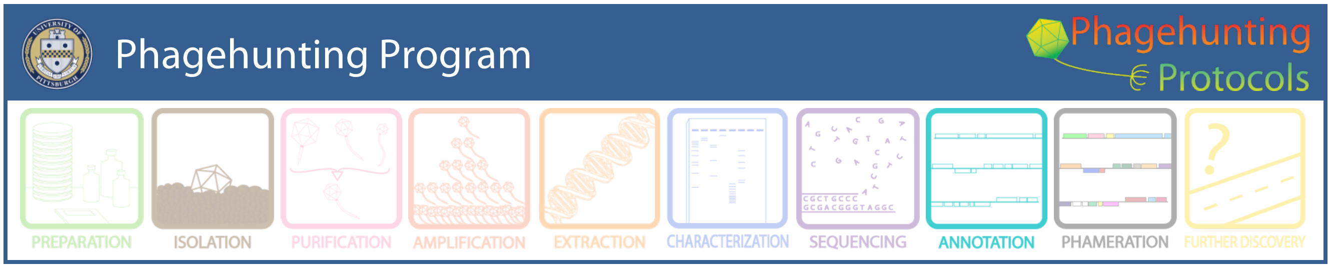


What's New for School Year 2014-15 in Phage Genome Annotation

Created by DJS December, 2014.

The What's New for 2015 includes changes in:

1. An additional file is required when submitting for QC
2. Notes required at submission
   a. An additional entry (ST)
   b. More specific explanations required, especially concerning the reporting of Functions
3. Guiding Principles Update
4. Shine-Dalgarno evaluation. This includes a new set of matrices for evaluating the relevance of the Shine-Dalgarno sequence
5. GeneMark
   a. Ease of Use
   b. Change of Output
6. DNA Master BLAST parameters
7. FASTA Output for ALL ORFs.
8. Starterator
9. Genome Announcements

General Note: Some of these changes will impact your classroom practice, so read carefully and don't hesitate to email with for further clarifcation.

1. **Additional File Required** (**Section 12**, pages 125 – 127)
   When you submit your final annotation for QC, we want you to submit two files. One of the files contains all of the info we have asked for in the past (with new additions, so read Section 12) labeled [YourPhageName]_Complete Notes.dnam5. The second file contains only reportable* functions (and notations sometimes) in the Notes section labeled [YourPhageName}_Final.dnam5.

   *Reportable function is the operative word here. You and your students will find lots of 'hints' to possible functions. Record those in your CompleteNotes file and hypothesize away as to the hows and whys a phage could use that function, but we are only going to report functions to NCBI that follow the guidelines that we have established in the annotation guide. Discernment is key!

2. **Notes required at submission** (**Section 9.6**, pages 103 -104)
   Once again, we are messing with the notes section of your annotations. Read the explanation of how to complete the notes for each gene call carefully! (p.104) Improve your functional annotation by carefully describing the supporting documentation for your assigned functions (**See Section 10**).

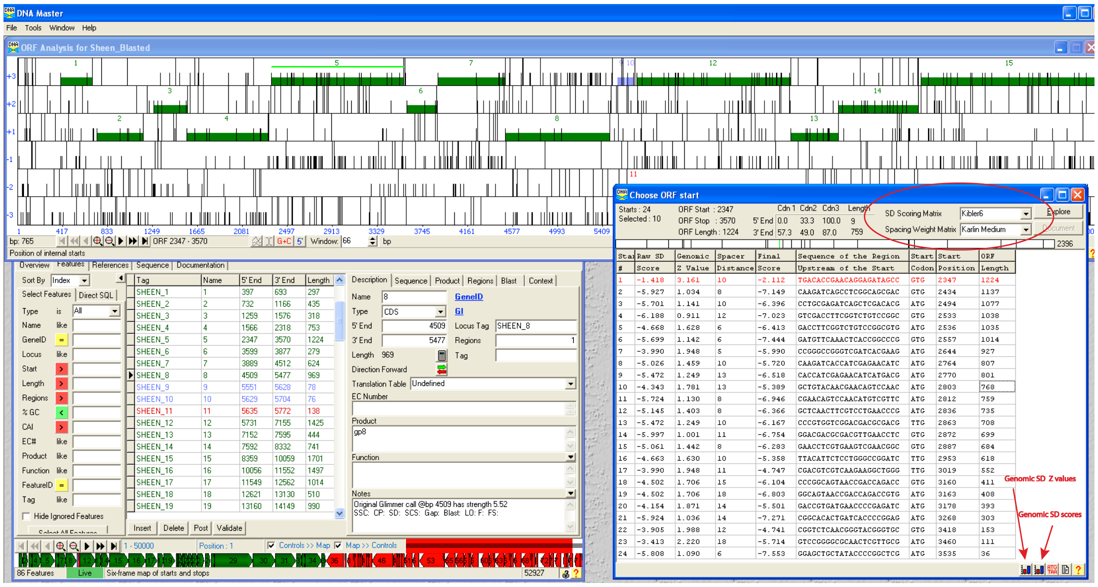
   We have a new program for you to use this year – Starterator. See #**8** of this What's New for more information! Just remember to include your Starterator data in the ST section of your notes.

3. **Guiding Principles Update (Section 7,** p. 64-65).
   The Guiding Principles are always worth reading each year, but this year we have revised them a bit. The main concept that we want to reinforce is that the evaluation of the Shine-Dalgarno score takes a back seat to everything else on the list. One of the considerations Welkin and I would like to showcase is #**10a**. This one states that the ribosome likes it when the genes have a 4 bp overlap. That overlap allows the ribosome to hop to the next gene with the least possible effort. If you have the option to select a start with a 4 bp overlap, we recommend that you do!

4. **How to Use the New Shine-Dalgarno Matrices (Section 8.4.2,** p. 74-75)
   This is the probably the change that will cause you to pause this year. We introduced the newly developed Shine Dalgarno scoring matrices last year, but this year we will insist that you use them. It is a bit daunting as you begin to use this, but relax, you can do it!

   Here is the skinny. In the Choose ORF start window, set the SD Scoring Matrix to Kibler 6 and the Spacing Weight Matrix to Karlin Medium. The new algorithms are based on the assumption that if you evaluated all of the starts across the genome – that includes the "real" starts, the additional (other) starts in an ORF with coding potential, and EVERY other possible start (for the frames with no coding potential) – you would get a normal distribution of 'scores'. The "real" starts would actually be found outside that normal distribution of (ridiculous) starts because they would be remarkably different than the random start sites included in this data set.

   There are 2 scores to evaluate: Raw SD score and Final score. These scores are based on the math as described by D. Kibler & S. Hampson (see references below). The Raw score compares the upstream sequence of a start to the SD sequence of E. coli – AGGAGGA. The Final score uses that score plus the spacing (which is 7 – 10 bp upstream of the start). In either case the best score is the largest score (which is the **least negative number).** The second number to help you evaluate the score is the Z score. You are looking for the Raw/Final score to be greater than 2 standard deviations from the mean.


DNA Master
File Tools Window Help
ORF Analysis for Sheen_Blasted
+3
+2
+1
-1
-2
-3
1 417 833 1249 1665 2081 2497 2913 3329 3745 4161 4577 4993 5409
Overview Features References Sequence Documentation
Select Features Direct SQL
Description Sequence Product Regions Blast Context
Insert Delete Post Validate
Select All Features
Controls >> Map
Map >> Controls
Explore
Document

bp: 765 ORF 2347 - 3570 Window: 66 bp

Position of internal starts

Sort By: Index

Type is All
Name like
GeneID =
Locus like
Start >
Length >
Regions >
% GC <
CAI >
EC# like
Product like
Function like
FeatureID =
Tag like

☐ Hide Ignored Features

Tag	Name	5' End	3' End	Length
SHEEN_1	1	397	693	297
SHEEN_2	2	732	1166	435
SHEEN_3	3	1259	1576	318
SHEEN_4	4	1566	2318	753
SHEEN_5	5	2347	3570	1224
SHEEN_6	6	3599	3877	279
SHEEN_7	7	3889	4512	624
SHEEN_8	8	4509	5477	969
SHEEN_9	9	5551	5628	78
SHEEN_10	10	5629	5704	76
SHEEN_11	11	5635	5772	138
SHEEN_12	12	5731	7155	1425
SHEEN_13	13	7152	7595	444
SHEEN_14	14	7592	8332	741
SHEEN_15	15	8359	10059	1701
SHEEN_16	16	10056	11552	1497
SHEEN_17	17	11549	12562	1014
SHEEN_18	18	12621	13130	510
SHEEN_19	19	13160	14149	990

Name: 8
Type: CDS
5' End: 4509
3' End: 5477
Length: 969
Direction Forward
Translation Table: Undefined
Locus Tag: SHEEN_8
Regions: 1
Tag:
EC Number:
Product: gp8
Function:
Notes: Original Glimmer call @bp 4509 has strength 5.52
SSC: CP: SD: SCS: Gap: Blast: LO: F: FS:

1 - 50000 Position : 1

86 Features Live Six-frame map of starts and stops 52927

Choose ORF start

Starts : 24
Selected : 10

ORF Start : 2347
ORF Stop : 3570
ORF Length : 1224

	Cdn 1	Cdn2	Cdn3	Length
5' End	0.0	33.3	100.0	9
3' End	57.3	49.0	87.0	759

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Karlin Medium

2396

Start #	Raw SD Score	Genomic Z Value	Spacer Distance	Final Score	Sequence of the Region Upstream of the Start	Start Codon	Start Position	ORF Length
1	-1.418	3.161	10	-2.112	TGACACCGAACAGGAGATAGCC	GTG	2347	1224
2	-5.927	1.034	8	-7.149	CAAGATCAGCCTCGGCAGCGAC	GTG	2434	1137
3	-5.701	1.141	10	-6.396	CCTGCGAGATCAGCTCGACACG	ATG	2494	1077
4	-6.188	0.911	12	-7.023	GTCGACCTTCGGTCTGTCCGGC	GTG	2533	1038
5	-4.668	1.628	6	-6.413	GACCTTCGGTCTGTCCGGCGTG	ATG	2536	1035
6	-5.699	1.142	6	-7.444	GATGTTCAAACTCACCGGCCCG	GTG	2557	1014
7	-3.990	1.948	5	-5.990	CCGGGCCGGGTCGATCACGAAG	ATG	2644	927
8	-5.026	1.459	10	-5.720	CAAGATCACCATCGAGAACATC	ATG	2764	807
9	-5.472	1.249	13	-6.518	CACCATCGAGAACATCATGACG	ATG	2770	801
10	-4.343	1.781	13	-5.389	GCTGTACAACGAACAGTCCAAC	ATG	2803	768
11	-5.724	1.130	8	-6.946	CGAACAGTCCAACATGTCGTTC	ATG	2812	759
12	-5.145	1.403	8	-6.366	GCTCAACTTCGTCCTGAACCCG	ATG	2836	735
13	-5.472	1.249	10	-6.167	CCCGTGGTCGGACGACGCGACG	TTG	2863	708
14	-5.997	1.001	11	-6.754	GGACGACGCGACGTTGAACCTC	GTG	2872	699
15	-5.061	1.442	8	-6.283	GAACCTCGTGAAGTCGAACGGC	GTG	2887	684
16	-4.663	1.630	10	-5.358	TTACATTCTCCTGGGCCGGATC	TTG	2953	618
17	-3.990	1.948	11	-4.747	CGACGTCGTCAAGAAGGCTGGG	TTG	3019	552
18	-4.502	1.706	15	-6.104	CCCGGCAGTAACCGACCAGACC	GTG	3160	411
19	-4.502	1.706	18	-6.803	GGCAGTAACCGACCAGACCGTG	ATG	3163	408
20	-4.154	1.871	14	-5.501	GACCGTGATGAACCCCGAGATC	ATG	3178	393
21	-5.924	1.036	14	-7.271	CGGCACACTGATCACCCCGGAG	ATG	3268	303
22	-3.905	1.988	12	-4.741	CGGTCTCAACGGGTACGGGTGC	GTG	3418	153
23	-3.413	2.220	18	-5.714	GTCCGGGGCGCAACTCGTTGCG	ATG	3460	111
24	-5.808	1.090	6	-7.553	GGAGCTGCTATACCCCGGCTCG	ATG	3535	36

Genomic SD Z values

Genomic SD scores

References for the Scoring Matrices:
http://www.ics.uci.edu/~kibler/pubs/Metmbs02.pdf
http://www.ics.uci.edu/~kibler/pubs/TR03.pdf

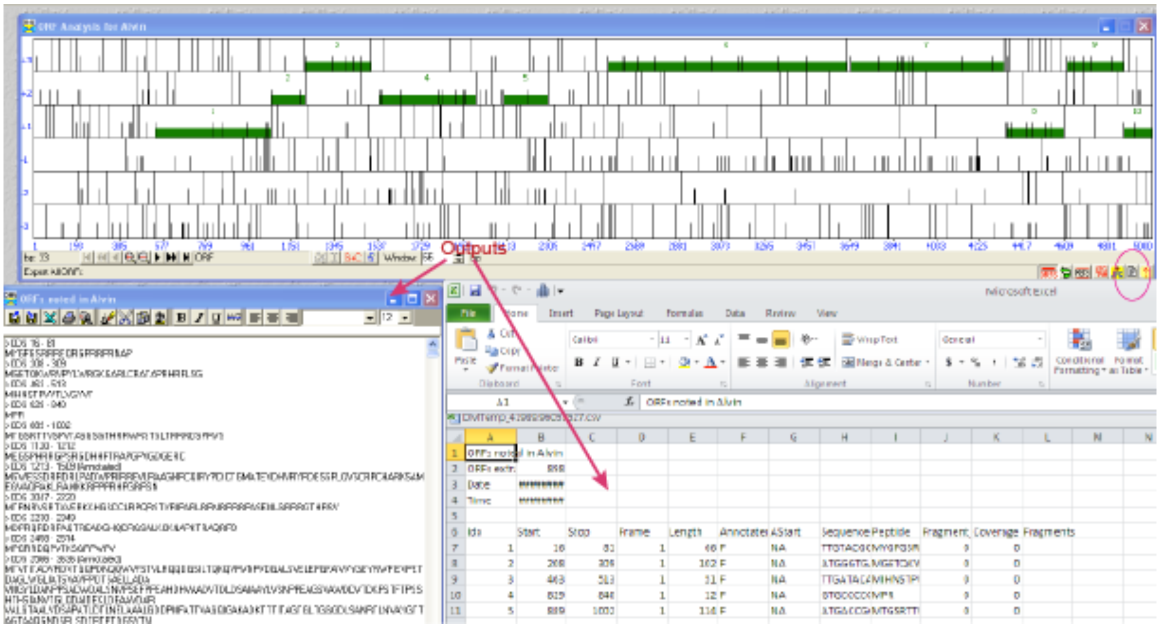
5. **GeneMark changes** (**Section 5.3**, pages 45 – 48)
   a. The good news: Dan has added three buttons to your sequenced phage page. Each button will automatically generate the pdf required for 3 GeneMark predictions:
      i. GeneMark heuristic – this is the graph of coding potential that matches the output from GeneMark provided to DNA Master. In this case, the program is trained on the sequence of the phage that you submit to it.
      ii. GeneMark smeg – this is the graph of coding potential when GeneMark is trained on the model, *M. smegmatis* mc^{2}155.
      iii. GeneMark TB – this is the graph of coding potential when GeneMark is trained on the model, *M. tuberculosis* HR37Rv.
   b. The not so good news: if you are like me, you really liked the landscape output of GeneMark. It seems that the only output that is available now is available in a portrait orientation. The really bad news is that GeneMark's servers were hacked this summer and all GeneMark programs were unavailable for a number of weeks this summer. Their latest and best algorithms are in place now, but they did not resurrect my favorite version (2.5).

6. **DNA Master BLAST parameters**

As long as we are on the subject of security issues on the web, NCBI changed some of their BLAST parameters too. Because of this Dr. Lawrence changed the rates of BLAST queries and their retrievals from NCBI. Please refer to **Section 4.5** (p. 37) for how to best BLAST an entire genome. The changes are required to meet specifications included in NCBI's revisions to Blast batch queries. The new settings have slowed down the queries and their retrievals. If you exceed the expected rates of queries, your requests will not go through.

Blasting is best accomplished in Off-Peak hours (9PM – 5AM). The information you get back from an NCBI BLAST is different than a PhagesDB BLAST or Phamerator search. Do not skip one for the other. When we BLASTEd at the In Silico workshop, only 1 of 40 worked the first time (late Monday evening)!

7. **FASTA Output for ALL ORFs** (Not found in the guide…. Yet!)
   If you have the opportunity to do some Mass Spec, you may want to run your data against all possible ORFs from your genome. You can use the small right hand button located in the Frames window. One button push will generate 2 outputs: 1) an Excel sheet listing all ORFs and 2) a list of FASTA sequences of all ORFs.



```
>CDS 16 - 81
[illegible]
>CDS 208 - 309
[illegible]
>CDS 463 - 513
[illegible]
>CDS 826 - 840
MPR
>CDS 889 - 1002
[illegible]
>CDS 1130 - 1212
[illegible]
>CDS 1213 - 1509 (Annotated)
[illegible]
>CDS 3017 - 2220
[illegible]
>CDS 2230 - 2249
[illegible]
>CDS 2490 - 2514
[illegible]
>CDS 2386 - 3535 (Annotated)
[illegible]
```

DMTemp_4[illegible]27.csv

	A	B	C	D	E	F	G	H	I	J	K	L
1	ORFs noted in Alvin											
2	ORFs extr.	898										
3	Date	########										
4	Time	########										
5												
6	Ids	Start	Stop	Frame	Length	Annotated	AStart	Sequence	Peptide	Fragment	Coverage	Fragments
7	1	16	81	1	66	F	NA	TTGTACGC	MYGFGSR	0	0	
8	2	208	309	1	102	F	NA	ATGGGTG	MGETGKV	0	0	
9	3	463	513	1	51	F	NA	TTGATACA	MIHNSTP	0	0	
10	4	819	846	1	28	F	NA	GTGCCCC	MPR	0	0	
11	5	889	1002	1	114	F	NA	ATGACCG	MTGSRTT	0	0	

8. **Starterator (Section 6.2**, p. 59 - 61)
   This is a new program written by a recent Pitt graduate, Marissa Pacey. This program allows you to compare the possible starts of all genes of the same pham. Input can include a phamerated gene, an unphamerated gene, a pham, or the whole phage genome. As you can imagine, the program will take a longer time to process the data for a whole phage genome than it will for a single gene. Starterator installation instructions are included with the other software installation instructions and on phagesDB. Use the help function of the program and the information included in **Section 6.2** (p. 59) to interpret results.

9. **Genome Announcements** are 500 word peer-reviewed articles that can be written for one or 100 genomes. We want to encourage you and your students to write a Genome Announcement for the phage(s) that you annotate this year! We have heard from the folks at Genome A and they are on board with submissions of any size. One of the benefits to you submitting your school's Genome Announcement is that you can more easily include

your students as authors. Please know that the QC of the annotation is still a required part of the process. Also know that we will be happy to review your announcement before submission if you want our help (and we would love to know that you are submitting it!).

Included are the “What’s New” from the previous years, because they contain good recommendations! Enjoy your bioinformatics semester!

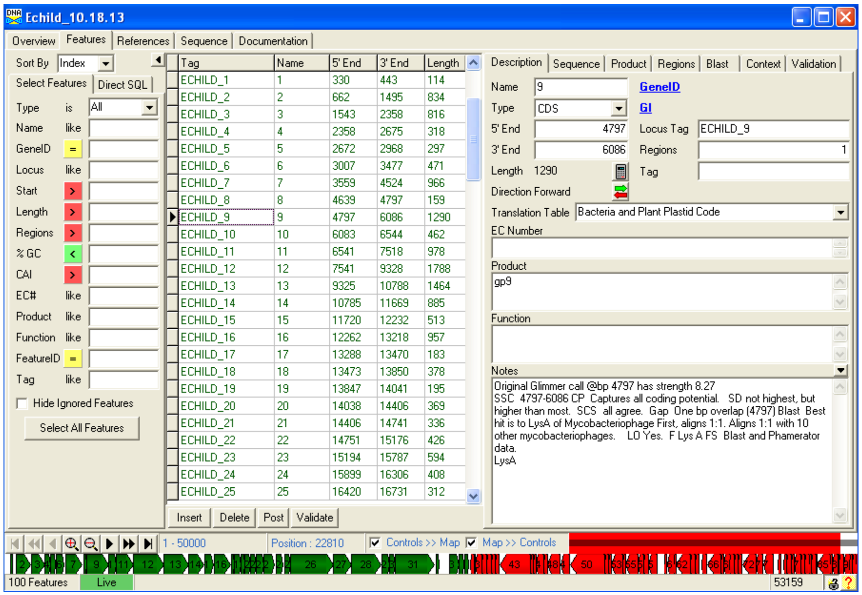
What's New for School Year 2013-14 in Phage Genome Annotation

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).



Sort By: Index

Type is All
Name like
GeneID =
Locus like
Start >
Length >
Regions >
% GC <
CAI >
EC# like
Product like
Function like
FeatureID =
Tag like

Tag	Name	5' End	3' End	Length
ECHILD_1	1	330	443	114
ECHILD_2	2	662	1495	834
ECHILD_3	3	1543	2358	816
ECHILD_4	4	2358	2675	318
ECHILD_5	5	2672	2968	297
ECHILD_6	6	3007	3477	471
ECHILD_7	7	3559	4524	966
ECHILD_8	8	4639	4797	159
ECHILD_9	9	4797	6086	1290
ECHILD_10	10	6083	6544	462
ECHILD_11	11	6541	7518	978
ECHILD_12	12	7541	9328	1788
ECHILD_13	13	9325	10788	1464
ECHILD_14	14	10785	11669	885
ECHILD_15	15	11720	12232	513
ECHILD_16	16	12262	13218	957
ECHILD_17	17	13288	13470	183
ECHILD_18	18	13473	13850	378
ECHILD_19	19	13847	14041	195
ECHILD_20	20	14038	14406	369
ECHILD_21	21	14406	14741	336
ECHILD_22	22	14751	15176	426
ECHILD_23	23	15194	15787	594
ECHILD_24	24	15899	16306	408
ECHILD_25	25	16420	16731	312

Name: 9
Type: CDS
5' End: 4797
3' End: 6086
Length 1290
Direction Forward
Locus Tag: ECHILD_9
Regions: 1
Tag:
Translation Table: Bacteria and Plant Plastid Code
EC Number:
Product: gp9
Function:
Notes:
Original Glimmer call @bp 4797 has strength 8.27
SSC 4797-6086 CP Captures all coding potential. SD not highest, but higher than most. SCS all agree. Gap One bp overlap (4797) Blast Best hit is to LysA of Mycobacteriophage First, aligns 1:1. Aligns 1:1 with 10 other mycobacteriophages. LO Yes. F Lys A FS Blast and Phamerator data.
LysA

1 - 50000 Position : 22810
100 Features Live 53159

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!

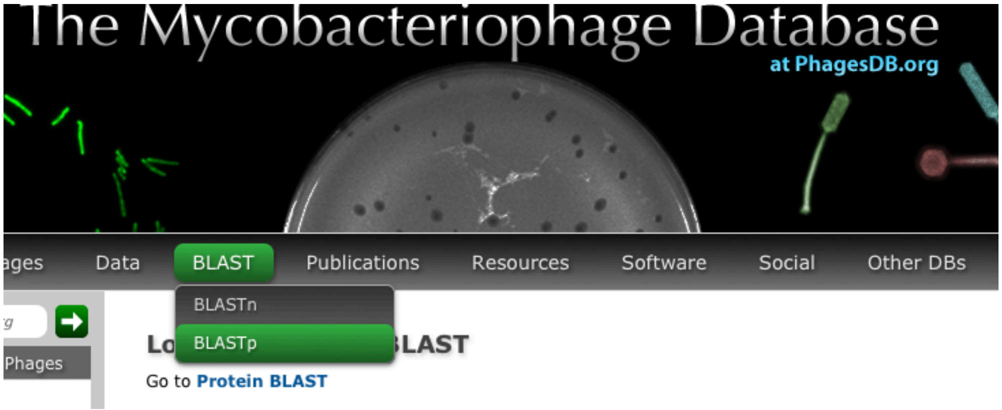
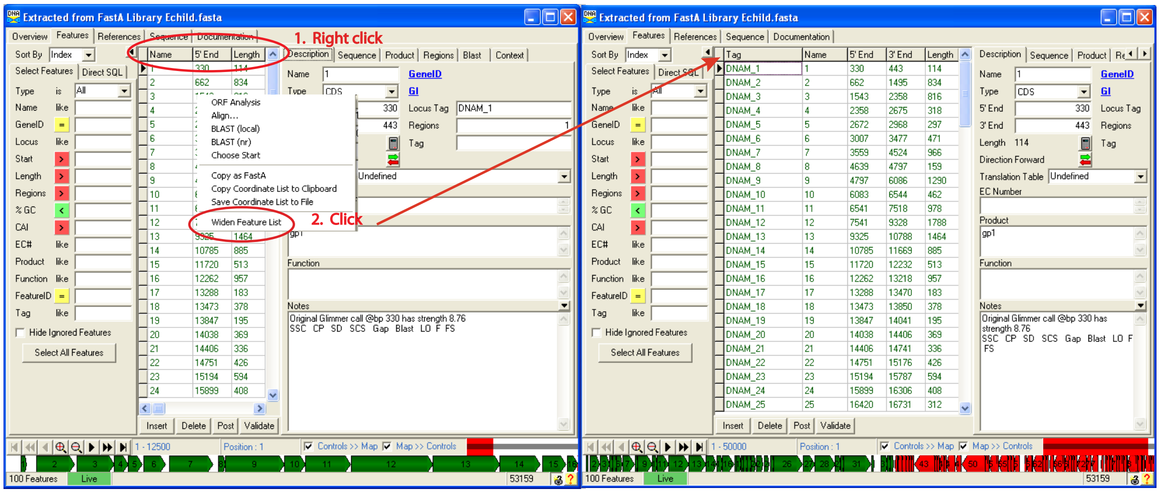


Figure 2

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**.



Sort By: Index

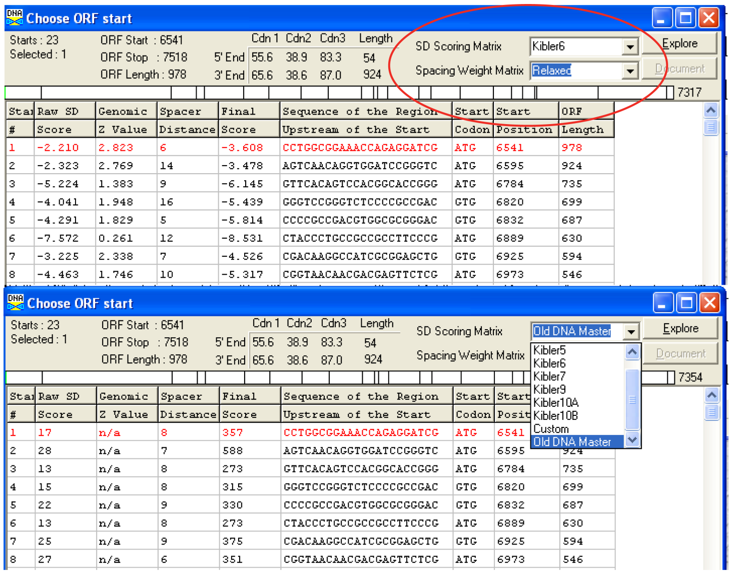
Notes: Original Glimmer call @bp 330 has strength 8.76
SSC CP SD SCS Gap Blast LO F FS

Tag	Name	5' End	3' End	Length
DNAM_1	1	330	443	114
DNAM_2	2	662	1495	834
DNAM_3	3	1543	2358	816
DNAM_4	4	2358	2675	318
DNAM_5	5	2672	2968	297
DNAM_6	6	3007	3477	471
DNAM_7	7	3559	4524	966
DNAM_8	8	4639	4797	159
DNAM_9	9	4797	6086	1290
DNAM_10	10	6083	6544	462
DNAM_11	11	6541	7518	978
DNAM_12	12	7541	9328	1788
DNAM_13	13	9325	10788	1464
DNAM_14	14	10785	11669	885
DNAM_15	15	11720	12232	513
DNAM_16	16	12262	13218	957
DNAM_17	17	13288	13470	183
DNAM_18	18	13473	13850	378
DNAM_19	19	13847	14041	195
DNAM_20	20	14038	14406	369
DNAM_21	21	14406	14741	336
DNAM_22	22	14751	15176	426
DNAM_23	23	15194	15787	594
DNAM_24	24	15899	16306	408
DNAM_25	25	16420	16731	312

Figure 3

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!



Starts : 23
Selected : 1
ORF Start : 6541
ORF Stop : 7518
ORF Length : 978

	Cdn 1	Cdn2	Cdn3	Length
5' End	55.6	38.9	83.3	54
3' End	65.6	38.6	87.0	924

SD Scoring Matrix: Kibler6
Spacing Weight Matrix: Relaxed

Star #	Raw SD Score	Genomic Z Value	Spacer Distance	Final Score	Sequence of the Region Upstream of the Start	Start Codon	Start Position	ORF Length
1	-2.210	2.823	6	-3.608	CCTGGCGGAAACCAGAGGATCG	ATG	6541	978
2	-2.323	2.769	14	-3.478	AGTCAACAGGTGGATCCGGGTC	ATG	6595	924
3	-5.224	1.383	9	-6.145	GTTCACAGTCCACGGCACCGGG	ATG	6784	735
4	-4.041	1.948	16	-5.439	GGGTCCGGGTCTCCCCGCCGAC	GTG	6820	699
5	-4.291	1.829	5	-5.814	CCCCGCCGACGTGGCGCGGGAC	GTG	6832	687
6	-7.572	0.261	12	-8.531	CTACCCTGCCGCCGCCTTCCCG	ATG	6889	630
7	-3.225	2.338	7	-4.526	CGACAAGGCCATCGCGGAGCTG	GTG	6925	594
8	-4.463	1.746	10	-5.317	CGGTAACAACGACGAGTTCTCG	ATG	6973	546

SD Scoring Matrix: Old DNA Master

Star #	Raw SD Score	Genomic Z Value	Spacer Distance	Final Score	Sequence of the Region Upstream of the Start	Start Codon	Start Posit	ORF Length
1	17	n/a	8	357	CCTGGCGGAAACCAGAGGATCG	ATG	6541	
2	28	n/a	7	588	AGTCAACAGGTGGATCCGGGTC	ATG	6595	924
3	13	n/a	8	273	GTTCACAGTCCACGGCACCGGG	ATG	6784	735
4	15	n/a	8	315	GGGTCCGGGTCTCCCCGCCGAC	GTG	6820	699
5	22	n/a	9	330	CCCCGCCGACGTGGCGCGGGAC	GTG	6832	687
6	13	n/a	8	273	CTACCCTGCCGCCGCCTTCCCG	ATG	6889	630
7	25	n/a	9	375	CGACAAGGCCATCGCGGAGCTG	GTG	6925	594
8	27	n/a	6	351	CGGTAACAACGACGAGTTCTCG	ATG	6973	546

Figure 4

Event Manager

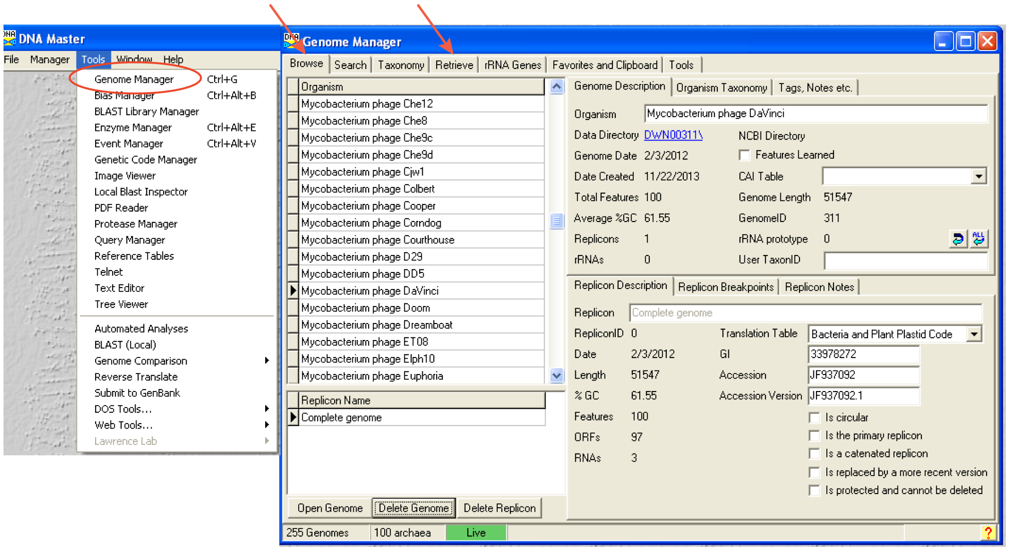
Though this isn't new or different, it is often overlooked. Go to Tools -> Event Manager. This window gives you more information when DNA Master is not working. Use the DNA Master Help Menu to evaluate the various events.

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→ 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.


DNA Master
File Manager Tools Window Help
Genome Manager Ctrl+G
Bias Manager Ctrl+Alt+B
BLAST Library Manager
Enzyme Manager Ctrl+Alt+E
Event Manager Ctrl+Alt+V
Genetic Code Manager
Image Viewer
Local Blast Inspector
PDF Reader
Protease Manager
Query Manager
Reference Tables
Telnet
Text Editor
Tree Viewer
Automated Analyses
BLAST (Local)
Genome Comparison
Reverse Translate
Submit to GenBank
DOS Tools...
Web Tools...
Lawrence Lab
Genome Manager
Browse Search Taxonomy Retrieve rRNA Genes Favorites and Clipboard Tools
Organism
Mycobacterium phage Che12
Mycobacterium phage Che8
Mycobacterium phage Che9c
Mycobacterium phage Che9d
Mycobacterium phage Cjw1
Mycobacterium phage Colbert
Mycobacterium phage Cooper
Mycobacterium phage Corndog
Mycobacterium phage Courthouse
Mycobacterium phage D29
Mycobacterium phage DD5
Mycobacterium phage DaVinci
Mycobacterium phage Doom
Mycobacterium phage Dreamboat
Mycobacterium phage ET08
Mycobacterium phage Elph10
Mycobacterium phage Euphoria
Replicon Name
Complete genome
Open Genome Delete Genome Delete Replicon
Genome Description Organism Taxonomy Tags, Notes etc.
Replicon Description Replicon Breakpoints Replicon Notes

Organism Mycobacterium phage DaVinci
Data Directory DWN00311\ NCBI 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 GenomeID 311
Replicons 1 rRNA prototype 0
rRNAs 0 User TaxonID

Replicon Complete genome
RepliconID 0 Translation Table Bacteria and Plant Plastid Code
Date 2/3/2012 GI 33978272
Length 51547 Accession JF937092
% GC 61.55 Accession Version JF937092.1
Features 100 □ Is circular
ORFs 97 □ Is the primary replicon
RNAs 3 □ Is a catenated 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.

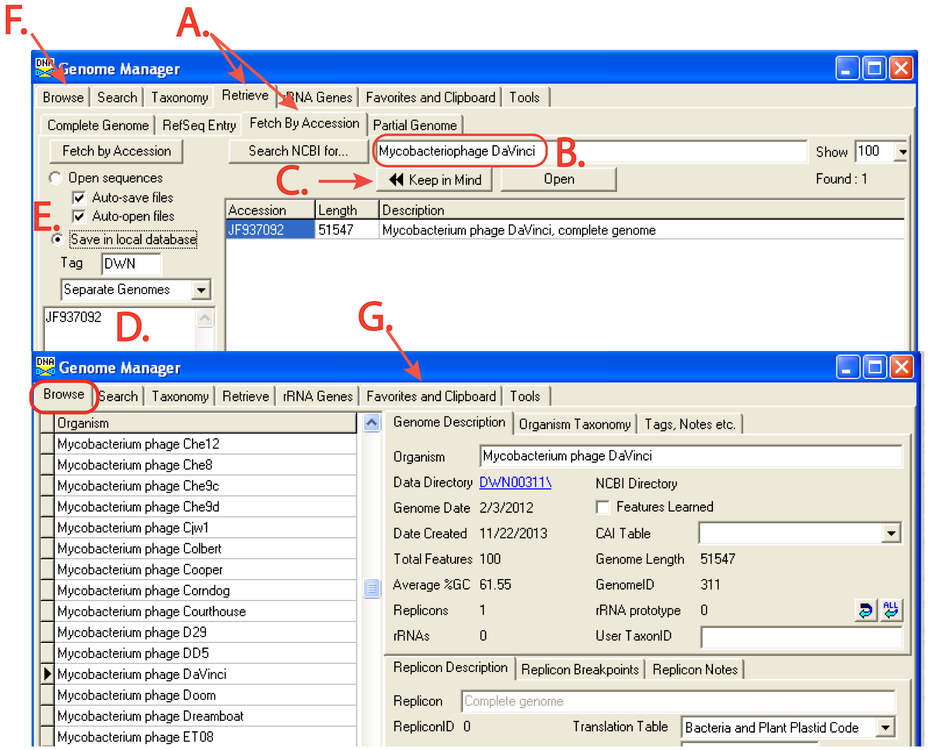
C. When you find the files of interest, click the "keep in Mind' button.

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.



Mycobacteriophage DaVinci

Show 100

Found : 1

○ Open sequences

☑ Auto-save files

☑ Auto-open files

● Save in local database

Tag DWN

Separate Genomes

JF937092

Accession	Length	Description
JF937092	51547	Mycobacterium phage DaVinci, complete genome

Organism
Mycobacterium phage Che12
Mycobacterium phage Che8
Mycobacterium phage Che9c
Mycobacterium phage Che9d
Mycobacterium phage Cjw1
Mycobacterium phage Colbert
Mycobacterium phage Cooper
Mycobacterium phage Corndog
Mycobacterium phage Courthouse
Mycobacterium phage D29
Mycobacterium phage DD5
Mycobacterium phage DaVinci
Mycobacterium phage Doom
Mycobacterium phage Dreamboat
Mycobacterium phage ET08

Organism Mycobacterium phage DaVinci

Data Directory DWN00311\ NCBI 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 GenomeID 311

Replicons 1 rRNA prototype 0

rRNAs 0 User TaxonID

Replicon Complete genome

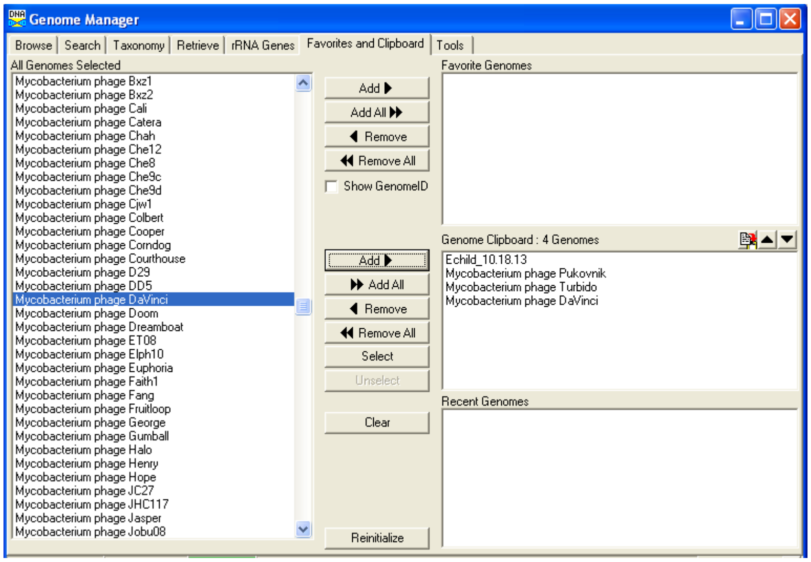
RepliconID 0 Translation Table Bacteria and Plant Plastid Code

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.



All Genomes Selected

Mycobacterium phage Bxz1
Mycobacterium phage Bxz2
Mycobacterium phage Cali
Mycobacterium phage Catera
Mycobacterium phage Chah
Mycobacterium phage Che12
Mycobacterium phage Che8
Mycobacterium phage Che9c
Mycobacterium phage Che9d
Mycobacterium phage Cjw1
Mycobacterium phage Colbert
Mycobacterium phage Cooper
Mycobacterium phage Corndog
Mycobacterium phage Courthouse
Mycobacterium phage D29
Mycobacterium phage DD5
Mycobacterium phage DaVinci
Mycobacterium phage Doom
Mycobacterium phage Dreamboat
Mycobacterium phage ET08
Mycobacterium phage Elph10
Mycobacterium phage Euphoria
Mycobacterium phage Faith1
Mycobacterium phage Fang
Mycobacterium phage Fruitloop
Mycobacterium phage George
Mycobacterium phage Gumball
Mycobacterium phage Halo
Mycobacterium phage Henry
Mycobacterium phage Hope
Mycobacterium phage JC27
Mycobacterium phage JHC117
Mycobacterium phage Jasper
Mycobacterium phage Jobu08

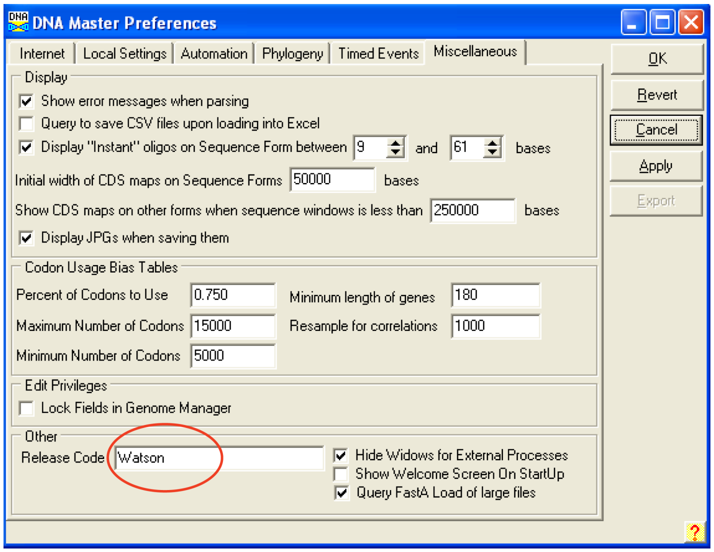
Genome Clipboard : 4 Genomes

Echild_10.18.13
Mycobacterium phage Pukovnik
Mycobacterium phage Turbido
Mycobacterium phage DaVinci

Figure 7

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).



Display
☑ Show error messages when parsing
☐ Query to save CSV files upon loading into Excel
☑ Display "Instant" oligos on Sequence Form between 9 and 61 bases
Initial width of CDS maps on Sequence Forms 50000 bases
Show CDS maps on other forms when sequence windows is less than 250000 bases
☑ Display JPGs when saving them

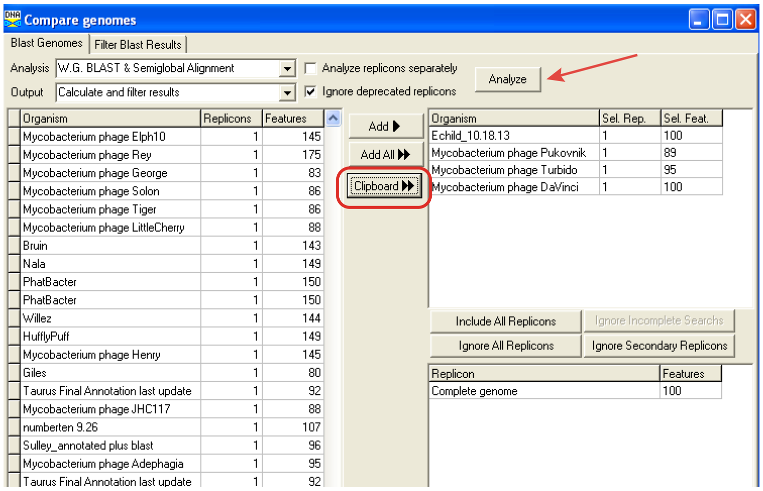
Codon Usage Bias Tables
Percent of Codons to Use 0.750
Minimum length of genes 180
Maximum Number of Codons 15000
Resample for correlations 1000
Minimum Number of Codons 5000

Edit Privileges
☐ Lock Fields in Genome Manager

Other
Release Code Watson
☑ Hide Widows for External Processes
☐ Show Welcome Screen On StartUp
☑ Query FastA Load of large files

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.



Analysis W.G. BLAST & Semiglobal Alignment

Output Calculate and filter results

□ Analyze replicons separately

☑ Ignore deprecated replicons

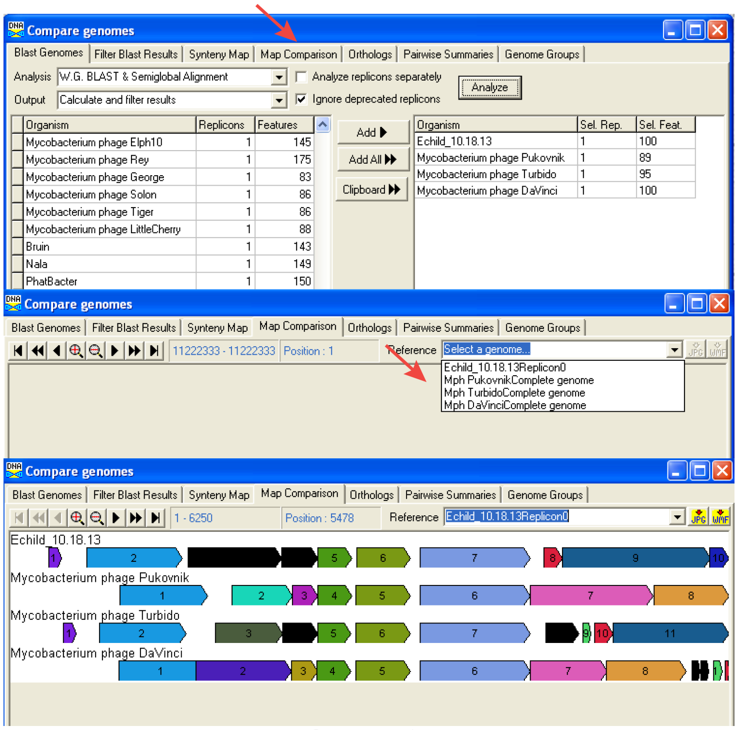
Organism	Replicons	Features
Mycobacterium phage Elph10	1	145
Mycobacterium phage Rey	1	175
Mycobacterium phage George	1	83
Mycobacterium phage Solon	1	86
Mycobacterium phage Tiger	1	86
Mycobacterium phage LittleCherry	1	88
Bruin	1	143
Nala	1	149
PhatBacter	1	150
PhatBacter	1	150
Willez	1	144
HufflyPuff	1	149
Mycobacterium phage Henry	1	145
Giles	1	80
Taurus Final Annotation last update	1	92
Mycobacterium phage JHC117	1	88
numberten 9.26	1	107
Sulley_annotated plus blast	1	96
Mycobacterium phage Adephagia	1	95
Taurus Final Annotation last update	1	92

Organism	Sel. Rep.	Sel. Feat.
Echild_10.18.13	1	100
Mycobacterium phage Pukovnik	1	89
Mycobacterium phage Turbido	1	95
Mycobacterium phage DaVinci	1	100

Replicon	Features
Complete genome	100

Figure 9

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 genomes

Blast Genomes | Filter Blast Results | Synteny Map | Map Comparison | Orthologs | Pairwise Summaries | Genome Groups

Analysis: W.G. BLAST & Semiglobal Alignment
Output: Calculate and filter results
□ Analyze replicons separately
☑ Ignore deprecated replicons

Analyze
Add ▶
Add All ▶▶
Clipboard ▶▶

Organism	Replicons	Features
Mycobacterium phage Elph10	1	145
Mycobacterium phage Rey	1	175
Mycobacterium phage George	1	83
Mycobacterium phage Solon	1	86
Mycobacterium phage Tiger	1	86
Mycobacterium phage LittleCherry	1	88
Bruin	1	143
Nala	1	149
PhatBacter	1	150

Organism	Sel. Rep.	Sel. Feat.
Echild_10.18.13	1	100
Mycobacterium phage Pukovnik	1	89
Mycobacterium phage Turbido	1	95
Mycobacterium phage DaVinci	1	100

Compare genomes

Blast Genomes | Filter Blast Results | Synteny Map | Map Comparison | Orthologs | Pairwise Summaries | Genome Groups

11222333 - 11222333 Position : 1 Reference: Select a genome...

Echild_10.18.13Replicon0
Mph PukovnikComplete genome
Mph TurbidoComplete genome
Mph DaVinciComplete genome

Compare genomes

Blast Genomes | Filter Blast Results | Synteny Map | Map Comparison | Orthologs | Pairwise Summaries | Genome Groups
JPG
WMF

1 - 6250 Position : 5478 Reference: Echild_10.18.13Replicon0

Echild_10.18.13
1 2 5 6 7 8 9 10
Mycobacterium phage Pukovnik
1 2 3 4 5 6 7 8
Mycobacterium phage Turbido
1 2 3 5 6 7 9 10 11
Mycobacterium phage DaVinci
1 2 3 4 5 6 7 8

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 http://exon.gatech.edu/heuristic_hmm2.cgi 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.