		
Mph TurbidoComplete genome											
Mph DaVinciComplete genome											
	enomes										
	senomes Les puis public		Mar Car				a :		1		
Blast Genomes	Filter Blast Results	Synteny Map	Map Con	ipanso	ri Urthologs	Pairwise	e Summaries	Genome Grou	aps		
₩ ≪		6250	Pos	ition : !	5478 R	eference	Echild_10.1	8.13Replicon0		•	JPG UMP
Echild 10.18.	13										
	2	_			5 6		7			9	10
Mycohacteriu	m phage Pukovnik										
	1		2		4 5		6		7		8
Mycobactoriu	m nhaga Turhida		- /	7		7		1			
A A A A A A A A A A A A A A A A A A A	ni priage Turbido	3		Ν.			7			11	
Musshant	n nhana Da <i>\i</i> rri		- 1			7	(1	* * * *		
wycobacteriu	m priage Davinci	-					0				
	1	2		· / ·	5	7	6	1		* >	
				-							

Figure 10

Note: This is a quick and easy way to compare your gene calls with genomes of your choice. It is especially helpful to be able to graphically display what is in a particular genome file. Depending on how many genomes you've added, and how long they are, this process can take quite a bit of time (an hour). For just a few genomes, it will be relatively quick—several minutes. To read more about this, look in the DNA Master Help files for "Genome Comparison".



What's New and Newly Recommended in the Annotation Guide

Created by djs November, 2012. Last updated on 1.22.2013, Revised 4.19.2103 Revised 10.22.2013 and still included because it is just good advice!

Phamerator News:

The user manual for Phamerator is located in Help menu. Check it for updates!

We recommend that you use the Notes Template: (This is not new, but we are still recommending it.)

This template can be added to the preferences setting and be added to the **Notes** window of all auto-annotated gene predictions The suggested template is **SSC: CP: SD: SCS: Gap: Blast: LO: F: FS:**.

- **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 possible gene for that ORF.
- **F** (Function): Gene Function
- **FS** (Function source): source for the function (see **Section 10**). If the function assignment comes from a Hatfull-approved map in the Appendix, please also enter it into the field labeled "Function" directly above the "Notes" field. Otherwise, only enter the putative functional assignment in the Notes.
- Anything else you think is important. In particular if you made a different choice than previous annotators have made in published genomes, and feel very strongly about your choice, this is the place to let us know. **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")

Caution: Do not add hard returns in the template of the **Notes** window because it takes up too much space and is a formatting problem.

GeneMark options:

1. It can be difficult for students to understand the data imported into DNA Master has the same value as the GeneMark *M*.*Tb* or *M*. *smegmatis* graphical data. You can get the same data (in a somewhat different graphic output – portrait vs. landscape representation) at <u>http://exon.gatech.edu/heuristic_hmm2.cgi</u> where you can run the GeneMark heuristic model on the web.

2. There is an updated version of GeneMark (2.8) for bacterial models. It can be found at http://exon.gatech.edu/gmhmm2_prok.cgi. This is relatively new and untested. Preliminary data suggests it will evaluate the genomes identically to GeneMark 2.5.

3. When using the web-based GeneMark against a model organism, the model organism you pick is based on the assumption that the phage and host have a relationship. That relationship may not be helpful when calling genes. I recommend running both the *M. smegmatis* and *M.Tb* and compare. One may provide better information than the other. If you do this, please send us a note with your evaluation.

Provisional Cluster Assignments:

No cluster assignment should be made on phagesdb.org before the genome sequence has been completed and evaluated. If you assign a cluster based on other data, it will be considered a provisional assignment.

Sequence Changes:

If you retrieve your sequence from phagesdb.org, your phage sequence will be 'finished' and in the correct orientation. If you are working on additional sequences or find that the sequence does need changed, DNA Master can make those changes without losing your database information for that genome. (Which means you won't lose the work you have done to this point.) You can reverse-complement the genome or add and subtract bases. However, changes made on the screen do not become part of the binary database file UNTIL you post them to that file. Posting in the sequencing window is accomplished by clicking on the **Raw** button.