Predicting Genes in Mycobacteriophages

December 5, 2016
2016 Bioinformatics Workshop Training
SEA-PHAGES Cohort 9
D. Jacobs-Sera
OBJECTIVES: Participants will be able to:

• Characterize and investigate mycobacteriophage genomes. This includes the details of gene calling evaluation, functional assignments, preparing a final product of a submission file that meets QC requirements. This year’s genome is Mycobacteriophagae CrystalP.

• Install and use the software for annotation/analyses of mycobacteriophages. The software includes DNA Master, Phamerator, and other web-based tools.

• Identify the basic concepts that underpin the bioinformatics of phage biology.

• Explore classroom implementation strategies
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....
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 bacteriophage genome sequences are in GenBank?

2500+

How many mycobacteriophage genomes are sequenced?

1201

How many mycobacteriophage genomes are published?

Tricky Question

Number in GenBank: 672
How many ATCGS are in a typical mycobacteriophage genome?

On average 70,000 base-pairs
Range 40,000 to 165,000 bps

What is the universal format for a sequence?

FASTA
>Echild complete sequence, 53159 bp including 10bp overhang (CGGTCGGTTA), Cluster A2
TGCGGCCGCCCATCTCTGTACGGGATTTTCCAAGTCTAGTCGGAGTCGCCGAGC
CGGCGCAGGAGCCGCTCAACCAGCCCTCTGTGCGCCCCCAGGAGCGACAGAT
CCCCGCTACGCAGGGTAGTATGTTATGGGCTAACTCGGCAAGACGCTCCTGAG
GCCGCGAGACACATGTCAACCAGGCTTTGATGTTATATTGACGACACGGGT
CGTTAAGAGGAACATGGGCTAGGTATGGGCTACCCAAACTTAGGATTCAAAA
CCAGTCCCTTGCCCCGCCGCTGTGCTGGTGTGCTGCTGCTCTTGGGCCGGCGGGG
CCAGGTCCAACCACGCAGGAGACACCACCATGAGATCATCCGCTGCTGCTC
GCCGGGGCGCTCGACTCTACTCTACTCTCGATCGCTCGCCGCCGGCGCTGGATG
GGAAATCTACGAGCGGTGGATGAGGACGAAATACCTCTCTATAGTATCTA
CGCACTTTGCTCGGTGGGCTAGTATCTCATGTAATCTAGTTATTGAG
GGCCTAAAAAGGCCGCAATAAGAACGCACAGCGCGGTCTCTCAAGAGACGCC
CACTAGGGCGCTCGAAGTAACTACGCCCTTTGAGGCCCGGTATCTGACC
GGCAACCGCGCGGTTTTCTCGGCAGCGCGGGCTCATAGAAGGG
GTGAGGCAACCGTGTCAGCGCTCTCGATCAGCTGCGCTGTGCGCTGAGC
CCGGAAAAGTTGCACTCTAGCTGACTCGACGATGTTTCGCAGCAGT
How do you make sense of the ATCGs?

Convert to genes

How do you convert ATCGs to Genes?

Codons
Code for Amino Acids, Starts, Stops
• Phages use the Bacterial Plastic code (NCBI: Table 11)

• 3 starts
  - ATG (methionine)
  - GTG (valine)
  - TTG (leucine)

• 3 stops (TAA, TAG, TGA)

• Space in-between: Open Reading Frame -- ORF
If there are 3 choices (frames) in the forward direction, how many are in the reverse direction?
Six Frame Translations
Features found in Mycobacteriophage genomes

- protein coding genes ✓
- tRNAs ✓
- tmRNAs
- AttP sites
- Terminators
- Frame shifts ✓
- ...

•
Gene Evaluations

For each feature we have 3 questions

• Is it a gene?
• What is its starts?
• What is its function?
Gene Evaluations

• We are always looking for the **supporting data**
• We 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 of 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 (ver. 3.02; iterated) predictions:

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

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.


GeneMark Output (trained on *M. tuberculosis*)
1. In any segment of DNA, typically only one frame in one strand is used for a protein-coding 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 sub-principles.

11. tRNA genes are not called precisely in the program embedded in DNA Master, and require extra attention. (Please refer to Section 9.5.)
Comparisons with what we already know

• Phamerator comparisons

• BLAST comparisons
  • At NCBI
  • ***At phagesDB***

• Starterator data “New & Improved!!!”
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, Mortoellus_Draft_95
- Track 4: Kamiyu_95, Athena_97, Daisy_94, Corofin_95
- Track 5: ChaChing_Draft_95, Yalhalom_Draft_91, Phlyer_95
- Track 6: NOZO_Draft_96, Compostia_Draft_98, Gadget_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, Yalhalom_Draft_91, NOZO_Draft_96, Compostia_Draft_98, Phlyer_95, OrangeOswald_94, Gadget_93, Mortoellus_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, Yalhalom_Draft_91, NOZO_Draft_96, Compostia_Draft_98, Phlyer_95, OrangeOswald_94, Gadget_93, Mortoellus_Draft_96, ChaChing_Draft_95
  Percent with start 3 called: 55.0%
Blast Comparisons
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 details, click to show alignments
### Sequences producing significant alignments:

Select: All None Selected: 0

<table>
<thead>
<tr>
<th>Description</th>
<th>Max score</th>
<th>Total score</th>
<th>Query cover</th>
<th>E value</th>
<th>Ident</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>gp12 [Mycobacterium phage Timshell]</td>
<td>1047</td>
<td>1047</td>
<td>99%</td>
<td>0.0</td>
<td>90%</td>
<td>AEJ92326.1</td>
</tr>
<tr>
<td>terminase [Mycobacterium phage Obama12]</td>
<td>1045</td>
<td>1045</td>
<td>99%</td>
<td>0.0</td>
<td>88%</td>
<td>YP_009007203.1</td>
</tr>
<tr>
<td>terminase [Mycobacterium phage Dhanush]</td>
<td>1045</td>
<td>1045</td>
<td>99%</td>
<td>0.0</td>
<td>88%</td>
<td>AGK87127.1</td>
</tr>
<tr>
<td>gp11 [Mycobacterium phage Flux]</td>
<td>1044</td>
<td>1044</td>
<td>99%</td>
<td>0.0</td>
<td>87%</td>
<td>AFL47832.1</td>
</tr>
<tr>
<td>terminase [Mycobacterium phage Nyxis]</td>
<td>1044</td>
<td>1044</td>
<td>99%</td>
<td>0.0</td>
<td>88%</td>
<td>YP_009005850.1</td>
</tr>
<tr>
<td>terminase [Mycobacterium phage HINdeR]</td>
<td>1043</td>
<td>1043</td>
<td>99%</td>
<td>0.0</td>
<td>89%</td>
<td>YP_008051863.1</td>
</tr>
<tr>
<td>gp10 [Mycobacterium phage Arturo]</td>
<td>1043</td>
<td>1043</td>
<td>99%</td>
<td>0.0</td>
<td>88%</td>
<td>AFU20464.1</td>
</tr>
<tr>
<td>terminase [Mycobacterium phage Kampy]</td>
<td>1043</td>
<td>1043</td>
<td>99%</td>
<td>0.0</td>
<td>87%</td>
<td>YP_009031871.1</td>
</tr>
<tr>
<td>gp11 [Mycobacterium phage Shaka]</td>
<td>1043</td>
<td>1043</td>
<td>99%</td>
<td>0.0</td>
<td>87%</td>
<td>AEF57321.1</td>
</tr>
<tr>
<td>gp11 [Mycobacterium phage Sabertooth]</td>
<td>1042</td>
<td>1042</td>
<td>99%</td>
<td>0.0</td>
<td>87%</td>
<td>AFU20551.1</td>
</tr>
<tr>
<td>terminase [Mycobacterium phage BelluTerra]</td>
<td>1040</td>
<td>1040</td>
<td>99%</td>
<td>0.0</td>
<td>87%</td>
<td>YP_009005569.1</td>
</tr>
<tr>
<td>gp11 [Mycobacterium phage iCleared]</td>
<td>1040</td>
<td>1040</td>
<td>99%</td>
<td>0.0</td>
<td>87%</td>
<td>AFL46618.1</td>
</tr>
<tr>
<td>gp11 [Mycobacterium phage Peaches]</td>
<td>1040</td>
<td>1040</td>
<td>99%</td>
<td>0.0</td>
<td>87%</td>
<td>YP_003358714.1</td>
</tr>
<tr>
<td>gp11 [Mycobacterium phage Wile]</td>
<td>1038</td>
<td>1038</td>
<td>99%</td>
<td>0.0</td>
<td>87%</td>
<td>YP_009014094.1</td>
</tr>
<tr>
<td>terminase [Mycobacterium phage QuinnKiro]</td>
<td>1023</td>
<td>1023</td>
<td>99%</td>
<td>0.0</td>
<td>86%</td>
<td>AIS73685.1</td>
</tr>
</tbody>
</table>
### Alignments

#### gp12 [Mycobacterium phage Timshel]
**Sequence ID:** ghAFJ93256.1  **Length:** 564  **Number of Matches:** 1

<table>
<thead>
<tr>
<th>Range 1: 4 to 564</th>
<th>GenPep</th>
<th>Graphics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Score</strong></td>
<td>Expect</td>
<td>Method</td>
</tr>
<tr>
<td>1047 bits (270.0)</td>
<td>504/562 (90%)</td>
<td>531/562 (94%)</td>
</tr>
</tbody>
</table>

**Query 5**
```
YLSTPFLPQQPFPKIFWPWVCLGCHSGWALPFTLGFVNLNLAEYRSAGPQGFPLB
YLQFLQPQPHIPFPWVCLGCHSGWALPFTLGFVNLNLAEYRSAGPQGFPLB
```
**Sbjct 4**
```
YLSTPFLPQQPFPKIFWPWVCLGCHSGWALPFTLGFVNLNLAEYRSAGPQGFPLB
YLQFLQPQPHIPFPWVCLGCHSGWALPFTLGFVNLNLAEYRSAGPQGFPLB
```
**Query 65**
```
QAÆLILWTVADGEAYRGGVLRKKGWNLIPCLCAALVLGCPVFASHIDFKPFPV
QAÆLILWTVADGEAYRGGVLRKKGWNLIPCLCAALVLGCPVFASHIDFKPFPV
```
**Sbjct 64**
```
QAÆLILWTVADGEAYRGGVLRKKGWNLIPCLCAALVLGCPVFASHIDFKPFPV
QAÆLILWTVADGEAYRGGVLRKKGWNLIPCLCAALVLGCPVFASHIDFKPFPV
```
**Query 115**
```
GKRMHAIATIAVSSQQTSSTPSLFWMISQKMTFYGGLVXNVKXVITYEDGRGIAAMTS
GKRMHAIATIAVSSQQTSSTPSLFWMISQKMTFYGGLVXNVKXVITYEDGRGIAAMTS
```
**Sbjct 124**
```
GKRMHAIATIAVSSQQTSSTPSLFWMISQKMTFYGGLVXNVKXVITYEDGRGIAAMTS
GKRMHAIATIAVSSQQTSSTPSLFWMISQKMTFYGGLVXNVKXVITYEDGRGIAAMTS
```
**Query 173**
```
SPASMEGHRPTVLINETQMVQGPDQGDVMDVQVGNVQSPIKGALCNAMIP
SPASMEGHRPTVLINETQMVQGPDQGDVMDVQVGNVQSPIKGALCNAMIP
```
**Sbjct 184**
```
SPASMEGHRPTVLINETQMVQGPDQGDVMDVQVGNVQSPIKGALCNAMIP
SPASMEGHRPTVLINETQMVQGPDQGDVMDVQVGNVQSPIKGALCNAMIP
```
**Query 215**
```
GVNVTVEAKYHDOQLSGKAVGDYLADEADPFTPVESIPSECEDPGYPEAGIALM
GVNVTVEAKYHDOQLSGKAVGDYLADEADPFTPVESIPSECEDPGYPEAGIALM
```
**Sbjct 214**
```
GVNVTVEAKYHDOQLSGKAVGDYLADEADPFTPVESIPSECEDPGYPEAGIALM
GVNVTVEAKYHDOQLSGKAVGDYLADEADPFTPVESIPSECEDPGYPEAGIALM
```
**Query 235**
```
DLGÆLFARQSGWYVLPLLEILGVNHLRTFVPTEVRSKQFLQVHESWIAPGQWDLRAVTD
+GLÆLFARQSGWYVLPLLEILGVNHLRTFVPTEVRSKQFLQVHESWIAPGQWDLRAVTD
```
**Sbjct 304**
```
DGÆLFARQSGWYVLPLLEILGVNHLRTFVPTEVRSKQFLQVHESWIAPGQWDLRAVTD
DGÆLFARQSGWYVLPLLEILGVNHLRTFVPTEVRSKQFLQVHESWIAPGQWDLRAVTD
```
**Query 345**
```
KAÆLQCDKQRTGFDFDSKDDTALVACRVSIGMFLIAKSNPDEHEEVPREDQAV
LATK+KQRTGFDFDSKDDTALVACRVSIGMFLIAKSNPDEHEEVPREDQAV
```
**Sbjct 364**
```
PLÆLKNHRRITGFDFDSKDDTALVACRVSIGMFLIAKSNPDEHEEVPREDQAV
```
**Query 415**
```
VRSÆFQGVRQQFGVRQADVEDFKEFAVYQVGRDFKKLKINATPGNPVQFAMDFQGQMFQKPDLC
VRSÆFQGVRQQFGVRQADVEDFKEFAVYQVGRDFKKLKINATPGNPVQFAMDFQGQMFQKPDLC
```
**Sbjct 424**
```
VRSÆFQGVRQQFGVRQADVEDFKEFAVYQVGRDFKKLKINATPGNPVQFAMDFQGQMFQKPDLC
VRSÆFQGVRQQFGVRQADVEDFKEFAVYQVGRDFKKLKINATPGNPVQFAMDFQGQMFQKPDLC
```
**Query 485**
```
ERFÆDVAIVHELHHQDQPLQLVQHNLANRHEFTYIDASIREKSDSKKKDAVACVALAF
ERFÆDVAIVHELHHQDQPLQLVQHNLANRHEFTYIDASIREKSDSKKKDAVACVALAF
```
**Sbjct 544**
```
ERFÆDVAIVHELHHQDQPLQLVQHNLANRHEFTYIDASIREKSDSKKKDAVACVALAF
```
**Query 545**
```
GAÆQGQFQMXXXRCAXSVRIR
```
**Sbjct 544**
```
GÆQGQFQMXXXRCAXSVRIR
```

### Download  
- GenPep  
- Graphics

#### terminase [Mycobacterium phage Obama12]
**Sequence ID:** refYP_009607203.1  **Length:** 565  **Number of Matches:** 1

<table>
<thead>
<tr>
<th>Range 1: 1 to 565</th>
<th>GenPep</th>
<th>Graphics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Score</strong></td>
<td>Expect</td>
<td>Method</td>
</tr>
<tr>
<td>1045 bits (270.0)</td>
<td>496/566 (88%)</td>
<td>537/566 (94%)</td>
</tr>
</tbody>
</table>

**Query 1**
```
MALEYLSITPFLPQQPFPKIFVWCLGCHSGWALPFTLGFVNLNLAEYRSAGPQGFPLB
```
**Sbjct 1**
```
MALEYLSITPFLPQQPFPKIFVWCLGCHSGWALPFTLGFVNLNLAEYRSAGPQGFPLB
```
**Query 61**
```
PTÆQARÎLWVADGEAYRGGVLRKKGWNLIPCLCAALVLGCPVFASHIDFKPFPV
PTÆQARÎLWVADGEAYRGGVLRKKGWNLIPCLCAALVLGCPVFASHIDFKPFPV
```
**Sbjct 61**
```
PTÆQARÎLWVADGEAYRGGVLRKKGWNLIPCLCAALVLGCPVFASHIDFKPFPV
PTÆQARÎLWVADGEAYRGGVLRKKGWNLIPCLCAALVLGCPVFASHIDFKPFPV
```
**Query 111**
```
GNVPKRHHWIAVAVQSSQQTNTLSLFVPVQISQMTQEtYDGLVXNVKXVITYEDGRGIAAMTS
```
**Sbjct 110**
```
GNVPKRHHWIAVAVQSSQQTNTLSLFVPVQISQMTQEtYDGLVXNVKXVITYEDGRGIAAMTS
GNVPKRHHWIAVAVQSSQQTNTLSLFVPVQISQMTQEtYDGLVXNVKXVITYEDGRGIAAMTS
```

**See 3 more titles**
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

```
ATATSPASMSNGRPTLVSNETQGNAVGFDGVTNGVMDVDEIENVSKIFGARKLAI
AHIPGNDTVAKAYDHQILLSGKAVDSTLQMYALEAPADTPVSIEPSKEKEDPGEYAE
AQLMGDELARGDSYWLPEELGLSVLTNPVETSRRKFQLNQVAHEDSWIAAPSEWDR
AVTDXKALAKQDDXITLGFDSKSDDWWLACVRSQSDLFLTSKWNPDYFNEEVRED
VDAVYRGRQHVRVNYEAYQDQRWDRFPRKRIAHQGTYGTRAPARMTQCK
ALDCEPRTVAVHHELHDGNFVLRQHVLNARHHTTYDATISINSEKSDSSKIDDAVCA
VLAFGARQDYQNSKRSRSGAKAYVINZ
```

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 define and scores. Click to show alignments

Color Key for Alignment Scores

<table>
<thead>
<tr>
<th>&lt;40</th>
<th>40-50</th>
<th>50-80</th>
<th>80-200</th>
<th>&gt;=200</th>
</tr>
</thead>
</table>

lc111
Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Score (bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sheen-Draft_14, function unknown, 566</td>
<td>1171</td>
<td>0.0</td>
</tr>
<tr>
<td>Timshel_12, Terminase., 564</td>
<td>1053</td>
<td>0.0</td>
</tr>
<tr>
<td>HINdeR_11, terminase, 564</td>
<td>1048</td>
<td>0.0</td>
</tr>
<tr>
<td>Obaman12-draft_10, function unknown, 565</td>
<td>1046</td>
<td>0.0</td>
</tr>
<tr>
<td>LHTSCC_11, Terminase, large subunit, 565</td>
<td>1046</td>
<td>0.0</td>
</tr>
<tr>
<td>Arturo_10, terminase, 565</td>
<td>1046</td>
<td>0.0</td>
</tr>
<tr>
<td>Phightern1804-Draft_10, function unknown, 565</td>
<td>1046</td>
<td>0.0</td>
</tr>
<tr>
<td>Eagle_10, Terminase, 565</td>
<td>1046</td>
<td>0.0</td>
</tr>
<tr>
<td>Camperdownii_Draft_10, function unknown, 565</td>
<td>1046</td>
<td>0.0</td>
</tr>
<tr>
<td>Flux_11, terminase, 565</td>
<td>1045</td>
<td>0.0</td>
</tr>
<tr>
<td>Dhanush_11, terminase, 565</td>
<td>1045</td>
<td>0.0</td>
</tr>
<tr>
<td>Wander_Draft_10, function unknown, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>TygerBlood-Draft_10, function unknown, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>TiroTheta9_11, terminase, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>TinaPeyge-Draft_gp10, function unknown, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>Shaka_11, terminase, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>Sabertooth_11, Terminase, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>Milliski_Draft_10, function unknown, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>Melvin_MELVIN_11, Terminase, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>MeeZee_11, Terminase, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>Medusa_11, terminase, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>Maverick-DRAFT_11, function unknown, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>Lemur_Draft_10, function unknown, 557</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>Kratark_Draft_10, function unknown, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>KFPoly_Draft_10, function unknown, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>Kampy-Draft_10, function unknown, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>Holli-draft_10, function unknown, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>HamSlice-Draft_10, function unknown, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>Gadost_Draft_10, function unknown, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>Funston_Draft_DRAFT_10, function unknown, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
<tr>
<td>Eris_Draft_10, function unknown, 565</td>
<td>1044</td>
<td>0.0</td>
</tr>
</tbody>
</table>
>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  YLSTEPPLLQPQPKHIPVWLCHEDGSWALPYYTLGWGVLNLAYVRSPPPAGGFPITPTE 64
  YL+ PLLQPQPKHIPVWLCHEDGSWALPYYTLGWGVLNLAYVRSPPPAGGFPITPTE 64

Sbjct: 4  YLNPGPLLQPQPKHIPVWLVHEGSWLAPARTLGWVNLAYVRSPPPAGGFPITPTE 63

Query: 65  QARFILWYAVDENGVYAYREGVLRMRHKGDPLCAAILALVEGVPVFASFSDKGNV 124
  QARFILWYAVDENGVYAYREGVLRMRHKGDPLCAAILALVEGVPVFASFSDKGNV 124

Sbjct: 64  QARFILWYAVDENGVYAYREGVLRMRHKGDPLCAAILALVEGVPVFASFSDKGNV 123

Query: 125  GKKHAAAAMITTSASVADQKNTFSLPFFMVIMSKMKTQYGLDVKVYTPFEDGGRKIEATS 184
  GKHAAAAMITTSASVADQKNTFSLPFFMVIMSKMKTQYGLDVKVYTPFEDGGRKIEATS 184

Sbjct: 124  GKKHAAAAMITTSASVADQKNTFSLPFFMVIMSKMKTQYGLDVKVYTPFEDGGRKIEATS 184

Query: 185  SPASMGRNPRLVNETQWGVGDPGNVMDVGAMVDIEGNSKPARKLAICNHAIP 244
  SPASMGRNPRLVNETQWGVGDPGNVMDVGAMVDIEGNSKPARKLAICNHAIP 244

Sbjct: 184  SPASMGRNPRLVNETQWGVGDPGNVMDVGAMVDIEGNSKPARKLAICNHAIP 243

Query: 245  GNTVAEKAHYDWTDDLGSKAVDQGLMYALEAPADTPSEIPSEKDEPEYEGAQLM 304
  GNTVAEKAHYDWTDDLGSKAVDQGLMYALEAPADTPSEIPSEKDEPEYEGAQLM 304

Sbjct: 244  GNTVAEKAHYDWTDDLGSKAVDQGLMYALEAPADTPSEIPSEKDEPEYEGAQLM 303

Query: 305  DGLEVARKSYPWLEYQCVSTKRKFLNQVNAHDESWIPAEWVLRVATD 364
  +GLEVARKSYPWLEYQCVSTKRKFLNQVNAHDESWIPAEWVLRVATD 364

Sbjct: 304  EGLEIRAKGYSWSLEPGCAVSTKRKFLNQVNAHDESWIPAEWVLRVATD 363

Query: 365  KALALQDDRITLFDDGSKSSDDWTALVACRVSQGMLFLISWNPEDYPHEPEEVPREDVDAV 424
  L+X+DRTLFDDGSKSSDDWTALVACRVSQGMLFLISWNPEDYPHEPEEVPREDVDAV 423

Sbjct: 364  PLFKLKDNRITLFDDGSKSSDDWTALVACRVSQGMLFLISWNPEDYPHEPEEVPREDVDAV 423

Query: 425  VRSARFOYVVDGFRAVRKEAEYDQWQDFKRLKNKNATPNPVAFDNRGQKFRALDC 484
  VRSARFOYVVDGFRAVRKEAEYDQWQDFKRLKNKNATPNPVAFDNRGQKFRALDC 484

Sbjct: 424  VRSARFOYVVDGFRAVRKEAEYDQWQDFKRLKNKNATPNPVAFDNRGQKFRALDC 483

Query: 485  ERFVDAVIELHLDHGNPNLQVRQNLARNRHPITYDAISIRKESDKSSKIDAAVCVALAF 544
  ERFVDAVIELHLDHGNPNLQVRQNLARNRHPITYDAISIRKESDKSSKIDAAVCVALAF 543

Sbjct: 484  ERFVDAVIELHLDHGNPNLQVRQNLARNRHPITYDAISIRKESDKSSKIDAAVCVALAF 543

Query: 545  GARQDVQMKSKHRSRQAKAIVR 566
  G+RQDY MSKKHR G AVIR 566

Sbjct: 544  GARQDVQMKSKHRSRQAKAIVR 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  YLSTEPPLLQPQPKHIPVWLCHEDGSWALPYYTLGWGVLNLAYVRSPPPAGGFPITPTE 64
  YL+ PLLQPQPKHIPVWLCHEDGSWALPYYTLGWGVLNLAYVRSPPPAGGFPITPTE 64

Sbjct: 4  YLNPGPLLQPQPKHIPVWLVHEGSWLAPARTLGWVNLAYVRSPPPAGGFPITPTE 63

Query: 65  QARFILWYAVDENGVYAYREGVLRMRHKGDPLCAAILALVEGVPVFASFSDKGNV 124
  QARFILWYAVDENGVYAYREGVLRMRHKGDPLCAAILALVEGVPVFASFSDKGNV 124

Sbjct: 64  QARFILWYAVDENGVYAYREGVLRMRHKGDPLCAAILALVEGVPVFASFSDKGNV 123

Query: 125  GKKHAAAAMITTSASVADQKNTFSLPFFMVIMSKMKTQYGLDVKVYTPFEDGGRKIEATS 184
  GKHAAAAMITTSASVADQKNTFSLPFFMVIMSKMKTQYGLDVKVYTPFEDGGRKIEATS 184

Sbjct: 124  GKKHAAAAMITTSASVADQKNTFSLPFFMVIMSKMKTQYGLDVKVYTPFEDGGRKIEATS 184

Query: 185  SPASMGRNPRLVNETQWGVGDPGNVMDVGAMVDIEGNSKPARKLAICNHAIP 244
  SPASMGRNPRLVNETQWGVGDPGNVMDVGAMVDIEGNSKPARKLAICNHAIP 244

Sbjct: 184  SPASMGRNPRLVNETQWGVGDPGNVMDVGAMVDIEGNSKPARKLAICNHAIP 243

Query: 245  GNTVAEKAHYDWTDDLGSKAVDQGLMYALEAPADTPSEIPSEKDEPEYEGAQLM 304
  GNTVAEKAHYDWTDDLGSKAVDQGLMYALEAPADTPSEIPSEKDEPEYEGAQLM 304

Sbjct: 244  GNTVAEKAHYDWTDDLGSKAVDQGLMYALEAPADTPSEIPSEKDEPEYEGAQLM 303

Query: 305  DGLEVARKSYPWLEYQCVSTKRKFLNQVNAHDESWIPAEWVLRVATD 364
  +GLEVARKSYPWLEYQCVSTKRKFLNQVNAHDESWIPAEWVLRVATD 364

Sbjct: 304  EGLEIRAKGYSWSLEPGCAVSTKRKFLNQVNAHDESWIPAEWVLRVATD 363

Query: 365  KALALQDDRITLFDDGSKSSDDWTALVACRVSQGMLFLISWNPEDYPHEPEEVPREDVDAV 424
  L+X+DRTLFDDGSKSSDDWTALVACRVSQGMLFLISWNPEDYPHEPEEVPREDVDAV 423

Sbjct: 364  PLFKLKDNRITLFDDGSKSSDDWTALVACRVSQGMLFLISWNPEDYPHEPEEVPREDVDAV 423

Query: 425  VRSARFOYVVDGFRAVRKEAEYDQWQDFKRLKNKNATPNPVAFDNRGQKFRALDC 484
  VRSARFOYVVDGFRAVRKEAEYDQWQDFKRLKNKNATPNPVAFDNRGQKFRALDC 484

Sbjct: 424  VRSARFOYVVDGFRAVRKEAEYDQWQDFKRLKNKNATPNPVAFDNRGQKFRALDC 483

Query: 485  ERFVDAVIELHLDHGNPNLQVRQNLARNRHPITYDAISIRKESDKSSKIDAAVCVALAF 544
  ERFVDAVIELHLDHGNPNLQVRQNLARNRHPITYDAISIRKESDKSSKIDAAVCVALAF 543

Sbjct: 484  ERFVDAVIELHLDHGNPNLQVRQNLARNRHPITYDAISIRKESDKSSKIDAAVCVALAF 543

Query: 545  GARQDVQMKSKHRSRQAKAIVR 566
  G+RQDY MSKKHR G AVIR 566

Sbjct: 544  GARQDVQMKSKHRSRQAKAIVR 564
Things to do often:

- Save .dnam5 file often
- Save .dnam5 file as a new name. (Then don’t save the old named one.)
Let’s get started!

1. Gather Data
2. Basic DNA Master functions
3. Gene Assignments
4. Functional Assignments
Mycobacteriophage CrystalP

Named for Crystal Petrone, Administrative Assistant

Genome length: 76483
Physical ends: 9 bp 3’ overhang
GC% content: 63.0
Gathering Data

• Obtain your genome (phagesdb.org)
• BlastN genome
• Use DNA Master to obtain Glimmer, GeneMark, and tRNA (Aragorn) data
• Obtain GeneMark data on web (trained on M. smeg and Mtb)
• BlastP NCBI using DNA Master
  • Additional tRNA data
  • Phamerator data
  • Starterator data
  • HHPred data
  • CDD data
For tonight after dinner

- We are using DNA Master
  - Current Build is 2478
  - Make sure to set Preferences
- Bookmark the following websites:
  - Mycobacteriophage Database & BLASTp
    [http://www.phagesdb.org](http://www.phagesdb.org)
  - NCBI BLASTp
  - GeneMark
    [http://exon.gatech.edu/genemark_prok_gms_plus.cgi](http://exon.gatech.edu/genemark_prok_gms_plus.cgi)