Predicting Genes in Actinobacteriophages

2022 Genomics Workshop Training SEA-PHAGES Cohort 15 Deborah Jacobs-Sera

والاستعادة فيعنني الشعائية العجر ومحاصبه المتحاذ فالمتجها والملحا وماحته متعملة فاستحط العري حالات أرضي فتحسب متدار والتتحد سيلتهم ري من من محمد المربي المسلح محمد المربي ا والقاد ويتوقع والمتار والايتناء ومحاوجتها ومحاد والمتحد والمتحاد الم Contraction and the second the state of the s and a second - ----and the second and the second s and have the contraction of the The second se --------------and the second sec and the second of the second s and the second s -----and the second s What are they?

How did they get to be that way?

and the second sec - interesting - management - first and the second second and the second - ----and a second and a s and the second in the second se and the second state and the s - marine and in the second the second se a second and a second and a second a s warmen and the second s and the second and the second and the second s مل منصبحه بر مستحص and the second s and a second and a second as a second a second a

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?



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
 - o ATG (methionine)
 - \circ GTG (valine)
 - TTG (leucine)
 - 3 stops (TAA, TAG, TGA)
 - Space in-between: Open Reading Frame -- ORF

ATGGACCTCTCGCCC ATGGACCCTCCTCGCCC TGGACCTCTCGCC.... GGACCTCTCGCC....

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 CCGTTAAGAGGACATGGCCTAGGTATGGCTACCCAAACTTAGATTCAAAA GCCGGGGCGCTCGCACTCATTCGCATCGCCGCCCGAGGCGGCTGGGATGC GGAAATCTACGAGCCGTGGGATGAGGACGAATACCTCCTATAGTGATCTA CGCCACTTGCTCGGTGGGTGTCAAGTGATACTCATGTATCTAGTTATTGA GGGCCTAAAGGCCCGAATAAGAGCCGCACAGGCGGCTCTCTAAGAGCGCC CACTAGGGCGCTCGAAGTAATACCGGCCTTGAGGGCCGGTTATCTGACCC GTGAGGCAACCGTGTACGGCACTCGCTCGAGTGCCTACTGGGCCTCGCAG CCGGGGGAAGTTCGACGTTCTGAACCTGCGGATGACGTTCCCGAGCACGTC

· · · · · 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 · · E · · M · · G · · S · · D · · L · · R · · F · · S · · P · · A · · L · · V · · D · · S · · R · · N · · G · · ¶ $\cdot Q + I + L + V + C + T + E + P + G + G + G + F + A + V + S + P + K + W + G + L + T + C + G + F + R + Q + H + L + L + I + P + E + M + G + \P$ $1 \cdots$ to carattered to construct the transformed and the transformed and the transformed and the transformed the transformation of transfo AAAACCAGACATGCCTTGGGCCCCCAAAGCGCCCAAAGGGGCTTTACCCCAGACTGGACGCCCAAAAGCGGTCGTGAACAACTAAGGGCTTTACCC · O · L · N · O · D · T · R · F · G · P · T · E · R · N · G · R · F · P · T · O · G · A · T · K · A · L · V · O · O · N · G · F · H · ¶ $(\mathbf{A} + \mathbf{S} + \mathbf{K} + \mathbf{P} + \mathbf{R} + \mathbf{Y} + \mathbf{P} + \mathbf{V} + \mathbf{R} + \mathbf{P} + \mathbf{N} + \mathbf{R} + \mathbf{P} + \mathbf{K} + \mathbf{G} + \mathbf{S} + \mathbf{I} + \mathbf{P} + \mathbf{D} + \mathbf{S} + \mathbf{R} + \mathbf{N} + \mathbf{E} + \mathbf{G} + \mathbf{A} + \mathbf{S} + \mathbf{T} + \mathbf{S} + \mathbf{E} + \mathbf{R} + \mathbf{F} + \mathbf{P} + \mathbf{F} +$ · · · · · S ·W··C··K··N··I··G··S··I··P $\cdot \mathbb{C} \cdot \mathbb{I} \cdot \mathbb{K} \cdot \mathbb{T} \cdot \mathbb{Q} \cdot \mathbb{V} \cdot \mathbb{S} \cdot \mathbb{G} \cdot \mathbb{P} \cdot \mathbb{P} \cdot \mathbb{K} \cdot \mathbb{A} \cdot \mathbb{T} \cdot \mathbb{E} \cdot \mathbb{G} \cdot \mathbb{F} \cdot \mathbb{H} \cdot \mathbb{P} \cdot \mathbb{R} \cdot \mathbb{V} \cdot \mathbb{Q} \cdot \mathbb{P} \cdot \mathbb{K} \cdot \mathbb{R}$ · E · V · T · M · P · P · V · P · K · D · P · S · V · R · A · 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··C··V··L··V··A··I··S··L··R··R··G··L··R··C··L··R··I··M··M··W··W¶ · 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 · · ¶ T -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-·····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··G¶ GAACGGCCTACCACAACGCACCGTAGGCAACTGCCACGCGACCACCTTACTGTAAACCCGCAGCGG CTCTGAGC · 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 · ¶ ·····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·¶ T <mark>D .</mark> V A M L Y N D F W T A D N A K A R A E A O V R L E K A · Y · · O · · 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 · E · · G · ¶ GATATCAACGGGTTGTTCCGTGTGGCGATGCTCTATAACGATTTTTGGACTGCCGATAACGCGAAGGCGCGCGGGGGGCTCAGGTTCGGTTGGAGAAGG CTATAGTTGCCCAACAAGGCACACCGCTACGAGATATTGCTAAAAACCTGACGGCTATTGCGCTTCCGCGCACGCCTCCGAGTCCAAGCCAACCTCTTCC · 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 · · F · · ¶ · P · Y · * · R · T · T · C · H · P · S · A · R · Y · R · N · K · S · O · R · Y · R · S · P · A · H · P · P · E · P · E · T · P · S · P ¶ T · T · L · I · M · G · R · I · R · W · L · V · A · V · W · S · G · R · L · R · R · P · R · I · R · L · R · L · R · G · R · S · G · G · S ¶ 401 ··· CCGACACTGATTATGGGACGAATCCGTTGGCTCGTCGCCGTTTGGAGTGGCGACGACGACGAGGATTCGAAGGGCTAAGGGGTCGAAGCGGCGGAA GACTAATACCCTGCTTAGGCAACCGAGCGGCAAACCTCACCGTCTAACTCCGCTGGCTCCTAAGCT $\cdot S \cdot \cdot V \cdot S \cdot \star \cdot P \cdot V \cdot F \cdot G \cdot N \cdot A \cdot R \cdot R \cdot R \cdot K \cdot S \cdot H \cdot C \cdot I \cdot S \cdot A \cdot V \cdot S \cdot S \cdot E \cdot F \cdot A \cdot L \cdot P \cdot D \cdot F \cdot R \cdot R \cdot F$. C. . C. . D. . T. . D. . F. . D. . C. . N · O .m..c..a..a..c.¶ T ········<mark>·S··E··A··A··P··V··C··P··P··E··P··G··D··D··P··R··L··K··L··V··T··*··</mark>R··P··Y··G··C··F··A··G··P··C··C··¶ L R L R R C A H R S L V T I L V * S L * P D G L M A V L O V P A V ¶ 501 ····GTCTGAGGCTGCGCCGGTGTGCCCACCGGAGCCTGGTGACGATCCTCGTTTGAAGCTTGTGACCGGCCTTATGGCCGGTGTTTGCAGGTCCCTGCTGT CAGACTCCGACGCGCCACACGGGTGGCCTCGGACCACTGCTAGGAGCAAACTTCGAACACTGGACGGCGGAATACCGACAAAACGTCCAGGGACGACA · 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 · ¶ $\cdots \cdots G \cdot F \cdot S \cdot V \cdot S \cdot V \cdot S \cdot V \cdot G \cdot S \cdot A \cdot G \cdot V \cdot \star \cdots L \cdot H \cdot \star \cdots G \cdot S \cdot D \cdot G \cdot V \cdot R \cdot P \cdot W \cdot L \cdot I \cdot V \cdot G \cdot A \cdot G \cdot R \cdot T \cdot S \cdot R \cdot \P$ - D - L - A - F - P - T - L - G - P - O - V - C - D - F - I - E - D - R - M - V - F - G - P - G - S - L - S - G - O - A - A - R - L - ¶ 601 ···· GGATTTAGCGTTCCTACGTTCGGGTCCGCAGGTGTGTGGGATCGTTCATGAGGATCGGATGGTGTTCGGCCCTGGCTCATTGTCGGGGCAGGCCGCACGTCTC ······CCTAAATCGCAAAGGATGCAACCCAGGCGTCCACACACTGAAGTAACTCCTAGCCTACCACAAGCCGGGACCGAGTAACAGCCCCGTCCGGCGTGCAGAG ¶

Six-Frame Translations



Ovechkin_Draft



Ovechkin













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 (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 27417 -3 28925 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 S NCBI **Microbial Genomes**

GLIMMER

TaxPlot BL

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

Collaborators gMap ProtMap

Microbial Genome Annotation Tools

HOME SEARCH SITE MAP

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.

Prokaryotic Projects

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.

Genome

in complete sequence, 52927 bp including 10 bp 3' overhang (CGGGCGGTAA), Cluster A7, Order 4, Window 96, Step 12, 2/27



GeneMark Output (trained on *M. tuberculosis*)

	DNA Master – 🗗 🗙						
	File Tools Window Help						
1 0		ORF Analysis for Jefe_Draft	- • •				
	1	4 7 8	観				
	+3						
			10				
1 1	+2						
1 1			9				
	+1						
1 2 36 57 86 95 100	-2						
1 1		╶╆┰╴╴╧┶╴╴╧┷┺┲╤╧╼┺╌╤╴╩╬╖╖╤╬╴╌╌╖╧╶╬╗╤╤╧╧╬┱╧╤┲╶╤╤╤╧╧╬╝╴╝╝╝╖╧╧╝╧╝╸╝┚┻╶╤╤╝╬╧╴╄╌╖╤┾╤╝╧╤╧═					
app. 140 app. 140 <th< th=""><th>-3 </th><th></th><th></th></th<>	-3						
Dr. 1101 Milet OFF. Stat. Stat. Chocs	1 173 345		4473 4645 4817 5000				
Concerned Concerned <t< td=""><th>bp: 1140 H 📢 🔍 🔍</th><td>3 ▶ ▶ ▶ 0RF 827 - 1126</td><td></td></t<>	bp: 1140 H 📢 🔍 🔍	3 ▶ ▶ ▶ 0RF 827 - 1126					
Universe Versee Vers							
Sort By Index Imag Name Find Image Description [Sequence] Product Begins Bits Concest OLL Choces OLE start Image Choces OLE start Image Concest OLE start Image	Uverview realules Reference	ences sequence Documentation	and the second second				
Select Feature Direct SQL UNAW_1 1 <th< td=""><th>Sort By Index 👻 💆</th><td>Tag Name 5'End 3'End Length A Description Sequence Product Regions Blast Context Choose ORF start</td><td></td></th<>	Sort By Index 👻 💆	Tag Name 5'End 3'End Length A Description Sequence Product Regions Blast Context Choose ORF start					
Type I I Color Color Color GI Name Bke DNAM_3 3 603 102	Select Features Direct SQL	L DNAM 1 1 1 645 645 Name 3 GenelD Starts 15 ORF Start 839 Cdn 1 Cdn 2 Cdn 3 Length SD Scoring Matrix	Kibler6 Kibler6				
Name Reg Divertion Start Start <t< td=""><th>Type is All 💌</th><td>▼ DNAM 3 3 838 1126 288 Type CDS ▼ GI</td><td>atrix Karlin Medium 💌 Document</td></t<>	Type is All 💌	▼ DNAM 3 3 838 1126 288 Type CDS ▼ GI	atrix Karlin Medium 💌 Document				
Genello ONAM_5 5 2820 295 244 3* End 1126 Regions Tag Tag Beach Data 10 Descent 2 Value Descent 2 Value <thd< td=""><th>Name like</th><td>DNAM 4 4 1123 2745 1623 5'End 839 Locus Tag DNAM_3</td><td></td></thd<>	Name like	DNAM 4 4 1123 2745 1623 5'End 839 Locus Tag DNAM_3					
Locus New ONAM_7 7 467 900 900 9497 409 Tg Stat OnAM_7 7 9457 9321 105 Faster 10 94 Longh 9 ONAM_8 8 3662 4422 741 93 Tg Faster 10 94 Longh 9 ONAM_9 9 4422 4748 327 66 457 450 CA 0 NAM_10 10 475 5122 378 C C Onam 10 475 5122 378 C C Onam C C Faster 10 94 Tag Tag <th <="" tag<="" td=""><th>GenelD =</th><td>DNAM 5 5 2820 3053 234 3'End 1126 Regions 1 States SD Genomic Spacer Final Sequence of the Begion Start S</td><td>tart ORF</td></th>	<th>GenelD =</th> <td>DNAM 5 5 2820 3053 234 3'End 1126 Regions 1 States SD Genomic Spacer Final Sequence of the Begion Start S</td> <td>tart ORF</td>	GenelD =	DNAM 5 5 2820 3053 234 3'End 1126 Regions 1 States SD Genomic Spacer Final Sequence of the Begion Start S	tart ORF			
Start DNAM, 7 7 3457 3621 105 Lengh DNAM, 9 8 362 422 474 277 450 Lengh DNAM, 9 8 362 CA 200 CA 200 CA 474 512 378 442 474 277 450 Co DNAM, 10 10 4745 5122 378 CA C Co Co <th< td=""><th>Locus like</th><td>DNAM_6 6 3050 3457 408 Length 288 M Ian</td><td>osition Length</td></th<>	Locus like	DNAM_6 6 3050 3457 408 Length 288 M Ian	osition Length				
Length DNAM_9 8 SB62 442 741 Pegine DNAM_9 9 442 748 277 743 741 741 745 752 744 353 V col DNAM_11 11 1515 5404 213 0 -6.954 22.935 10 -6.954 22.940 353 353 2.359 2.0 -6.954 AGGAATCOCCACAGE CACOUCACING CTGA GE GTE 758 358 V col No.M 12 1515 6403 677 353 365 2.359 10 -2.982 AGGAATCOCCCACOUCACING CTGA GE GTE 758 358 Product Imation 11 15 6403 6774 353 365 2.983 11 -7.158 Conduct GTE 758 358 2.993 10 -2.982 AGGACACOUCACOUCACINAGE 744 353 Product Imation	Start >	DRAM_7 7 3457 3621 165 Direction Forward T Feature ID 94	77 450				
Column 0 NAM_10 10 44/2 4/48 327 748 327 748 328 748 328 749 2.099 10 -4.043 CCATTOCCATORGETOCALCOC CTG 669 488 3 6.0 0 0 0 66 512 378 748 353 CA 0 0 0 0 66 511 411 0 66 512 378 359 2 -4.061 TAGORGETOCACACCE COCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	Length	DNAM_8 8 3682 4422 741	80 447				
Product 0 NAM_11 11 512 3/8 12 -4.061 TRADGOTECOCCARCECUTATE TO TO 794 395 CA 0 0 NAM_112 12 5404 213 0 -2.889 12 -4.061 TRADGOTECOCCARCECUTATE TO TO 795 359 CA 0 0 NAM_112 12 5404 213 0 -2.884 9.05.07 67009CTOCCERCECUTATE TO TO 795 359 Poduct 0 NAM_115 15 6403 674.7345 5600 6.888 11 -7.116 6005 GCCACCTOCCACCECUTOT ACCOCCIDECOCCACCUTTOT ATC 685 282 Poduct 0 NAM_115 15 6403 674.7345 5600 6.488 11 -7.158 2824 AGRACEANCECUTOT ACCACCTOCCACCUTOT ATC 685 282 Function 0 NAM_17 17 744 936 6.438 1.477 1.438 1.471 6.437 7.441 839 282 Function 0 NAM_17 17 745 785 7.441 838 1.477 1.4383 1.477 1.438 <th>Daniana D</th> <td>DNAM 9 9 4422 4748 327 TRINSION TO BE DECEMBER 200 00 00 00 00 00 00 00 00 00 00 00 00</td> <td>89 438</td>	Daniana D	DNAM 9 9 4422 4748 327 TRINSION TO BE DECEMBER 200 00 00 00 00 00 00 00 00 00 00 00 00	89 438				
X UC Drvdm_1 1	Hegions	DNAM_10 10 4/45 5122 3/8 CC Walleer	34 393				
CA Diffuent 1: 1:	% GC <	DNAM_11 11 0132 0404 213 5 -4.299 1.888 9 -5.074 GGAACGCCACGAGGCGAC GTG 7	58 369				
ECH New DNAM_14 14 6176 6408 231 Product like DNAM_15 15 6403 6477 345 248 10 248 24			34 333				
Product Is Function Functi	EC# like	DNAM 14 14 5175 5405 231	29 200				
Function like DNAM_116 16 6840 7385 546 FeatureD D	Product like	DNAM_15 15 6403 6747 345 Function	45 282				
FeaturelD ONAM_17 17 7451 762 171 Tag like ONAM_18 18 764 18 1.547 11 -6.575 CATCTACTACCACGTACQACGACCC CTC 944 183 ONAM_19 19 6013 6201 189 Dismand Element al 678 Original Element al 678 Original Element al 678 0.566 12 -7.775 CTCGGACGACGACGACGCOC ATC 1034 93 Select AI Features ONAM_22 22 10518 10516 99 - - - - - - - - - - - - - 1034 93 -	Function like	DNAM_16 16 6840 7385 546	78 249				
Tag NRA 10 7624 0016 333 Topological 10 -4.637 CTALACCTACACCENCE<	FeatureID =	DNAM_17 17 7451 7621 171	44 183				
Hide Ignored Features DNAM_13 13 8013 8013 8013 8013 8013 8013 8013 8013 8013 8013 8013 189 staf al 873 13 -2.071 2.930 16 -3.847 CCTGAACAGAGAGCATCGAACGOG ATG 1034 93 Mide Ignored Features DNAM_22 21 10091 10537 447 -0.011 -2.930 16 -3.846 CCTGAACAGAGAGCATCGAACGOG ATG 1039 18 Invest Delete Pox Validate -	Tag like	DAAM_18 18 7624 8016 333 Totes 1 2 -3.942 2.021 10 -4.637 CTACGACGACGGGGGGGGGGCCC GTG 9	50 177				
1 1 0 94/2 0009 1058 1057 447 Select AI Features 0NAM_22 22 10519 10561 99 1 1 5 -2. 624 2. 661 8 -3.846 TCTCCTCGCCATCCTCTCCCCCG ATG 109 18 Intert Delete Post Valdate - - - 8 -3.846 TCTCCTCGCCATCCTCTCCCCG ATG 109 18 Intert Delete Post Valdate - <	Hide Impored Features	DNAM 19 19 80/3 8201 189 Ungma timmer call gop out as strength 3.10, Genewaik Calls 1 3 -2.071 2.930 16 -3.867 CCTGGAGGAGGCCATCGAGGGG ATG 1	034 93				
Select Al Features Universe 1 21 100001 100007 1447 Insert Delete Post Validate Validate Validate M 4(10) Postion: 9155 If Controls >> Map I/r Map >> Controls Map /r Map >> Controls) The ignored reduces	DNAM_20 20 8427 0034 bb8	097 30				
Inset Delete Post Validate H 4<	Select All Features	DNAM 22 22 10515 10507 1	109 18				
Inset Delete Poxt Validate H 4							
H 44 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4		Inset Delete Post Validate					
		N - 50000 Poulion 9155 C Controls >> Mac 7 Max >> Controls	termine the second				
	MILE HANDER		the state				
	91 Features ive Swift		and the state of t				
cession / / Favorite Genome Recent File 😥 ALT DD	Accession V	Favorite Genome Recent File 👷 🏨 DD					

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

stn blastp bla	stx tblastn tblastx
Enter Query a	iequence
Enter accession r	Iumber(s), gl(s), or FASTA sequence(s) 🤐 Clear Query subrange 🐭
ENERGIANGY CARGENOU AVERGLAT DELEARAN NOVES SUBBELINGYNA REMLEL I SOMNEDYD ATPONEVAE DMRODEN DSSKETDAAVEARLAE	
Or, upload file	Browse No file selected.
Job Title	
	Frite a developing the foregoing DL 107 seconds 0
Align two or m	ore sequences 🥹
Choose Searc	ch Set
Database	Non-redundant protein sequences (nr)
Organism	Exclude +
optional	Enter organism common name, binomial, or tax id. Only 20 too taxa will be shown.
Exclude Optional	Models (XM/XP) Uncultured/environmental sample sequences
Entrez Query	You The Create custom database
optional	Enter an Entrez query to limit search 🤢
Program Sele	ction
Algorithm	blastp (protein-protein BLAST)
	O PSI-BLAST (Position-Specific Iterated BLAST)
	O PHI-BLAST (Pattern Hit Initiated BLAST)
	 DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
	Choose a BLAST algorithm 😡



Distribution of 100 Blast Hits on the Query Sequence 🤬



Sequences producing significant alignments:

Select: All None Selected:0

A-4	🕻 Alignments 🖥 Download 🖂 GenPept Graphics Distance tree of results Multiple alignment 🗘						
	Description	Max score	Total score	Query cover	E value	Ident	Accession
C	gp12 [Mycobacterium phage Timshel]	1047	1047	99%	0.0	90%	AEJ92326.1
E	terminase [Mycobacterium phage Obama12]	1045	1045	99%	0.0	88%	YP_009007203.1
C	terminase [Mycobacterium phage Dhanush]	1045	1045	99%	0.0	88%	AGK87127.1
E	gp11 [Mycobacterium phage Flux]	1044	1044	99%	0.0	87%	AFL47832.1
E	terminase [Mycobacterium phage Nyxis]	1044	1044	99%	0.0	88%	YP_009005850.1
C	terminase [Mycobacterium phage HINdeR]	1043	1043	99%	0.0	89%	YP_008051863.1
C	gp10 [Mycobacterium phage Arturo]	1043	1043	99%	0.0	88%	AFU20464.1
E	terminase [Mycobacterium phage Kampy]	1043	1043	99%	0.0	87%	YP_009031871.1
E	gp11 [Mycobacterium phage Shaka]	1043	1043	99%	0.0	87%	AEF57321.1
C	gp11 [Mycobacterium phage Sabertooth]	1042	1042	99%	0.0	87%	AFU20551.1
E	terminase [Mycobacterium phage BellusTerra]	1040	1040	99%	0.0	87%	YP_009005569.1
e	gp11 [Mycobacterium phage ICleared]	1040	1040	99%	0.0	87%	AFL46618.1
E	gp11 [Mycobacterium phage Peaches]	1040	1040	99%	0.0	87%	YP_003358714.1
E	gp11 [Mycobacterium phage Wile]	1038	1038	99%	0.0	87%	YP_009014094.1
E	terminase [Mycobacterium phage QuinnKiro]	1023	1023	99%	0.0	86%	AIS73685.1

Alignments

Bownload v GenPept Graphics

gp12 [Mycobacterium phage Timshel]

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

Score		Expect	Method		Identities	Positives	Gaps
1047 b	bits(270	7) 0.0	Compositional n	natrix adjust.	504/562(90%)	531/562(94%)	1/562(0%)
Query	5	LSTEPLL	POPPHKIGPVWLCH	EDGSWALPKYT	LGWGVLNWLAEYVE	RSPAGGGPFIPTLE	64
Sbjct	4	LNPGPLL	POPPHKIGPVWQVH	EDGSWALPART	LGWGVLNWLAEYVF	RSPAGGGVFIPTLE	63
Query	65	ARFILWW	YAVDENGVYAYREG	VLRRMKGWGKD	PLCAAIALVELCG	PVAFSHWDEKGNPV	124
Sbjct	64	ARFILWW	YAVDERGNYAYREG	CLRRMKGWGKD	PLCAAIALVELCG	PVAFSHWDLDGSPV	123
Query	125	KRRHAAW	ITIAAVSQDQTKNT	FSLFPVMISKQ	MKTEYGLDVNKFVI	IYTEDGGRIEAATS	184
Sbjct	124	SKPRHAAW	ITVAAVSQDQTKNT	FSMFPVMISKK	MKVDYGLDVNKFV	IYSEEGGRIEAATS	183
Query	185	SPASMEGNI SPASMEGNI	RPTLVIENETQWWG	VGPDGNVNDGV	AMDDVIEGNVSKI	PGARKLAICNAHIP	244
Sbjct	184	PASMEGN	RPTLVIENETQWWG	VGPDGNVNDGP	AMDDVIEGNVSKI	PGARKLAICNAHIP	243
Query	245	SNDTVAEK	AYDHWODILSGKAV	DTGLMYDALEA	PADTPVSEIPSEK	EDPEGYEAGIAQLM	304
Sbjct	244	SNDTVAEK	AYDHWODILSGKAV	DTGIMYDALEA	PADTPVSEIPSEK	EDPEGYEKGIAQLM	303
Query	305	GLEVARG	DSYWLPLEEILGSV	LNTRNPVTESR	RKFLNQVNAHEDS	NIAPSEWDRLAVTD	364
Sbjct	304 1	EGLEIARG	DSYWLPLDEIMGSV	LNTRNPVTESR	RKFLNQVNAHEDSV	WIAPAWWDRLALTD	363
Query	365	L+K+	DRITLGFDGSKSDD DRITLGFDGSKSDD	WTALVACRVSD	GMLFLIKSWNPEDY GMLFL+ WNP DY	PHEEVPREDVDAV	424
Sbjct	364	PLFKLKKN	DRITLGFDGSKSDD	WTALVACRVSD	GMLFLLDKWNPNDY	PHDEVPRDEVDAV	423
Query	425	RSAFORY	DVVGFRADVKEFEA	YVDQWGRDFKR	KLKINATPGNPVAL	PDMRGQTKRFALDC	484
Sbjct	424	RSAFORY	DVVGFRADVKEFEA	YVDQWGRDFKR	KLKVNATPGNPIA	FDMRGQTKRFALDC	483
Query	485	ERFVDAVI	EHELHHDGNPVLRQ	HVLNARRHPTT	YDAISIRKESKDS:	SKKIDAAVCAVLAF	544
Sbjct	484	ERFLDAVI	EKELWHDQNPVLRQ	HVLNARRHPTT	FDAISIRKESKDS	SKKIDAAVCAVLAF	543
Query	545	ARODYOM	SKKHRSGAKAVIIR	566			
Sbict	544	SRODYLM	SKKHRGGG-AVIVR	564			

Bownload v 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			V Next Match	Previous Mat
Score		Expect	Method		Identities	Positives	Gaps
1045	bits(27	03) 0.0	Compositional	matrix adjust.	496/566(88%)	537/566(94%)	1/566(0%)
Query	1	MALEYLSTE M+L	PLLPOPPHKIGP	WLCHEDGSWAL	PKYTLGWGVLNWLA P+ TLGWG+LNWLA	AEYVRSPAGGGPFI A+YVRSPAGGGPF+	60
Sbjct	1	MSLANHHPV	PLLPQPPHKIGP	VWQVREDGSWHL	PERTLGWGILNWLA	AKYVRSPAGGGPFL	60
Query	61	PTLEQARFI	LWWYAVDENGVY	AYREGVLRRMKG	WGKDPLCAAIALVI	ELCGPVAFSHWDEK	120
Sbjct	61	PTLEQARFI	LWWYAVDERGVY	AYREGVLRRMKG	WGKDPLCAAIALAB	ELCGPVAFSHWDLE	120
Query	121	GNPVGKRRE	AAWITIAAVSQL	QTKNTFSLFPVM	ISKOMKTEYGLDV	KEVIYTEDGGRIE	180

The Mycobacteriophage Database						
Home Phages I	Data BLAST Publications Resources Software Social Other DBs About					
Search PhagesDB.org	Local Protein BLAST					
Recently Added Phages	Go to Nucleotide BLAST					
Bryanna	This tool will run a local BLAST search against our protein databases. This includes all proteins from the					
RadDad	most recent Phamerator update. Proteins marked "Draft" are from auto-annotated files.					
Dyappa	Choose program to use and database to search					
Daishi	Program blastp Database Mycobacteriophage Proteins as of Feb 07, 2014					
Recently Modified Phages	Enter sequence below in FASTA format					
HINdeR	AATSSPASMEGNRPTLVIENETOWWGVGPDGNVNDGVAMDDVIEGNVSKIPGARKLAICN					
Ferb	AHIPGNDTVAEKAYDHWQDILSGKAVDTGLMYDALEAPADTPVSEIPSEKEDPEGYEAGI AQLMDGLEVARGDSYNLPLEEILGSVLNTRNPVTESRRKFLNOVNAHEDSWIAPSEWDRL					
Biaggio	AVTDKALALQKDDRITLGFDGSKSDDWTALVACRVSDGMLFLIKSWNPEDYPHEEVPRED					
Cgwhichard	ALDCERFVDAVIEHELHHDGNPVLRQHVLNARRHPTTYDAISIRKESKDSSKKIDAAVCA					
Filch						
Recently Finished Phages						
Cambiare (G)	Browse No file selected.					
AlanGrant (B4)	Set subsequence: From To					
Corofin (B3)						
Baee (B5)	Clear sequence BLAST					

Distribution of 100 Blast Hits on the Query Sequence



	Score E
Sequences producing significant alignments:	(bits) Value
Shoon-Draft 14 function unknown 566	1171 0.0
Timphel 12 Terminane 564	1052 0.0
WINdop 11 torringso 564	1055 0.0
Obama12 draft 10 function unknown 565	1046 0.0
LARGOC 11 Terrinano largo gubunit 565	1046 0.0
historia terminase, large subunit, 565	1046 0.0
Phickter 1004 Draft 10 function unknown Eff	1046 0.0
Phighter1804-Draft 10, function unknown, 565	1046 0.0
Lagie 10, Terminase, 565	1046 0.0
Camperdownii Draft 10, function unknown, 565	1046 0.0
Flux II, terminase, 565	1045 0.0
Dhanush II, terminase, 565	1045 0.0
wander Draft 10, function unknown, 565	1044 0.0
TygerBlood-Drait 10, function unknown, 565	1044 0.0
TiroThetay II, 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
al	1011 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%)
Ouerv: 5 YLSTEPLLPOPPHKIGPVWLCHEDGSWALPKYTLGWGVLNWLAEYVRSPAGGGPFIPTLE 64
          YL+ PLLPQPPHKIGPVW HEDGSWALP TLGWGVLNWLAEYVRSPAGGG FIPTLE
Sbjct: 4 YLNPGPLLPOPPHKIGPVWQVHEDGSWALPARTLGWGVLNWLAEYVRSPAGGGVFIPTLE 63
Query: 65 QARFILWWYAVDENGVYAYREGVLRRMKGWGKDPLCAAIALVELCGPVAFSHWDEKGNPV 124
          QARFILWWYAVDE G YAYREG LRRMKGWGKDPLCAAIALVELCGPVAFSHWD G+PV
Sbjct: 64 OARFILWWYAVDERGNYAYREGCLRRMKGWGKDPLCAAIALVELCGPVAFSHWDLDGSPV 123
Ouery: 125 GKRRHAAWITIAAVSQDQTKNTFSLFPVMISKQMKTEYGLDVNKFVIYTEDGGRIEAATS 184
          GK RHAAWIT+AAVSQDQTKNTFS+FPVMISK+MK +YGLDVNKFVIY+E+GGRIEAATS
Sbjct: 124 GKPRHAAWITVAAVSQDQTKNTFSMFPVMISKKMKVDYGLDVNKFVIYSEEGGRIEAATS 183
Query: 185 SPASMEGNRPTLVIENETQWWGVGPDGNVNDGVAMDDVIEGNVSKIPGARKLAICNAHIP 244
           SPASMEGNRPTLVIENETQWWGVGPDGNVNDG AMDDVIEGNVSKIPGARKLAICNAHIP
Sbjct: 184 SPASMEGNRPTLVIENETOWWGVGPDGNVNDGPAMDDVIEGNVSKIPGARKLAICNAHIP 243
Query: 245 GNDTVAEKAYDHWQDILSGKAVDTGLMYDALEAPADTPVSEIPSEKEDPEGYEAGIAQLM 304
          GNDTVAEKAYDHWODILSGKAVDTG+MYDALEAPADTPVSEIPSEKEDPEGYE GIAOLM
Sbjct: 244 GNDTVAEKAYDHWQDILSGKAVDTGIMYDALEAPADTPVSEIPSEKEDPEGYEKGIAQLM 303
Ouery: 305 DGLEVARGDSYWLPLEEILGSVLNTRNPVTESRRKFLNOVNAHEDSWIAPSEWDRLAVTD 364
          +GLE+ARGDSYWLPL+EI+GSVLNTRNPVTESRRKFLNOVNAHEDSWIAP+ WDRLA+TD
Sbjct: 304 EGLEIARGDSYWLPLDEIMGSVLNTRNPVTESRRKFLNQVNAHEDSWIAPAWWDRLALTD 363
Query: 365 KALALQKDDRITLGFDGSKSDDWTALVACRVSDGMLFLIKSWNPEDYPHEEVPREDVDAV 424
              L+K+DRITLGFDGSKSDDWTALVACRVSDGMLFL+ WNP DYPH+EVPR++VDAV
Sbjct: 364 PLFKLKKNDRITLGFDGSKSDDWTALVACRVSDGMLFLLDKWNPNDYPHDEVPRDEVDAV 423
Query: 425 VRSAFORYDVVGFRADVKEFEAYVDOWGRDFKRKLKINATPGNPVAFDMRGQTKRFALDC 484
          VRSAFORYDVVGFRADVKEFEAYVDOWGRDFKRKLK+NATPGNP+AFDMRGOTKRFALDC
Sbjct: 424 VRSAFQRYDVVGFRADVKEFEAYVDQWGRDFKRKLKVNATPGNPIAFDMRGQTKRFALDC 483
Query: 485 ERFVDAVIEHELHHDGNPVLRQHVLNARRHPTTYDAISIRKESKDSSKKIDAAVCAVLAF 544
          ERF+DAVIE EL HD NPVLRQHVLNARRHPTT+DAISIRKESKDSSKKIDAAVCAVLAF
Sbjct: 484 ERFLDAVIEKELWHDONPVLROHVLNARRHPTTFDAISIRKESKDSSKKIDAAVCAVLAF 543
Query: 545 GARQDYQMSKKHRSGAKAVIIR 566
          G+RODY 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+ PLLPOPPHKIGPVW HEDGSWALP TLGWGVLNWLAEYVRSPAGGGPFIPTLE
Sbjct: 4 YLNPGPLLPQPPHKIGPVWQVHEDGSWALPARTLGWGVLNWLAEYVRSPAGGGPFIPTLE 63
Query: 65 QARFILWWYAVDENGVYAYREGVLRRMKGWGKDPLCAAIALVELCGPVAFSHWDEKGNPV 124
          OARFILWWYAVDE G YAYREG LRRMKGWGKDPLCAAIALVELCGPVAFSHWD G+PV
Sbjct: 64 QARFILWWYAVDERGNYAYREGCLRRMKGWGKDPLCAAIALVELCGPVAFSHWDLDGSPV 123
```

Phamerator map



Starterator data

Note: In the above figure, vellow 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 904, Corolin 95
 Track 5: ChaChing, Draft 95, Yahalom Draft 91, Phlyer 95
 Track 5: NOZO Draft 96, Compostia_Draft 98, Gadjet 93, OrangeOswald_94
 Track 7: Chardler, 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, Kamivu 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, Philyer 95, OrangeOswald_94, Gadjet_93, Mortcellus_Draft_96, ChaChing_Draft_95,

Percent with start 3 called: 55.0%



Pham 4391



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.co m/guiding-principles



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

Dutline Introduction



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?