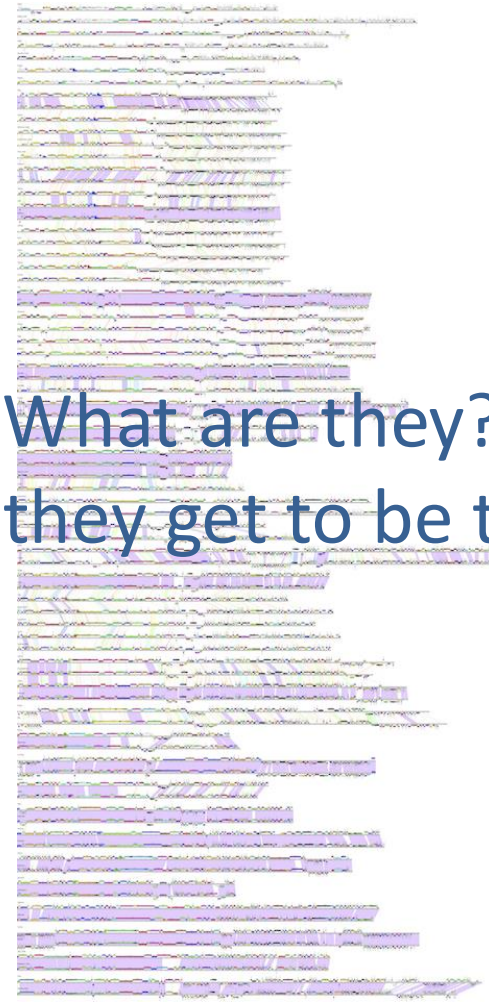
Predicting Genes in Actinobacteriophages

2024 Phage Genomics Workshop Training

SEA-PHAGES Cohort 17

Deborah Jacobs-Sera



What are they?

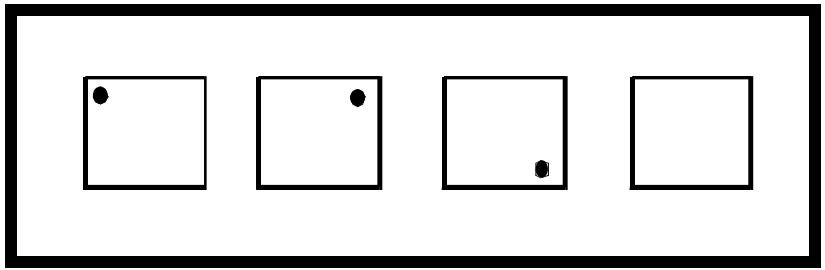
How did they get to be that way?

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?

5081

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?

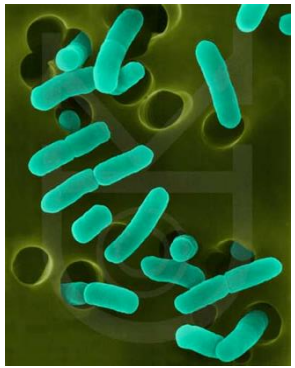
FASTA

How do you make sense of the nucleotide sequence?

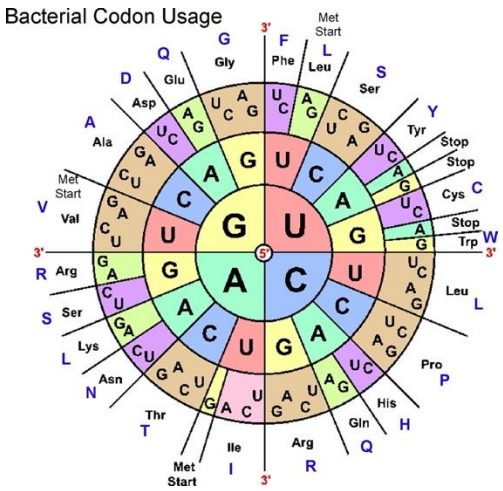
Convert to genes

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 ….

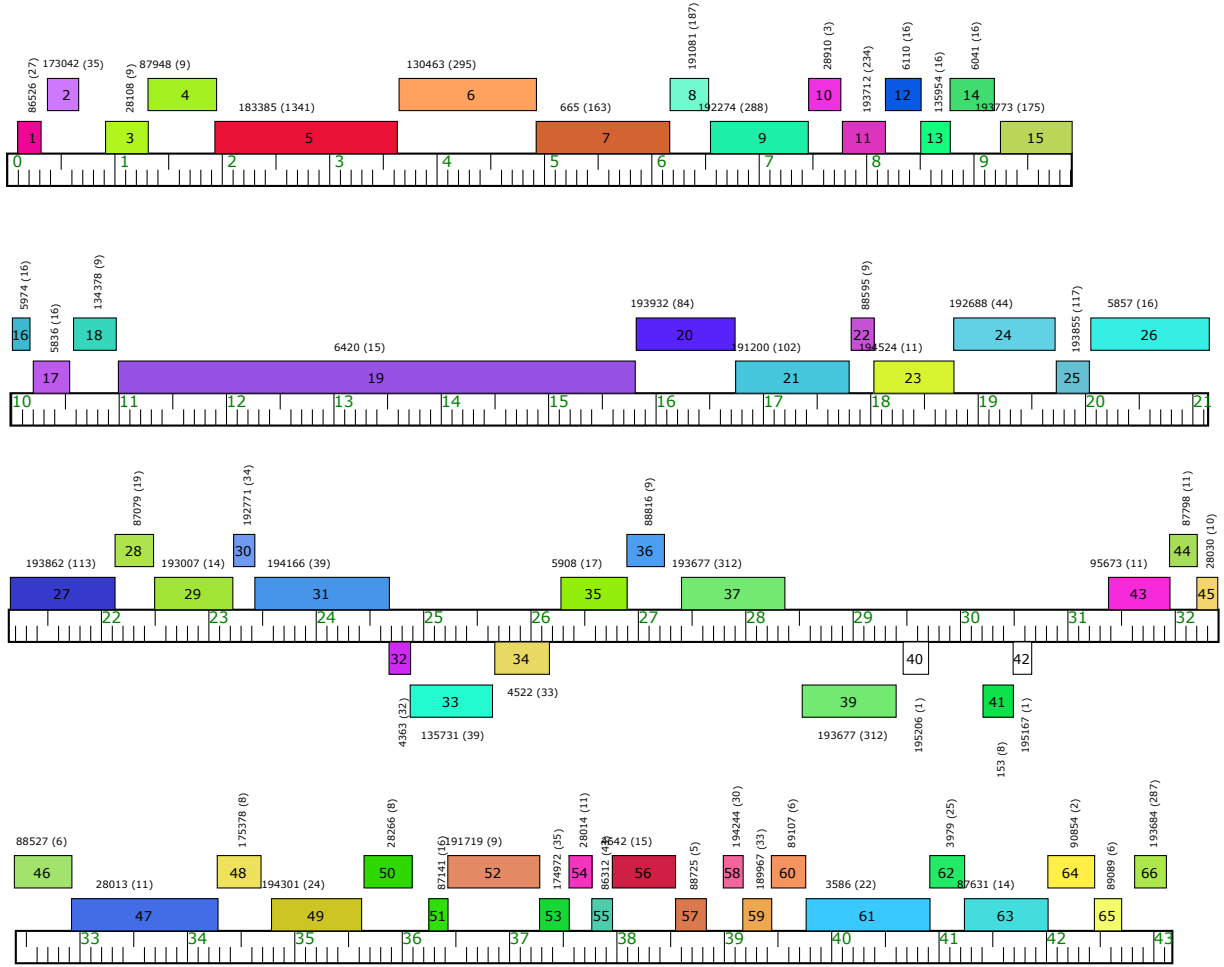
GGA CCT CTC GCC ….

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

QuinnAvery_Draft (FF)



How to find predictable genes?

Use programs that predict coding potential.

- Programs use math to detect coding potential
  - 2 programs widely used: Glimmer & GeneMark
  - These are algorithms that use interpolated Markov models for patterns in the order of DNA's nucleotides
  - Use a sample of the provided genome, find largest ORF, look for patterns and apply it to the target (your provided genome)

GLIMMER

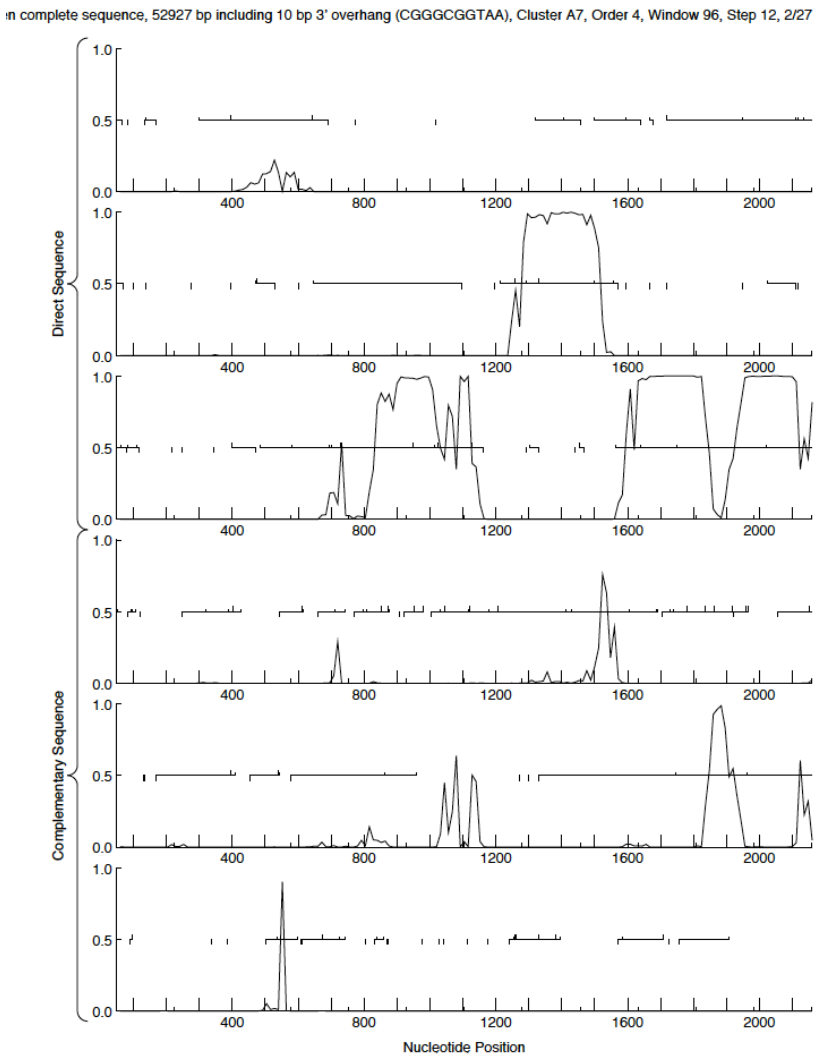
```
GLIMMER (ver. 3.02; iterated)  predictions:
 orfID      start     end  frame  score
--------    -----    -----  --    -----
>Sheen complete sequence, 52927 bp including 10 bp 3' overhang (CGGGCGGTAA), Cluster A7
orf00001      732     1166  +3    11.17
orf00002     1259     1576  +2    14.04
orf00004     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

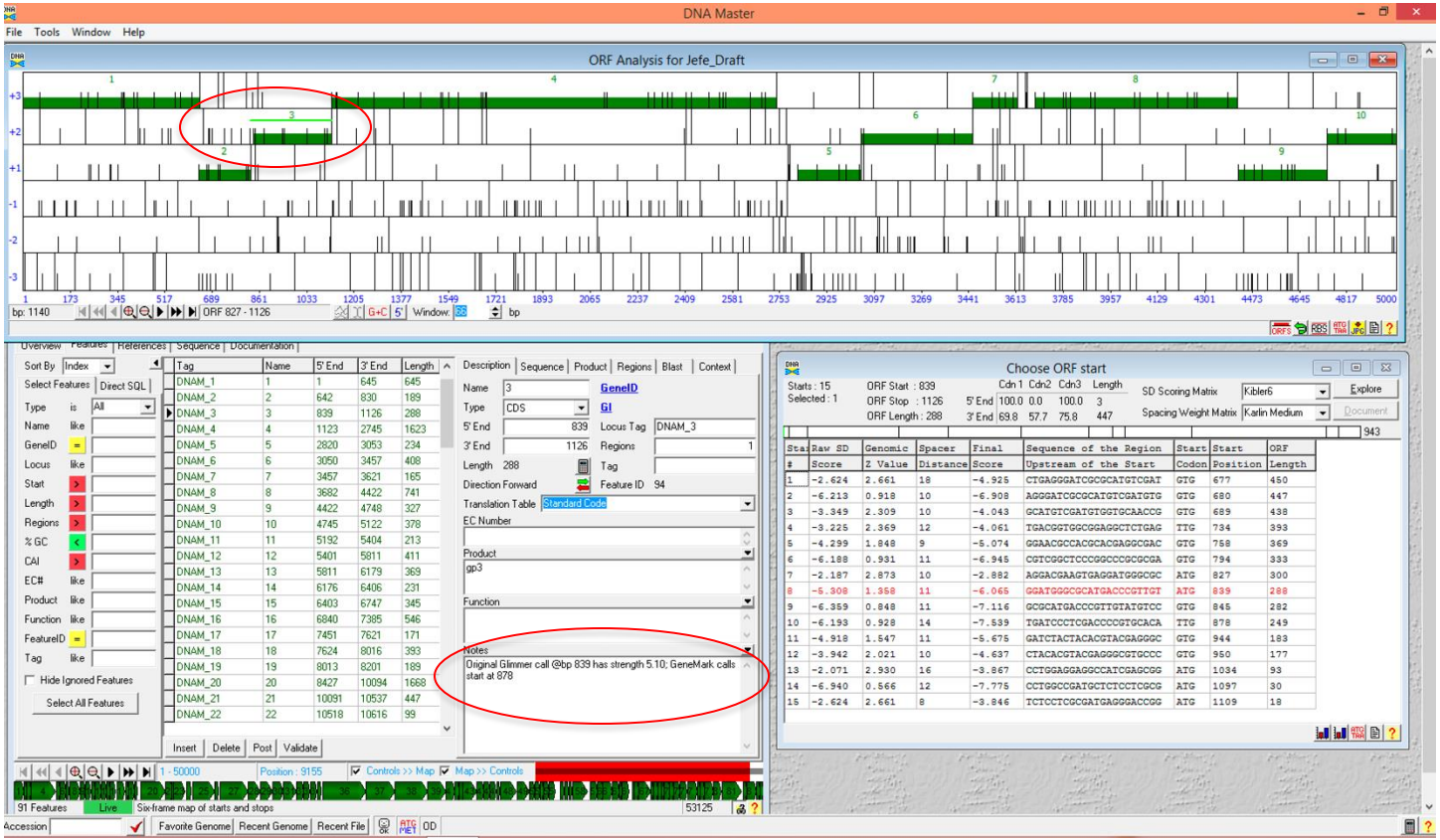
Microbial Genome Annotation Tools

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

GUIDING PRINCIPLES
OF BACTERIOPHAGE GENOME ANNOTATION

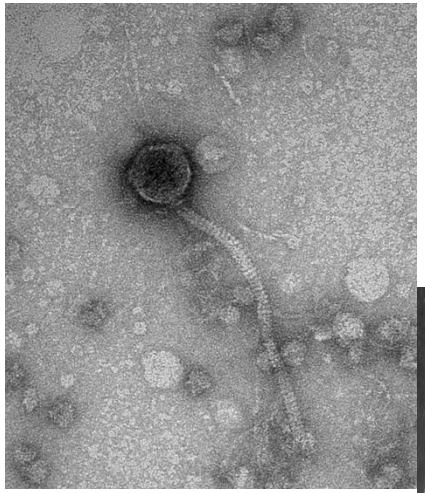
- Found in “Phage Annotation, Genomics and Data Interpretation” Section of the Bioinformatics Guide
- 15 Key Directives
- Read for tomorrow
  https://seaphagesbioinformatics.helpdocsonline.com/guiding-principles

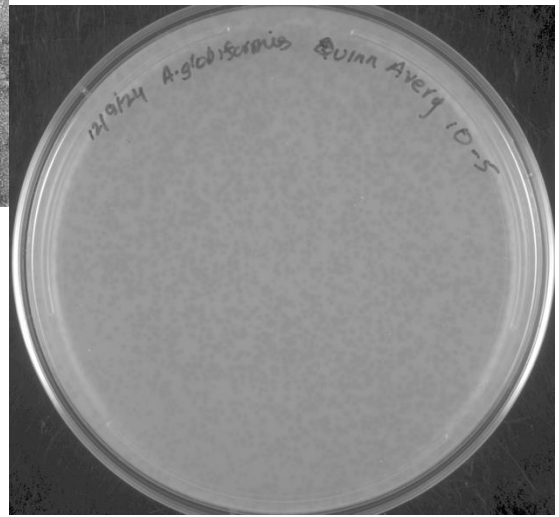
Let’s get started!

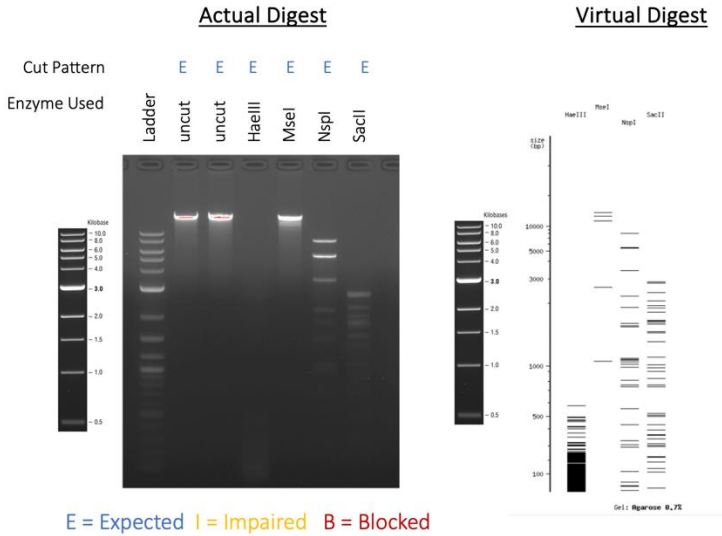
1. Gather Data
2. Auto-annotate in DNA Master
3. Gene Calling
4. Functional Assignments

Arthrobacter phage QuinnAvery

Bacteriophage: QuinnAvery
Cluster: FF
Host Bacterium: *Arthrobacter globiformis* B-2979







Found by Jennifer Ingram and worked on by R. Cass, G, Asuresh, H. Gesinski, A. Nene at the Phage Discovery Workshop (17A), HHMI.
Lysogen data:
https://qubeshub.org/publications/5037/1

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

Once you have Build 2705, turn off updates. Go to Preferences -> Timed Events -> **unclick** the first entry "Automatic checks for DNA Master updates ...."