Predicting Genes in Actinobacteriophages

2022 Genomics Workshop Training

SEA-PHAGES Cohort 15

Deborah Jacobs-Sera

What are they?

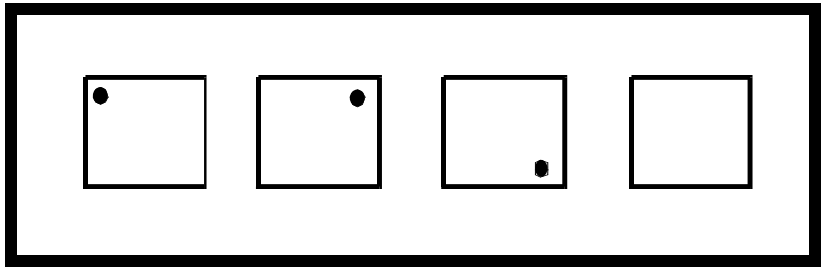
How did they get to be that way?

It is all about finding the patterns…

Since the beginning of time, woman (being human) has tried to make order and sense out of her surroundings. Gene annotation and analysis is just a primal instinct to make order.

Young children, as they prepare to enter school, are tested to see if they are ready by recognizing patterns, a form of making order.

1. Where will the dot appear in the 4th box?



Remember, everything you need to know, you learned in kindergarten….

Make-Believe or Putative

Remember, you are working in the putative gene world. All gene **predictions** are made with the best evidence to date. Most of that evidence is computational (bioinformatic), not experimental. Tomorrow's data may give us better evidence, but your prediction today is the best it can be … today! Make good predictions following a consistent approach. Let these predictions lead to experimentation that can provide the evidence to improve future predictions.

How many phage genome sequences are in GenBank?

~~13390~~ No longer countable

How many actinobacteriophage genomes are sequenced?

4153

How many As, Cs, Ts, and Gs are in a mycobacteriophage genome?

On average: ~70,000 base-pairs
Range: ~40,000 to ~165,000 bp

What is the universal format for a sequence?

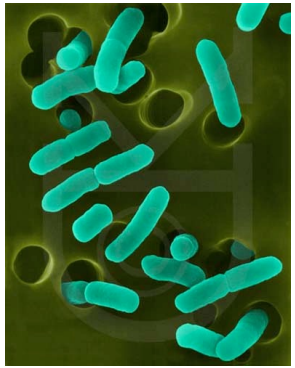
FASTA

How do you make sense of the nucleotide sequence?

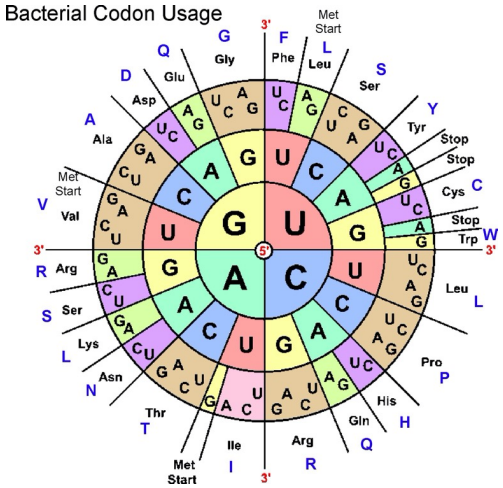
Convert to genes

How do you convert ATCGs to genes?

Codons
Code for Amino Acids, Starts, Stops



www.cen.ulaval.ca



- Phages use the Bacterial and Plant Plastid code (NCBI: Table 11)
- 3 starts
  - ATG (methionine)
  - GTG (valine)
  - TTG (leucine)
    - 3 stops (TAA, TAG, TGA)
    - Space in-between: Open Reading Frame -- ORF

ATGGACCTCTCGCCC

ATG GAC CTC TCG CCC

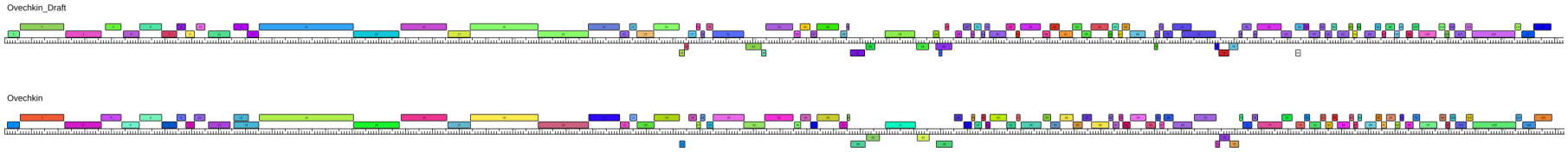
TGG ACC TCT CGC ….

GGA CCT CTC GCC ….

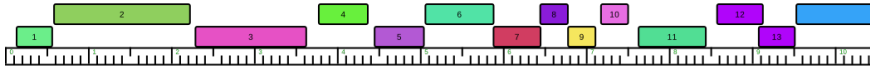
If there are 3 choices (frames) in the forward direction, how many are in the reverse direction?

>Echild complete sequence, 53159 bp including
10bp overhang (CGGTCGGTTA), Cluster A2
TGCGGCCGCCCCATCCTGTACGGGTTTCCAAGTCGATCGGAGTCCCGAGC
CGGCGCAGGAGCGCCTCACCCAGCCTCTGTGCGCCCCCAGGACGCAAGAT
CCCCGCTCACGCGGGTAGTTGTATGGGCTAATCGGCAAACGGCCTCTGAG
GCCGCGAGACCAATGTCACACCAGGTGGTGGATGTTATTGACGCACGCGT
CCGTTAAGAGGACATGGCCTAGGTATGGCTACCCAAACTTAGATTCAAAA
CCAGTCCCCTGGCCCCCGTCGTCGGTGTTGCCGCTCCTGGCGGCGCGGGG
CCAGGTCCACCACCGCAGGGAGGACCGCCATGAAGATCATCCGCTCGCTC
GCCGGGGCGCTCGCACTCATTCGCATCGCCGCCCGAGGCGGCTGGGATGC
GGAAATCTACGAGCCGTGGGATGAGGACGAATACCTCCTATAGTGATCTA
CGCCACTTGCTCGGTGGGTGTCAAGTGATACTCATGTATCTAGTTATTGA
GGGCCTAAAGGCCCGAATAAGAGCCGCACAGGCGGCTCTCTAAGAGCGCC
CACTAGGGCGCTCGAAGTAATACCGGCCTTGAGGGCCGGTTATCTGACCC
GGCAACCGCCGGGTCTTCTGCCGCGCCCAGTGGCGCGGCTCATAGAAGGG
GTGAGGCAACCGTGTACGGCACTCGCTCGAGTGCCTACTGGGCCTCGCAG
CCGGGGAAGTTCGACGTTCTGAACCTGCGGATGACGTTCCCGAGCACGTC

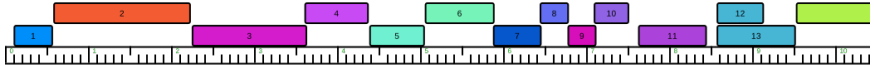
Six-Frame Translations



Ovechkin_Draft

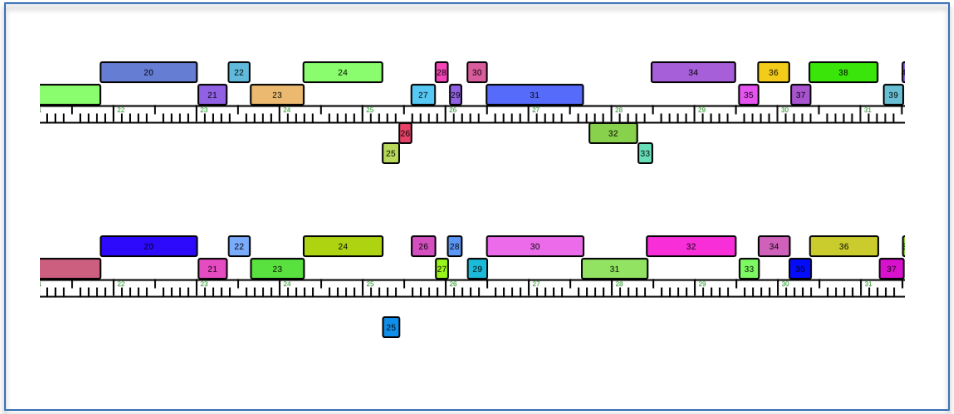


Ovechkin



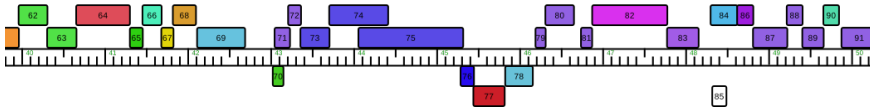
Ovechkin_Draft

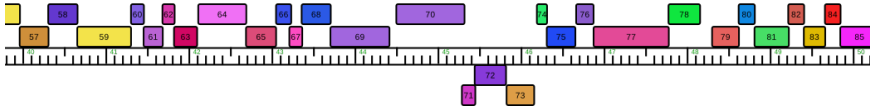
Ovechkin

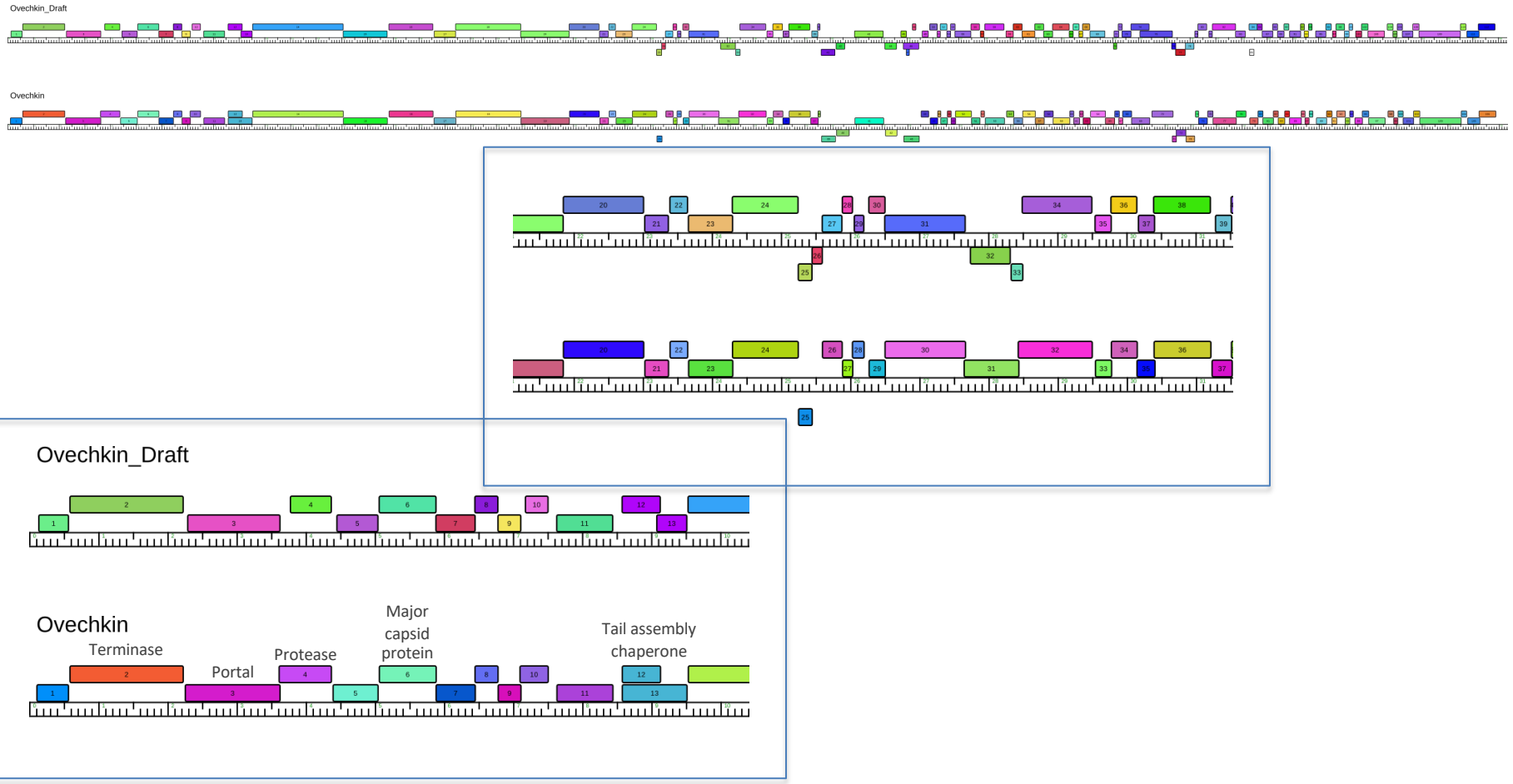


Ovechkin_Draft

Ovechkin






Ovechkin_Draft
Ovechkin
Ovechkin_Draft
Ovechkin
Terminase
Portal
Protease
Major capsid protein
Tail assembly chaperone

Gene Evaluations

For each feature we have 3 questions.

- Is it a gene?
- What is its start?
- What is its function?

For each question, what is the supporting data?

Gene Evaluations

- We are always looking for the **supporting data.**
- DNA master is our genome editor
- We (it) use 2 programs, Glimmer and GeneMark, to identify coding potential.
- We use Phamerator output for a visual representation of gene and nucleotide similarity.
- We use the Guiding Principles to remind us all of the parameters.
- As we evaluate, we can:
  - Add a gene
  - Delete a gene
  - Change a gene start

Supporting Data #1: Coding Potential
Glimmer and GeneMark

- Use Hidden Markov Models to identify coding potential
- Use **a sample** of the genome
- Identify longest ORFS in that sample
- Calculate patterns in the nucleotides:

  2 at a time, 4 at a time

GLIMMER

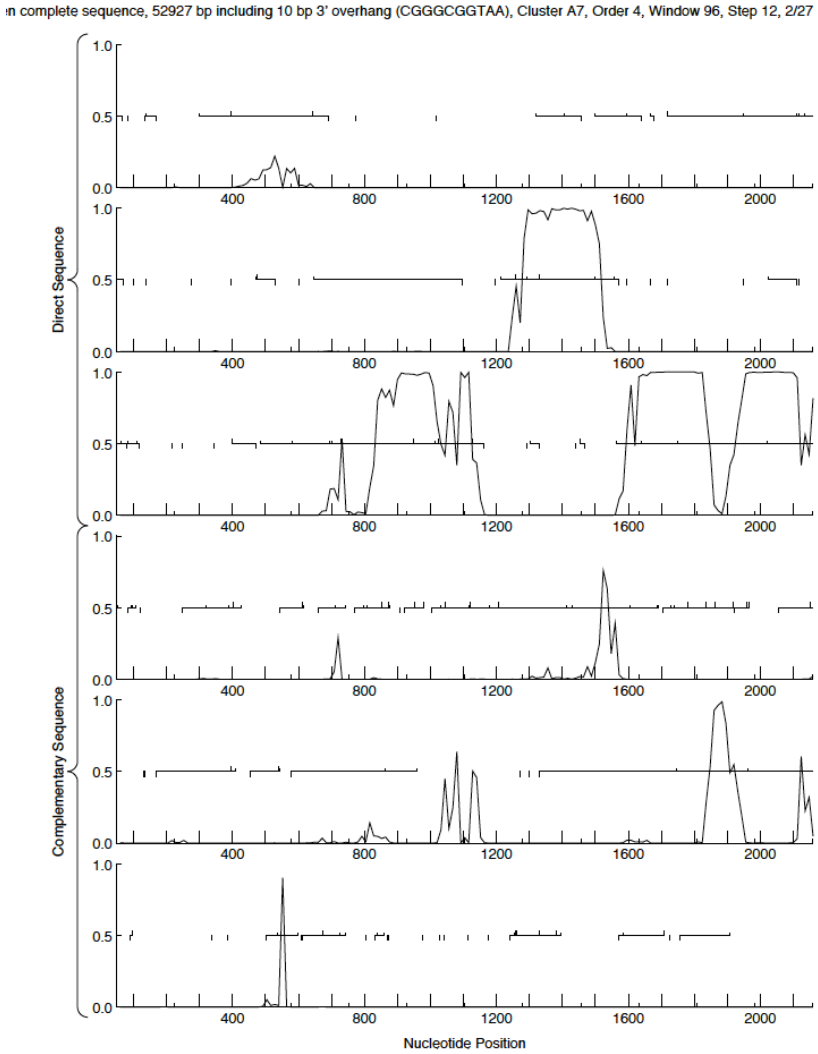
```
GLIMMER (ver. 3.02; iterated)  predictions:
 orfID      start     end  frame  score
--------    -----    -----  --    -----
>Sheen complete sequence, 52927 bp including 10 bp 3' overhang (CGGGCGGTAA), Cluster A7
orf00001      732     1166  +3    11.17
orf00002     1259     1576  +2    14.04
orf00004     1566     2318  +3    11.01
orf00006     2347     3570  +1    10.85
orf00007     3599     3877  +2     1.93
orf00008     3889     4512  +1    10.39
orf00009     4509     5477  +3     5.52
orf00011     5731     7155  +1    12.91
orf00012     5772     5635  -1     2.63
orf00013     7152     7595  +3    10.63
orf00014     7592     8332  +2     5.98
orf00016     8359    10059  +1    11.16
orf00018    10056    11552  +3    15.98
orf00020    11549    12562  +2    11.29
orf00021    12621    13130  +3    13.44
orf00022    13160    14149  +2    18.56
orf00023    14229    14390  +3     9.03
orf00025    14394    14768  +3    11.22
orf00026    14765    14920  +2     2.58
orf00028    14917    15300  +1    11.86
orf00029    15303    15647  +3    10.29
orf00030    15660    16109  +3     7.67
orf00032    16124    16708  +2    15.52
orf00033    16821    17186  +3    12.50
orf00035    17354    17614  +2     5.52
orf00037    17618    20998  +2    11.78
orf00038    21003    22982  +3    15.20
orf00041    22979    24781  +2    16.51
orf00042    24798    25265  +3     6.64
orf00043    25298    25588  +2     6.78
orf00044    25593    27047  +3    13.74
orf00045    27051    27377  +3     7.90
orf00047    28925    27417  -3     7.82
orf00048    29214    29071  -1    14.71
orf00049    29802    29311  -1     3.17
orf00050    29936    29799  -3     9.82
orf00051    30417    30229  -1    13.15
```

http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi

Microbial Genome Annotation Tools

GLIMMER is a system for finding genes in microbial DNA, especially the genomes of bacteria, archaea, and viruses. GLIMMER (Gene Locator and Interpolated Markov ModelER) uses interpolated Markov models to identify coding regions.

- Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER, *Nucleic Acids Research* 27:23 (1999), 4636-4641.
- Salzberg S, Delcher A, Kasif S, White O. Microbial gene identification using interpolated Markov models, *Nucleic Acids Research* 26:2 (1998), 544-548.



GeneMark Output (trained on *M. tuberculosis*)

DNA Master

File Tools Window Help

ORF Analysis for Jefe_Draft

bp: 1140 | ORF 827 - 1126 | G+C | 5' | Window: 30 bp

Overview | Features | References | Sequence | Documentation

Sort By: Index

Select Features | Direct SQL

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

Select All Features

Tag	Name	5' End	3' End	Length
DNAM_1	1	1	645	645
DNAM_2	2	642	830	189
DNAM_3	3	839	1126	288
DNAM_4	4	1123	2745	1623
DNAM_5	5	2820	3053	234
DNAM_6	6	3050	3457	408
DNAM_7	7	3457	3621	165
DNAM_8	8	3682	4422	741
DNAM_9	9	4422	4748	327
DNAM_10	10	4745	5122	378
DNAM_11	11	5192	5404	213
DNAM_12	12	5401	5811	411
DNAM_13	13	5811	6179	369
DNAM_14	14	6176	6406	231
DNAM_15	15	6403	6747	345
DNAM_16	16	6840	7385	546
DNAM_17	17	7451	7621	171
DNAM_18	18	7624	8016	393
DNAM_19	19	8013	8201	189
DNAM_20	20	8427	10094	1668
DNAM_21	21	10091	10537	447
DNAM_22	22	10518	10616	99

Insert | Delete | Post | Validate

Description | Sequence | Product | Regions | Blast | Context

Name: 3 GeneID

Type: CDS GI

5' End: 839 Locus Tag: DNAM_3

3' End: 1126 Regions: 1

Length: 288 Tag:

Direction Forward Feature ID: 94

Translation Table: Standard Code

EC Number

Product: gp3

Function

Notes: Original Glimmer call @bp 839 has strength 5.10; GeneMark calls start at 878

1 - 50000 | Position : 9155 | Controls >> Map | Map >> Controls

91 Features | Live | Six-frame map of starts and stops | 53125

Accession | Favorite Genome | Recent Genome | Recent File | OD

Choose ORF start

Starts : 15
Selected : 1

ORF Start : 839
ORF Stop : 1126
ORF Length : 288

	Cdn 1	Cdn2	Cdn3	Length
5' End	100.0	0.0	100.0	3
3' End	69.8	57.7	75.8	447

SD Scoring Matrix: Kibler6 | Explore

Spacing Weight Matrix: Karlin Medium | Document

943

Sta #	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	-2.624	2.661	18	-4.925	CTGAGGGATCGCGCATGTCGAT	GTG	677	450
2	-6.213	0.918	10	-6.908	AGGGATCGCGCATGTCGATGTG	GTG	680	447
3	-3.349	2.309	10	-4.043	GCATGTCGATGTGGTGCAACCG	GTG	689	438
4	-3.225	2.369	12	-4.061	TGACGGTGGCGGAGGCTCTGAG	TTG	734	393
5	-4.299	1.848	9	-5.074	GGAACGCCACGCACGAGGCGAC	GTG	758	369
6	-6.188	0.931	11	-6.945	CGTCGGCTCCCGGCCCGCGCGA	GTG	794	333
7	-2.187	2.873	10	-2.882	AGGACGAAGTGAGGATGGGCGC	ATG	827	300
8	-5.308	1.358	11	-6.065	GGATGGGCGCATGACCCGTTGT	ATG	839	288
9	-6.359	0.848	11	-7.116	GCGCATGACCCGTTGTATGTCC	GTG	845	282
10	-6.193	0.928	14	-7.539	TGATCCCTCGACCCCGTGCACA	TTG	878	249
11	-4.918	1.547	11	-5.675	GATCTACTACACGTACGAGGGC	GTG	944	183
12	-3.942	2.021	10	-4.637	CTACACGTACGAGGGCGTGCCC	GTG	950	177
13	-2.071	2.930	16	-3.867	CCTGGAGGAGGCCATCGAGCGG	ATG	1034	93
14	-6.940	0.566	12	-7.775	CCTGGCCGATGCTCTCCTCGCG	ATG	1097	30
15	-2.624	2.661	8	-3.846	TCTCCTCGCGATGAGGGACCGG	ATG	1109	18

Supporting Data #2:

- DNA Master – our genome editor
- Draft Annotation
  - Glimmer
  - GeneMark
- Refinement of Draft Annotation
  - Blast
    - NCBI
    - PhagesDB
  - HHPred Suite
  - Comparative Data
    - Phamerator
    - Starterator

Blast Comparisons

BLAST®

Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ blastp suite

blastn | blastp | blastx | tblastn | tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

From

To

Or, upload file

Browse... No file selected.

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database

Non-redundant protein sequences (nr)

Organism
Optional

Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional

Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query
Optional

Create custom database

Enter an Entrez query to limit search

Program Selection

Algorithm

- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

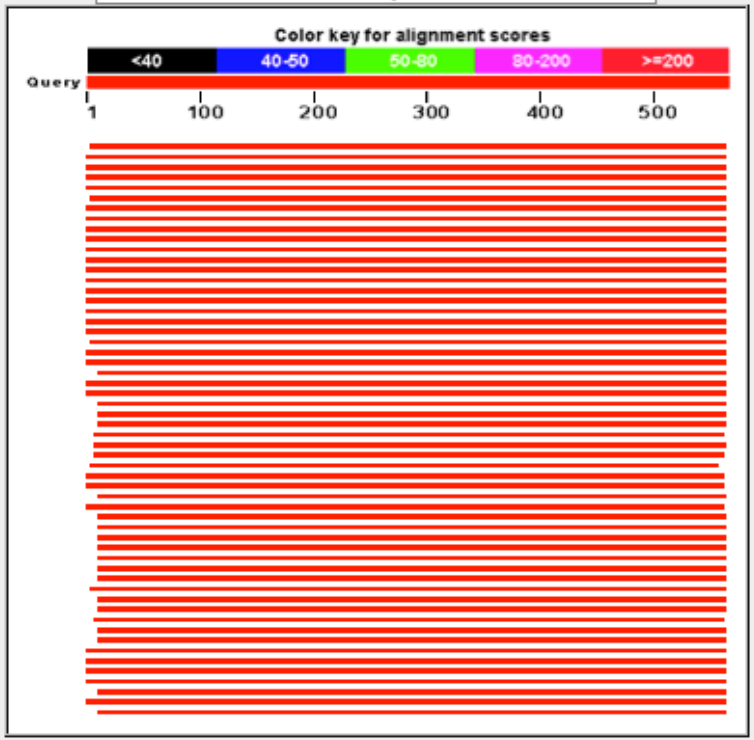
Show results in a new window

Algorithm parameters

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 100 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments



Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
gp12 [Mycobacterium phage Timshel]	1047	1047	99%	0.0	90%	AEJ92326.1
terminase [Mycobacterium phage Obama12]	1045	1045	99%	0.0	88%	YP_009007203.1
terminase [Mycobacterium phage Dhanush]	1045	1045	99%	0.0	88%	AGK87127.1
gp11 [Mycobacterium phage Flux]	1044	1044	99%	0.0	87%	AFL47832.1
terminase [Mycobacterium phage Nyxis]	1044	1044	99%	0.0	88%	YP_009005850.1
terminase [Mycobacterium phage HINdeR]	1043	1043	99%	0.0	89%	YP_008051863.1
gp10 [Mycobacterium phage Arturo]	1043	1043	99%	0.0	88%	AFU20464.1
terminase [Mycobacterium phage Kampy]	1043	1043	99%	0.0	87%	YP_009031871.1
gp11 [Mycobacterium phage Shaka]	1043	1043	99%	0.0	87%	AEF57321.1
gp11 [Mycobacterium phage Sabertooth]	1042	1042	99%	0.0	87%	AFU20551.1
terminase [Mycobacterium phage BellusTerra]	1040	1040	99%	0.0	87%	YP_009005569.1
gp11 [Mycobacterium phage ICleared]	1040	1040	99%	0.0	87%	AFL46618.1
gp11 [Mycobacterium phage Peaches]	1040	1040	99%	0.0	87%	YP_003358714.1
gp11 [Mycobacterium phage Wile]	1038	1038	99%	0.0	87%	YP_009014094.1
terminase [Mycobacterium phage QuinnKiro]	1023	1023	99%	0.0	86%	AIS73685.1

Alignments

Download GenPept Graphics

gp12 [Mycobacterium phage Timshel]

Sequence ID: gb|AEJ92326.1| Length: 564 Number of Matches: 1

Range 1: 4 to 564 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
1047 bits(2707)	0.0	Compositional matrix adjust.	504/562(90%)	531/562(94%)	1/562(0%)

```
Query  5    YLSTEPLLPQPPHKIGPVWLCHEDGSWALPKYTLGWGVLNWLAEYVRSPAGGGPFIPTLE  64
            YL+  PLLPQPPHKIGPVW  HEDGSWALP  TLGWGVLNWLAEYVRSPAGGG FIPTLE
Sbjct  4    YLNPGPLLPQPPHKIGPVWQVHEDGSWALPARTLGWGVLNWLAEYVRSPAGGGVFIPTLE  63

Query  65   QARFILWWYAVDENGVYAYREGVLRRMKGWGKDPLCAAIALVELCGPVAFSHWDEKGNPV  124
            QARFILWWYAVDE G YAYREG LRRMKGWGKDPLCAAIALVELCGPVAFSHWD  G+PV
Sbjct  64   QARFILWWYAVDERGNYAYREGCLRRMKGWGKDPLCAAIALVELCGPVAFSHWDLDGSPV  123

Query  125  GKRRHAAWITIAAVSQDQTKNTFSLFPVMISKQMKTEYGLDVNKFVIYTEDGGRIEAATS  184
            GK RHAAWIT+AAVSQDQTKNTFS+FPVMISK+MK +YGLDVNKFVIY+E+GGRIEAATS
Sbjct  124  GKPRHAAWITVAAVSQDQTKNTFSMFPVMISKKMKVDYGLDVNKFVIYSEEGGRIEAATS  183

Query  185  SPASMEGNRPTLVIENETQWWGVGPDGNVNDGVAMDDVIEGNVSKIPGARKLAICNAHIP  244
            SPASMEGNRPTLVIENETQWWGVGPDGNVNDG AMDDVIEGNVSKIPGARKLAICNAHIP
Sbjct  184  SPASMEGNRPTLVIENETQWWGVGPDGNVNDGPAMDDVIEGNVSKIPGARKLAICNAHIP  243

Query  245  GNDTVAEKAYDHWQDILSGKAVDTGLMYDALEAPADTPVSEIPSEKEDPEGYEAGIAQLM  304
            GNDTVAEKAYDHWQDILSGKAVDTG+MYDALEAPADTPVSEIPSEKEDPEGYE GIAQLM
Sbjct  244  GNDTVAEKAYDHWQDILSGKAVDTGIMYDALEAPADTPVSEIPSEKEDPEGYEKGIAQLM  303

Query  305  DGLEVARGDSYWLPLEEILGSVLNTRNPVTESRRKFLNQVNAHEDSWIAPSEWDRLAVTD  364
            +GLE+ARGDSYWLPL+EI+GSVLNTRNPVTESRRKFLNQVNAHEDSWIAP+ WDRLA+TD
Sbjct  304  EGLEIARGDSYWLPLDEIMGSVLNTRNPVTESRRKFLNQVNAHEDSWIAPAWWDRLALTD  363

Query  365  KALALQKDDRITLGFDGSKSDDWTALVACRVSDGMLFLIKSWNPEDYPHEEVPREDVDAV  424
                L+K+DRITLGFDGSKSDDWTALVACRVSDGMLFL+  WNP DYPH+EVPR++VDAV
Sbjct  364  PLFKLKKNDRITLGFDGSKSDDWTALVACRVSDGMLFLLDKWNPNDYPHDEVPRDEVDAV  423

Query  425  VRSAFQRYDVVGFRADVKEFEAYVDQWGRDFKRKLKINATPGNPVAFDMRGQTKRFALDC  484
            VRSAFQRYDVVGFRADVKEFEAYVDQWGRDFKRKLK+NATPGNP+AFDMRGQTKRFALDC
Sbjct  424  VRSAFQRYDVVGFRADVKEFEAYVDQWGRDFKRKLKVNATPGNPIAFDMRGQTKRFALDC  483

Query  485  ERFVDAVIEHELHHDGNPVLRQHVLNARRHPTTYDAISIRKESKDSSKKIDAAVCAVLAF  544
            ERF+DAVIE EL HD NPVLRQHVLNARRHPTT+DAISIRKESKDSSKKIDAAVCAVLAF
Sbjct  484  ERFLDAVIEKELWHDQNPVLRQHVLNARRHPTTFDAISIRKESKDSSKKIDAAVCAVLAF  543

Query  545  GARQDYQMSKKHRSGAKAVIIR  566
            G+RQDY MSKKHR G  AVI+R
Sbjct  544  GSRQDYLMSKKHRGGG-AVIVR  564
```

Download GenPept Graphics

terminase [Mycobacterium phage Obama12]

Sequence ID: ref|YP_009007203.1| Length: 565 Number of Matches: 1

▶ See 3 more title(s)

Range 1: 1 to 565 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
1045 bits(2703)	0.0	Compositional matrix adjust.	496/566(88%)	537/566(94%)	1/566(0%)

```
Query  1    MALEYLSTEPLLPQPPHKIGPVWLCHEDGSWALPKYTLGWGVLNWLAEYVRSPAGGGPFI  60
            M+L      PLLPQPPHKIGPVW   EDGSW LP+ TLGWG+LNWLA+YVRSPAGGGPF+
Sbjct  1    MSLANHHPVPLLPQPPHKIGPVWQVREDGSWHLPERTLGWGILNWLAKYVRSPAGGGPFL  60

Query  61   PTLEQARFILWWYAVDENGVYAYREGVLRRMKGWGKDPLCAAIALVELCGPVAFSHWDEK  120
            PTLEQARFILWWYAVDE GVYAYREGVLRRMKGWGKDPLCAAIAL ELCGPVAFSHWD +
Sbjct  61   PTLEQARFILWWYAVDERGVYAYREGVLRRMKGWGKDPLCAAIALAELCGPVAFSHWDLE  120

Query  121  GNPVGKRRHAAWITIAAVSQDQTKNTFSLFPVMISKQMKTEYGLDVNKFVIYTEDGGRIE  180
            GNPVGK RHAAWITIAAVSQDQTKNTFSLFPVMISK++KT+YGLDVNKFVIY+E GGRIE
```

The Mycobacteriophage Database

at PhagesDB.org

Home | Phages | Data | BLAST | Publications | Resources | Software | Social | Other DBs | About

Search PhagesDB.org

Recently Added Phages

- Savanna
- Bryanna
- RadDad
- Dyanna
- Daishi

Recently Modified Phages

- HINdeR
- Ferb
- Biaggio
- Cgwhichard
- Filch

Recently Finished Phages

- Cambiare (G)
- AlanGrant (B4)
- Corofin (B3)
- Baee (B5)

Local Protein BLAST

Go to **Nucleotide BLAST**

This tool will run a local BLAST search against our protein databases. This includes all proteins from the most recent Phamerator update. Proteins marked "Draft" are from auto-annotated files.

Choose program to use and database to search

Program blastp Database Mycobacteriophage Proteins as of Feb 07, 2014

Enter sequence below in **FASTA** format

```
AATSSPASMEGNRPTLVIENETQWWGVGPDGNVNDGVAMDDVIEGNVSKIPGARKLAICN
AHIPGNDTVAEKAYDHWQDILSGKAVDTGLMYDALEAPADTPVSEIPSEKEDPEGYEAGI
AQLMDGLEVARGDSYWLPLEEILGSVLNTRNPVTESRRKFLNQVNAHEDSWIAPSEWDRL
AVTDKALALQKDDRITLGFDGSKSDDWTALVACRVSDGMLFLIKSWNPEDYPHEEVPRED
VDAVVRSAFQRYDVVGFRADVKEFEAYVDQWGRDFKRKLKINATPGNPVAFDMRGQTKRF
ALDCERFVDAVIEHELHHDGNFVLRQHVLNARRHPTTYDAISIRKESKDSSKKIDAAVCA
VLAFGARQDYQMSKKHRSGAKAVIIRZ
```

Or load it from disk

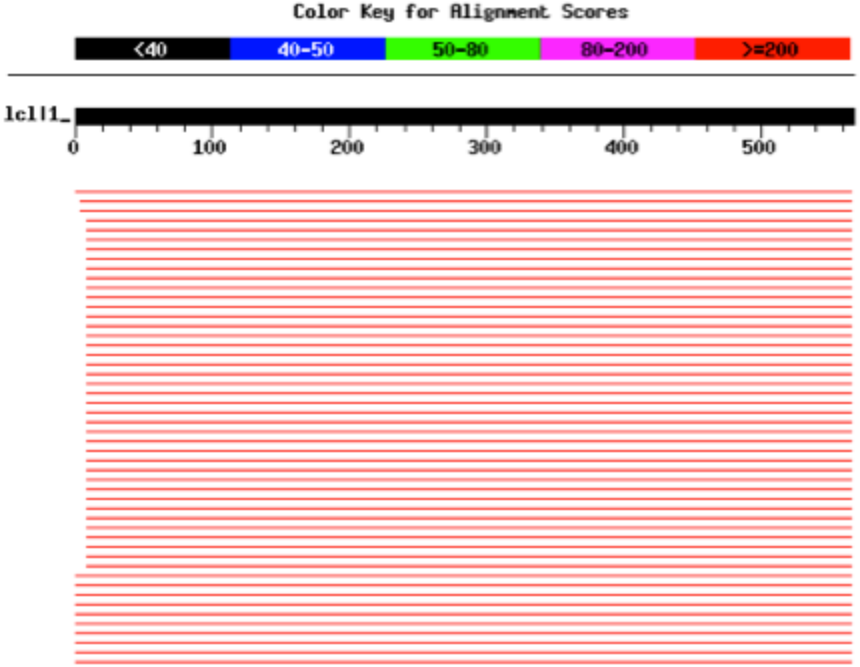
Browse... No file selected.

Set subsequence: From To

Clear sequence BLAST

Distribution of 100 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments

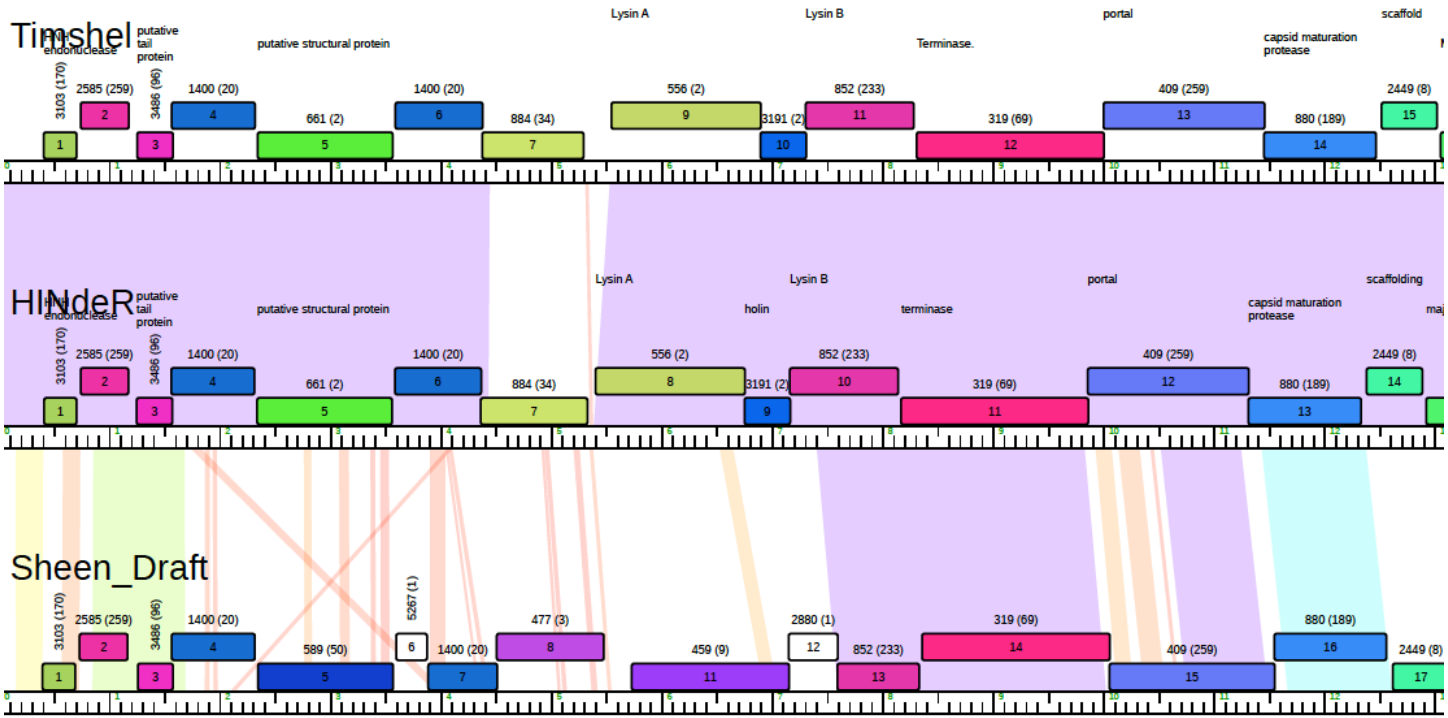


Sequences producing significant alignments:	Score (bits)	E Value
Sheen-Draft_14, function unknown, 566	1171	0.0
Timshel_12, Terminase., 564	1053	0.0
HINdeR_11, terminase, 564	1048	0.0
Obama12-draft_10, function unknown, 565	1046	0.0
LHTSCC_11, Terminase, large subunit, 565	1046	0.0
Arturo_10, terminase, 565	1046	0.0
Phighter1804-Draft_10, function unknown, 565	1046	0.0
Eagle_10, Terminase, 565	1046	0.0
Camperdownii_Draft_10, function unknown, 565	1046	0.0
Flux_11, terminase, 565	1045	0.0
Dhanush_11, terminase, 565	1045	0.0
Wander_Draft_10, function unknown, 565	1044	0.0
TygerBlood-Draft_10, function unknown, 565	1044	0.0
TiroTheta9_11, terminase, 565	1044	0.0
TinaFeyge-Draft_gp10, function unknown, 565	1044	0.0
Shaka_11, terminase, 565	1044	0.0
Sabertooth_11, Terminase, 565	1044	0.0
Millski_Draft_10, function unknown, 565	1044	0.0
Melvin_MELVIN_11, Terminase, 565	1044	0.0
MeeZee_11, Terminase, 565	1044	0.0
Medusa_11, terminase, 565	1044	0.0
Maverick-DRAFT_11, function unknown, 565	1044	0.0
Lemur_Draft_10, function unknown, 557	1044	0.0
Kratark_Draft_10, function unknown, 565	1044	0.0
KFPoly_Draft_10, function unknown, 565	1044	0.0
Kampy-Draft_10, function unknown, 565	1044	0.0
Holli-draft_10, function unknown, 565	1044	0.0
HamSlice-Draft_10, function unknown, 565	1044	0.0
Gadost_Draft_10, function unknown, 565	1044	0.0
Funston_Draft_DRAFT_10, function unknown, 565	1044	0.0
Eris_Draft_10, function unknown, 565	1044	0.0

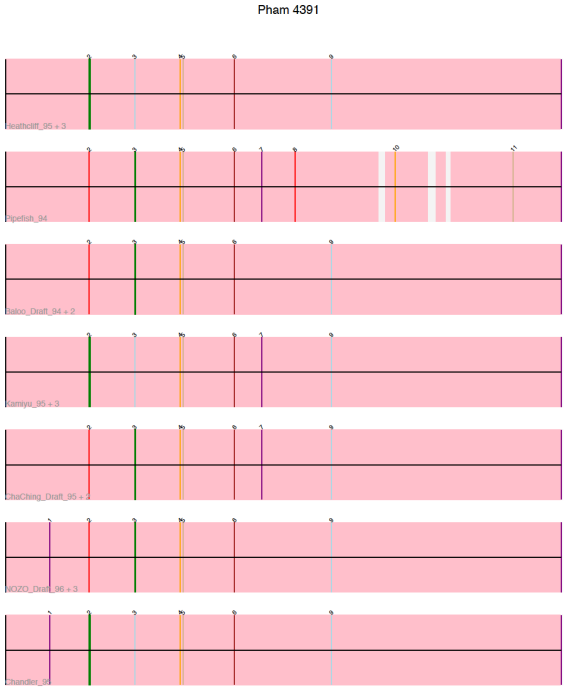
```
>Timshel_12, Terminase., 564
         Length = 564

 Score = 1053 bits (2724), Expect = 0.0
 Identities = 504/562 (89%), Positives = 531/562 (94%), Gaps = 1/562 (0%)

Query: 5   YLSTEPLLPQPPHKIGPVWLCHEDGSWALPKYTLGWGVLNWLAEYVRSPAGGGPFIPTLE 64
           YL+  PLLPQPPHKIGPVW  HEDGSWALP  TLGWGVLNWLAEYVRSPAGGG FIPTLE
Sbjct: 4   YLNPGPLLPQPPHKIGPVWQVHEDGSWALPARTLGWGVLNWLAEYVRSPAGGGVFIPTLE 63

Query: 65  QARFILWWYAVDENGVYAYREGVLRRMKGWGKDPLCAAIALVELCGPVAFSHWDEKGNPV 124
           QARFILWWYAVDE G YAYREG LRRMKGWGKDPLCAAIALVELCGPVAFSHWD  G+PV
Sbjct: 64  QARFILWWYAVDERGNYAYREGCLRRMKGWGKDPLCAAIALVELCGPVAFSHWDLDGSPV 123

Query: 125 GKRRHAAWITIAAVSQDQTKNTFSLFPVMISKQMKTEYGLDVNKFVIYTEDGGRIEAATS 184
           GK RHAAWIT+AAVSQDQTKNTFS+FPVMISK+MK +YGLDVNKFVIY+E+GGRIEAATS
Sbjct: 124 GKPRHAAWITVAAVSQDQTKNTFSMFPVMISKKMKVDYGLDVNKFVIYSEEGGRIEAATS 183

Query: 185 SPASMEGNRPTLVIENETQWWGVGPDGNVNDGVAMDDVIEGNVSKIPGARKLAICNAHIP 244
           SPASMEGNRPTLVIENETQWWGVGPDGNVNDG AMDDVIEGNVSKIPGARKLAICNAHIP
Sbjct: 184 SPASMEGNRPTLVIENETQWWGVGPDGNVNDGPAMDDVIEGNVSKIPGARKLAICNAHIP 243

Query: 245 GNDTVAEKAYDHWQDILSGKAVDTGLMYDALEAPADTPVSEIPSEKEDPEGYEAGIAQLM 304
           GNDTVAEKAYDHWQDILSGKAVDTG+MYDALEAPADTPVSEIPSEKEDPEGYE GIAQLM
Sbjct: 244 GNDTVAEKAYDHWQDILSGKAVDTGIMYDALEAPADTPVSEIPSEKEDPEGYEKGIAQLM 303

Query: 305 DGLEVARGDSYWLPLEEILGSVLNTRNPVTESRRKFLNQVNAHEDSWIAPSEWDRLAVTD 364
           +GLE+ARGDSYWLPL+EI+GSVLNTRNPVTESRRKFLNQVNAHEDSWIAP+ WDRLA+TD
Sbjct: 304 EGLEIARGDSYWLPLDEIMGSVLNTRNPVTESRRKFLNQVNAHEDSWIAPAWWDRLALTD 363

Query: 365 KALALQKDDRITLGFDGSKSDDWTALVACRVSDGMLFLIKSWNPEDYPHEEVPREDVDAV 424
               L+K+DRITLGFDGSKSDDWTALVACRVSDGMLFL+  WNP DYPH+EVPR++VDAV
Sbjct: 364 PLFKLKKNDRITLGFDGSKSDDWTALVACRVSDGMLFLLDKWNPNDYPHDEVPRDEVDAV 423

Query: 425 VRSAFQRYDVVGFRADVKEFEAYVDQWGRDFKRKLKINATPGNPVAFDMRGQTKRFALDC 484
           VRSAFQRYDVVGFRADVKEFEAYVDQWGRDFKRKLK+NATPGNP+AFDMRGQTKRFALDC
Sbjct: 424 VRSAFQRYDVVGFRADVKEFEAYVDQWGRDFKRKLKVNATPGNPIAFDMRGQTKRFALDC 483

Query: 485 ERFVDAVIEHELHHDGNPVLRQHVLNARRHPTTYDAISIRKESKDSSKKIDAAVCAVLAF 544
           ERF+DAVIE EL HD NPVLRQHVLNARRHPTT+DAISIRKESKDSSKKIDAAVCAVLAF
Sbjct: 484 ERFLDAVIEKELWHDQNPVLRQHVLNARRHPTTFDAISIRKESKDSSKKIDAAVCAVLAF 543

Query: 545 GARQDYQMSKKHRSGAKAVIIR 566
           G+RQDY MSKKHR G  AVI+R
Sbjct: 544 GSRQDYLMSKKHRGGG-AVIVR 564
```

```
>HINdeR_11, terminase, 564
         Length = 564

 Score = 1048 bits (2709), Expect = 0.0
 Identities = 501/562 (89%), Positives = 530/562 (94%), Gaps = 1/562 (0%)

Query: 5   YLSTEPLLPQPPHKIGPVWLCHEDGSWALPKYTLGWGVLNWLAEYVRSPAGGGPFIPTLE 64
           YL+  PLLPQPPHKIGPVW  HEDGSWALP  TLGWGVLNWLAEYVRSPAGGGPFIPTLE
Sbjct: 4   YLNPGPLLPQPPHKIGPVWQVHEDGSWALPARTLGWGVLNWLAEYVRSPAGGGPFIPTLE 63

Query: 65  QARFILWWYAVDENGVYAYREGVLRRMKGWGKDPLCAAIALVELCGPVAFSHWDEKGNPV 124
           QARFILWWYAVDE G YAYREG LRRMKGWGKDPLCAAIALVELCGPVAFSHWD  G+PV
Sbjct: 64  QARFILWWYAVDERGNYAYREGCLRRMKGWGKDPLCAAIALVELCGPVAFSHWDLDGSPV 123
```

Phamerator map



Starterator data



Note: In the above figure, yellow indicates the location of called starts comprised solely of computational predictions (i.e. auto-annotations by Glimmer/GeneMark), green indicates the location of called starts with at least 1 manual gene annotation.

Pham 4391 Report

This analysis was run 11/27/16.

Pham number 4391 has 20 members, 6 are drafts.

Phages represented in each track:
- Track 1 : Heathcliff_95, Bernardo_92, Akoma_96, Audrey_95
- Track 2 : Pipefish_94
- Track 3 : Baloo_Draft_94, Phaedrus_90, Mortcellus_Draft_96
- Track 4 : Kamiyu_95, Athena_97, Daisy_94, Corofin_95
- Track 5 : ChaChing_Draft_95, Yahalom_Draft_91, Phlyer_95
- Track 6 : NOZO_Draft_96, Compostia_Draft_98, Gadjet_93, OrangeOswald_94
- Track 7 : Chandler_95

Summary of Final Annotations (Info on gene starts based on numbers in diagram):

The start number called the most often in the published annotations is start number 2, it was called in 9 of the 14 non-draft genes in the pham.

Genes that call this "Most Annotated" start:

- Heathcliff_95, Kamiyu_95, Akoma_96, Bernardo_92, Athena_97, Chandler_95, Audrey_95, Daisy_94, Corofin_95,

Genes that have the "Most Annotated" start but do not call it:

- Pipefish_94, Baloo_Draft_94, Phaedrus_90, Yahalom_Draft_91, NOZO_Draft_96, Compostia_Draft_98, Phlyer_95, OrangeOswald_94, Gadjet_93, Mortcellus_Draft_96, ChaChing_Draft_95,

Genes that do not have the "Most Annotated" start:

-

Summary by start number:

- Start number 2 is called in: Heathcliff_95, Kamiyu_95, Akoma_96, Bernardo_92, Athena_97, Chandler_95, Audrey_95, Daisy_94, Corofin_95,

Percent with start 2 called: 45.0%

- Start number 3 is called in: Pipefish_94, Baloo_Draft_94, Phaedrus_90, Yahalom_Draft_91, NOZO_Draft_96, Compostia_Draft_98, Phlyer_95, OrangeOswald_94, Gadjet_93, Mortcellus_Draft_96, ChaChing_Draft_95,

Percent with start 3 called: 55.0%

GUIDING PRINCIPLES
OF BACTERIOPHAGE GENOME ANNOTATION

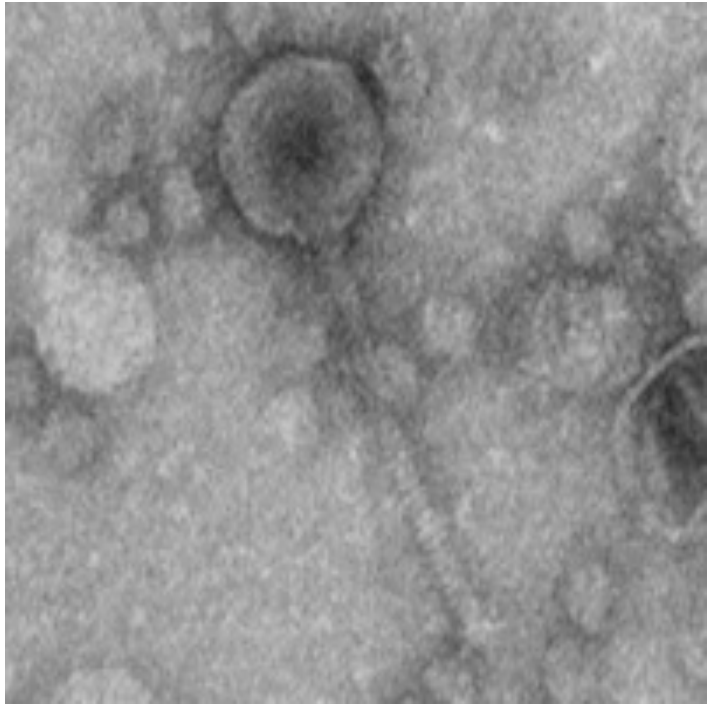
- Found in “Phage Annotation, Genomics and Data Interpretation” Section of the Bioinformatics Guide
- 15 Key Directives
- Read for tomorrow
  https://seaphagesbioinformatics.helpdocsonline.com/guiding-principles

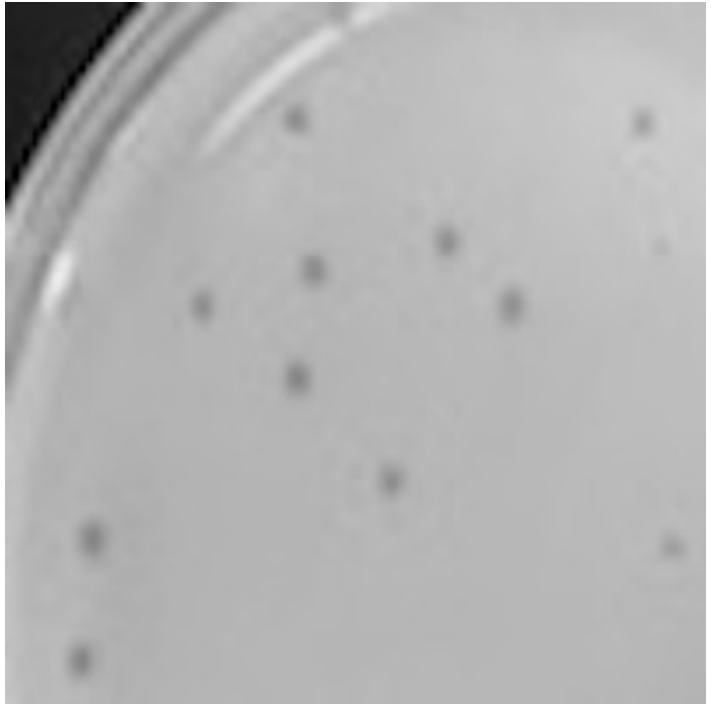

H2O

Let's get started!

1. Gather Data
2. Auto-annotate in DNA Master
3. Gene Calling
4. Functional Assignments

Microbacterium Jefe





Tonight's Tasks:

Annotation Outline

- Outline Introduction
- Setting Up Your Computer
  - DNA Master
    - Installing DNA Master ✓
    - Updating DNA Master ✓
    - Setting DNA Master preferences
  - Websites to bookmark
- Surveying Your Genome
  - Retrieving your genome sequence
  - Comparing your genome's sequence ✓
  - Comparing your genome's genes ✓
  - Clustering your genome
- Gathering Data
  - Creating a coding potential graph
  - Making a Phamerator map
  - Guiding principles of phage annotation ← Tonight's reading
- Automatically Annotating Your Genome
  - Auto-annotation
    - Complete Genome Blastp in DNA Master

DNA Master
Current Build
2705

Tips for DNA Master Files

- Save .dnam5 file often
- Save .dnam5 file as a new name. (Then don’t save the old named one.)
- If working in a virtual machine, be mindful about closing/shutting down
- Did I mention to save your files often using a new name?