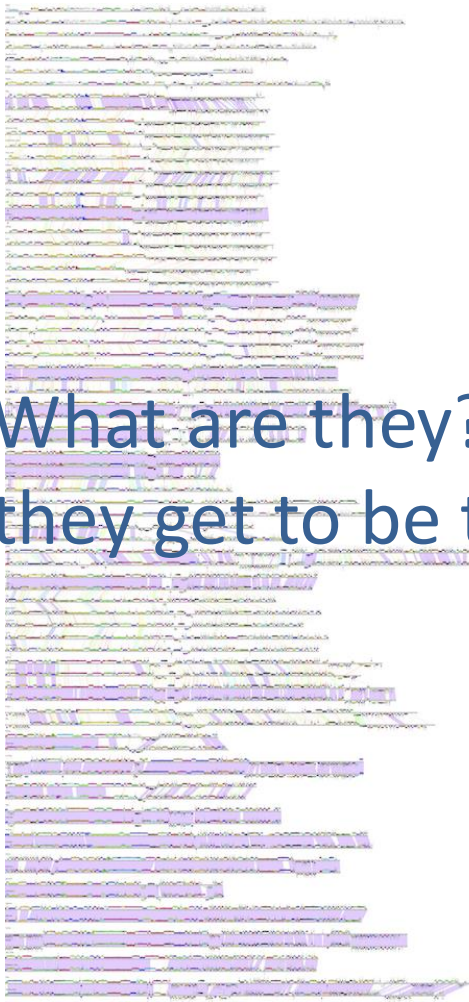


Predicting Genes in Actinobacteriophages

2024 Phage Genomics Workshop Training

SEA-PHAGES Cohort 17

Deborah Jacobs-Sera



What are they?

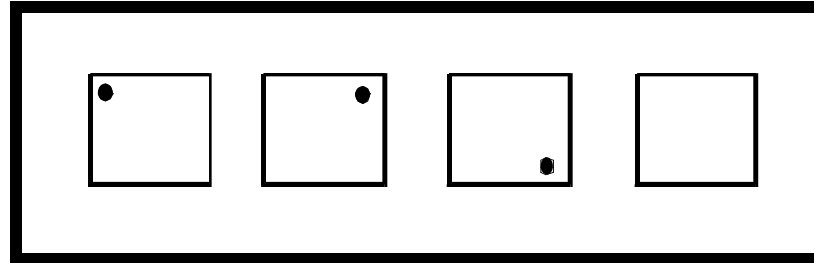
How did they get to be that way?

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

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?

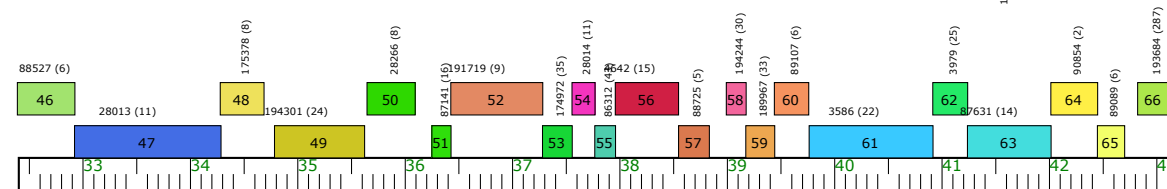
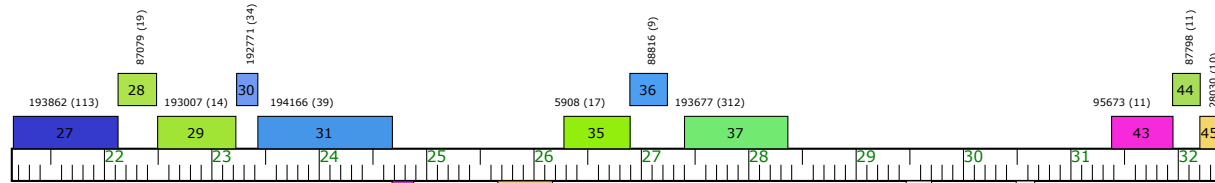
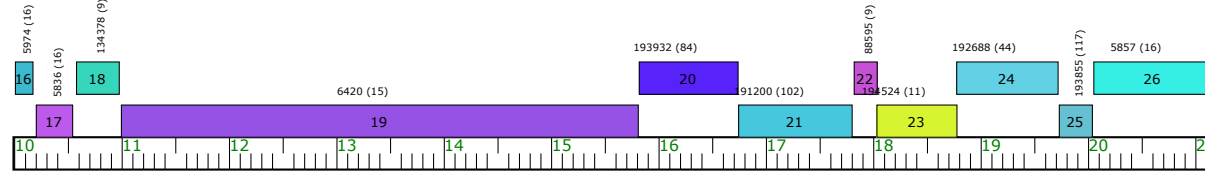
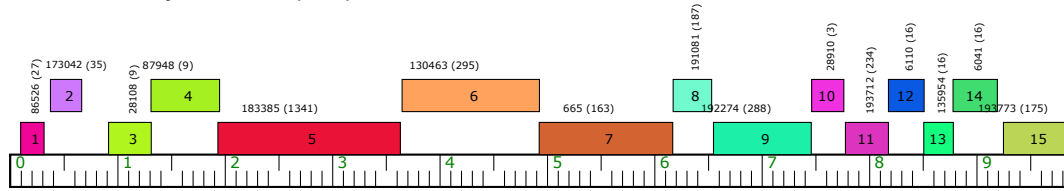
>Echild complete sequence, 53159 bp including
10bp overhang (CGGTCGGTTA), Cluster A2

TGCGGCCGCCCATCCTGTACGGGTTTCCAAGTCGATCGGAGTCCCGAGC
CGGCGCAGGAGCGCCTCACCCAGCCTCTGTGCGCCCCCAGGACGCAAGAT
CCCCGCTCACGCGGGTAGTTGTATGGGCTAATCGGCAAACGGCCTCTGAG
GCCGCGAGACCAATGTCACACCAGGTGGTGGATGTTATTGACGCACGCGT
CCGTTAAGAGGACATGGCCTAGGTATGGCTACCCAACTTAGATTCAAAA
CCAGTCCCCTGGCCCCCGTCGTCGGTGTTGCCGCTCCTGGCGGGCGGGGG
CCAGGTCCACCACCGCAGGGAGGACCGCCATGAAGATCATCCGCTCGCTC
GCCGGGGCGCTCGCACTCATTGCGATCGCCGCCCGAGGGCGGCTGGGATGC
GGAAATCTACGAGCCGTGGGATGAGGACGAATACCTCCTATAGTGATCTA
CGCCACTTGCTCGGTGGGTGTCAAGTGATACTCATGTATCTAGTTATTGA
GGCCTAAAGGCCCGAATAAGAGCCGCACAGGCGGCTCTCTAAGAGCGCC
CACTAGGGCGCTCGAAGTAATACCGGCCTTGAGGGCCGGTTATCTGACCC
GGCAACCGCCGGGTCTTCTGCCGCGCCAGTGGCGCGGCTCATAGAAGGG
GTGAGGCAACCGTGTACGGCACTCGCTCGAGTGCCTACTGGGCCTCGCAG
CCGGGGAAGTTCGACGTTCTGAACCTGCGGATGACGTTCCCGAGCACGTC

Six-Frame Translations

```
C R F W S V R N P G V S R F P R N G V * P A V F A S T C * F P K W E
A D F G L Y G T R G F R G F P P E M G S D L R F S P A L L V D S R N G
Q I L V C T E L P G G F A V S P K W G L T C G F R Q H L L I P E M G
1 TGCAGATTTGGCTGTACGGAAACCGGGGGTTTCGGGTTCCCGGAATGGGCTGACCTGCGGTTTCGCCAGCAGCTTGGTATTCCCGAATGGG
ACGCTTAAACACAGATGCCTTTGGGCCCCAAAGCCAAAGGGGTTACCCAGACTGGACGCCAAAGCGGCTCGTGAACAATGAAGGGCTTACCC
O L N Q D T R F G P T E R N G R F P T O G A T K A L V Q N G F H
A E K P R Y P V R P R N K G S I P D S R R N E G A S S S E R P P
S C I K T Q V S G P P K A T E G F H P R R V Q P K R W C K N I P G S I P
E V I M P P V P K D P S V R A R R R N K S A T R A T L S A D H D V V
R K S S C H L Y L K I L L C V L V A I S L R R G L R C L R I M M W W
G S H H A T C T * R S F C A C S S Q * V C D A G Y V V C G S * C G
101 AGGAAGTCATCATGCCACCTGACTAAAGATCCTTCTGTGCGTGTCTGCGCAATAGTCTGCCAGCGGGGTACGTTGTCTGGGATCATGATGGT
TCCTTCATAGTACGGTGGACATGGATTTCTAGGAAGACAGCCAGGACCGCTTATCAGACGCTGCCCGGATGCAACAGACCGCTAGTACTACACCA
S S T M M G G T G L S G E T R A R R L L D A V R A V N D A S * S T E T
L F D D D H W R Y R F I R R H T S T A I L R R R P S R O R R I H I H
P L * * A V Q V * L D K Q A H E D C Y T Q S A P * * T T Q P D H P
A P D L P D G V A W H P L T V R W W N D I W A S P M A P E Y T D S
L Q T C R M V L R G I R * R C A G G M T F G R R R W P P R S T Q T R
G S R L A G W C C V A S V D G A L V E * H L G V A D G P G V H R L C
201 GGCTCCAGACTTGGCGGATGGTGTGGTGGCATCGTTCAGCGTGGCTGGTGAATGACATTTGGCGTGGCGGATGGCCCCGGAGTACACAGACTCG
CCGAGGCTGAACGGCTACACAAACGACCGTAGGCAACTGCCAGCGACCTTACTGTAACCCGAGCGGCTACCGGGGCTCATGGGCTGAGG
A G S K G S P T A H C G N V T R Q H F S M Q A D G I A G S Y V S E
H S W V Q R I T N R P M R Q R H A P P I V N P R R R H G R L V C V R
P E L S A P H H Q T A D T S P A S T S H C K P T A S P G P T C L S
D I N G L F R V A M L Y N D F W T A D N A K A R A E A Q V R L E K A
I S T G C S V W R C S I T I F G L P I T R R R V R R L R F G W R
Y Q R V V P C G D A L * R F L D C R * R E G A C G G S G S V G B G
301 GATATCAACGGTGTTCGTTGCGGATGCTATACGATTTTGGACTCCGATAACCGAAGCGCGTGGCGGCTCAGGTCGGTGGAGAGC
CTAAGTGGCCAAAGGCACACCGGTACGAGATATGCTAAAAAAGCTGACCGCTATTGGCTTCCGGCAGCGCTCCGAGTCCAAAGCTTCTTC
S I L P N N R T A I S * L S K Q V A S L A F A R A S A * T R N S P
I D V P Q E T H R H E I V I K P S G I V R L R T R L S L N P Q L L
P Y * R T T G H P S A R Y R N K S Q R Y R S P A H P P E P T P S P
D T D Y G T N P L A R R R L E W O I E A T E D S K A K G S K R R K
P T L I M G R I R W L V A V W S G R L R R P R I R R L R G R S G G S
R H * L W D E S V G S P F G V A D * C D R G F E F G V E A A E
401 CCGACTGATTATGGGACGATCCGTTGGCTCGTCCCGTTTGGAGTGGCAGATTGAGGCGACCGAGGATTCGAAGGCTAAGGGCTCGAAGCGGGGAA
GGCTGACGATACCTCCTTAGGCAACCGGACCGGCAAACTGCGCTAATCCCGGCTCTAGGCTCCGATTCGAGTTCGAGGCTCCCGGCTT
A S V S * P V F G N A R R K S H C I S A V S S E F A L P D F R E
G V S I I P R I R Q S T A T Q L P L N L R G L I R L S L P R L P P
R C Q N H S S D T P E D G N P T A S Q P S R P N S P * P T S A A S
S E A A P V C P P E P G D D P R L K L V T * R P Y G F A G V C C
L R L R R C A H R S P L V T I L V * S L * P D G L M A V I Q V P A V
V * G C A G V P T G A W * R S S F E A C D L T A L W L F P C R S I L W
501 GTCTGAGGCTGGCCGGTGGCCACCGGAGCTGGTACGATCCTCGTTGAAGCTTGTGACCTGACCGGCTTATGGCTGTGGTGGCTGCTG
CAGACTCCGACCGGCCACAGGGTGGCTCGGACCACTGCTAGGAGCAAACTTCGAACACTGGACTGCCGAATACCGACAAAAGCTCCAGGAGCAC
D S A A G T H G G S G P S S G R K F S T V Q R G * P Q K A P G Q Q
L R L S R R H A W R L R T V I R T Q L K H G S P R I A T K C T G A T
T Q P Q A P T G V P A Q H R D E N S A Q S R V A K H S N Q L D R S
G F S V S Y V G S A G V * L H * G S D G V R P W L I V G A G R T S R
D L A F P T L G P Q V C D F I E D R M V F G P G S L S G Q A A R L
I * R F L R W V R R C V T S L R I G W C S A L A H C R G R P H V S
601 GGATTCAGCTTTCTACTGTTGGTCCGCGAGTGTGACTTCATTGAGGATCGGATGGTGTTCGGCCCTGGCTCATTGTGGGGGAGCCGACCTCTC
CCTAAATCGAAAGGATGCAACCCAGGCTCCACACACTGAAGTAACTCTAGCCTACCAAGCGGGGACCGAGTAAACAGCCCGCTCCGGCTGCAGAG
```

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 (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

```

GLIMMER

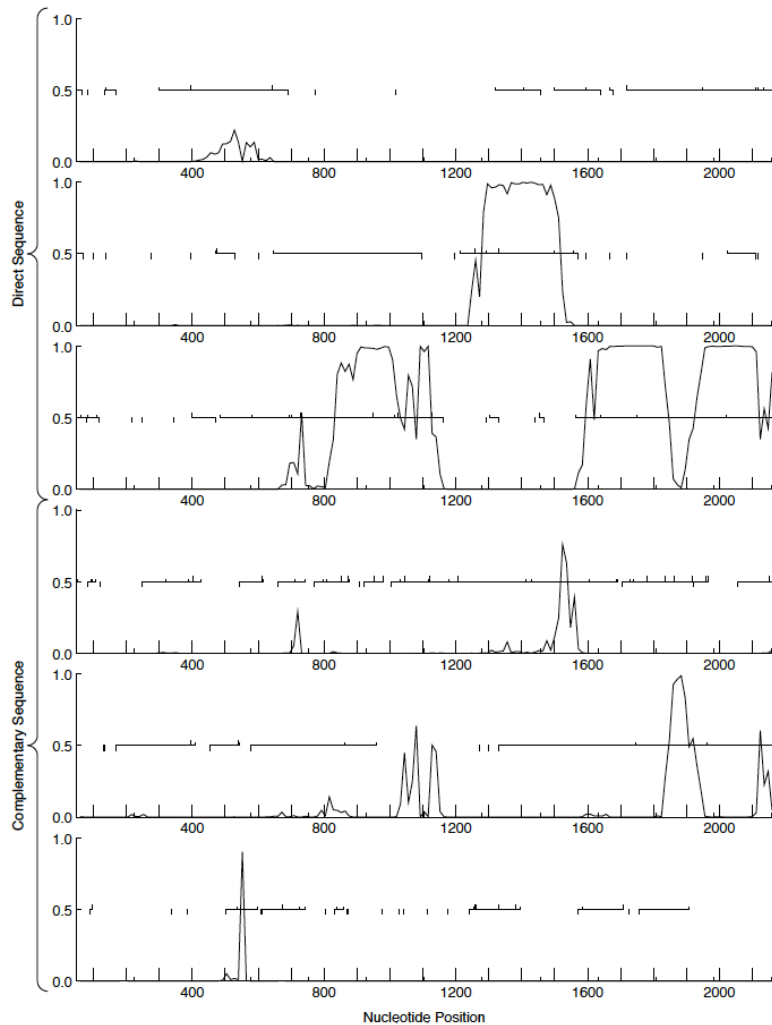


http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi

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 Model(ER)) 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 173 345 517 689 861 1033 1205 1377 1549 1721 1893 2065 2237 2409 2581 2753 2925 3097 3269 3441 3613 3785 3957 4129 4301 4473 4645 4817 5000

ORF 827 - 1126

Overview Features | References | Sequence | Documentation

Sort By Index

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

Select Features: Direct SQL

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

Insert Delete Post Validate

Choose ORF start

Start: 15 ORF Start: 839 Cdn1 Cdn2 Cdn3 Length
 Selected: 1 ORF Stop: 1126 5' End: 100.0 0.0 100.0 3
 ORF Length: 288 3' End: 69.8 57.7 75.8 447

SD Scoring Matrix Kibler6 Explore
 Spacing Weight Matrix Karlin Medium Document

Sta	Raw SD	Genomic	Spacer	Final	Sequence of the Region	Start	Start	ORF
#	Score	Z Value	Distance	Score	Upstream of the Start	Codon	Position	Length
1	-2.624	2.661	18	-4.925	CTGAGGATCGCCGATGTGAT	GTG	677	450
2	-6.213	0.918	10	-6.908	AGGGATCGCCATGTGATGTG	GTG	680	447
3	-3.349	2.309	10	-4.043	GCATGTGATGTGTTGCAACCG	GTG	689	438
4	-3.225	2.369	12	-4.061	TGACGTTGCCGAGGCTCTGAG	TTG	784	393
5	-4.299	1.848	9	-5.074	GGAACCCACCCACAGGCGCAC	GTG	768	369
6	-6.188	0.931	11	-6.945	CGTCGCTCCCGCCCGCCGCA	GTG	794	333
7	-2.187	2.873	10	-2.882	AGGACGAGTGGAGTGGCGCC	ATG	827	300
8	-5.308	1.358	11	-6.065	GGATGGGCGCATGACCCGTTT	ATG	839	288
9	-6.359	0.848	11	-7.116	CGCCATGACCCGTTGATGTCC	GTG	845	282
10	-6.193	0.928	14	-7.539	TGATCCCTGACCCCGTGACCA	TTG	878	249
11	-4.918	1.547	11	-5.675	GATCTACTACAGTADGAGGCC	GTG	944	183
12	-3.942	2.021	10	-4.637	CTACAGTACAGGCGCGTCCCG	GTG	950	177
13	-2.071	2.930	16	-3.867	CCTGGAGGAGCCATGAGCGGG	ATG	1034	93
14	-6.940	0.566	12	-7.775	CCTGGCGCATGCTCTCTCGGG	ATG	1097	30
15	-2.624	2.661	8	-3.846	TCTCTCCGATGAGGACCGGG	ATG	1109	19

Notes: Original Glimmer call @bp 839 has strength 5.10. GeneMark calls start at 678

Position: 9155

31 Features Live Six-frame map of starts and stops

Accession Favorite Genome Recent Genome Recent File ATG ME1 OD

GUIDING PRINCIPLES OF BACTERIOPHAGE GENOME ANNOTATION

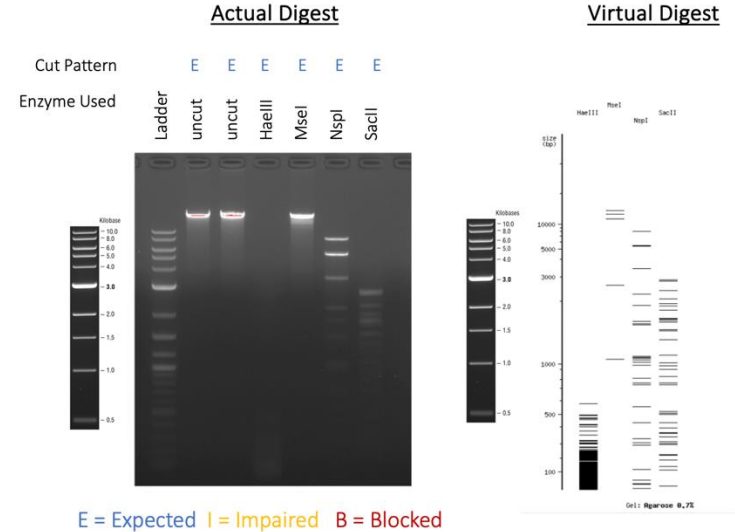
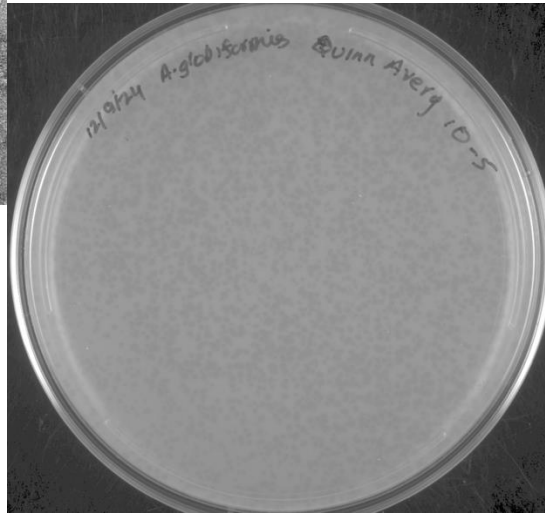
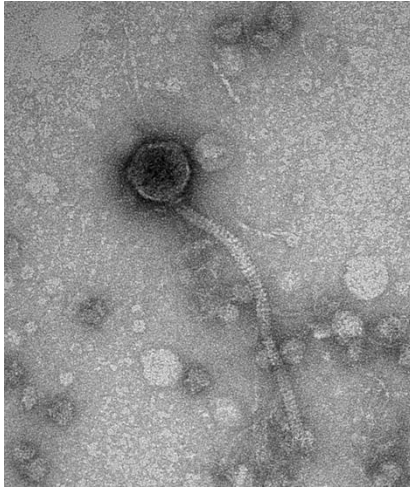
- Found in “Phage Annotation, Genomics and Data Interpretation” Section of the Bioinformatics Guide
- 15 Key Directives
- Read for tomorrow
<https://seaphagesbioinformatics.helpdocsonline.com/guiding-principles>

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 ←
- Automatically Annotating Your Genome
 - Auto-annotation ←

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 ..."

← Tonight's reading

Complete Genome Blastp in DNA Master ←

